Universiteit Antwerpen

Faculteit Wetenschappen
Departement Fysica

MEASUREMENT OF THE ELECTROWEAK PRODUCTION OF A Z BOSON IN ASSOCIATION WITH TWO JETS WITH THE CMS DETECTOR

METING VAN DE ELECTROZWAKKE PRODUCTIE VAN EEN Z BOSON IN ASSOCIATIE MET TWEE JETS MET DE CMS DETECTOR

Proefschrift voorgelegd tot het behalen van de graad van
doctor in de wetenschappen: Fysica
aan de Universiteit Antwerpen
te verdedigen door
Tom Cornelis

Promotor: Prof. Dr. Pierre Van Mechelen

Antwerpen, 2015

CONTENTS

ACRONYMS

BDT	Boosted Decision Tree
BEH	Brout-Englert-Higgs
BLF	Branching Logarithmic Fit
CDF	Cumulative Distribution Function
CJV	Central Jet Veto
CHS	Charged Hadron Subtraction
CMS	Compact Muon Solenoïd
DY	Drell-Yan
ECAL	Electromagnetic Calorimeter
EW	Electroweak
GSF	Gaussian Sum Filter
HCAL	Hadronic Calorimeter
HB	Hadron Barrel
HE	Hadron Endcap
HF	Hadron Forward
HLT	High Level Trigger
HO	Hadron Outer
JES	Jet Energy Scale
JER	Jet Energy Resolution
JPT	Jet-plus-track
KDE	Kernel Density Estimator
KF	Kalman Filter
L1	Level-1 Trigger
LHC	Large Hadron Collider
LEP	Large Electron-Positron Collider
LO	Leading Order
MC	Monte Carlo
ME	Matrix Element
MLP	Multilayer Perceptron
NLO	Next-to-leading Order
NNLO	Next-to-next-to-leading Order
PDF	Probability Density Function
PF	Particle Flow
PV	Primary Vertex
QCD	Quantum Chromodynamics
QED	Quantum Electrodynamics
ROC	Receiver Operating Characteristic
SM	Standard Model
UE	Underlying Event
VBF	Vector Boson Fusion
VBS	Vector Boson Scattering

INTRODUCTION

The Standard Model of particle physics is one of the most successful theories of the past century. It summarizes our current understanding of all known elementary particles found in nature and the interactions between them with the exception of gravity. Since the discovery of the Higgs boson in 2012, all predictions by the Standard Model have been experimentally confirmed. Nevertheless, there are still many open questions in the Standard Model which cause us to believe our theory is not yet complete and new physics could occur in conditions not probed yet in past experiments.

The Large Hadron Collider provides us with proton-proton collisions at energies never seen before, and collects huge amounts of new data. This allows us to study the Standard Model from a new angle, as some processes are becoming experimentally accessible for the first time. One such process is Weak Vector Boson Fusion, where from each of the protons a W- or Z-boson is emitted from one of its quarks, which subsequently fuse with each other. The two quarks form a distinctive event signature of two particle jets found in the forward and backward region of the detector respectively, while the fusion products of the vector boson interaction are found in the central part of the detector. In this thesis, we focus on the situation where two W-bosons fuse into a well known Z-boson ($WW \rightarrow Z$), resulting in an event signature of a Z-boson and two jets (Zjj). It is, however, not possible to extract a pure $WW \rightarrow Z$ signal from data, as it interferes with other Zjj processes mediated by purely electroweak interactions. Fortunately, it is still possible to apply and test some of the typical Vector Boson Fusion search strategies, and we succeeded in measuring the electroweak production of Zjj and compared its signal strength with the Standard Model prediction. The analyses presented in this thesis are therefore an important benchmark for studies or searches in other Vector Boson Fusion topologies within and beyond the Standard Model, of which the Vector Boson Fusion Higgs production ($WW \rightarrow H$ or $ZZ \rightarrow H$) is the most well known.

This thesis starts with an overview of the Standard Model in Chapter 1, including the phenomenology of proton-proton collisions and how they are simulated. The signal process is discussed in detail in Chapter 2, along with an overview of the background processes. Chapters 3 and 4 describe the Large Hadron Collider and CMS, the experimental setup used to perform the measurements in this thesis. Many different analyses are carried out with this experimental setup, and the object reconstruction algorithms presented in Chapter 5 are developed and shared by a large collaboration.

An important new tool to separate our signal from background processes, is the quark-gluon jet discrimination tool described in Chapter 6. Even though development of this tool started as a small side-project for this thesis, its great potential for many other analyses caused us to upgrade it to an official CMS tool. Hence, the chapter also includes a discussion of more recent updates to the tool, to be applied in future CMS analyses. The quark-gluon jet discrimination tool and its performance on data was earlier documented in a public analysis summary:

- CMS Collaboration, "Performance of quark/gluon discrimination using pp collision data at $\sqrt{s} = 8$ TeV ",
  CMS-PAS-JME-13-002

Chapters 7 and 8 presents the measurement of the electroweak production of a Z boson in association with two jets using proton-proton collisions at a centre-of-mass energy of $\sqrt{s} = 7$ and 8 TeV respectively. These results describe the very first measurements of these kind, and were the smallest cross sections measured at the CMS detector at the time of their publications:

- CMS Collaboration, "Measurement of the hadronic activity in events with a Z and two jets and extraction of the cross section for the electroweak production of a Z with two jets in pp collisions at $\sqrt{s} = 7$ TeV",
  JHEP, vol. 1310, p.062, 2013

- CMS Collaboration, "Measurement of electroweak production of two jets in association with a Z boson in proton-proton collisions at $\sqrt{s} = 8$ TeV",
  Eur.Phys.J., vol. C75, no. 2, p.66, 2015

The thesis then proceeds with Chapter 9 where the measurements of the hadronic activity in Zjj events are described, which have been carried out for the $\sqrt{s} = 7$ and 8 TeV analyses. Finally, a summary and outlook for future follow-up analyses is given in Chapter 10.

1

THE STANDARD MODEL OF PARTICLE PHYSICS

The Standard Model (SM) of particles and interactions is a theory which successfully explains all known elementary particles and most of the phenomena in elementary particle physics. It describes all known fundamental interactions, except gravitation. The contribution of gravitation is negligible to high energy particle physics experiments, due to its strength being multiple orders of magnitude weaker compared to the other fundamental interactions. The SM is a quantum field theory, and this chapter therefore starts with a short introduction on quantum fields and gauge theories, which are the needed ingredients to describe the interactions and particle content of the SM. At the end of this chapter, it is described how the SM can be studied using proton-proton collisions and how Monte Carlo (MC) event generators provide useful simulations to compare the data with theory predictions.

1.1 QUANTUM FIELDS: FERMIONS AND BOSONS

The mathematical description of the SM is a quantum field theory [1] which means its fundamental objects are quantum fields which are defined at all points in space-time. Particles are considered to be the excitations of these fields, and each type of particle is the quantum of a corresponding field. Fields could be classified according to its spin, which is related to how the field behaves under transformations in space-time. A scalar field (spin $s = 0$) has a scalar operator associated to each point in space-time, and is invariant under space-time transformations. A vector field ($s = 1$) associates a vector to each point in space-time, and is therefore affected by a rotation of the coordinate system: a vector needs a full rotation of 2π to turn back into its original direction. In addition to scalar and vector fields, one can also identify spinor fields ($s = 1/2$), which need two full rotations to turn back into its original direction. The spin of a field can be seen as an additional (discrete) degree of freedom: when a particle with spin s is measured along an axis it can have a projection quantum number of $-s, -s+1, \ldots, s-1, s$. Hence, there are $2s+1$ possibilities for this projection quantum number. A spin-1/2 field for example, has two quantum states at each point in space-time, which interchange under a rotation of 2π.

Even though the SM only describes scalar, spinor and vector fields, higher half-integer and integer values are allowed in relativistic quantum field theories, and

could occur in theories beyond the SM or in composed particles. Fields with half-integer and integer spin are respectively called fermionic and bosonic fields. Their corresponding particles are called fermions and bosons, and a system of identical fermions or bosons behaves differently under the exchange of two particles. Consider the wave function $\Psi(x_1^\mu, \sigma_1, x_2^\mu, \sigma_2)$ for a system of two identical particles, i.e. two excitations of the same field, having positions x_1^μ and x_2^μ and spin states σ_1 and σ_2 respectively. This wave function is asymmetric under the exchange of two fermions, while it is symmetric for bosons:

$$\Psi(x_1^\mu, \sigma_1, x_2^\mu, \sigma_2) = (-1)^{2s}\Psi(x_2^\mu, \sigma_2, x_1^\mu, \sigma_1) \tag{1}$$

If both particles are indistinguishable particles, i.e. they share the same space-time coordinate and spin state, equation 1 reduces to $(-1)^{2s} = 1$, which can never be fulfilled in the case of fermions. Hence, two identical fermions can not occupy the same quantum state simultaneously, known as the Pauli-exclusion principle.

1.1.1 *Lagrangian formalism*

A relativistic field theory is mathematically expressed by its Lagrangian density $\mathcal{L}$, which is a function of the fields ϕ_a and their derivatives $\partial_\mu\phi_a$. The action S can be constructed by integrating the Lagrangian density over space-time:

$$\mathcal{S} = \int \mathcal{L}\left(\phi_a(x), \partial_\mu\phi_a(x)\right) \mathrm{d}^4x \tag{2}$$

The action describes all possible trajectories for a system evolving from one given configuration to another. The principle of stationary action postulates that a system evolves along a path which is an extremum, i.e. $\delta S = 0$. The equations of motion are therefore described by the Euler-Lagrange equations

$$\partial_\mu\left(\frac{\partial\mathcal{L}}{\partial_\mu\phi_a}\right) - \frac{\mathcal{L}}{\partial\phi_a} = 0 \tag{3}$$

which can be retrieved for each field ϕ_a in the Lagrangian.

1.1.2 *A free spinor field*

Ordinary matter is built out of particles described by spin 1/2 fields. In the absence of interactions, a free spinor field with mass m is described by the Dirac Lagrangian[1]:

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi \tag{4}$$

1 In this work we will adopt natural units in which the speed of light and the Planck constant are set to unity: $c = \hbar = 1$

in which γ^μ are the Dirac matrices[2] and $\bar{\psi} \equiv \psi^\dagger \gamma^0$ with $\psi^\dagger$ being the hermitian conjugate of the 4-component Dirac spinor ψ. Using the principle of stationary action $\delta S = 0$, the Dirac equation of motion is retrieved:

$$(i\gamma^\mu \partial_\mu - m)\psi = 0 \tag{5}$$

which has both positive and negative energy solutions. The negative energy solutions can be interpreted as an anti-particle: a particle with the same mass, but opposite charges. Both the fermion f and the anti-fermion $\bar{f}$ are the result from the same underlying field.

A 4-component Dirac spinor could be decomposed in its left and right-handed components:

$$\psi = \left(\frac{1-\gamma^5}{2}\right)\psi + \left(\frac{1+\gamma^5}{2}\right)\psi = \psi_L + \psi_R \tag{6}$$

in which we defined

$$\begin{aligned} \psi_L &= \left(\frac{1-\gamma^5}{2}\right)\psi \\ \psi_R &= \left(\frac{1+\gamma^5}{2}\right)\psi \end{aligned} \tag{7}$$

with $\gamma^5 = i\gamma^1\gamma^2\gamma^3\gamma^4$. We will see that the weak interaction only acts on the left-handed component of the Dirac spinor.

1.2 GAUGE SYMMETRIES

The SM postulates that interactions between particles are determined by underlying local gauge symmetries. Symmetries are transformations of the field $\psi \mapsto \psi'$ which describe the same physical situation. This is possible if the symmetries are represented by unitary groups, i.e. $UU^{-1} = 1$, so that the expectation values of the quantum-mechanical observables are invariant:

$$|\psi|^2 \mapsto |U\psi|^2 = \psi^\dagger U^{-1} U\psi = |\psi|^2 \tag{8}$$

If a symmetry acts in the same way at every point in space-time, the symmetry is global and leaves the Lagrangian invariant. According to Noether's theorem, each continuous global symmetry leads to the existence of a conservation law, and it conserved currents allow us to define a charge for the particles. It is however possible to construct a local symmetry, which depends on the position in space-time. Such a local symmetry leads to additional terms in the Lagrangian which break the invariance of the Lagrangian under this symmetry. This invariance can

2 The Dirac matrices satisfy $\gamma^\mu\gamma^\nu + \gamma^\mu\gamma^\nu = 2g^{\mu\nu}$ where $g^{\mu\nu}$ is the metric tensor. This requirement ensures the Dirac equation of motion is consistent with the relativistic energy-momentum relation $E^2 = p^2 + m^2$.

be restored by introducing additional gauge fields, which couple to the conserved currents of the global symmetry. In this section, we will first describe a simple U(1) gauge symmetry, before generalizing to more complex SU(N) groups.

1.2.1 *U(1) gauge symmetry*

Consider a theory that is invariant under a transformation of the Dirac spinor through a phase rotation $\alpha(x)$:

$$\psi(x) \mapsto e^{i\alpha(x)}\psi(x) \tag{9}$$

Because the phase $\alpha(x)$ varies from point to point in space-time, the U(1) symmetry is said to be local. Because of this local dependency, the derivative from the Dirac spinor transforms accordingly as

$$\partial_\mu\psi(x) \mapsto e^{i\alpha(x)}\partial_\mu\psi(x) + (i\partial_\mu\alpha(x))e^{i\alpha(x)}\psi(x) \tag{10}$$

In order to keep the Lagrangian invariant under this $U(1)$ transformation, we need to replace the original space-time derivative ∂_μ by a modified derivative D_μ that transforms covariantly under the same phase transformation:

$$D_\mu\psi(x) \mapsto e^{i\alpha(x)}D_\mu\psi(x) \tag{11}$$

The modified derivative can be constructed by introducing a new vector field $A_\mu(x)$,

$$D_\mu \equiv \partial_\mu + ieA_\mu(x) \tag{12}$$

which cancels the unwanted term in equation 10 if it transforms as

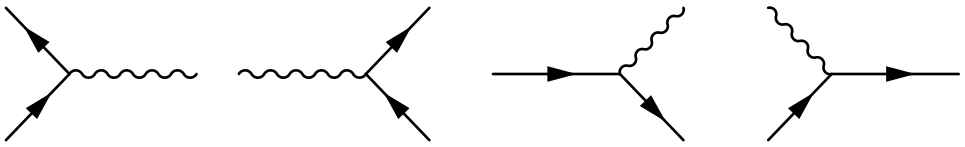
$$A_\mu(x) \mapsto A_\mu(x) - \frac{1}{e}\partial_\mu\alpha(x) \tag{13}$$

Here we also introduced the coupling constant e, which is a free parameter in the theory whose value should be established by experiment. Because the introduced gauge field A_μ is a vector field, its associated particles are bosons. Replacing ∂_μ by D_μ in equation 4 yields a Lagrangian which is invariant for transformation under the local U(1) symmetry:

$$\begin{aligned}\mathcal{L} &= \bar{\psi}(i\gamma^\mu D_\mu - m)\psi \\ &= \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - e\bar{\psi}\gamma^\mu\psi A_\mu\end{aligned} \tag{14}$$

The last term in this Lagrangian can be rewritten as $j^\mu A^\mu$ with j^μ the conserved Noether current[3]. This additional term therefore expresses the coupling with strength e between the conserved fermion current and the new vector field. The resulting interaction vertices, shown in Figure 1, have therefore two fermion legs and one boson leg.

3 Using Noether's theorem, one can show $-e\bar{\psi}\gamma^\mu\psi$ is a conserved current of the global U(1) phase invariance



(a) annihilation (b) pair creation (c) radiation (d) absorption

Figure 1: Feynman diagrams for interaction vertices in a U(1) gauge theory, associated with the $e\bar{\psi}\gamma^\mu\psi A_\mu$ term in the Lagrangian. In these diagrams, the arrow of time flows from left to right. Fermions are represented with solid lines with an arrow going forward in time, anti-fermions have their arrow of time going backwards in time. The gauge boson is represented with a wavy line. At each vertex, an incoming (ψ) and outgoing ($\bar{\psi}$) (anti-)fermion interact with a gauge boson.

To have a full description of the dynamics of both the fermion and gauge field in one Lagrangian, we need to add a kinetic term for a vector boson field: a locally invariant term that depends on A_μ and its derivatives but not on ψ. Because the covariant derivative of ψ is invariant under the U(1) transformation, the same holds for the second covariant derivative and hence for the commutator:

$$[D_\mu, D_\nu]\psi(x) \mapsto e^{i\alpha(x)}[D_\mu, D_\nu]\psi(x) \tag{15}$$

However the commutator itself is not a derivative:

$$\begin{aligned}[D_\mu, D_\nu]\psi &= [\partial_\mu, \partial_\nu]\psi(x) + ie\left([\partial_\mu, A_\nu] - [\partial_\nu, A_\mu]\right)\psi - e^2[A_\mu, A_\nu]\psi \\ &= ie(\partial_\mu A_\nu - \partial_\nu A_\mu)\psi \qquad (16)\\ &= ieF_{\mu\nu}\psi \qquad (17)\end{aligned}$$

Here we retrieve $F_{\mu\nu}$, also known as the electromagnetic field tensor, which is therefore also invariant under the U(1) transformation. Using the electromagnetic field tensor, one can construct a kinetic term for the gauge field:

$$-\frac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{18}$$

which is also gauge invariant.

It is not possible to add a mass term of the form $m^2 A_\mu A^\mu$ as it would transform in ways that cannot be compensated to obtain gauge invariance. Hence, the gauge boson has to remain massless. In summary, if we postulate that a fermion field obeys a local U(1) gauge symmetry, there must exist a massless spin 1 boson. We can therefore identify this Lagrangian with Quantum Electrodynamics (QED), in which the fermion and boson describe the electron and photon respectively:

$$\mathcal{L}_{\text{QED}} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - e\bar{\psi}\gamma^\mu\psi A_\mu \tag{19}$$

1.2.2 *Gauge symmetry for non-abelian groups*

The U(1) group has one generator, but the procedure can easily be generalized to groups with non-commuting generators as was proposed by Yang and Mills [2]. Instead of a single Dirac field, we consider a multiplet of N Dirac fields

$$\psi = \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \\ \vdots \\ \psi_N(x) \end{pmatrix} \tag{20}$$

which transform into each other under a local SU(N) symmetry:

$$\psi(x) \mapsto e^{i\alpha^a(x)t^a}\psi(x) \tag{21}$$

The t^a are the $N^2 - 1$ generators of the symmetry group, which are the Pauli matrices in SU(2) and the Gell-Mann matrices in SU(3). The transformation of the covariant derivative associated with this transformation is:

$$D_\mu \equiv \partial_\mu - igA^a_\mu t^a \tag{22}$$

and contains one new vector field for each independent generator of the local symmetry. The vector fields have to transform as:

$$A^a_\mu \mapsto A^a_\mu + \frac{1}{g}\partial_\mu \alpha^a + f^{abc}A^b_\mu \alpha^c \tag{23}$$

in which the structure constants f^{abc} are defined by the commutation relations of the matrices t^a:

$$[t^a, t^b] = if^{abc}t^c \tag{24}$$

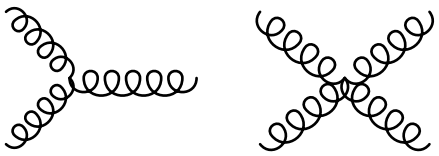
Similar to the U(1) case, we can define the field strength tensor as the commutator of the covariant derivative:

$$\begin{aligned} igF^a_{\mu\nu}t^a &= [D_\mu, D_\nu] \\ \Leftrightarrow F^a_{\mu\nu} &= \partial^\mu A^\nu - \partial^\nu A^\mu + gf^{abc}A^b_\mu A^c_\nu \end{aligned} \tag{25}$$

Differently from the U(1) situation, the field tensor in the non-abelian case contains an additional term which indicates the vector fields are self-interacting. Because of this term, the $F^a_{\mu\nu}F^{a\mu\nu}$ term in the Lagrangian will expand in triplet and quartic terms of gauge bosons, resulting in vertices as shown in Figure 2.

In summary, the Lagrangian for a multiplet of fermion fields which is invariant under $SU(N)$ transformations and interact with the $N^2 - 1$ gauge bosons is given by:

$$\mathcal{L}_{\text{Yang-Mills}} = -\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} - \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi + ig\bar{\psi}\gamma^\mu A^a_\mu t^a \psi \tag{26}$$



(a) (b)

Figure 2: Feynman diagrams for self-interaction of SU(N) gauge bosons

Similar to the U(1) case, one can use Noether's theorem to find the global symmetry currents $g\bar{\psi}\gamma^\mu t^a \psi$ of the fermion field. The interaction term in the Lagrangian can again be written using these currents as $j^{a\mu} A^a_\mu$, so that one can easily see how each of the $N^2 - 1$ gauge bosons correspond couples to a global conserved current, resulting in similar vertices as seen in Figure 1. The $N^2 - 1$ currents in the gauge boson representation are equivalent to assigning N charges in the fermion representation, i.e. one charge for each fermion field in the multiplet. In this view, the $j^{a\mu}$ represents combinations of charge and anti-charge which are carried over by the gauge bosons.

1.2.3 *Renormalization and the running coupling constant*

From the Lagrangian, it is possible to calculate the probability amplitude of a particle interaction. The calculations could be elegantly presented by the use of Feynman diagrams, which describe the perturbative contribution to the amplitude by a particular class of particle paths. Figure 3 shows the first-order and two examples of second-order contributions to the fermion scattering process. For each additional vertex in the Feynman diagram, a factor $\alpha = \frac{g^2}{4\pi}$ is introduced in the scattering amplitude, while the virtual particles (i.e. the particles described by the internal lines in the Feynman diagram) contribute to the total scattering amplitude through their propagators which depend on their momentum.

Perturbation theory assumes the coupling constant is small enough ($\alpha \ll 1$) such that the higher-order contributions could be treated as smaller corrections to the leading-order contribution. However, in these higher-order diagrams, the momentum of the particles involved in the loops are not uniquely determined by the momenta of the incoming an outgoing particles. Therefore, one needs to consider all possible momenta in the propagators, introducing integrals over momentum space which run from zero to infinity. These integrals lead to divergences which can be handled through renormalization. The renormalization procedure hides the divergences by absorbing them in a redefinition of the coupling constant and the mass of the fermion (which also enters in the fermion propagator). Because the higher-order contributions are dependent on the total momentum transfer, renormalization introduces a renormalization scale μ^2, the point at which the subtractions

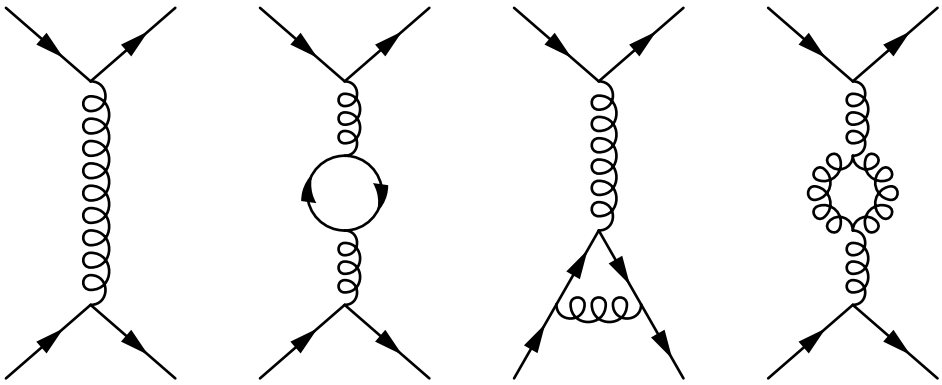


Figure 3: Feynman diagrams for a fermion scattering process. In addition to the first-order diagram (a) in which the incoming fermions are simply scattered by exchange of a gauge boson, many higher-order loop contributions exist which contribute to the probability amplitude of the process. Diagrams of type (b) and (d), in which the second-order contribution involves vertices of the type $\bar{\psi}\gamma^\mu\psi A_\mu$ occur in both U(1) and SU(N) gauge theories. Diagrams of type (d) involve the self-interaction of gauge bosons and only occur in SU(N) theories.

which remove the divergences are performed. As a result, the renormalization coupling also depends on μ^2. This scale dependence is described by the β-function:

$$\beta(\alpha) = \mu^2 \frac{\partial \alpha(\mu)}{\partial \mu^2} = \frac{\partial \alpha(\mu)}{\partial \ln \mu^2} \tag{27}$$

Even though α depends on μ^2, the β-function is independent of the renormalization scale, and only a function of the coupling itself. It can therefore easily be calculated through a perturbation series, and is in the lowest order for a U(1) theory given by:

$$\beta(\alpha) = \frac{2\alpha^2}{3\pi} n_f + \mathcal{O}(\alpha^3) \tag{28}$$

with n_f the number of fermions which could form a loop contribution, which is one (the electron) in the case of QED. This positive β-function tells us that the coupling increases for increasing momentum scale, and thus for interactions at smaller distances. For non-abelian SU(N) theories, the first-order β-function is given by

$$\beta(\alpha) = \frac{\alpha^2}{2\pi}\left(-\frac{11}{3}N + \frac{2}{3}n_f\right) + \mathcal{O}(\alpha^3) \tag{29}$$

In the non-abelian case, a negative contribution to the β-function is added as a result of the self-interaction of the gauge boson, which will dominate if $11N > 2n_f$. In this situation, the coupling decreases for higher momentum scale or shorter distances, a property called asymptotic freedom.

1.3 INTERACTIONS IN THE STANDARD MODEL

1.3.1 *The electroweak interaction*

The Electroweak (EW) theory, first proposed by Glashow, Salam and Weinberg [3, 4, 5], gives a unified description of the electromagnetic and weak interaction. It is generated by the gauge group $\mathrm{SU}(2)_L \otimes \mathrm{U}(1)_Y$ and we can therefore construct a covariant derivative of the form

$$D_\mu = \partial_\mu - ig_2 W_\mu^a t^a - ig_1 \frac{Y}{2} B_\mu \tag{30}$$

in which we introduced three vector fields W_μ^a corresponding with the $\mathrm{SU}(2)_L$ group and one vector field B_μ for the $\mathrm{U}(1)_Y$ group. The fermions which transform under the the $\mathrm{SU}(2)_L$ symmetry can be grouped in flavor doublets. The three $\mathrm{SU}(2)_L$ gauge bosons correspond with the 3 conserved currents and only couple to the left-handed components of the fermion fields, hence the subscript L in the notation of the gauge group. Interactions between the two flavors in a doublet are mediated by the W^1 and W^2 bosons, while the W^0 couples with two fermions of the same flavor. This is reflected in the isospin charge T_3, where the W^0 boson has no weak isospin, while the W^1 and W^2 bosons carry weak isospin $T_3 = \pm 1$. The two flavors in the doublet have has $T_3 = 1/2$ and $T_3 = -1/2$ respectively. The right-handed particles carry no weak isospin charge, and do not interact with the $SU(2)_L$ gauge bosons.

The three $\mathrm{SU}(2)_L$ bosons are however not found independently and mix with the $U(1)_Y$ gauge bosons in order to form the W-boson, Z-boson and photon fields through the following linear combinations:

$$\begin{aligned} W_\mu^\pm &= \frac{W_\mu^1 \pm i W_\mu^2}{\sqrt{2}} \\ Z_\mu^0 &= \cos\theta_W W_\mu^0 - \sin\theta_W B_\mu \\ A_\mu &= \sin\theta_W W_\mu^0 + \cos\theta_W B_\mu \end{aligned} \tag{31}$$

The Weinberg angle θ_W defines the mixing between the neutral $\mathrm{SU}(2)_L$ gauge field and the $\mathrm{U}(1)_Y$ gauge field. As a result it links the coupling constants of the two symmetry groups:

$$\tan\theta_W = \frac{g_1}{g_2} \tag{32}$$

The quantum number conserved via the B_μ exchange, is called the weak hypercharge Y. The electromagnetic charge is retrieved as a simple sum out of the weak isospin and hypercharge using the Gell-Mann-Nishijima formula [6, 7, 8]:

$$Q = T_3 + \frac{Y}{2} \tag{33}$$

Because left and right-handed particles carry a different weak isospin charge, the weak hypercharge should also be different for them in order to end up with the same electric charge as their left-handed partners. The photon is a mix of the W^0_μ and B_μ bosons, and we can therefore expect it to couple to $W^\pm$ bosons. Indeed, the linear combination of W^1_μ and W^2_μ result in a positively and negatively charged W boson which couple to the photon.

Because the $\mathrm{SU}(2)_L$ theory act only on left-handed fermions, the mass term in equation 4 cannot exist: the left- and right-handed components of the fermion field transform differently under $\mathrm{SU}(2)_L$ and the mass term would spoil gauge invariance.

1.3.2 *Electroweak symmetry breaking*

From experiments, it is well established that the SM fermions and the W- and Z-bosons have mass. But our gauge theories did not allow us to introduce mass terms for them. A model to dynamically generate mass for these particles was proposed by Brout, Englert and Higgs [9, 10, 11]. This is achieved by a spontaneous symmetry breaking of the $\mathrm{SU}(2)_L \otimes \mathrm{U}(1)_Y$ symmetry using a doublet of complex scalar fields for which the ground state does not respect the symmetry:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{34}$$

The isospin doublet has weak hypercharge $Y = 1$, and thus the upper component with $T_3 = 1/2$ has electric charge $+1$ while the $T_3 = -1/2$ component has no electric charge.

The Lagrangian describing this weak isospin doublet contains a kinetic and potential term:

$$\begin{aligned} \mathcal{L}_{\text{scalar}} &= \frac{1}{2}(\partial_\mu \Phi)(\partial^\mu \Phi) - V(\Phi) \\ &= \frac{1}{2}(\partial_\mu \Phi)(\partial^\mu \Phi) - \mu^2 \Phi^2 - \lambda \Phi^4 \end{aligned} \tag{35}$$

The potential $V(\Phi)$ contains a mass term μ^2 and a parameter λ describing the strength of the field's self interaction. In order to ensure a global minimum for the potential term, λ is required to be positive. The mass term μ^2 can take negative or positive values. If μ^2 is positive, the global minimum is at $\Phi = 0$ and the symmetry is respected. But in the case of $\mu^2 < 0$, the potential has no longer a minimum at $\Phi = 0$, which is now a local maximum. Instead we have a degenerate vacuum, given by an infinite number of minima at

$$\Phi_0^2 = \frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2} \tag{36}$$

After the symmetry breaking, the electric charge is still conserved and the photon is still massless. Hence, we have to choose a ground state which is invariant under $U(1)_{em}$:

$$\Phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \tag{37}$$

Once can rewrite Φ using this ground state as

$$\Phi_0 = e^{\frac{i\xi_a t^a}{2}} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{38}$$

using three fields ξ_a which parametrize the degenerate vacuum state. We are now free to choose the unitary gauge with $\xi_a = 0$, which removes these unphysical fields. As a result, these degrees of freedom are added to the weak gauge bosons, which consequently become massive. The remaining field $h(x)$ is the Brout-Englert-Higgs (BEH) field, which has a scalar particle associated to it: the Higgs boson. The kinetic term for the scalar doublet is then given by

$$\begin{aligned} \mid D_\mu \Phi \mid^2 &= \left| \left(\partial_\mu - i g_2 W_\mu^a t^a - i g_1 \frac{Y}{2} B_\mu \right) \Phi \right|^2 \\ &= \frac{1}{2} \left| \begin{pmatrix} \partial_\mu - \frac{i}{2} g_2 W_\mu^0 - \frac{i}{2} g_1 B_\mu & -\frac{i}{2} g_2 \left(W_\mu^1 - i W_\mu^2 \right) \\ -\frac{i}{2} g_2 \left(W_\mu^1 + i W_\mu^2 \right) & \partial_\mu + \frac{i}{2} g_2 W_\mu^0 - \frac{i}{2} g_1 B_\mu \end{pmatrix} \begin{pmatrix} 0 \\ v + h \end{pmatrix} \right|^2 \\ &= \frac{1}{2} \left| \partial_\mu h \right|^2 + \frac{g_2^2 v^2}{8} \left| W_\mu^1 - i W_\mu^2 \right|^2 + \frac{v^2}{8} \left| g_2 W_\mu^0 - g_1 B_\mu \right|^2 + \mathcal{O}(h) \end{aligned} \tag{39}$$

Here we can identify a kinetic term for the BEH field, while the terms of order $\mathcal{O}(h)$ yield the interactions between the Higgs boson and the electroweak gauge bosons. In addition, we find new terms in the Lagrangian which do not depend on the BEH field. Instead they give mass to linear combinations of the W_μ and B_μ bosons. The second term in (39) gives mass to the charged W bosons:

$$\frac{g_2^2 v^2}{8} \left| W_\mu^1 - i W_\mu^2 \right|^2 = \frac{m_W^2}{2} \left[\left(W_\mu^+ \right)^2 + \left(W_\mu^- \right)^2 \right] \tag{40}$$

The third therm in (39) represents the linear combination of W_μ^0 and B_μ to form the massive Z boson:

$$\begin{aligned} Z_\mu &= \frac{1}{g_2^2 + g_1^2} (g_2 W_\mu^0 - g_1 B_\mu) \\ A_\mu &= \frac{1}{g_2^2 + g_1^2} (g_1 W_\mu^0 - g_2 B_\mu) \end{aligned} \tag{41}$$

Comparing with (31), we can rewrite the Weinberg angle as

$$\cos \theta_W = \frac{g}{\sqrt{g_2^2 + g_1^2}} \tag{42}$$

The mass terms for the W- and Z bosons can be easily identified in (39):

$$\begin{aligned} m_W &= \frac{gv}{2} \\ m_Z &= \frac{v}{2}\sqrt{g_2^2 + g_1^2} \end{aligned} \tag{43}$$

The masses of the W- and Z-bosons are therefore related as $m_W = m_Z \cos\theta_W$. Since the $SU(2)_L \otimes U(1)$ symmetry is broken to $U(1)_{\text{em}}$, the weak isospin and weak hypercharge are no longer conserved. It is broken is such a way that the combination corresponding to the electric charge is still conserved.

The same scalar doublet can also be used to generate masses for the fermions through Yukawa couplings. For example, the generation of mass for the down component of a weak isospin doublet is given by

$$\begin{aligned} \mathcal{L}_{\text{d,yukawa}} &= -G_d \left[\begin{pmatrix} \bar{u}_L & \bar{d}_L \end{pmatrix} \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} d_R + \bar{d}_R \begin{pmatrix} \phi^+ & \phi^0 \end{pmatrix} \begin{pmatrix} u_L \\ d_L \end{pmatrix} \right] \\ &= -G_d \left[\bar{d}_L (v+h) d_R + \bar{d}_R (v+h) d_L \right] \\ &= -\frac{G_d}{\sqrt{2}} \left(v\bar{d}d + h\bar{d}d \right) \\ &= -m_d \left(\bar{d}d + \frac{1}{v}\bar{d}d \right) \end{aligned} \tag{44}$$

in which we have the mass m_d and an interaction with the Higgs boson with a coupling proportional to its mass. In a similar way one can generate a mass for the upper component of the doublet using a the conjugate doublet with opposite weak hypercharge [12]:

$$\Phi_c = i\sigma_2 \Phi^* = \begin{pmatrix} -\bar{\phi}^0 \\ \phi^- \end{pmatrix} \xrightarrow{\textit{breaking}} \frac{1}{\sqrt{2}} \begin{pmatrix} v + h(x) \\ 0 \end{pmatrix} \tag{45}$$

1.3.3 *The strong interaction*

Quantum Chromodynamics (QCD) is the gauge theory describing the strong interaction. The theory follows a $SU(3)_c$ symmetry and is therefore described by

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4} G^a_{\mu\nu} G^{\mu\nu a} + \bar{\psi}(i\gamma^\mu D_\mu - m)\psi \tag{46}$$

in which $G^{\mu\nu}$ takes the form of equation 25, ψ are triplets of fermion fields, and the covariant derivative is given by

$$D_\mu = \partial_\mu - i g_3 G^a_\mu \lambda^a \tag{47}$$

with λ^a the 8 generators of $SU(3)$. The $SU(3)_c$ theory therefore leads to eight gauge bosons, the gluons, which interact with the fermion fields taking part in the

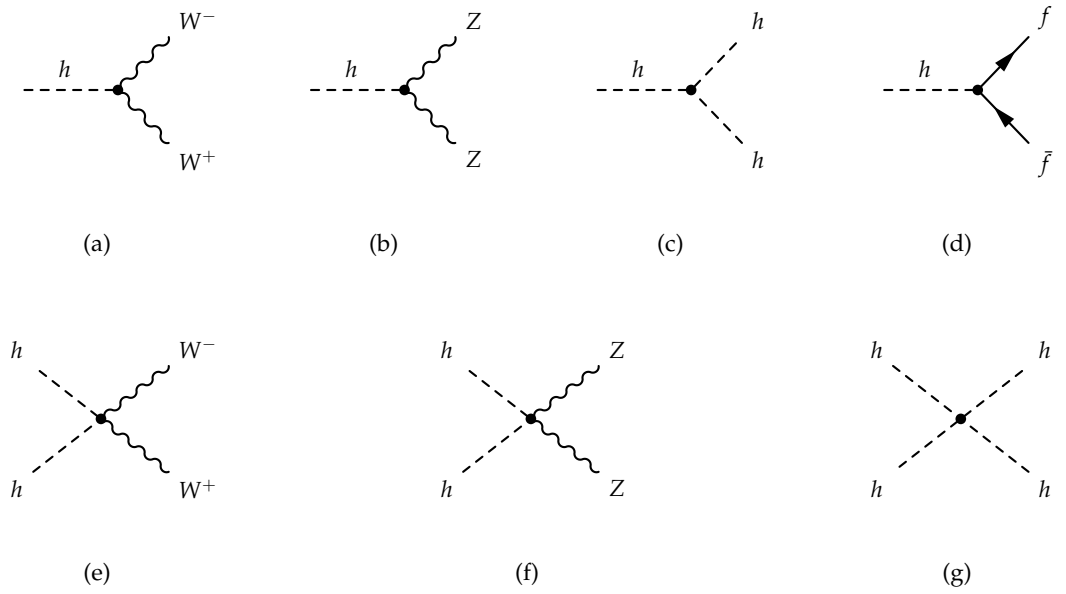


Figure 4: Interaction vertices involving the Higgs boson: the kinetic term of the scalar doublets yield terms involving hWW, hZZ, $hhWW$ and $hhZZ$. The $\lambda\Phi^4$ term in the Lagrangian results in triple and quartic self-interactions of the Higgs boson. When introducing mass terms for the fermions through the BEH field, the interaction between the Higgs boson and the fermion is also added.

QCD interaction, called quarks. Each of the three quark fields in the multiplet is represented by a color charge: green (g), red (r), blue (b), and corresponding anti-colors $\bar{g}$, $\bar{r}$ and $\bar{b}$. The eight gluons carry the combinations of color and anti-color charge.

As explained in section 1.2.3, a non-abelian gauge theory can become asymptotic free if $11N > 2n_f$. In the SM, there are 6 fermion multiplets which take part in the QCD interaction, and this requirement is indeed fulfilled. Figure 5 shows the running of the coupling, which has also consequences for the validity of perturbation theory: long distance or low momentum processes could not be described anymore by the means of perturbation theory since α_s is too big in this region. We can therefore classify QCD processes in two regimes:

- *Hard* scattering processes, characterized by a large momentum transfer between the colliding partons. Those could be accurately described by perturbation theory in which the involved quarks and gluons are treated as free particles.

- The *soft* QCD regime, with interactions at low momentum exchange and long distances for which the strong coupling rapidly increases. This leads to the confinement of quarks into *hadrons*, colorless bound states of quarks. These bound states could be mesons, which are combinations of a quark and an

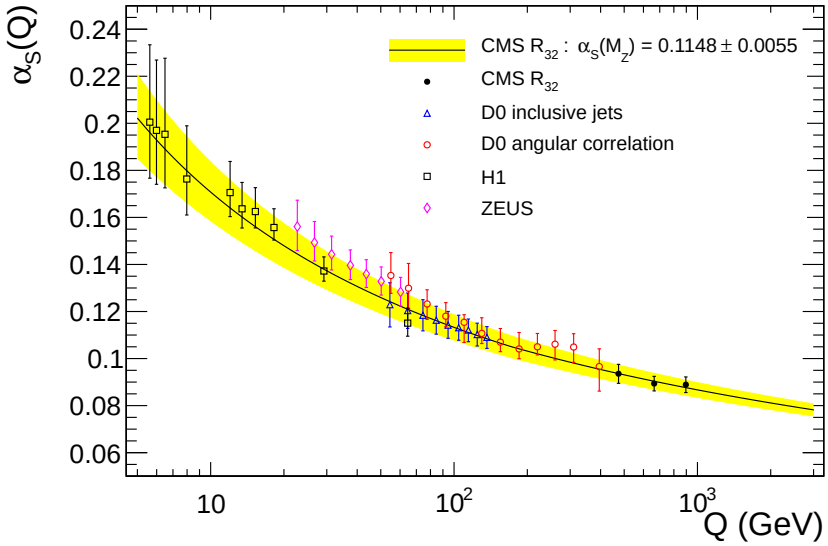


Figure 5: The running of the strong coupling constant α_s (black line) and its total uncertainty, as a function of the transverse momentum scale Q, calculated by evolving the α_s measurement by the CMS experiment and evolved towards lower energies where it is in good agreement with the measurements of other experiments. [13]

anti-quark ($g\bar{g}$, $r\bar{r}$ or $b\bar{b}$), or they could be baryons, in which three quarks or three anti-quarks are grouped together, each of different color (grb or $\bar{g}\bar{r}\bar{b}$).

1.4 PARTICLES OF THE STANDARD MODEL

A simplified sketch of the particles in the SM is shown in figure 6. We have already introduced the gauge bosons when discussing the electroweak and strong interactions: the photon (γ), W^+, W^-, Z and the 8 gluons. The SM also includes a scalar boson, the Higgs boson, which is the particle associated with the BEH field. In the former sections we discussed that fermions could transform as triplets under the strong interaction, doublets under the weak interaction and singlets under the $U(1)_Y$ interaction. The fermions do not have to interact with all the forces, and we can therefore classify them as quarks, which carry color charge, and leptons, which do not experience the strong interaction. As explained in section 1.1.2, each fermion has an anti-particle, which has the same properties but opposite charges. Fermions are found in three generations with increasing mass: for each quark and lepton from the first generation, there is one in the second and third generation which has identical charges, but with a different mass, and therefore different coupling to the BEH field.

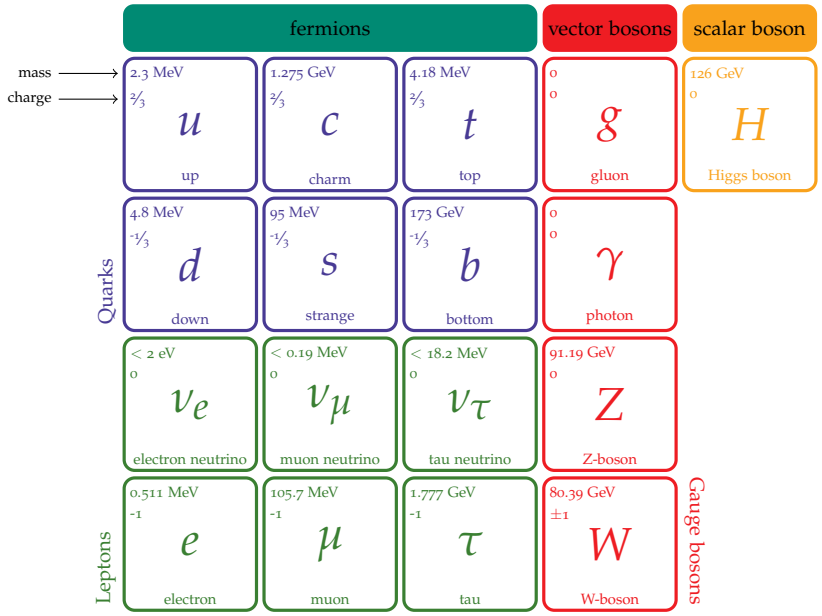


Figure 6: Elementary particles included in the SM. Each fermion has an anti-particle with opposite charges. The quarks come in three different colors (red, green and blue). There are 8 gluons carrying combinations of color and anti-color charge. There are two W-bosons: one postively and one negatively charged.The mass values are taken from [14].

1.4.1 *Leptons*

The three generations of leptons consists of electrons (e), muons (μ), taus (τ) and their associated neutrinos (ν_e, ν_μ and ν_τ). The e,μ and τ have negative electric charge and are able to interact with the photon, as opposed to their neutrinos which have no electric charge. The left-handed components of these particles form $SU(2)_L$ doublets which could transform into each other by radiating or absorbing a W-boson:

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L ; \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L ; \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L$$

The right-handed components of the electron, muon and tau carry no weak isospin charge and do therefore only experience electromagnetism. In addition to the left-handed neutrino, one can expect a hypothetical right-handed neutrino to exist. The right-handed neutrino would not be mixed with the left-handed neutrino as it does not couple to the Higgs field. Because it does also not couple with any of the (known) gauge bosons, it is not interacting with the three forces in the SM.

1.4.2 *Quarks*

Every generation of quarks has two types of quark flavors, an up-type and a down-type, for which the left-handed components again form a doublet for the $SU(2)_L$ interaction:

$$\begin{pmatrix} u \\ d \end{pmatrix}_L ; \begin{pmatrix} c \\ s \end{pmatrix}_L ; \begin{pmatrix} t \\ b \end{pmatrix}_L$$

As was the case for the leptons, the difference between the electric charges of the two flavors in one doublet has to be equal to 1. There are no neutral quarks though: the up-type quarks have charge $2/3$ while the down-type quarks have charge $-1/3$. This means both up and down flavors are interacting with the photon, and are known to have a right-handed component. Because the quarks experience the strong interaction, each quark comes in three colors forming a triplet under $SU(3)_c$.

Baryons and mesons have, due to the fractional charges of their components, always integer charge. Examples of baryons are the proton, comprising two up-quarks and one down-quark, and the neutron, comprising one up-quark and two down-quarks. The most common mesons are the neutral pion ($u\bar{u}$ or $d\bar{d}$) and the charged pions ($u\bar{d}$ and $d\bar{u}$). These valence quarks, which contribute to the quantum numbers of the hadron, are surrounded by virtual gluons and quark-antiquark pairs which are continuously emitted and absorbed by the valence quarks through the strong interaction. Each of the components in the hadron, called *partons*, carries a fraction x of the total momentum of the hadron.

1.5 HADRON COLLISIONS

High energy physics processes could be studied at particle accelerators, in which the particle beams are brought into collision, initiating interactions from the SM or beyond. The studies in this thesis have been carried out on proton-proton collisions provided by the Large Hadron Collider (LHC), which is described in further detail in Chapter 3.

The pp-collisions could be subdivided in elastic and inelastic collisions, for which the later category could be further subdivided in diffractive and non-diffractive collisions:

- *Elastic scatterings* are interactions of the type $p + p \rightarrow p + p$, where only momentum is exchanged between the incoming protons, and no new particles are produced. As the two protons stay intact, elastic scatterings do not involve an exchange of quantum numbers between the protons.

- *Diffractive processes* are similar to elastic scatterings in the sense that they do not exchange quantum numbers between the protons, but here the transferred moment causes one or both of the protons to break up. In addition to these single-diffractive ($p + p \rightarrow p + X$) and double diffractive ($p + p \rightarrow X + X$) processes, there is a third diffractive topology called central diffraction which leaves both protons invariant but causes an excited system between them ($p + p \rightarrow p + X + p$), also known as central exclusive production. Because there is again no exchange of quantum numbers between the protons, central systems or objects are only allowed if they have neutral quantum numbers.

- In the final category, *non-diffractive inelastic* processes, none of the protons survive the collision, and at least one parton of each colliding proton interact with each other. Most of these interactions occur with low (soft) momentum transfer, but sometimes the involved partons carry a large momentum fraction, resulting in a hard scattering event. Hard QCD interactions could be calculated using perturbation theory, which allow to predict the production rates and event properties accurately. These are also the most interesting processes for physics analysis, as the large amount of energy available allows the creation of heavy particles (i.e. Z,H,...). A sketch of a typical hard pp-collision is shown in Figure 7. Initial-state and final-state radiation originates from respectively the ingoing and outgoing partons of the hard scattering. This radiation give rise to parton showers, as described in the next section, before reaching the non-perturbative regime where hadronization sets in. The colored remnants of the protons involve additional radiation and hadronization in the event, forming the Underlying Event (UE). Sometimes, one or more additional hard interactions could occur in the same pp-collision, resulting in a multi-parton interaction.

1.5.1 *Formation of hadronic jets*

When a high-energy gluon or quark is produced in high-energy collisions, it will reduce its energy by emitting additional partons: gluons and quarks can emit a gluon, or a gluon can split into two quarks. The radiated partons are mostly soft, i.e. they carry a small fraction of the initial parton's momentum and are therefore emitted at small angles. These partons can in turn emit other partons, and this avalanche effect creates a parton shower in which the resulting partons are found in a rather narrow cone. This parton shower development stops when the partons reach the non-perturbative regime, at an energy of about 1 GeV, for which the strong coupling constant becomes too large to use perturbation theory. At this point, the colored partons are clustered into colour-singlet hadrons, a process called hadronization. The initial parton coming from the hard scattering will therefore be represented by a collimated spray of energetic hadrons, called a jet.

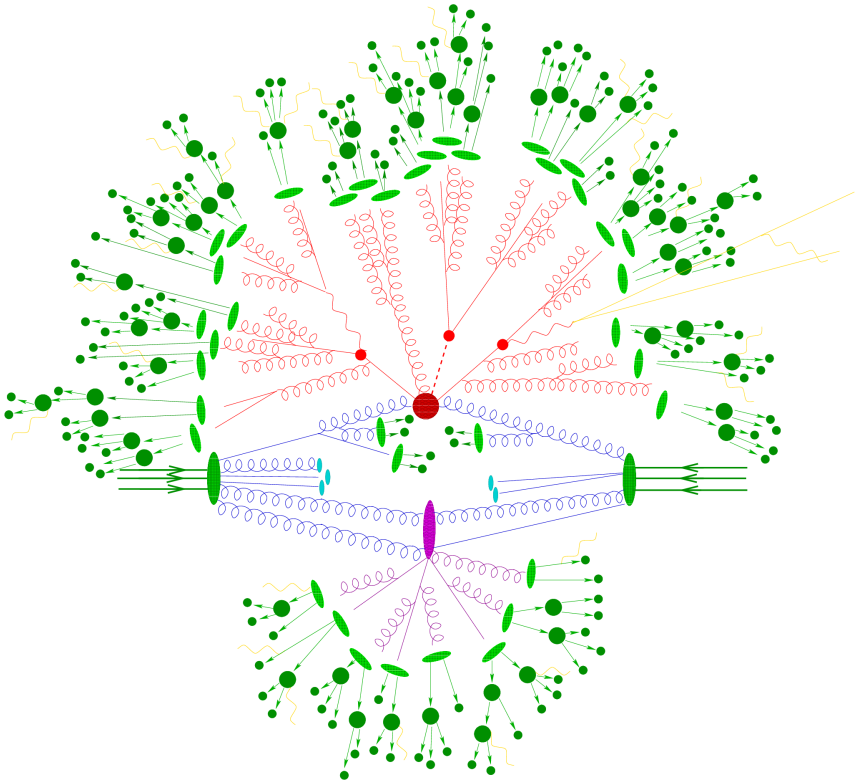


Figure 7: Sketch of a non-diffractive inelastic hadron collision. One parton of each incoming proton takes part in the hard scattering (dark red), while the proton remnants (cyan) form the underlying event, including a multi-parton interaction (violet). Final state radiation on the outgoing partons of the hard scattering (light red) and outgoing partons of the secondary interactions (violet) result in parton showers which hadronize (light green), and unstable hadrons further decay into stable hadrons (dark green). Photon radiation occurs at any stage (yellow). [15]

In high energy particle physics experiments, one needs a jet definition, a set of rules for how to group particles into jets and how to assign a momentum to the resulting jet. Ideally, each jet in the event originates from a parton in the hard event. However, the initial parton may radiate a hard parton, which will have a large angle with respect to the initial parton. It will therefore develop its own parton shower, resulting in an additional jet. On the other hand, if a highly boosted[4] W or Z decays into two partons, their partons will be collimated and could lead to a single jet. The jet definitions used in this work are further described in 5.5. In chapter 6, it is described how the substructure in a jet can be used to discriminate between quark-induced and gluon-induced jets.

1.6 MONTE CARLO SIMULATIONS

In order to compare the experimental observations to the theoretical predictions of the SM, MC event generators [16] are used to simulate the physics processes and the particles therein produced.

1.6.1 *Hard scattering*

The probability that a particular scattering process occurs is given by its cross section σ, which is expressed in barns[5] in nuclear and high-energy physics. The cross section of a hard inelastic scattering process $AB \to X$ can be computed using the *factorization* technique, in which the perturbative description of a hard scattering parton-parton interaction is separated from the non-perturbative contribution (i.e. the confinement in the hadrons) to the process:

$$\begin{aligned}\sigma &= \sum_{a,b} \int_0^1 \mathrm{d}x_a \mathrm{d}x_b \int f_a^A(x_a, \mu_F) f_b^B(x_b, \mu_F) \mathrm{d}\hat{\sigma}_{ab\to n}(\mu_F, \mu_R) \\ &= \sum_{a,b} \int_0^1 \mathrm{d}x_a \mathrm{d}x_b \int \mathrm{d}\Phi_n f_a^A(x_a, \mu_F) f_b^B(x_b, \mu_F) \\ &\quad \times \frac{1}{2x_a x_b s} \mid \mathcal{M}_{ab\to n} \mid^2 (\Phi_n, \mu_F, \mu_R) \end{aligned} \tag{48}$$

where

- f_a^A are the parton distribution functions, describing the probability to find a parton of type a (e.g. a gluon or a specific quark flavor) carrying a momentum fraction x in a hadron A. The f_a^A are in the non-perturbative regime, and

4 An object is boosted when it is produced with high kinetic energy with respect to its rest mass, causing its decay products to travel further in the same direction.

5 $1\mathrm{b} = 10^{-24}\mathrm{cm}^2$

could therefore not be calculated by the means of perturbation theory. However, they are independent of the process under consideration and hence universal to all high energy physics experiments. Hence, the f_a^A are provided by several experimental collaborations, based on data from previous and current experiments. The f_a^A are also dependent on the momentum scale at which the hadron is probed: the value of the coupling constant α_s affects gluon emission and gluon splitting processes, which in turn affect the quark and gluon parton distributions. A factorization scale μ_F is introduced at which the f_a^A are extracted. Emissions with transverse momenta below μ_F are absorbed into the f_a^A, while emissions at higher transverse momenta are supposed to be calculated perturbatively. Fortunately it is possible to describe the dependence of these splitting processes as a function of μ_F, resulting in the DGLAP evolution equations [17, 18, 19] which allow to extrapolate the parton distribution functions from one scale to another.

- $\hat{\sigma}_{ab\to n}$ is the parton-level cross section of the interaction of initial state partons a and b, producing the final state n. Because only the two incoming partons are involved in the hard scattering, the center-of-mass energy[6] $\sqrt{s}$ of the pp-interaction gets reduced to a momentum transfer $Q = \sqrt{x_a x_b s}$. It depends on the momenta given by the final state phase space Φ_n and on the factorization and renormalization scales. It could be reduced to the formulation of the Matrix Element (ME) $\mathcal{M}_{ab\to X}$, which is a sum over the Feynman diagrams contributing to the scattering amplitude. Note that the ME enters the cross section formula as its square. As a result, the total cross section of different processes leading to the same final state is not simply the sum of the cross sections of the individual contributing processes, and positive or negative interference effects should be taken into account. Once the ME is calculated, MC techniques are used to sample the phase space Φ_n and obtain a set of generated events.

The sums over parton types a and b in equation 48 reflects how same a final state X can be created by different combinations of incoming parton types, hence one needs to include all those contributions, and integrate over their possible momentum fractions x_a and x_b.

1.6.2 *Parton showering and hadronization*

The ingoing and outgoing partons of the hard scattering process will further emit initial- and final-state radiation, creating parton showers. Because we cannot calculate ME up to arbitrarily order, a parton shower algorithm is used to approximate

6 The center-of-mass frame is defined as the frame where the total momentum vanishes, i.e. $\sum \vec{p}_i = 0$, and therefore defines the available energy in a particle collision. The center-of-mass energy can be calculated in any frame as $E_{CM} = \sqrt{s} = \sqrt{(p_1 + p_2)}$ with p_1 and p_2 the four momenta of the colliding protons.

these higher-order real emissions, evolving from the high momentum scale of the hard process down to the soft scale, of order 1 GeV, where non-perturbative confinement effects set in. Parton shower algorithms are based on the calculation of the *Sudakov form factor* [20], describing the probability for a parton that it does not split into other partons when evolving from one scale to another smaller scale. By using this formalism iteratively, one can generate a sequence of parton splittings, simulating final-state parton showers by evolving towards lower scales. The initial-state radiation is simulated by evolving backwards towards a higher-x parton in the parton distribution function. Parton showers are a good approximation in the soft and collinear limit, but they are not well suited to describe hard partons emitted at large angles.

When the parton shower reaches the non-perturbative regime, the partons need to be confined into hadrons. Two phenomenological models are used to simulate how the partons turn into hadrons:

- The Lund string model, implemented in PYTHIA, is based on the linear confinement of partons. The QCD potential between two quarks forming a color singlet can be approximated as $V(r) = \kappa r$ for large distances, which can be interpreted as a color string connecting the two quarks. When these two quarks q and $\bar{q}$ move apart from their production vertex the potential will rapidly grow. As a result it will become energetically favourable to create a new $q'\bar{q}'$ pair between those partons, screening the color charge of the initial partons. The initial color singlet is now replaced by two independent color singlets $q\bar{q}'$ and $q'\bar{q}$. This process repeats until the invariant masses of the color pairs are of the order of a hadron mass.

- The cluster model, used by HERWIG++, is based on the preconfinement of parton showers. In the final stage of the parton shower, the gluons are forced into $q\bar{q}$ pairs, and the quarks are clustered into colorless groups. The clusters, often having high invariant mass, are further decayed to smaller mass scales suitable to form hadrons.

1.6.3 *Description of MC generator programs*

Some of the most used Leading Order (LO) generators are PYTHIA [21, 22] and HERWIG++ [23, 24], which are multi-purpose generators who do not only calculate the hard process, but also provide the parton showering, QED radiation, hadron decays, and the contribution from the UE. They have the advantage of describing the complete final state of the event at hadron level, but provide only a limited set of ME processes. The non-perturbative nature of the UE implies the need to describe it using phenomenological models. These models are based on different kinds of assumptions and have free parameters which need to be tuned to data. Throughout this thesis, PYTHIA is the most used multi-purpose generator and uses

tune Z2 for the analysis done on pp collisions with $\sqrt{s} = 7$ TeV and tune Z2* for the analysis at $\sqrt{s} = 8$ TeV [25].

Specific ME generators exist which provide more flexibility for the ME calculations, such as generating additional hard emissions on the ME level, which are typically underestimated by the parton shower algorithms. An example of such a ME is MADGRAPH [26], which can then easily be interfaced to PYTHIA which further treats the event with the parton showering, UE and hadronization. When a ME generator is interfaced to the parton shower, a new difficulty arises. The emission of an additional parton can now be obtained in two ways: either the ME included the additional parton emission, or it could be emitted by the parton shower. To avoid this double counting, a proper matching between the ME and parton shower is needed. The MADGRAPH samples used in this thesis are matched to PYTHIA using the MLM procedure [27, 28].

Some ME generators like POWHEG [29, 30, 31, 32] provide a description of the ME at Next-to-leading Order (NLO) level, which means it adds one loop contributions to the process. This allows for a more precise cross section estimate and a better control of the uncertainties. Other programs which provide NLO calculations are MCFM [33] and VBFNLO [34, 35, 36].

2

ELECTROWEAK PRODUCTION OF A Z BOSON IN ASSOCIATION WITH TWO JETS

2.1 SIGNAL PROCESS

2.1.1 *Vector boson fusion*

In $p\,p$ collisions, Vector Boson Fusion (VBF) happens when two electroweak[1] vector bosons, radiated from a quark in each of the two colliding protons, interact with each other. The two quarks are typically scattered away from the beam axis and inside the detector acceptance where they reveal as hadronic jets. The distinctive feature of VBF processes is therefore the presence of two energetic jets, often referred to as the tagging jets, which are found at small angles with respect to the proton beam axis on opposite sides of the detector. As a result, VBF processes are characterized with a large pseudorapidity[2] separation $\Delta\eta_{jj}$ between the two tagging jets, as well as a large dijet invariant mass[3] M_{jj}. A VBF process does not involve QCD interactions, hence no color flow is exchanged between the two quarks. Therefore, the hadronization process will mainly occur between the tagging jets and their proton remnants while few hadronic activity is expected in the central region of the detector. Similar to diffractive processes, this leads to a rapidity-gap [37], a region with no hadrons except those produced by the central system. It is useful to express the rapidity of the objects with respect to the rapidity center of the tagging jets, i.e. by using the Zeppenfeld variable [38]

$$y^* = y - \frac{y_{j_1} + y_{j_2}}{2} \tag{49}$$

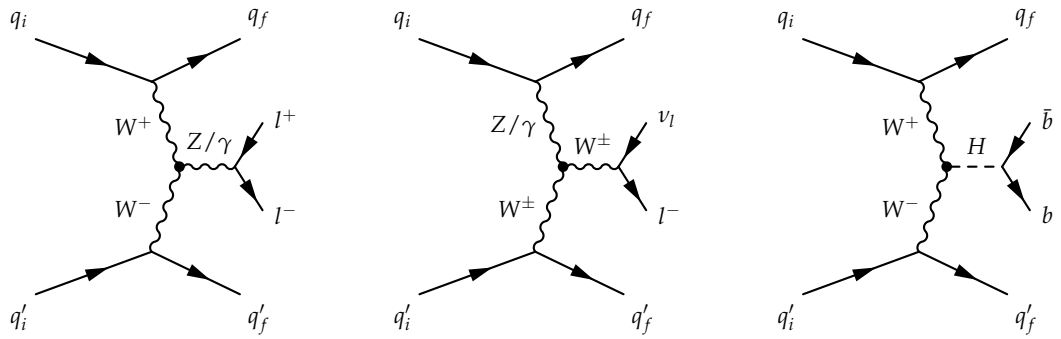
or its value divided by the rapidity separation:

$$z^* = \frac{y^*}{\Delta y_{jj}} = \frac{1}{2}\frac{2y - (y_{j_1} + y_{j_2})}{y_{j_1} - y_{j_2}} \tag{50}$$

1 Even though gluons are also vector boson particles, VBF refers only to the fusion of two electroweak bosons, i.e. involving W^+,W^-,Z and γ.

2 As explained in 4.1.1, the pseudorapidity η and rapidity y are coordinates measuring the angle of a particle or a jet with respect to the transverse plane on the beamline, where the two sides of the beamline correspond to plus and minus infinity. If both jets end up close to the beamline in opposite directions, a large pseudorapidity separation can be expected.

3 The invariant mass of a system is the mass in the rest frame of the system in consideration. It can be calculated in any frame as $M^2 = (\sum E)^2 - (\sum \vec{p})^2$, which becomes large for a system of two energetic, opposite oriented jets.



(a) $WW \to Z$ (b) $WZ \to W$ (c) $WW \to H$

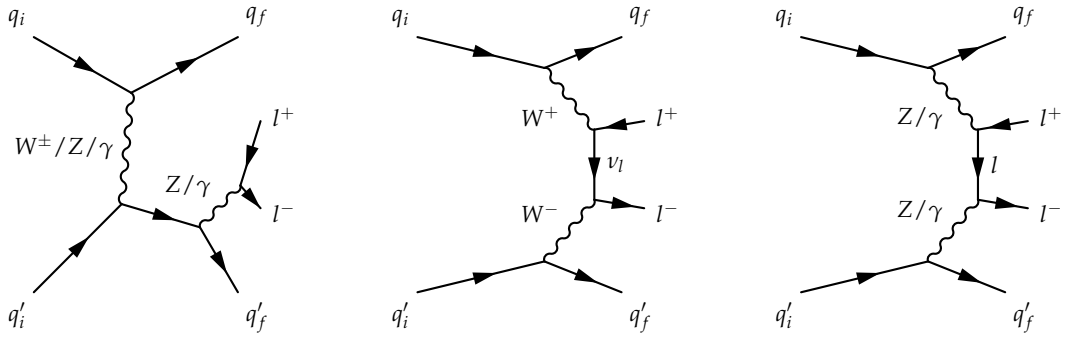
Figure 8: Vector boson fusion processes in the SM: two weak bosons fusing into (a) a Z-boson or photon, (b) a W-boson, (c) a Higgs-boson.

Using this reference, objects found at the same rapidity as one of the tagging jets have $z^* = \pm 0.5$. The additional hadronic activity is supposed to be suppressed in the central region, and would be found mostly at $|z^*| > 0.5$. On the other hand, the central objects produced by the VBF system are expected to be found at $|z^*| < 0.5$.

Different interactions are possible between the colliding electroweak bosons, and a wide range of VBF processes can be defined, all sharing the same typical VBF dijet signature described above. In this thesis, we will focus on the production of a Z-boson through the VBF process, e.g. $W^+W^- \to Z/\gamma$, in which the massive Z subsequently decays to a dielectron or dimuon pair. Even though only a small fraction[4] of the Z-bosons decays through electrons and muons, those objects provide a clear signature in the detector, which allows for an easy reconstruction of the dilepton invariant mass. By selecting a mass window around the nominal Z-boson mass, one can select a region in the dilepton invariant mass spectrum where the Z-resonance dominates over the contribution of the photon, and other dilepton resonances are avoided.

The clean signature of the VBF-Z process makes it ideally suited to establish the VBF signature and probe the rapidity gaps in those events [38, 39] which can be used to study the impact of a Central Jet Veto (CJV) to select VBF processes [40]. The VBF-Z process has also been suggested as a way to probe anomalous triple gauge couplings [41], i.e. beyond the SM contributions to the WWZ vertex. The techniques and knowledge we learn from studying the VBF-Z process, pave the way for the study of other processes with the VBF topology. One such process is the VBF Higgs production, which could benefit from cross-checks and validations against the VBF-Z process [42]. Hence, some similarities (for example the application of a quark-gluon discriminator tool) can be found between this analysis and the search for the Higgs boson produced through VBF [43, 44]. The VBF-Z topology can also function as an important production mechanism in supersymmetry models [45, 46], where

4 A Z-boson couples to all charged and/or left-handed fermions with mass below $m_Z/2$, with a factor relative to the square of their coupling $T_3 - Q \sin\theta_W$. From this it can be derived the dielectron and dimuon channels account each for about 3.36% of the Z-boson decays.



(a) Z-bremsstrahlung (b) multi-peripheral (c) multi-peripheral

Figure 9: Other contributions to the pure electroweak production of $lljj$

a supersymmetric fermion pair is produced by the decay of the Z, or through a multi-peripheral mechanism similar to the one described in the next section.

2.1.2 *Electroweak production of the $lljj$ final state*

The production of a Z-boson through VBF is a purely electroweak process of order $\mathcal{O}(\alpha_{EW}^4)$ which results into a final state of two leptons and two jets. However, there are other processes of $\mathcal{O}(\alpha_{EW}^4)$ leading to the same final $lljj$ state. These processes, shown in Figure 9, share much of the characteristics of the VBF process. Moreover, large destructive interference terms occur between the different $\mathcal{O}(\alpha_{EW}^4)$ contributions, making the combined cross section much lower compared to what would be the VBF production cross section if the other processes did not occur in nature. It is therefore impossible to isolate and define a pure VBF signal. Instead, we have to take all contributions to the pure electroweak production of $lljj$ into account, defining EW $lljj$ as our signal process.

In addition to the VBF-Z process, the other pure EW contributions to $lljj$ are given by

- The Z-strahlung process, where an additional Z-boson is radiated off the incoming or outgoing quarks of a process in which another weak vector boson is exchanged between the quarks. The ME contribution of this process has opposite sign with respect to the VBF process, resulting in the destructive interference terms.

- Multi-peripheral processes, which are non-resonant productions of $lljj$ where two weak vector bosons, one of each incoming quark, transform into leptons by exchanging a lepton between them. Even though the selection on the dilepton invariant mass makes their contribution relatively small compared to the diagrams involving a Z-boson, multi-peripheral processes are still important

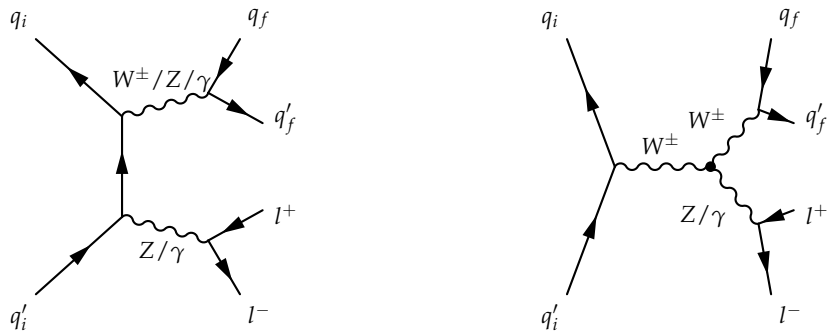


Figure 10: Example Feynman diagrams of diboson production resulting in the $lljj$ final state

as they share all of the characteristics of VBF processes. Those processes will therefore still contribute when very pure VBF selections are applied (for example a very high dijet invariant mass cut).

- Also the diboson pair production VZ with $V \to jj$ and $Z \to ll$, shown in Figure 10, is a pure EW process of $\mathcal{O}(\alpha_{EW}^4)$. However, the two jets are result of the decay of a weak vector boson, and will therefore be produced close to each other. The very distinctive kinematical properties from the diboson process will make it possible to treat it as a background, which will easily be reduced when requiring a large dijet invariant mass.

This thesis presents the measurement of EW Zjj process[5] by the CMS experiment using pp collisions at $\sqrt{s} = 7$ TeV [47] and $\sqrt{s} = 8$ TeV [48]. The same measurement has also been carried out by the ATLAS collaboration [49]. In a very similar way, one can do a measurement of the EW Wjj production, and a first measurement by CMS became available recently [50]. Both collaborations have also studied Vector Boson Scattering (VBS) [51, 52], through the EW production of same-sign $WWjj$, using $\sqrt{s} = 8$ TeV collisions, applying very similar analysis strategies as for the EW Zjj measurements.

2.1.3 *Simulation of signal process*

The EW Zjj process is simulated at LO using MADGRAPH 1.5 interfaced with PYTHIA 6 for parton showering and hadronization. The parton distribution functions used for the MADGRAPH event generation are provided by the *CTEQ6L1* collaboration [53], and the factorization and renormalization scales are both fixed to the Z-boson

5 As in the title of this thesis, we often denote our signal process by its dominant contribution EW Zjj instead of the more correct EW $lljj$.

mass ($\mu_R = \mu_F = m_Z$). The requirement $M_{jj} > 120$ GeV applied at the parton level reduces the contribution of diboson processes to negligible level.

A NLO calculation [54] of the EW Zjj cross section has been implemented in the VBFNLO program, which uses the $CT10$ [55] parton distribution function with $\mu_F = \mu_R = m_Z$. The cross sections as predicted by MADGRAPH and VBFNLO are given in Table 1, together with the kinematic region for which they are calculated.

	7 TeV	8 TeV
dijet invariant mass	$M_{jj} > 120$ GeV	
dilepton invariant mass	$M_{ll} > 50$ GeV	
transverse momentum of the tagging jets	$p_{Tj} > 25$ GeV	
pseudorapidity of the tagging jets	$\mid \eta_j \mid < 4$	$\mid \eta_j \mid < 5$
σ_{LO}(EW $lljj$) (MADGRAPH)	162 fb	208 fb
σ_{LO}(EW $lljj$) (VBFNLO)	157 fb	213 fb
σ_{NLO}(EW $lljj$) (VBFNLO)	166 fb	219 fb

Table 1: Defined kinematic region (at parton level) for the signal and resulting cross section per lepton flavor

2.2 BACKGROUND PROCESSES

2.2.1 *Drell-Yan background*

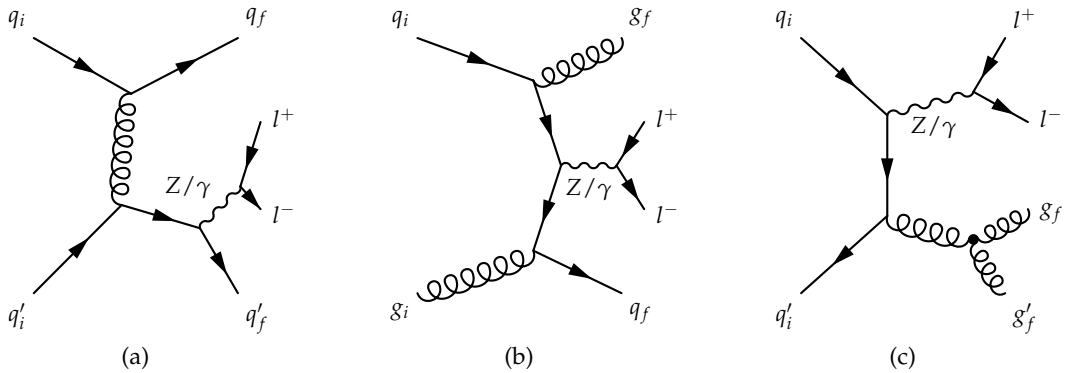


Figure 11: A sample of example diagrams which contribute to Drell-Yan plus two jets production

The production of two leptons in association with two jets in pp-collisions is dominated by Drell-Yan (DY) plus two jets processes, of which some example diagrams are shown in Figure 11 These proceed through a mixture of EW and strong processes of order $\mathcal{O}(\alpha_{EW}^2 \alpha_{QCD}^2)$, and occur much more often than pure electroweak processes because of the higher coupling constant for QCD processes. Because DY

processes share the same final state of two jets and two leptons which are decay products from a Z boson, it is the most difficult background to deal with. Fortunately, DY jets tend to be more central and less energetic, and often the jets originate from gluons instead of quarks. This will give us the possibility to exploit the VBF characteristics to separate our signal from this huge DY background.

The DY background is again generated at ME element using MADGRAPH 1.5 which includes up to four partons generated from QCD. The MADGRAPH events are interfaced with PYTHIA 6 for parton showering and hadronization, using the MLM prescription for the matching between ME and parton shower. In addition to the LO description by MADGRAPH, the MCFM program is used for a NLO description of the DY process at parton level. It uses the dynamic scale $\mu_0 = \sum_{i=1}^{n} p_T^i$ with $n = 4, 5$ final state partons is used with the factorization and renormalization scales set equal, $\mu_F = \mu_R = \mu_0$. The theoretical cross section of the DY process with $M_{ll} > 50$ GeV is computed at the Next-to-next-to-leading Order (NNLO) using FEWZ [56], which is used as an overall normalization factor to the simulated events.

Some of the DY diagrams share the same initial- and final-state particles and quantum numbers with diagrams in our pure EW signal process, and could therefore induce chromo-electroweak interferences [57]. The chromo-electroweak interference effects are found to be rather small, and the 7 TeV analysis has therefore made the assumption those could be neglected and treated the signal and DY background as completely independent processes. For the 8 TeV analysis, the interference is evaluated using a MADGRAPH 1.5 prediction in which three separate samples were generated: one of pure signal, one of pure background and one including both contributions. Figure 12 shows the differential cross section of the three samples as a function of the dijet invariant mass at parton-level. Because of the overwhelming background of DY Zjj at the lower M_{jj} values, the interference term could have a strong impact on the EW Zjj cross section measurement. Towards larger M_{jj} values, the interference term becomes negligible. An empirical function is fitted to this MADGRAPH prediction, in order to estimate the interference as a function of M_{jj}:

$$\frac{\sigma_{\text{EW}} + \sigma_{\text{int}}}{\sigma_{\text{EW}}} = 12.7733 + \frac{1773.74}{M_{jj}} - \frac{151127}{M_{jj}^2} + \frac{4049780}{M_{jj}^3} - 0.00044359 \cdot M_{jj} - \frac{88.2666}{\log M_{jj}} \tag{51}$$

This parametrization will be used to estimate the contribution of the interference to the measurement of σ_{EW}.

2.2.2 *Diboson backgrounds*

As stated above, diboson production of WZ and ZZ are considered as backgrounds even though they result in the same final state $lljj$ and proceed through purely

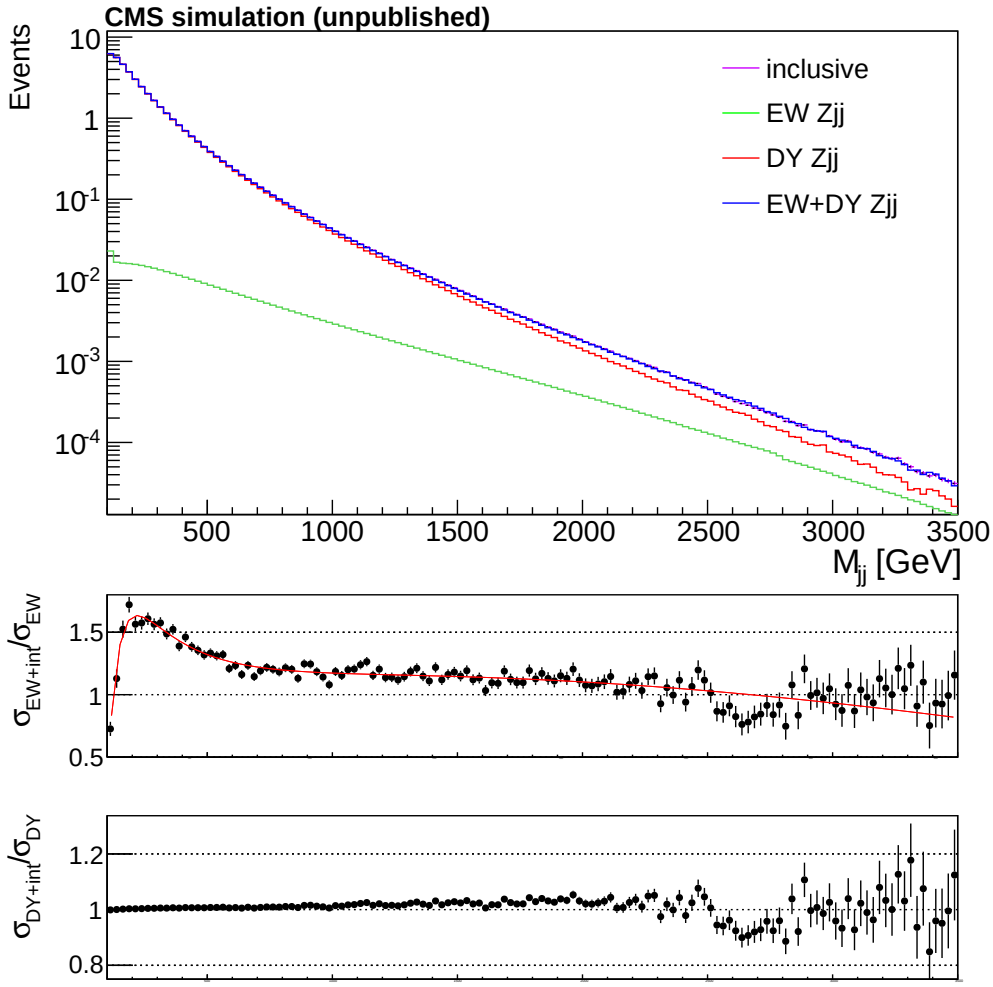


Figure 12: Dijet invariant mass distribution at parton-level for the exclusive EW Zjj and DY Zjj processes and for the inclusive case. The simple sum of EW Zjj and DY Zjj processes is shown for comparison. The bottom panels compare the effect of the interference between DY Zjj and EW Zjj with the exclusive EW Zjj (center) and DY Zjj (bottom) processes.

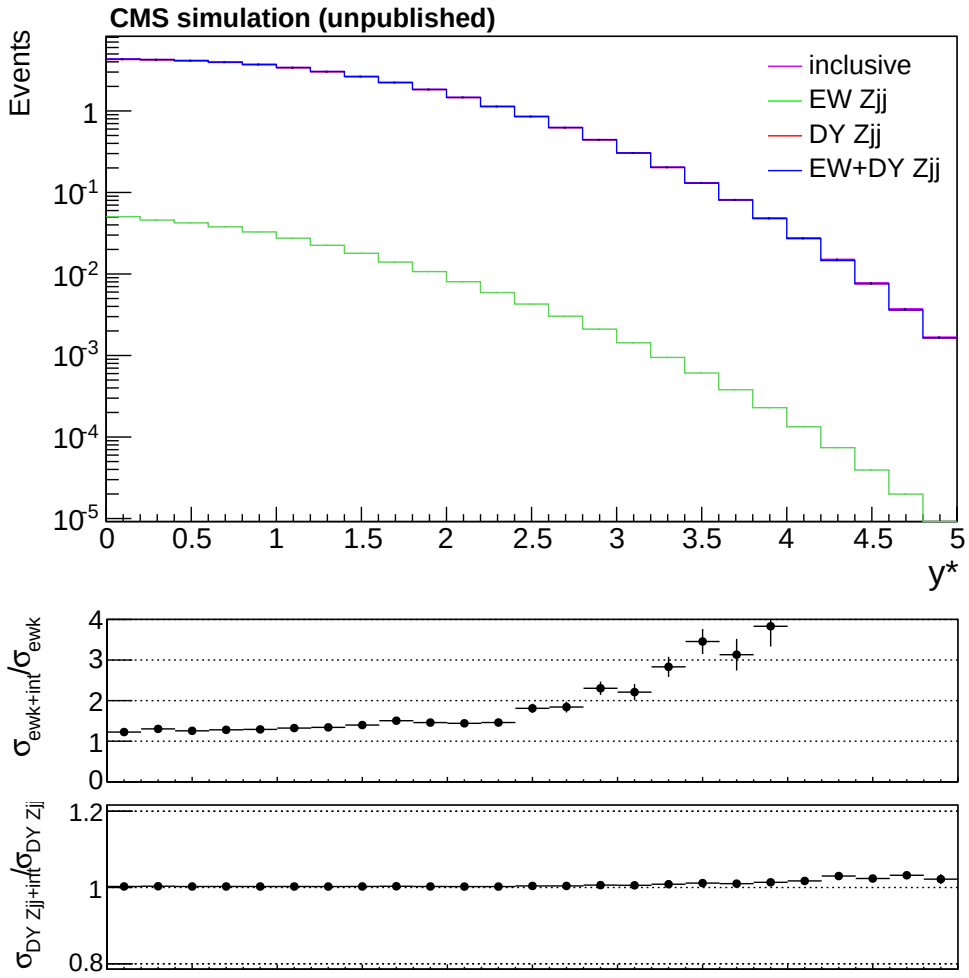


Figure 13: The Zeppenfeld variable for the Z-boson, $|y^*_Z|$, at parton-level for the exclusive EW Zjj and QCD Zjj processes and for the inclusive case. The simple sum of EW Zjj and QCD Zjj processes is shown for comparison. The bottom panels compare the effect of the interference between QCD Zjj and EW Zjj with the exclusive EW Zjj (center) and QCD Zjj (bottom) processes.

electroweak interactions. The selection of events with a large dijet invariant mass will, however, reduce these backgrounds to negligible levels. The full generation of diboson events are provided by PYTHIA 6. The events are normalized to a NLO cross section derived using MCFM and .

2.2.3 *Ditop background*

Another background is caused by $t\bar{t}$ events, where each of the top quarks decays into a b-quark and a W-boson, and the W-bosons subsequently decay leptonically. This results in $ll\nu_l\nu_l jj$, from which the neutrinos do not interact with the detector and hence we are left with a $lljj$ signature in the detector. The ditop background is simulated using MADGRAPH including up to three extra partons, with the top decaying in both the leptonic and hadronic decay channels, and is interfaced with PYTHIA using the MLM prescription. It is normalized to a NNLO cross section computed by TOP++ [58].

2.2.4 *Residual backgrounds*

For completeness, a few residual backgrounds are added. Even though they do not result in the final state, some events could pass our selections due to additional radiations or by faking other objects in the detector (e.g. a jet could be wrongly identified as an electron). These other backgrounds include a W+jets, single top and QCD multijet processes. The complete set of used MC samples and their cross sections is given in Table 5 and 7.

3

THE LARGE HADRON COLLIDER

The proton-proton collisions on which the analyses in these thesis are carried out, were provided by the LHC [59], a particle collider located at CERN, inside the 26.7 km tunnel which formerly hosted the Large Electron-Positron Collider (LEP) collider. It accelerates and circulates bunches of protons in two counter rotating beams, which are brought into collisions in four interaction points. The LHC is the world's most powerful accelerator capable of accelerating the protons to an energy of up to 7 TeV, leading to a center-of-mass energy of $\sqrt{s} = 14$ TeV. However, the data for this thesis was collected during Run I in which the LHC was operating at $\sqrt{s} = 7$ and 8 TeV.

3.1 LAYOUT AND DESIGN

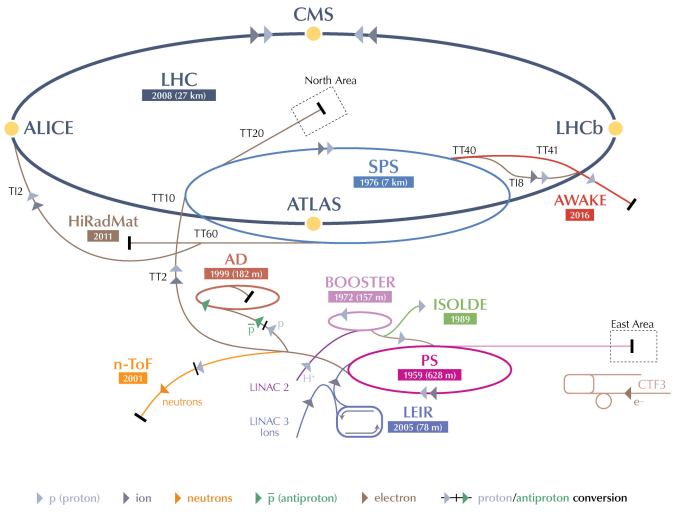


Figure 14: Schematic overview of the CERN accelerator complex [60]

Before the bunches are injected in the LHC, they have to pass through a series of pre-accelerators, shown in Figure 14. The accelerator chain starts with a hydrogen gas source which is passed through an electric field in order to ionize the hydrogen atoms, leaving only the protons to enter a linear accelerator (LINAC 2, to be replaced by LINAC 4 in 2017) where they are accelerated to 50 MeV. The protons are subsequently chained to the Booster, Proton Synchrotron (PS) and Super Proton Synchrotron (SPS) which are circular accelerators used to increase the energy

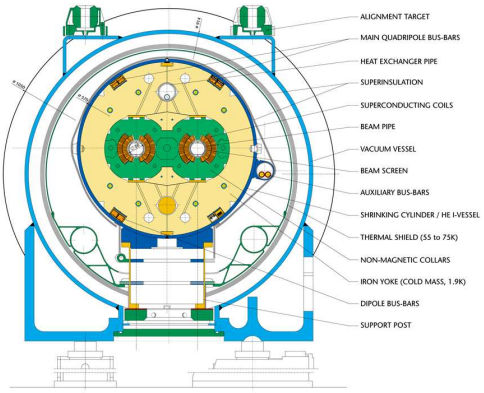


Figure 15: Schematic cross section of a superconducting LHC dipole magnet [61]

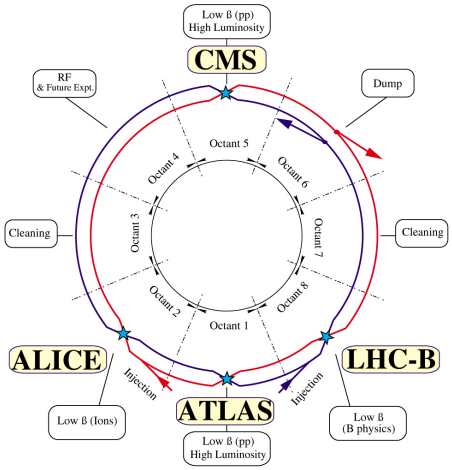


Figure 16: The four interaction points at the LHC [62]

to 1.4, 28 and 450 GeV respectively. Finally, the protons are injected in the two vacuum beam pipes of the LHC, in which they are further accelerated to the operating beam energy. The protons are accelerated using oscillating electric fields in the Radio Frequency (RF) cavities. Because the electric field oscillates, the beam cannot be a continuous stream of protons, and protons travelling out of phase are decelerated, so that the protons become organized in bunches. The circular accelerators use dipole magnets to bend the beam with their magnetic field. A schematic cross section of a LHC dipole magnet is shown in figure 15. Quadrupole and higher order magnets are used to focus the proton beams. Once the operational energy is reached, the beams are brought in collision. As only a few protons in each bunch collide, the bunches will keep circulating and cross again at the next interaction point. The intensity of the beams, and hence the rate of collisions, will gradually decrease. After a run of several hours, the beams are dumped by redirecting them into absorber material. The whole process is then restarted in order to create new proton bunches with full intensity.

The LHC can also operate with lead-ion ($^{208}Pb^{82+}$) beams. The lead-ions are first accelerated by LINAC 3 and transformed to short dense bunches suitable for the LHC by the Low Energy Ion Ring (LEIR), before moving to the PS after which they follow the usual accelerator chain to the LHC. The lead-ions are accelerated up to energies of 2.76 TeV/nucleon, resulting in heavy ion collisions with $\sqrt{s} = 1.15$ PeV.

3.2 LHC EXPERIMENTS

There are four interaction regions in which proton beams cross each other and are brought into collision. These 4 interaction points, shown in Figure 16, provide collisions to seven experiments:

- ALICE (A Large Ion Collider Experiment) [63] is a dedicated heavy-ion detector aiming to study the physics of strongly interacting matter at very high energy densities, where a phase of matter called the quark-gluon plasma forms. It is therefore designed to cope with the very large multiplicities of particles in these events.

- ATLAS (A Toroidal LHC Apparatus) [64] is one of the two general-purpose experiments at the LHC. The general-purpose experiments are used in both pp and ion collision modes, and study a wide range of topics including SM precision measurements and searches for new physics.

- CMS (Compact Muon Solenoid) [65] is the other general-purpose experiment, discussed in detail in Chapter 4.

- LHCb (LHC Beauty) [66] aims to study the rare decays of beauty and charm hadrons as well as looking for new physics through precise measurement of CP violating processes. Because b-hadrons tend to decay at forward angles, LHCb is designed as a single arm forward spectrometer.

- LHCf (LHC Forward) [67] consists of two small calorimeters, each one placed 140m away from the ATLAS interaction point. Its purpose is to study forward production of neutral particles in pp collisions at extremely low angles. The results will be used to verify hadronic shower models used in the study of ultra-high energy cosmic rays.

- MoEDAL (Monopole and Exotics Detector at the LHC) [68], is a small experiment searching for magnetic monopoles and new physics with highly-ionizing particle signatures. It also searches for massive stable slow-moving particles which appear in theories beyond the SM. It is placed around the same interaction point used by LHCb.

- TOTEM (Total Elastic and diffractive cross section Measurement) [69] consists of a set of small detectors placed within 220m around the CMS interaction point. Its goal is to measure the total pp cross section as well as studying the proton structure by looking at elastic scattering.

3.3 LUMINOSITY

The probability amplitude of a process is given by its cross section σ and allows us to predict the number of events per unit time generated by a process:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = L\sigma \tag{52}$$

where L the machine luminosity. Integrating the luminosity with respect to time yield the integrated luminosity

$$\mathcal{L} = \int L\mathrm{d}t \tag{53}$$

which allows us to calculate the predicted number of events $N = \sigma\mathcal{L}$ of a given process in a dataset.

The luminosity is proportional to the revolution frequency f of a proton around the LHC ring, the number of bunches k_b, and the square of the number of protons N_p contained in each bunch. Assuming identical a Gaussian beam distribution, the luminosity is given by:

$$L = \frac{f k_B N_p^2}{4\pi\sigma_x\sigma_y} F \tag{54}$$

$$= \frac{f k_B N_p^2 \gamma_r}{4\pi\epsilon_n\beta^*} F \tag{55}$$

where σ_x and σ_y characterize the Gaussian transverse beam profiles in the horizontal and vertical directions, and F is a geometric reduction factor (≤ 1) due to the crossing angle at the interaction point. The transverse beam size can be expressed as $\sigma_x\sigma_y = \frac{\epsilon_n\beta^*}{\gamma_r}$, in which γ_r is the Lorentz factor, ϵ_n is the normalized transverse emittance and β^* which gives the value of the β function at the collision point. The β function is also known as the amplitude function, and describes the envelope around all the particles in the beam. By focusing the beams at the collision point, the β^* value is minimized and a higher luminosity is obtained.

3.4 PILE-UP INTERACTIONS

The LHC is designed to deliver a very high peak luminosity $L = 10^{34}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$. This has the advantage to improve the rate of collisions, so that rare processes can be produced more abundantly and studied. The high luminosity comes at the price of having multiple proton-proton interactions in the same bunch crossing. This effect is called pile-up and an hard interaction event is therefore accompanied with soft additional interactions, which cause extra tracks and energy deposits in the detector. Pile-up events originate from their own interaction vertex, and a good vertex reconstruction is therefore important to distinguish particles originating from the

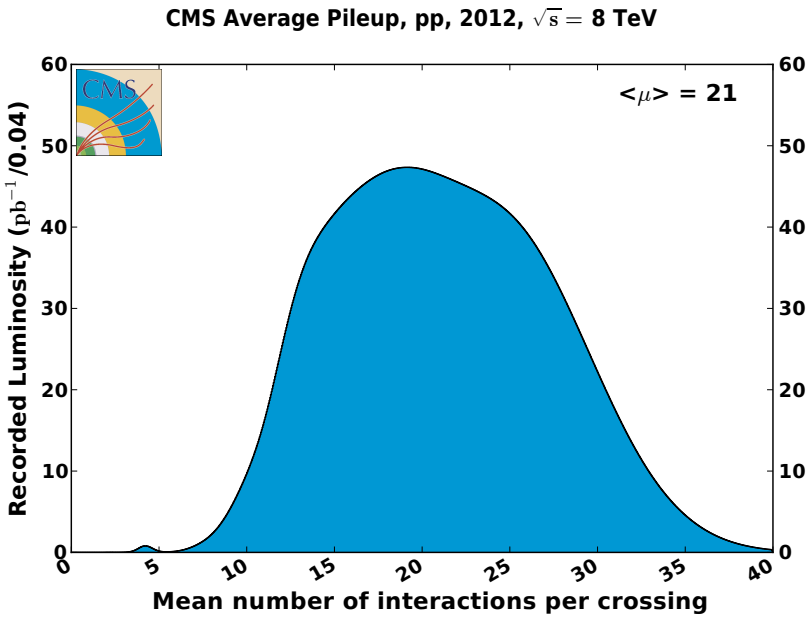


Figure 17: Mean number of interactions per bunch crossing at CMS, as measured during the $\sqrt{s} = 8$ TeV run in 2012

hard interaction from those originating in pile-up interactions. The pile-up distribution in a dataset can be calculated from the measured bunch-by-bunch luminosity. For MC simulations, the simulation of the sample of interest is overlaid with events from minimum-bias simulation (i.e. a simulation describing both soft and hard interactions). This happens after the detector simulation (as described in 4.8) where the detector hits of the minimum-bias sample is simply superimposed with those of the main interaction. The number of additional generated interactions follows an assumed distribution, expected to be close to the true pile-up distribution as measured in data. The MC events are then reweighted in order to fully match the true pile-up distribution of any given data sample.

In addition to the in-time pile-up for a given bunch crossing, one can also have out-of-time pile-up which originates from pp collisions in the previous and subsequent bunch crossings. This is due to the slower response speed of the subdetectors compared with the fast bunch crossing rate, so that successive bunch crossings cause overlapping energy deposits.

3.5 RUN PERIODS

The LHC produced its first proton-proton collisions in November 2009 at $\sqrt{s} = 900$ GeV, keeping the injected protons at the energy of the SPS. A few months later, in March 2010, the protons were accelerated to 3.5 TeV, half of the design energy. This embarked the start of Run I (2010-2013), in which the LHC was first operated at

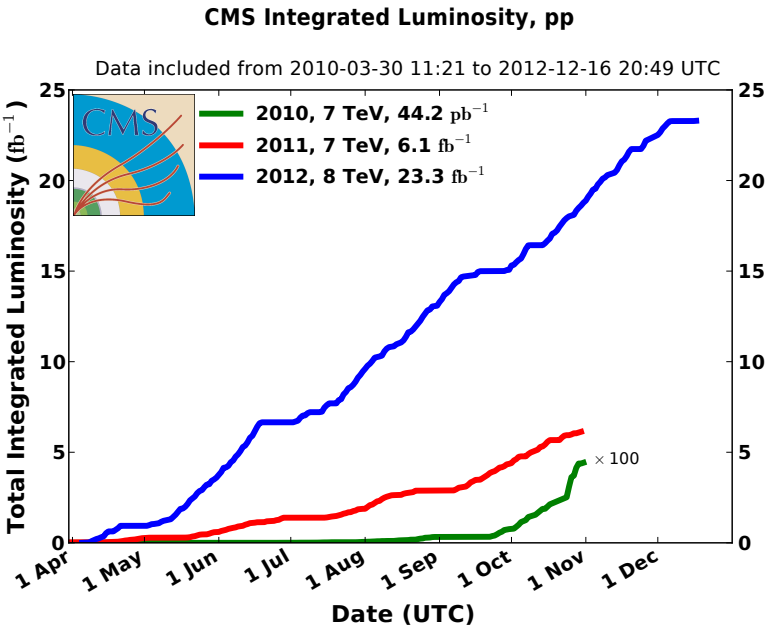


Figure 18: Cumulative luminosity versus time delivered to CMS for pp-collisions. The integrated luminosity of 2010 is multiplied with a factor 100 for easy comparison.

$\sqrt{s} = 7$ TeV (2010 and 2011), delivering more than 6.1 fb^{-1} in the CMS interaction point. While the first beams consisted of only a few bunches in 2010, this was raised to a maximum of 1380 bunches with 50 ns between successive bunch crossings, resulting in a peak luminosity exceeding $4 \cdot 10^{33}$ cm^{-2}s^{-1}. For the 2012 data run, the beam energies were raised to 4 TeV, resulting in a center-of mass energy of $\sqrt{s} = 8$ TeV. The peak luminosity exceeded $7 \cdot 10^{33}$ cm^{-2}s^{-1}, and a total of 23.3 fb^{-1} was delivered to CMS, as shown in Figure 18. The analyses described in this thesis makes use of the data collected in 2011 and 2012.

After a long shutdown, the LHC is brought back in operation in 2015 at a center-of-mass energy $\sqrt{s} = 13$ TeV. In this second run, the bunch crossing rate will be reduced to 25 ns, allowing the luminosity to achieve its design value, but also increasing the effect of out-of-time pile-up.

4

THE COMPACT MUON SOLENOID

4.1 INTRODUCTION

The CMS detector is designed to perform wide variety of measurements. It therefore consists of different layers constructed within an onion-like design, shown in Figure 19. Every layer takes a cylindrical shape in which the components parallel to the beam line are called the barrel regions, and components closing the detector on both sides are usually referred as the endcaps. The inner layer is a silicon based tracker, which aims to reconstruct the trajectories of all particles traversing the detector. It is surrounded by the electromagnetic calorimeter, which measures the energy of electrons and photons. The hadronic calorimeter is used to measure the energy of hadrons. The outermost layer is the muon system, which is able to measure the direction and momenta of the muons with great precision. Between the hadronic calorimeter and the muon system, a superconducting magnet is located, which is capable of reaching a magnetic field of 4.0 T in its contained volume. The large bending power provided by the solenoidal magnetic field, operating at 3.8 T during Run I, is needed to bend the tracks of charged particles in the transverse plane, which allows for a precise measurement of the charge and momentum of these particles. A 12000 tonne yoke made of steel is added in three barrel layers and three endcap disks on each side. The yoke increases the homogeneity of the magnetic field and reduces the stray field by returning the magnetic flux of the solenoid. The tracker and calorimeter barrel layers are installed inside the solenoid, hence they have to be very compact. The muon detectors are installed between the different layers of the iron return yoke.

4.1.1 *Coordinate conventions*

The origin of the CMS coordinate system is taken at the nominal collision point, with the x-axis pointing towards the centre of the LHC ring, the y-axis pointing vertically upwards, and the z-axis pointing along the direction of the counterclockwise rotating proton beam. The azimuthal angle ϕ is measured from the x-axis in the xy plane. The polar angle θ is measured from the z-axis, but is translated into the more convenient pseudorapidity η, defined as

$$\eta \equiv -\ln \tan \frac{\theta}{2} \tag{56}$$

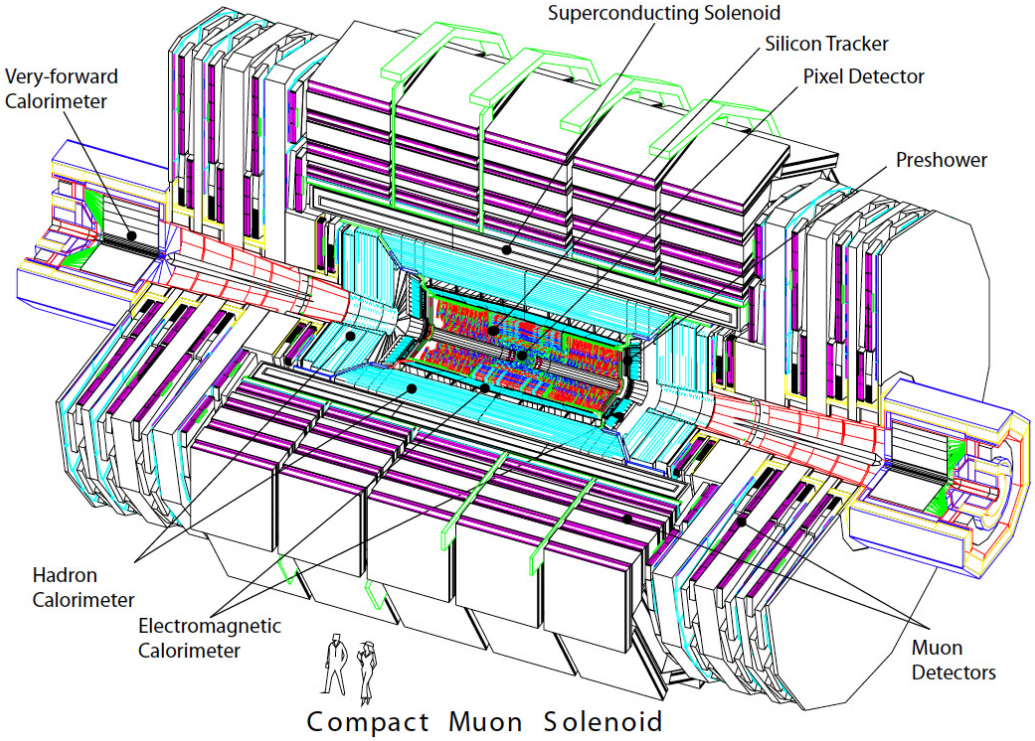


Figure 19: A perspective view of the CMS detector [65]

which is 0 in the direction perpendicular to the beam axis and reaches plus or minus infinity in the direction of the beam axis. The CMS detector covers the pseudorapidity range $-6.6 < \eta < 5.2$. The numerical values of the pseudorapidity differ only slightly from the rapidity, defined as

$$y \equiv \frac{1}{2} \ln \frac{E + p_z}{E - p_z} \tag{57}$$

and both are identical for massless particles. The advantage of using rapidity instead of the polar angle, is that its differences are invariant with respect to Lorentz boosts along the beam axis.

The initial momentum of the colliding partons along the beam axis is not known, momentum and energy are therefore only conserved in the transverse plane. The projection of momentum and energy in the transverse plane are denoted as p_T and E_T respectively. Particles which leave the detector undetected (e.g. neutrinos) could be revealed by the missing energy $\not{E}_T$ in the transverse plane.

4.2 TRACKING SYSTEM

The first layer around the interaction point is the tracking system, designed to provide a precise and efficient measurement of the charged particle trajectories

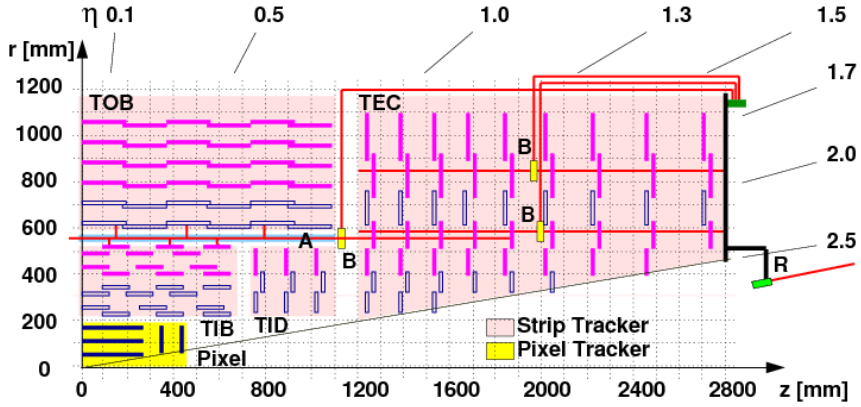


Figure 20: Sketch of the tracking system in CMS. Pixel and strip modules are shown in a quarter of the longitudinal plane. Red lines represent single modules, blue lines represent double modules. [70]

emerging from the LHC collisions. The magnitude and direction of these curved trajectories allow us to deduce the momentum and charge of these particles. Furthermore, the tracker is able to reconstruct vertices with great precision, which is needed to identify the secondary vertex of the decays of long-lived heavy particles, but also helps in distinguishing tracks from the primary vertex of interest from the many tracks originating in pile-up events. Due to the high pile-up and a short time between bunch crossings, a high granularity and fast response is required to process the large number of tracks.

The tracker has a total length of 5.8 m and a diameter of 2.5 m and has a coverage up to a pseudorapidity of $|\eta| < 2.5$. A schematic overview of its geometry is shown in Figure 20. Within 10 cm of the interaction point, a pixelated detector is needed to cope with the high particle flux. Around the pixel detector, narrow silicon micro strips are used. Both the pixel and micro strip detectors are silicon-based. When charged particles pass through these pixels and strips, they cause ionization in the silicon, creating electron-hole pairs. When an electric field is applied, these electrons and holes in the silicon drift towards the electrodes, at which a signal can be measured.

The pixel detector consist of three barrel layers at radii between 4.4 and 10.2 cm from the beam axis, and two endcaps at 34.5 and 46.5 cm in $\pm z$ enclosing the barrel layers at opposite sides. The silicon pixel size (100×150 μm^2 in $r - \phi$ and z) is chosen to achieve a good resolution and to keep the detector occupancy at a maximum of 1%. The strip tracker is composed of silicon micro strips and is divided in an inner region, with 4 barrel layers and 3 endcap layers, and an outer region, with 6 barrel layers and 9 endcap disks. The size of the strips ranges between 10 cm $\times$ 80μm in the most inner parts, to 25 cm $\times$ 183 μm at the most outer parts where the particle flux is lower. The larger cell size has the advantage of reducing the number of read-out channels. On the other hand, the electronics noise is a linear function of the strip length. In order to keep a good signal to noise ratio of well above 10, thicker silicon sensors have to be used in the outer region (500 μm as

opposed to 320 μm for the inner tracker). Some of the layers carry a second microstrip detector, tilted about 5.73° with respect to each other, in order to measure the second coordinate, respectively z and r for the barrel and endcap. Within a layer, each module is shifted slightly in z or r with respect to its neighbouring modules, allowing them to overlap. In this way gaps in the acceptance are avoided.

4.3 ELECTROMAGNETIC CALORIMETER

When electrons and positrons passes through a medium, they emit photons in the electric fields around the nuclei, a process called Bremsstrahlung. Photons traversing trough the electric fields are converted into electron-positron pairs, which in turn could emit new photons. An avalanche effect in which the number of involved particles rapidly grow, also known as an electromagnetic shower, is created. However, at each step the average energy of the particles decreases, and fewer of the photons have sufficient energy for particle pair production. Eventually, the energy of the electrons, positrons and photons are absorbed by the scintillation medium, which re-emits the absorbed energy in the form of scintillation light. This behaviour is exploited in the CMS Electromagnetic Calorimeter (ECAL), which consist of scintillation crystals with a photon detector attached to measure the scintillation light. The energy of the particle initiating the shower is reconstructed from the total deposited energy in the ECAL, which is proportional to the measured light. Muons are more massive than electrons, and are therefore much less affected by Bremsstrahlung. As a result, the muons go right through the ECAL, without initiating electromagnetic showers. Hadrons also interact less in the ECAL medium, and loose only a fraction of their energy in the ECAL.

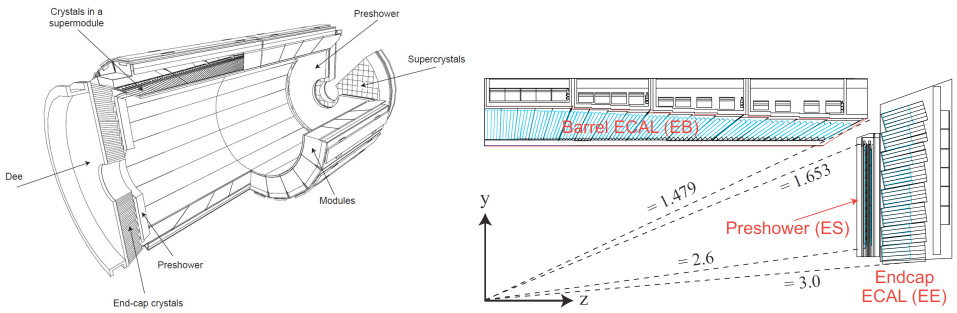


Figure 21: Layout of the CMS ECAL showing the arrangement of crystals in the barrel and endcap, with the preshower in front of the endcap. A longitudinal section of one quadrant is shown on the right. [65]

The probability of the Bremsstrahlung and pair production processes depends on the atomic number Z of the traversed medium. Scintillators which have a high atomic number Z, have stronger electric fields around their nuclei, and electromagnetic showers will develop more quickly in those atoms compared to atoms with low Z. In order to construct a compact calorimeter, a high stopping power is re-

quired and a scintillator with high density and high Z is preferred. Lead tungstate ($PbWO_4$) crystals are therefore a good choice for the CMS ECAL. In addition, it has a fast response time, emitting about 80% of its scintillation light within 25ns, and good radiation tolerance. However, its scintillation light output is rather low, which asks for highly efficient photodetectors placed at the rear of the crystal.

The barrel section covers the pseudorapidity interval of $|\eta| < 1.479$. The size of the crystals is chosen to have a granularity of about 0.0174 (i.e. 1°) in both ϕ and η. The crystals are tilted 3° with respect to the line of the nominal vertex direction. The scintillation (and Čerenkov) light produced in the crystals is detected by silicon avalanche photodiodes. The endcaps cover the range between pseudorapidity 1.479 and 3, and the light is read out by vacuum phototriodes. The endcaps are also equipped with a preshower detector, covering $1.653 < |\eta| < 2.6$ with a much higher granularity compared with the endcap crystals. Because of the high granularity it is able to distinguish neutral pions, decaying into two closely spaced photons, from prompt photons. It also improves the identification of electrons against minimum ionizing particles, and enhances the position determination of photons and electrons. The preshower is composed of two planes of lead, in which the electromagnetic shower is induced, followed by silicon strip sensors which measure the electron-positron pairs from the shower.

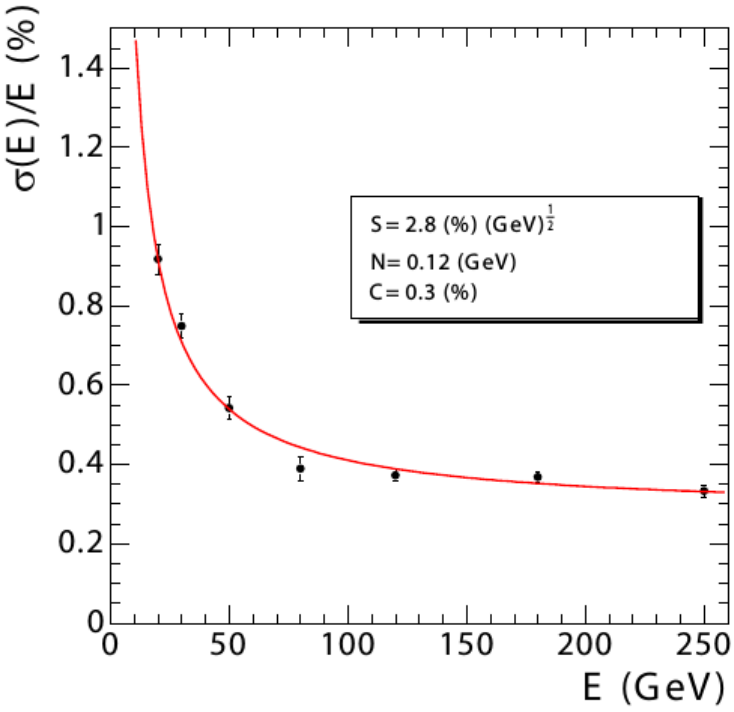


Figure 22: ECAL energy resolution as a function of electron energy, measured from a beam test. The stochastic (S), noise (N) and constant (C) terms of the fit are given. [65]

The energy resolution in the barrel, measured during beam tests [71], can be parametrized as a function of the energy:

$$\frac{\sigma_E}{E} = \frac{2.8\%}{\sqrt{E}} \bigoplus \frac{12\%}{E} \bigoplus 0.3\% \tag{58}$$

The first term corresponds with the stochastic contribution due to fluctuations in the lateral shower development and in the energy released in the pre-shower. The noise due to electronics, digitization and pile-up is represented in the second term. The last term is due to calibration errors, energy leakage from the back of the crystals and the non-uniformity of the longitudinal light collection. As shown in Figure 22, the electron energy resolution is below 1% for all energies above 20 GeV, and dropping below 0.4% for the highest energies.

4.4 HADRONIC CALORIMETER

The Hadronic Calorimeter (HCAL) measures the energy of neutral and charged hadrons, and is similar in concept as the ECAL. When hadrons enter a medium they initiate a hadronic cascade through strong interactions with the nuclei. Even though these interactions also occur in the ECAL, the hadronic showers develop much slower than the electromagnetic showers, and hadrons reach the HCAL where most of their energy is deposited. As opposed to the ECAL, which is a homogeneous detector, the HCAL is a sampling calorimeter which means it uses alternating layers of high density absorber material, in which hadronic showers develop fast, and layer of scintillators converting the absorbed energy into scintillation light.

The HCAL consists of four sub-detectors, for which their location in CMS is shown in Figure 23. The Hadron Barrel (HB) covers the pseudorapidity range up to $|\eta| < 1.3$, while Hadron Endcap (HE) covers the endcap region $1.3 < |\eta| < 3$. The absorber layers are made of Brass with plastic scintillators in between. The granularity is 0.087×0.087 radians in the barrel region and varies from 0.087×0.087 to 0.165×0.350 in the endcaps. The scintillation light is collected by a set of embedded wavelength-shifting fibres which carry the light to the read-out system. The Hadron Outer (HO) detector is an additional layer of scintillators of the barrel calorimeter placed outside of the magnet, in order to collect the energy of from penetrating hadron showers leaking through the barrel detector which eventually interact with the high density material composing the magnet.

Finally, coverage between $2.9 < |\eta| < 5$ is provided by the HF detector, using steel as the absorber and quartz (SiO_2) fibres. The signal originates from Čerenkov light in the quartz fibers, embedded parallel to the beam axis in the absorber, which is then channeled by the fibres to photomultipliers. The fibres are of two different lengths, the longer ones running over the full depth (165 cm) over the detector and the shorter ones starting 22 cm from the front of the detector. This allows to

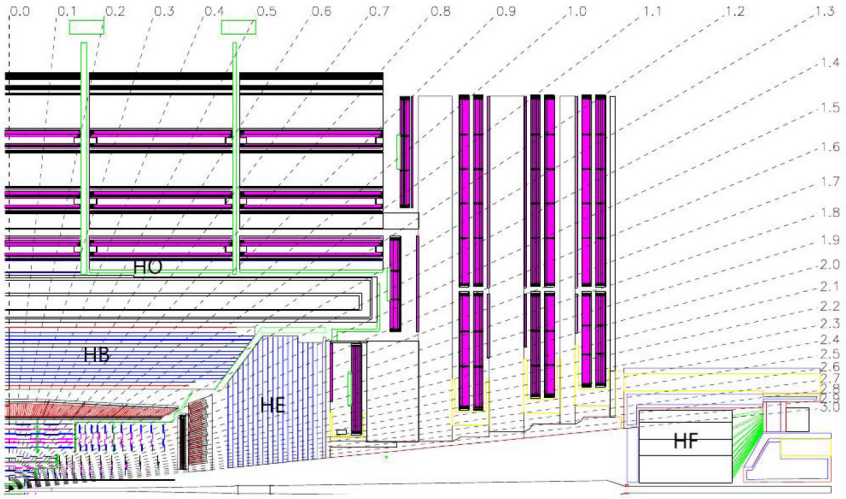


Figure 23: Longitudinal view of the CMS detector showing the locations of the HB, HE, HF and HO detectors [65]

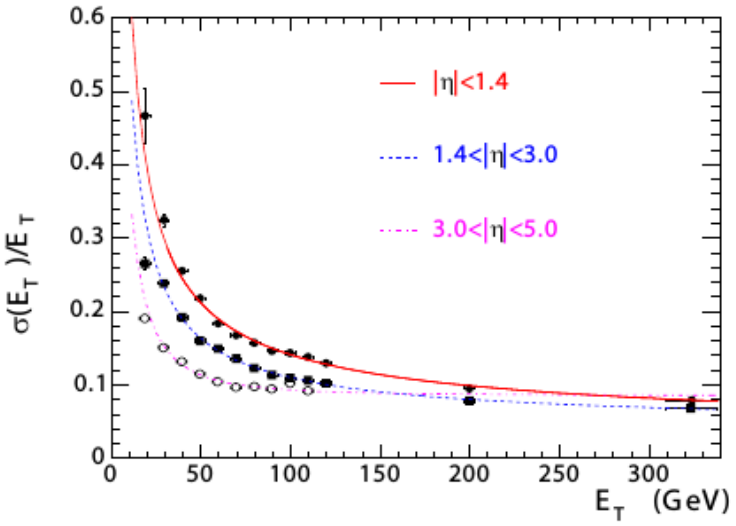


Figure 24: Jet transverse energy resolution as a function of the transverse energy for the barrel, endcap and forward region [65]

distinguish between electromagnetic and hadronic showers, as the electromagnetic showers have mostly finished before reaching the short fibres. The HF design leads to narrower and shorter hadronic showers. It is therefore ideally suited for the forward region which will experience high particle fluxes.

The energy resolution for the reconstructed jet transverse energy in the HCAL is shown in Figure 24 for the barrel, endcap and forward regions. At low energies around 30 GeV, the resolution is about 30%, but this improves for increasing energy to below 10%, in all regions of the detector.

4.5 FORWARD DETECTORS

In addition to the calorimeters described above, two other calorimeters are installed at the very forward regions of CMS, where they can measure energies at angles very close to the beamline. The Centauro And Strange Object Research (CASTOR) detector is installed at the $z < 0$ side of the interaction point and covers the pseudorapidity range $-6.6 < \eta < -5.2$. It is constructed from tungsten plates for the absorber layers and quartz plates in which Čerenkov light is generated. The Zero Degree Calorimeter (ZDC) consist of two detectors installed about 140 m away on each side of the interaction point. The detectors are made from tungsten with embedded quartz fibres and cover the region with $|\eta| > 8.3$. The forward detectors are not used in this work.

4.6 MUON SYSTEM

The muon system is the outer layer of the CMS detector. As muons can penetrate several layers of material without interacting, they are not stopped by the calorimeters. We can therefore track them in the muon system, shown in Figure 25, composed of three types of gaseous detectors placed between the return yoke layers. When a muon traverse a gas chamber, it creates electron-ion pairs in the gas. In the presence of an electric field, created by applying a voltage potential, these electrons drift to the positively charged anode whereas the ionized gas moves to the negatively charged cathode. If the electric field is strong enough, the electrons can ionize other atoms in the gas. This results in an electron avalanche which amplifies the current. The movement of the charges towards the anode induce an electric signal which can be read out.

In the barrel region ($|\eta| < 0.9$), where the muon rate is low and the magnetic field uniform and mostly contained in the return yoke, drift tubes are used. A drift tube has one wire in the middle acting as the anode. When a charged particle ionizes the gas, the electrons drift to the wire. The moving charges from the electrons induce a fast signal on the wire. The drift tubes are found in 4 detector stations alternated

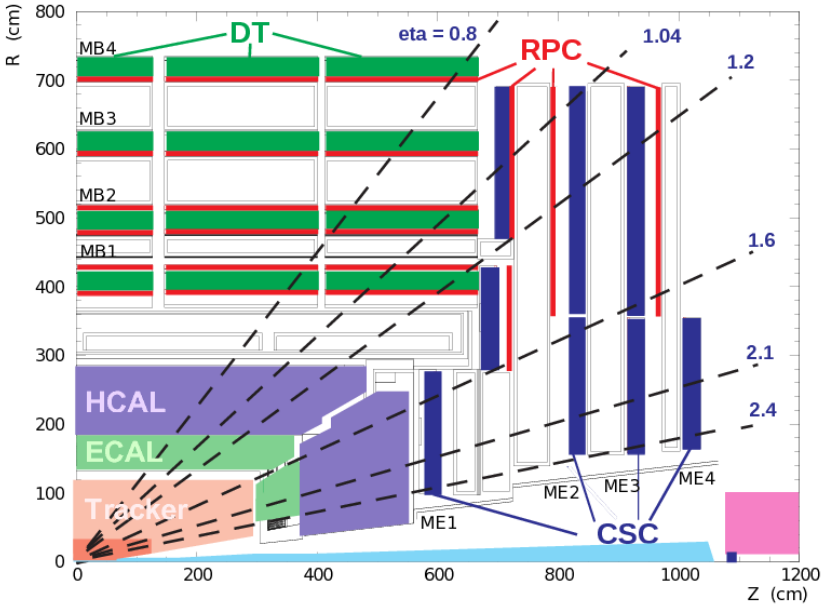
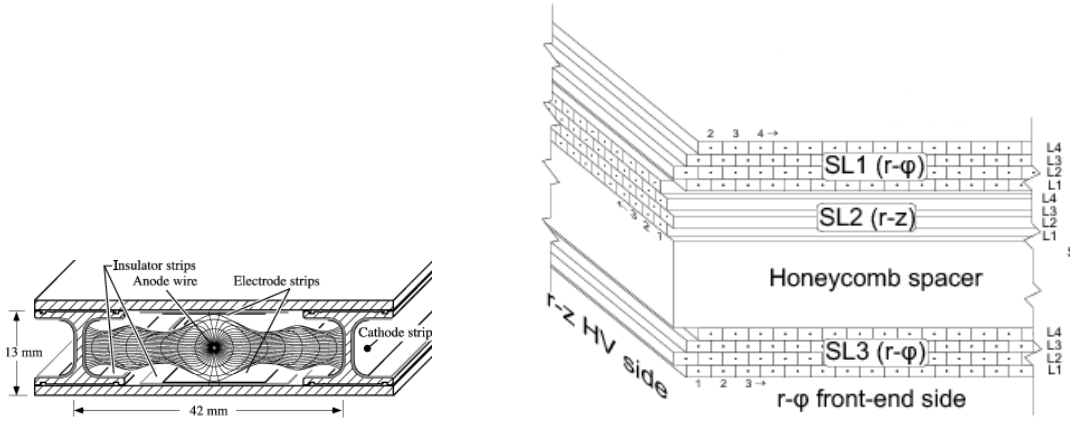


Figure 25: Longitudinal cross section of one quarter of the CMS muon system [65]

with the segmented return yoke. Each station is arranged in 2 or 3 superlayers, each consisting of 4 layers of drift tubes. The superlayers are orthogonal to each other, in which each superlayer focuses on the measurement in the direction of ϕ or z.

In the endcap region ($0.9 < |\eta| < 2.4$), where particle rates are higher and the magnetic field is non-uniform and large, cathode drift chambers are used. Cathode drift chambers have a trapezoidal shape and consist of a cathode plane divided in strips and multiple anode wires orthogonal to the cathode strips. The cathode strips run radially outward, measuring the ϕ coordinate, while the anode wires are optimized for bunch crossing identification and also provide a measurement in the η direction.

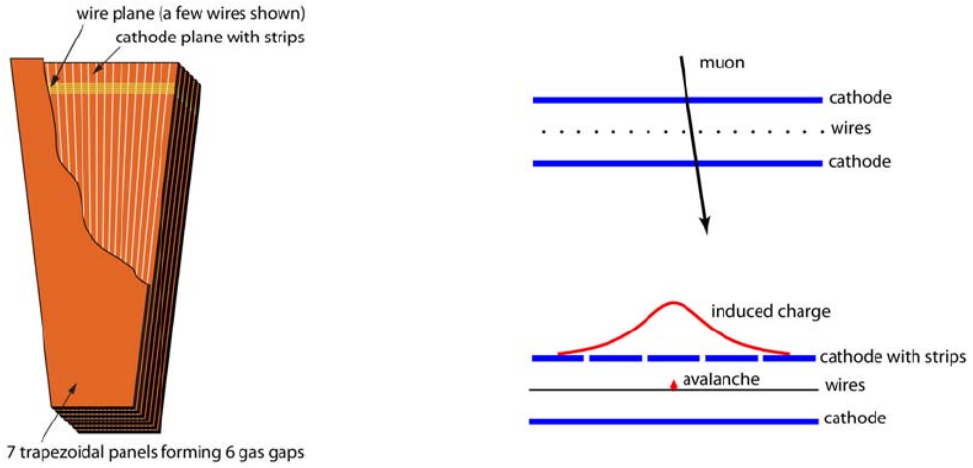
Finally, resistive plate chambers are used in both barrel and endcap regions up to $|\eta| < 2.1$. Resistive plate chambers consists of two gas gaps, each consisting of two parallel plates of which one is the anode and the other the cathode. Due to the electric field in the gap, an electron avalanche is created when an ionizing particle crosses the detector. The movement of the charges induces a signal on the read-out strips which are placed between the two gaps. Resistive plate chambers have a coarser position resolution compared with the other two muon detectors, but their fast response allows to assign a muon track unambiguously to the correct bunch crossing.



(a) Drift tube cell

(b) Drift tube chamber

Figure 26: Cross section of a drift tube cell, and layout of the cells in a drift tube chamber, showing the arrangement of the three superlayers, each containing 4 layers of drift tubes. The superlayers are attached to a honeycomb panel which acts as the support structure. [72]



(a) Cathode strip chamber

(b) Schematic view of one layer

Figure 27: Cathode strip chambers consist of 6 layers with the cathode strips of constant $\Delta\phi$ running radially outwards and anode wires running across. In addition to the signal on the wire, the movement of the positive ions towards the cathode plane induces a charge pulse in the strips. [65]

4.7 TRIGGER AND DATA ACQUISITION

The LHC provides proton-proton collisions at high interaction rates. It is impossible to store and process the large amount of data generated by all the collisions. The vast majority of events are however soft collisions, and the interesting events are very rare. An efficient trigger system is developed to lower the rate of acquired events to a manageable level, while still retaining most of the rare signal events. For more common processes a prescale is applied which means only a fraction of the events are stored.

The CMS trigger consist of two decisional levels. The Level-1 Trigger (L1) is composed of dedicated electronics and makes a quick decision about which events are kept for further processing. It reduces the initial event rate of about 1 GHz (at the design luminosity of the LHC) to an output rate of 100 kHz. The L1 decision to keep or discard an event has to be made within 3.2 μs. The full data is temporarily stored in pipelines of processing elements while waiting for the L1 decision. Because the tracker algorithms are too slow to fit into the allowed L1 decision time, the L1 trigger relies solely on the information of the muon system and the calorimeters.

If an events is accepted by the L1 trigger, the full readout of the event is initiated a passed on to the High Level Trigger (HLT) in which the event rate gets reduced by the HLT to about $300 - 600$ Hz. The HLT is software based trigger running at a computer farm and has access to the complete raw data which can be used to reconstruct basic physics objects. Events passing the HLT are recorded permanently for further physics analysis.

4.8 DETECTOR SIMULATION

The particles which are generated in MC events are passed on to a detector simulation, in order to have a description in terms of detector signals, as is the case with the real data. The data and simulations could then use the same reconstruction methods, allowing for a direct comparison of the objects at detector level. The detector simulation is based on a full description of the CMS detector geometry, implemented in the GEANT4 [73] toolkit. Each particle in the event are propagated through the different detector layers, including the dead material regions due to cables and support structures, while simulating all possible physical interactions which result in detector deposits or energy losses. In addition, secondary particles are generated which originate from the interactions between the particles and the detector material. For each subdetector, the response is accurately simulated providing an output signal in the same format as the output available in real data.

5

OBJECT RECONSTRUCTION

Particles traversing the detector induce track hits or energy deposits in the different detector layers of CMS. As shown in Figure 28, every type of particle has a different signature. A technique called Particle Flow (PF) [74] is used to reconstruct and identify all stable particles in the event: electrons, muons, photons, charged hadrons and neutral hadrons. Neutrinos do not leave a trace in the detector, and are revealed to the missing energy $\not{E}_T$ in the transverse plane. The PF technique combines information from all subdetectors, resulting in a more precise determination of the momenta of the particles. From the individual particles, higher-level objects like jets and the $\not{E}_T$ can be constructed. In this chapter, we discuss the track and vertex reconstruction, followed by an overview of the different physics objects.

5.1 TRACKS AND VERTEX RECONSTRUCTION

Due to the high pile-up in LHC collisions, the tracker is expected to be traversed by about 1000 particles at each bunch crossing. Also the bunch crossing before and after can contribute because of the finite time resolution of the detector. The reconstruction of tracks is therefore a challenging task, and fake tracks could be formed by a combination of unrelated hits. Track reconstruction in CMS [75] happens through an iterative procedure, in which the initial iterations search for the tracks which are the easiest to find, i.e. high p_{T} tracks produced near the interaction region. Once a hit is identified with a track, it is excluded from the subsequent iterations. In this way the combinatorial complexity is reduced from the next iterations in which the reconstruction of more difficult classes of tracks is attempted. Each iteration starts with a seed using only 2 or 3 hits, giving an initial estimate of the track parameters, i.e. the position and direction vectors and an initial estimate of the transverse momentum. The seed is chosen in the innermost layers of tracker as the high granularity of the pixel detectors ensures a lower occupancy and better estimates for the initial parameters. The seed trajectories are extrapolated along the expected flight path of a charged particle, searching for additional hits to be associated with the track. This is done layer by layer using a Kalman Filter (KF): the track parameters at each detector layer are used to find compatible measurement in the next detector layer, forming combinatorial trees of track candidates. When a compatible hit is found, the track parameters are updated taking the new hit into account before extrapolating to the next layer. At the end of each iteration, quality

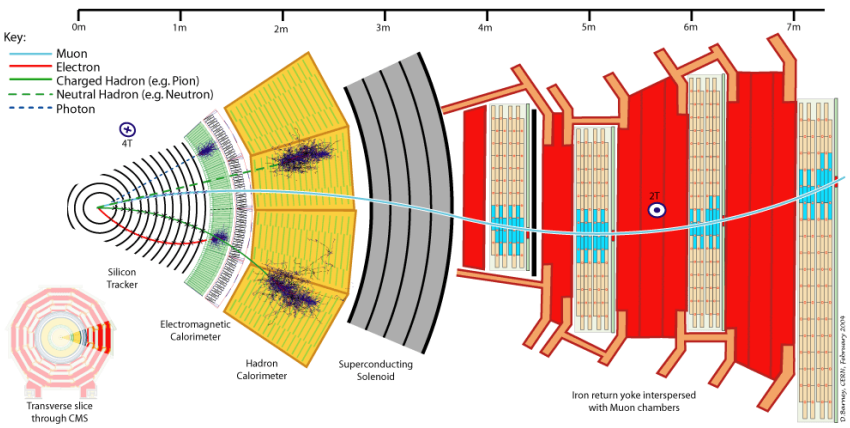


Figure 28: Transverse slice of the CMS detector showing the trajectories of particles and their hits in the different subdetectors. A track can be reconstructed for charged particles. Electrons and photons deposit their energy in the ECAL, whereas hadrons are stopped in the HCAL. The tracks of muons in constructed from hits in the tracker and the muon system [65]

flags are set based on the compatibility with the interaction region and whether their fit yielded a good χ^2 per degree of freedom. The quality criteria depend on the number of layers in which a track has at least one hit. The quality flags are loose, tight and high-purity, providing progressively more stringent requirements. Tracks failing for the loose quality flag are discarded.

The tracker only identifies charged particles, which are subject to the homogeneous magnetic field oriented along the z-axis. Hence, the reconstructed track follows a helix trajectory with its axis parallel to the z-axis. This trajectory can be fully characterized by 5 parameters describing its position, direction and momentum at the point of closest approach to the beam axis:

- the transverse impact parameter d_{xy} is the distance of closest approach from the helix trajectory to the beam axis.
- the longitudinal impact parameter d_z is the distance along the z-axis to the nominal interaction point located at the center of CMS
- the azimuthal angle ϕ of the momentum vector
- the angle θ between the momentum vector and the z-axis
- the transverse momentum p_T

Using the reconstructed tracks, the positions of the vertices can be fitted. There are two kind of vertices: the primary vertices are the locations associated with the different pp collisions in the same bunch crossing, secondary vertices are the result of decays of long-lived particles originating from a primary vertex. Tracks

produced promptly in the interaction region are clustered on the basis of their z coordinate at their point of closest approach to the centre of the beam spot. The primary vertices are required to pass quality criteria for the vertex fit and should have a maximum distance to the nominal interaction point along the beam axis of 24 cm. The main event Primary Vertex (PV), used as a reference for the reconstructed objects in the event, is chosen to be the one with the largest $\sum_i p_{T,i}^2$ where the sum runs over all the tracks used in the vertex fit.

5.2 PARTICLE FLOW EVENT RECONSTRUCTION

The PF event reconstruction starts from three fundamental elements: tracks from the charged particles in the silicon tracker, calorimeter clusters, and muon tracks in the muon system. These are linked together in blocks which could be identified as the different stable particles.
A specific calorimeter clustering algorithm is developed for the PF event reconstruction. The clustering is performed separately for the barrel and endcap subdetectors in the ECAL and HCAL, and for the preshower. No clustering is performed for the HF where each cell gives rise to exactly one cluster. The clustering algorithm starts with seeds which are found as local maxima above a given energy threshold. From the seed, topological clusters are grown by aggregating cells which exceed a given energy threshold and have at least one side in common with a cell already in the cluster. The thresholds are taken at two standard deviations above the electronics noise level in the calorimeter. For each seed within the topological clusters, exactly one PF cluster is constructed. The energy and position of each PF cluster are determined using an iterative procedure, starting from the seed positions as an initial estimate for the PF cluster positions. The energy of each cell within the topological cluster is shared between the different PF clusters based on the distance between the cell and the cluster. For the next iterations, the positions of the PF clusters are recomputed as the center-of-gravity of the seed cell and its neighbour cells. The procedure is repeated until the positions do not move by more than a small fraction of the position resolution.

The tracks and calorimeter clusters need to be connected to each other by a link algorithm to fully reconstruct each single particle, while getting rid of any possible double counting from the different detectors. Tracks are extrapolated from their last measured hit in the tracker to the calorimeters, where it is linked to any given cluster if the extrapolated position is within the cluster boundaries. The cluster envelope can be enlarged up to the size of a cell in each direction in order to account for the presence of gaps between the calorimeter cells, for the uncertainty on the position of the shower maximum, and for the effect of multiple scattering for low-momentum charged particles. Links between calorimeters are established when the cluster position in the more granular calorimeter is within the cluster envelope in the less granular calorimeter. The distance in the (η,ϕ) plane between the two linked elements quantifies the quality of the link. Links between tracks in

the silicon tracker and muon tracks in the muon system are established when the global fit between two tracks returns an acceptable χ^2. When several tracks from the silicon tracker can be linked to a muon track, only the global muon track with the smallest χ^2 is retained.

Elements which are directly or indirectly linked are grouped together in blocks. Due to the fine granularity in the CMS subdetectors, each block typically only contains a few elements. For complex events, the number of blocks will increase, while the number of elements in each block remains the same, making the performance of the algorithm independent from the event complexity. The blocks are used as simple inputs to built the list of individual particles. First, PF muons and PF electrons are identified as explained in 5.3 and 5.4. The tracks and clusters associated with the electrons and muons are removed from the block. Tighter quality criteria are applied on the remaining tracks: tracks for which the relative uncertainty on the p_T is smaller than the expected relative calorimeter energy resolution for charged hadrons are rejected. Each of the remaining tracks give rise to a PF charged hadron. If the energy in the calorimeter is compatible with the track momentum within uncertainties, the charged hadron momentum is redefined by a fit of the measurements in the tracker and the calorimeter. It could also be that the calibrated energy of the closest ECAL and HCAL clusters linked to the track(s) is significantly larger compared to what is expected from the momenta in the tracker. If this excess is larger than the expected calorimeter energy resolution, a PF photon or PF neutral hadron are defined. Also calorimeter clusters which are not linked to a track give rise to PF photons or PF neutral hadrons.

5.3 MUONS

In addition to the tracks from the inner tracker, standalone-muon tracks are independently reconstructed in the muon system. The hits within each drift tube chamber and cathode strip chamber are geometrically matched to form track segments. The track segments in the innermost muon chambers are taken as seeds to built a standalone-muon track towards the outer layers using the Kalman Filter technique. Muons can be reconstructed [76] from the combination of these two tracks using two approaches:

- *Global-muon reconstruction*, starting from the standalone-muon track which is matched to an inner track. Hits from both the inner track and standalone-muon track are combined into a global-muon track, using a Kalman Filter to update its parameters. For large transverse momenta, the global-muon fit improves the resolution compared to the tracker-only fit.

- *Tracker-muon reconstruction*, in which tracker-tracks with $p_T > 0.5$ GeV and total momentum $p > 2.5$ GeV are extrapolated to the muon system, taking into account its expected trajectory through the magnetic field including average

	tight muon (7 TeV)	tight muon (8 TeV)
global muon	yes	
tracker muon	yes	
χ^2/ndf in global-muon fit	< 10	
number of stations	> 1	
number of hits in muon system	> 0	
pixel hits	> 0	
tracker layers	> 8	> 5
d_z	–	< 5 mm
d_0	< 2 mm	

Table 2: Muon identification criteria for the tight muon selection

expected energy losses and multiple Coulomb scattering in the detector material. If the track can be matched to a muon segment the track qualifies as a tracker-muon. Only muon segments for which there is no other tracker-track that forms a better match are considered. Because only a single muon segment is needed, the tracker-muon reconstruction is more efficient for muons with low momenta.

Due to the very high efficiency in reconstructing tracks in both the tracker and the muon system, most muons are reconstructed by both approaches (i.e. sharing the same tracker-track), and are merged into one single muon candidate. Muons which are only reconstructed as standalone-muons have worse momentum resolution and are more likely to be the result of cosmic rays, hence they are not used in physics analyses.

In this work, muons are only selected if they pass the *tight* identification criteria, listed in Table 2. The identification criteria suppress backgrounds from the hadronic punch-through (leakage of hadronic showers into the muon system) and muons from pion decay in flight. Tight muons are required to be constructed both as global-muon and tracker-muon. The global-muon trajectory fit should have a reduced χ^2 (i.e. χ^2 divided by the number of degrees of freedom) less than 10, and should include at least one muon chamber hit. Furthermore it should have muon segments in at least 2 of the 8 muon stations[1], To guarantee a good p_{T} measurement, a minimal number of tracker layers is required. Upper values for the transverse and longitudinal impact parameters d_{xy} and d_z with respect to the main PV further suppresses cosmic muons, muons from decays in flight and tracks from pile-up.

In addition to the identification requirements, we require the muons to be isolated. The isolation requirement reduces the backgrounds of muons embedded in jets,

1 There are four barrel and four endcap muon stations

which are often the result from semi-leptonic decays of b or c quarks. The loose track-based relative isolation requirement is given by

$$I = \frac{\sum p_{T,i}}{p_T(\text{tot})} < 0.1 \tag{59}$$

in which the sum runs over all additional charged tracks within $\Delta R < 0.3$ around the muon, and $p_T(\text{tot})$ is the total p_T in this cone, including the muon. The additional charged tracks considered in the isolation variable must be consistent with the main PV of the event, and the isolation requirement is therefore insensitive to the presence of other pile-up interactions.

5.4 ELECTRONS

Electrons are reconstructed [77, 78] by associating a track in the silicon detector with a cluster of energy in the ECAL. Two complementary approaches are used: one algorithm starts with a tracker-driven seeding, the other uses an ECAL-driven seeding. The ECAL-driven seeding is more suitable for high p_T and isolated electrons, while the tracker-driven seeding which is better suited to low p_T electrons and electrons inside jets.

The ECAL-driven algorithm starts by the reconstruction of superclusters, a group of one or more clusters of energy deposits, in the ECAL. The electrons have to travel through the tracker material in which they radiate Bremsstrahlung photons. As these photons are not bend by the magnetic field, the energy reaches the ECAL with a significant spread in ϕ. Hence, the supercluster algorithm searches for a large spread in ϕ while looking for a narrow width along the η coordinate. The superclusters with $E_T > 4$ GeV are matched to track seeds in the pixel detector, from which the electron tracks are built. In order to deal with the Bremsstrahlung emission, a dedicated tracker algorithm has been developed, using a Gaussian Sum Filter (GSF) instead of the standard KF algorithm. The KF algorithm assumes the energy loss of a charged particle traversing a thin layer of material is Gaussian, while the Bremsstrahlung is highly non-Gaussian. When a photon is emitted, the next hit in the tracker is often far away from the position expected by the KF algorithm. As a result the pattern recognition could stop following the electron path, leading to shorter tracks. The GSF algorithm describes the energy loss of electrons by a mixture of Gaussian distributions, and is able to fit the sudden curvature radius change, caused by the Bremsstrahlung photon emission. However, the GSF track reconstruction cannot be run on all tracks, due to its high computing time consumption.

The tracker-based seeding relies on tracks reconstructed by the KF algorithm, which are submitted to a pre-identification [79] stage to find potential electron tracks and reject tracks from hadrons. Electron tracks reconstructed by the KF algorithm are usually shorter, and could be matched to a PF cluster in the ECAL. This information is combined using a multivariate approach, a Boosted Decision Tree (BDT), to

perform the pre-identification. The discriminating power of the tracker variables is also strengthened by refitting the tracks using a light [2] version of the GSF algorithm. Identified electron tracks are refitted with the full GSF algorithm.

Most electrons are found by both the ECAL-driven and tracker-driven algorithm. The GSF tracks from both the ECAL-driven and the tracker-driven trajectory seeds are merged and used by the PF reconstruction algorithm, in which they are linked to PF-clusters in the ECAL. The energy taken away by possible Bremsstrahlung photons are found by extrapolating straight line tangents to the electron track from each tracker layer to the ECAL.

The electron seeds are not always the result of isolated electrons from a PV, but could be caused by background sources, mainly originating from photon conversions (i.e. photons converting into electron-positron pairs in the tracker material) or from jets misidentified as electrons. Electron identification algorithms are used to discriminate between real and fake electrons. Variables that provide discrimination power for electron identification are:

- $\Delta\eta_{\text{trk-SC}}$ and $\Delta\phi_{\text{trk-SC}}$ describing how well the track matches the supercluster by extrapolating the track direction at the vertex to ECAL, assuming no radiation.

- the cluster shape covariance $\sigma_{\eta\eta}$, representing the spread of the electron deposit in the ECAL along the η coordinate. It helps in discriminating real electrons from jets which have a larger spread.

- the ratio H/E which is the energy deposited in the HCAL behind the ECAL supercluster, divided by the energy deposited in the ECAL. If this ratio is high, a substantial fraction of energy was deposited in the HCAL, which results in a higher change of the electron candidate being faked by a jet.

- the impact parameters d_0 and d_z, which is the distance to the main PV in the transverse and longitudinal direction respectively.

- $|1/E_{\text{SC}} - 1/p|$ with E_{SC} the energy of the supercluster and p the track momentum at the point of closest approach to the vertex

- the number of missing hits N_{miss} in the tracker, which is higher for electrons from photon conversions

- the conversion fit probability, quantifying the probability a track pair could be fitted to a common (photon) vertex [78]

2 The number of Gaussians in the mixture is reduced with respect to the standard GSF algorithm, increasing its speed so that it can be run for all tracks

	WP90 (7 TeV)		loose (8 TeV)	
	barrel	endcap	barrel	endcap
$\Delta\eta_{\text{trk-SC}}$	0.007	0.009	0.007	0.009
$\Delta\phi_{\text{trk-SC}}$	0.8	0.7	0.150	0.100
$\sigma_{\eta\eta}$	0.01	0.03	0.01	0.03
H/E	0.12	0.15	0.12	0.10
d_0	–		0.02 cm	
d_z	–		0.2 cm	
$\mid 1/E_{\text{SC}} - 1/p \mid$	–		0.05 GeV^{-1}	
N_{miss}	–		1	
conversion-fit probability	–		10^{-6}	

Table 3: Maximum values for electron identification variables at the WP90 and loose working points

Different working points for a simple cut-based identification are in use within the CMS collaboration. In this work, the *WP90* working point was used for the 7 TeV analysis, which corresponds with an efficiency of selecting 90% of the prompt electrons above $p_{\text{T}} > 20$ GeV. In the 8 TeV analysis, the cuts were slightly updated to the *loose* working point, the typical working point for a $Z \to ee$ analysis. Both working points are described in Table 3. Different cuts are used for the barrel and endcap region. Electrons in the barrel-endcap transition region $1.4442 < \eta < 1.5560$ are excluded as the full clusters cannot be reconstructed within this region.

In this thesis, the isolation requirement applied for electrons, which also helps in reducing the background from jets misidentified as electrons, is exactly the same as for the muons:

$$I = \frac{\sum p_{T,i}}{p_T(\text{tot})} < 0.1 \tag{60}$$

5.5 JET RECONSTRUCTION

5.5.1 *Jet algorithms*

Jet algorithms [80] provide a set of rules describing how particles are grouped into jets and which momentum should be assigned to the jet. These algorithms can be split in two categories: cone algorithms and sequential recombination algorithms. The simplest cone algorithm sums the momenta of all particles found within a cone of radius R in (η, ϕ) around an initial seed particle. This can be extended to iterative-cone algorithms, in which the direction of the resulting sum is used as a

new seed, and the procedure is iterated until the direction of the resulting cone is stable. Usually one takes the hardest particle in the event as a seed. A new seed is taken from the remaining particles and jets are constructed until no particles are left. One problem with cone algorithms is that they are collinear unsafe: the splitting of the hardest particle into a nearly collinear pair could promote another particle in the event to become the first seed for jet construction (i.e. if $p_1 > p_2$ but p_2 is larger than the pair produced by p_1), resulting in a different final set of jets. Another problem is how to deal with overlapping jets. Iterative cone algorithms are also infrared unsafe: the emission of an extra soft particle can cause the iterative process to find a new stable cone, which could again result in a different set of hard jets.

Sequential recombination algorithms use a bottom-up approach in which the algorithm repeatedly recombines the closest pair of particles according to some distance measure. The typical sequential recombination algorithm proceeds through the following steps:

- For each pair of entities i and j (which could be particles, tracks or calorimeter clusters), the distance d_{ij} as well as the beam-jet distance d_{iB} for each entity are introduced by:

$$d_{ij} = \min(p_{Ti}^{2p}, p_{Tj}^{2p}) \frac{\Delta R_{ij}^2}{R^2} \tag{61}$$

$$d_{iB} = p_{Ti}^{2p} \tag{62}$$

  in which $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and p_{Ti}, y_i and ϕ_i are respectively the transverse momentum, rapidity and azimuthal angle of i. The radius parameter R represents the cone width used for clustering and p is a parameter to govern the relative power of the energy versus geometrical (ΔR_{ij}) scales.

- From all the distances d_{ij} and d_{iB}, the smallest is taken. If the smallest is a d_{ij}, the two particles i and j are combined into a new entity k, summing up their four-momenta:

$$\begin{aligned} \mathbf{p}_k &= \mathbf{p}_i + \mathbf{p}_j \\ E_k &= E_i + E_j \end{aligned} \tag{63}$$

  The algorithm then restarts from the first step, in which the initial entities i and j are replaced by k.

- If the smallest distance of the event is of the form d_{iB} (i.e. there is no other entity left within R), the entity i is removed from the event and is considered as a final jet. The algorithm is then repeated until no entities are left in the event.

The parameter p in equation 62 governs the relative power of the energy versus geometrical scales. By choosing a value for this parameter, one can retrieve the k_T-algorithm ($p = 1$), Cambridge-Aachen algorithm ($p = 0$) or the anti-k_T algorithm

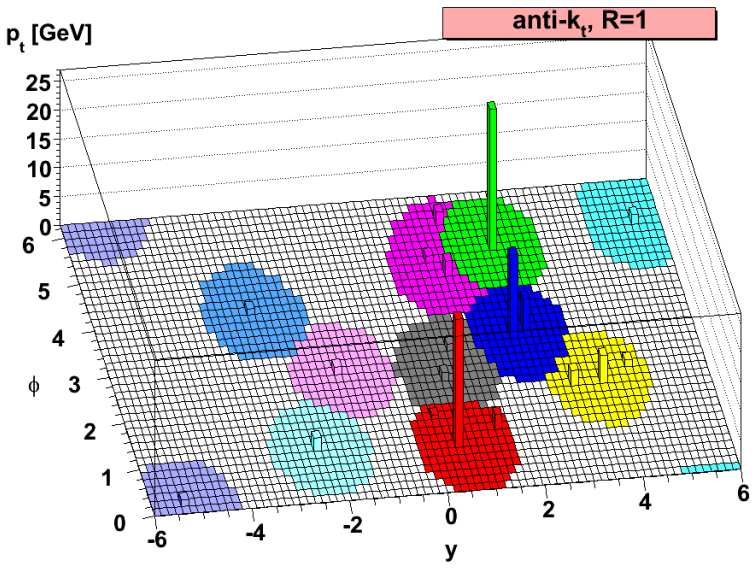


Figure 29: Illustration of jets obtained with the anti-k_T algorithm, showing the active areas in $\eta - \phi$ space which tend to be circular [81]

[81] for which $p = -1$. In this work, we will use jets clustered by anti-k_T algorithm in which the distance parameters reduce to

$$d_{ij} = \min\left(\frac{1}{p_{Ti}^2}, \frac{1}{p_{Tj}^2}\right)\frac{\Delta R_{ij}^2}{R^2} \tag{64}$$

$$d_{iB} = \frac{1}{p_{Ti}^2} \tag{65}$$

In this case the d_{ij} will be determined by the momentum of the hardest entity in the pair and the ΔR_{ij}. The jets will therefore grow around seeds of hard seeds, and soft particles will tend to cluster with hard ones before they cluster among themselves. As a result, the emission of particles do not modify the shape of the jet and the algorithm is infrared safe. Because the distance measure involves a combination of energy and angle, the jet grows collinear safe: collinear particles are combined together right at the beginning of the algorithm. If no hard neighbours are found within a distance of $2R$ of the seed, the jet will simply accumulate all soft particles within a circle of radius R, resulting in a perfectly conical jet. If another hard seed is within $R < \Delta R_{ij} < 2R$ there will be two hard jet, for which the hardest will be the most conical and the softer jet will miss the part overlapping with the hardest jet. If both hard particles are found within a distance R they will form one single jet.

5.5.2 *Jet types*

Within CMS, different approaches are used to reconstruct jets [82]:

- *Particle-flow jets* are reconstructed from all particle candidates produced by the PF algorithm, without energy threshold. In order to reduce the dependency on pile-up events, PF jets can be constructed with the Charged Hadron Subtraction (CHS) technique in which constituents originating from pile-up vertices (i.e. charged hadrons for which the track is associated to a PV other than the main PV) are discarded. In this thesis PF jets are used, clustered by the anti-k_T algorithm with distance parameter $R = 0.5$.

- *Calorimeter jets* are reconstructed from the energy deposits in the calorimeters, using towers formed by a combination of one or more HCAL cells and the geometrically corresponding ECAL crystals.

- *JPT jets* improve the measurement of the calorimeter jets by incorporating tracking information, according to the JPT algorithm [83]. Tracks are associated to each jet based on the separation between the jet axis and the track momentum vector, measured at the PV, in $\eta - \phi$ space. Tracks pointing within the jet cone at the calorimeter surface are used to correct the energy of the calorimeter deposits. The momentum of tracks bending out of the jet cone are simply added to the jet energy. The JPT algorithm corrects both the energy and direction of the calorimeter jet.

- *Track-jets* [84] are reconstructed from tracks of charged particles, measured in the tracker. By clustering only tracks which are associated to the same PV, track-jets do not have contributions from pile-up events. Furthermore, track-jets can go down to very low p_T jets and have an excellent angular resolution. In this thesis, a study with track-jets is described in 9.2.

5.5.3 *Jet energy scale corrections*

The raw energy of the reconstructed PF jet need to be corrected for additional energy deposits caused by pile-up interactions and electronic noise. On the other hand, the presence of material in front of the calorimeter or segmentation gaps between subdetectors could leave part of the jet energy unobserved.

Several levels of Jet Energy Scale (JES) [85] corrections are applied to the raw jet energy in order to obtain an energy that is closer to the energy of generator-level jets, i.e. jets reconstructed from generator particles with the same jet algorithm. In each level the four-momentum of the jet is scaled by a scale factor which depends on various jet related quantities (p_T, η, etc.).

5.5.3.1 *Offset correction*

The offset correction is applied to subtract the energy not associated with the main PV. This excess energy includes contributions from both in-time and out-of-time pile-up interactions, and from electronic noise in the calorimeters. The susceptibility of a jet to this diffuse noise is given by the jet area [86]. The jet area is calculated by adding a very large number of infinitely soft ghost particles to the event. The extent of the region in which these ghost particles are clustered to the jet defines the jet area A_j. Assuming a roughly uniform distribution of the diffuse noise, the change in transverse momentum of the jet will be proportional to its jet area. The diffuse noise is characterized by ρ, the average transverse momentum per unit area in $\eta - \phi$ space. Because pile-up is different for each event, ρ is calculated on an event-by-event basis. During Run I, $\rho_{k_T 6}$ was calculated [87] using PF jets clustered with a k_T-algorithm with $R = 0.6$. It is defined as

$$\rho_{k_T 6} = \text{median}\left[\frac{p_{T,j}}{A_j}\right] \tag{66}$$

where j runs over all jets. The k_T-algorithm has the tendency to naturally cluster a uniform background of soft particles into a large sample of soft pile-up jets, and is therefore an appropriate choice for the estimation of the diffuse noise. By taking the median, we ensure ρ is not biased by the presence of the hard jets.

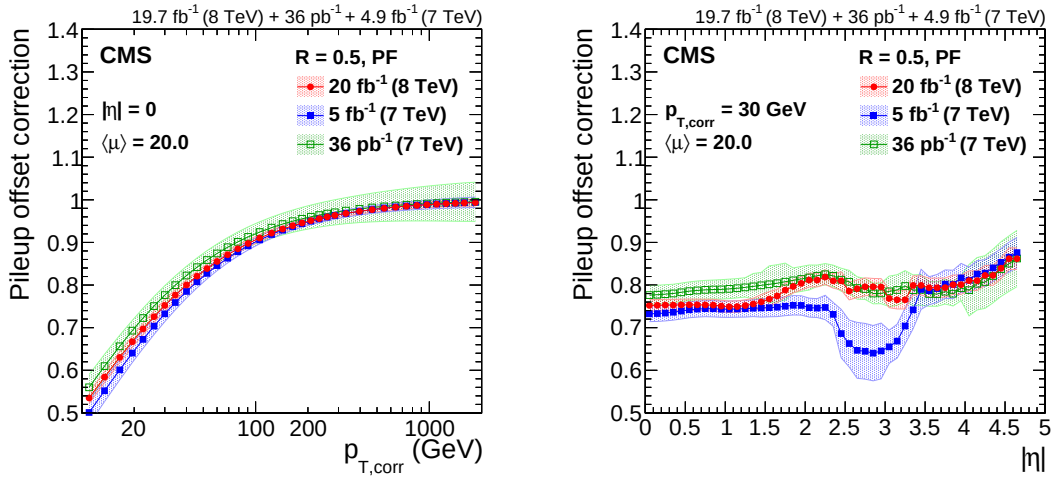
In Run II, ρ_{FG} will be calculated in a more efficient way directly from the PF candidates using a fixed grid in $\eta - \phi$ space. The average p_T is calculated for each block of $\eta - \phi$ in the grid, and the median is again taken as the estimate for the diffuse noise.

The basic *jet area method* describes an offset contribution which only depends on the area of the jet, and does not depend on the direction or transverse momentum of the jet. However, detector effects such as detection thresholds, noise and inefficiencies vary as a function of η and deposited energy, causing the offset to be non-uniform. Hence, CMS uses the *hybrid jet area method* instead of simply subtracting $\rho \cdot A_j$, for which the correction formula is given by[3]

$$p_T^{\text{offset}} = p_T^{\text{raw}} - [\rho_0(\eta) + \rho \cdot \beta(\eta) \cdot (1 + \gamma(\eta) \cdot \log p_T^{\text{raw}})] \cdot A_j \tag{67}$$

in which the parameters $\rho_0(\eta)$, $\beta(\eta)$ and $\gamma(\eta)$ are introduced for the required shaping of the offset versus η. These parameters are determined from a QCD multijet simulation, by reconstructing the same events with and without pile-up, and matching the reconstructed jets between these samples. An additional residual correction is applied on the data, estimated by comparing the offset energy in data to simulation. This is achieved by measuring the energy in a random placed cone in a randomly triggered event sample (a *zero-bias* sample), which has no energy deposits from hard interactions.

3 The logarithmic dependence on p_T quantified by the parameter $\gamma(\eta)$ was only added for the 8 TeV analyses



(a) Offset correction as a function of p_T

(b) Offset correction as a function of η

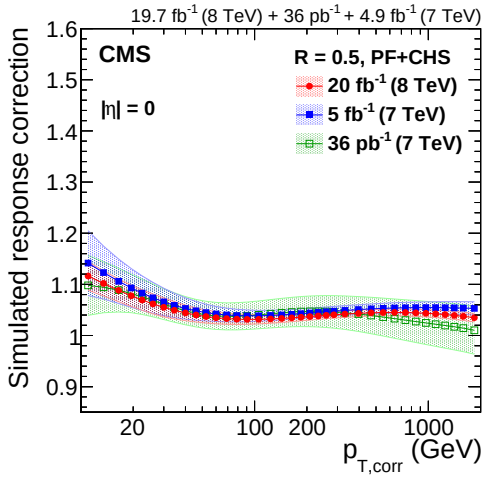
Figure 30: Pile-up offset corrections, with systematic uncertainty band, for AK5PF jets. The corrections are shown for the 2010 (green), 2011 (blue) and 2012 (red), extrapolated to the average 2012 pile-up conditions (an average of 20 pile-up interactions per event). [85]

5.5.3.2 *Simulation-based correction*

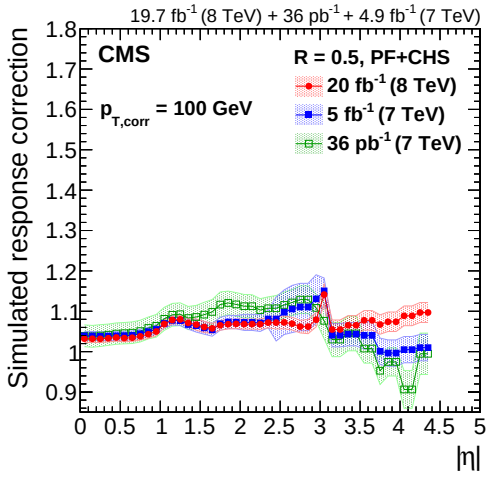
After removing the offset contribution, the energy of the jet needs further corrections for the non-uniform η and non-linear p_T response in the detector response. These effects are mostly mitigated by correction factors determined from studies on Monte-Carlo simulated jets, in which the energy of the reconstructed jets were matched with the energy of their corresponding generator-level jets. In this way, scale factors are derived in bins of η and p_T.

5.5.3.3 *Residual data-based correction*

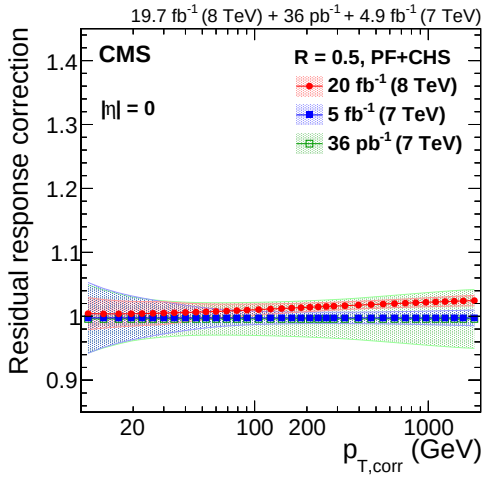
On top of the simulation-based correction, small residual corrections in η and p_T are applied only on data. The response of all jets in the event are corrected relatively to the jets in CMS barrel region ($|\eta| < 1.3$). The barrel region is chosen as the reference region because the detector is it is more uniform (leading to a smaller dependency on η) and a more precise response. In addition, it provides jets up to the highest transverse momenta. The corrections are determined using a high statistics dijet sample, in which one of the two jets is found in the barrel region and is used as the reference object, while the other jet is free to scan the whole detector. Because of momentum conservation, we can assume both jets to be balanced in the transverse plane. Deviations from this assumed p_T balance are used to derive corrections in bins of η.



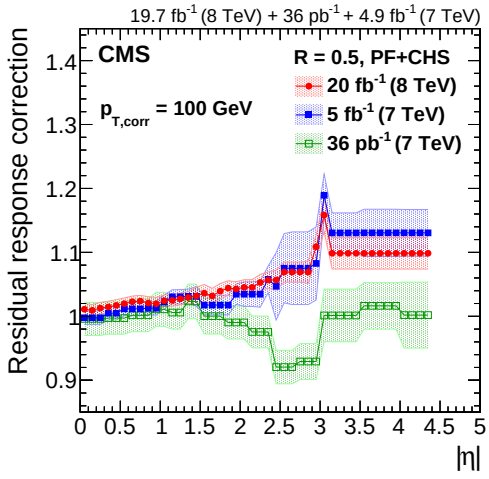
(a) Simulated response as a function of p_T



(b) Simulated response as a function of η



(c) Residual response as a function of p_T



(d) Residual response as a function of η

Figure 31: Simulated and residual response corrections, with systematic uncertainty band, for AK5PFCHS jets. The corrections are shown for the 2010 (green), 2011 (blue) and 2012 (red). [85]

The p_T scale is further corrected, using $\gamma + j$ and $Z + j$ samples for which the jet is found in the barrel region. Since the photons, electrons and muons can be measured much more accurately than jets, the p_T balance in those events can be used to derive an absolute jet energy scale correction. This correction is applied on all jets in data.

5.5.4 *Jet energy resolution*

The Jet Energy Resolution (JER) [85], i.e. the accuracy of the jet energy measurement, is relatively poor compared to the resolution of many other physics objects. Measurements of the JER are derived in a data-driven way from dijet and $\gamma/Z + j$ data events using the p_T balance of their corresponding final state objects.

Measurements have shown that the JER is worse for jets in data compared to the jets from MC simulation. The corrected transverse momenta of the jets in MC could be smeared in order to obtain a similar p_T resolution as found in data.

$$p_T^{\text{corr}} \rightarrow \max\left[0, p_T^{\text{gen}} + c \cdot \left(p_T^{\text{corr}} - p_T^{\text{gen}}\right)\right] \tag{68}$$

where c is the η-dependent core resolution scaling factor, i.e. the measured ratio between the resolution in data and simulation. The smearing is only applied to reconstructed jets which are well matched to a generator jet.

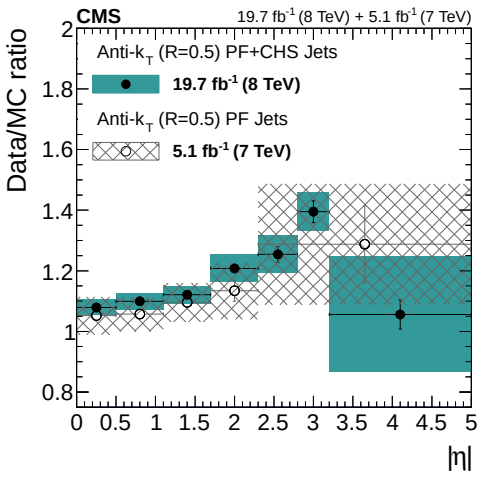


Figure 32: The JER data/MC scale factors as a function of η, shown for both 2011 and 2012 runs. [85]

5.5.5 *Jet identification*

For the 8 TeV analysis, jet identification criteria were applied in order to reject fake jets originating from calorimeter and readout electronics noise, while retaining the

vast majority of the jets ($> 99\%$). Jets passing the loose working point [88] for jet identification have

- at least two constituents
- a neutral hadron fraction of maximum 0.99, i.e. the energy deposited by the neutral particles in the HCAL is less than 99% of the total total raw energy
- a neutral electromagnetic fraction of maximum 0.99, i.e. the energy deposited by neutral particles in the ECAL is less than 99% of the total raw jet energy

Jets within the tracker acceptance ($|\eta| < 2.4$) are additionally required to have

- at least one charged particle
- a non-zero charged hadron fraction
- a charged electromagnetic fraction of maximum 0.99

For the quark-gluon jet discriminator studies described in section 6, the tight working point was used for which the maximum thresholds for the neutral hadron fraction, neutral electromagnetic fraction and charged electromagnetic fraction are lowered to 0.90.

6

QUARK-GLUON JET DISCRIMINATION

The tagging jets in VBF processes are always originating from quarks jets. On the other hand, the Zjj events from the DY background contain both quark and gluon jets. More generally, many physics analyses at the LHC are dealing with signal processes in which the jets are induced by quarks while backgrounds are often gluon-jet enriched. The ability to distinguish between those quark and gluon induced jets could significantly enhance the efficiency to extract the signal processes. One can exploit the substructure of a jet to separate between jets originating from quarks or gluons. In the past, these showering and fragmentation differences where measured at LEP [89, 90, 91, 92, 93] and the Tevatron collider [94, 95].

Quark-gluon jet discrimination tools have been studied in CMS since the start of Run I [96], originally focused only on the central region of the detector and for jets with high transverse momenta ($p_{\mathrm{T}} > 100$ GeV). The tool used for the EW Zjj measurement presented in this thesis, was the first one to extend quark-gluon jet discrimination to forward region and to soft jets down to $p_{\mathrm{T}} > 20$ GeV. This tool is a 5-variable likelihood algorithm, used in both $\sqrt{s} = 7$ and 8 TeV analyses. The same tool was also applied for the Higgs boson search in the VBF $H \to b\bar{b}$ channel [43]. Later on, a 3-variable likelihood was developed for general use within CMS, validated on data and documented in a public document [97]. This chapter starts with a general overview of the discriminating variables, structure of discrimination tools, and the estimation of the systematic uncertainty. It will also cover the subtle differences between the 5-variable and 3-variable likelihood, as well as some slight improvements to be expected for Run II.

6.1 OBJECT DEFINITION

The quark-gluon discriminator tools have been developed for use with PF jets reconstructed by the anti-k_T algorithm with distance parameter $R = 0.5$, for both CHS and non-CHS. This was the most generally used jet object definition within CMS for Run I. For Run II, the default configuration will take distance parameter $R = 0.4$, hence this cone size is used for the preliminary studies at $\sqrt{s} = 13$ TeV. Jet energy corrections have been applied as described in section 5.5.3. For simulation, the same jet clustering algorithm is applied to the stable generator particles in the event, thus defining generator jets. As noted in section 5.5.5, in this chapter we require jets to pass the tight jet identification criteria.

In order to develop and study a quark-gluon discriminator in simulation, one needs to assign a flavor to each reconstructed jet. We define the jet flavor with an empirical approach, in which the reconstructed jet is matched to a generator particle in the following way:

- if no generator jet is found within $\Delta R < 0.3$ from the reconstructed jet, we consider the jet as a pile-up jet
- if a generator jet is present, the reconstructed jet is matched to a generator parton[1] in $\Delta R < 0.3$; if multiple partons are found, the parton closest in ΔR is chosen.
- if no parton is found, the jet is considered 'undefined'.
- the flavor of the matched parton is assigned to the jet

The quark-gluon discriminator tool will in the first place aim to discriminate between gluons and light quark-jets (u,d or s). The hadronization properties of c and b-jets are different from light-quark jets, and behave more gluon-like, especially for jets with low transverse momentum.

6.2 CLASSIFIER VARIABLES

6.2.1 *Description of the variables*

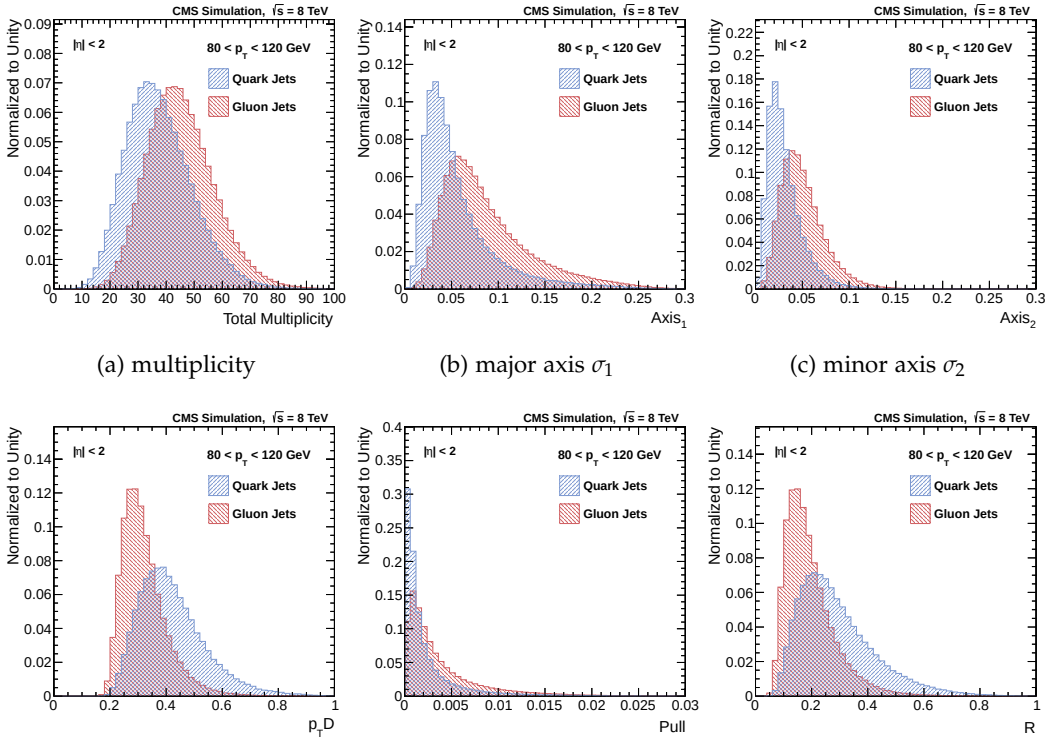
6.2.1.1 *Multiplicity*

The simplest and best studied variable one can use to discriminate quark and gluon jets, is the jet multiplicity. From the $SU(3)_c$ generators, one can calculate the color factors C_A, C_F and T_R given by

$$\begin{aligned} \delta_{ik} C_F &= t^a_{ij} t^a_{jk} && \Rightarrow C_F = \frac{N_c^2 - 1}{2N_c} = \frac{4}{3} \\ \delta^{ac} C_A &= f^{acd} f^{bcd} && \Rightarrow C_A = N_c = 3 \\ \delta^{ab} T_R &= t^a_{ij} t^b_{ij} && \Rightarrow T_R = \frac{N_f}{2} \end{aligned} \tag{69}$$

which are proportional to the probability of a quark emitting a gluon $q \to qg$, a gluon emitting a gluon $g \to gg$ and a gluon splitting into two quarks $g \to$

1 Only outgoing particles of the hardest subprocess in the event are considered, i.e. before they experience fragmentation. This corresponds with status code 3 in PYTHIA 6, status code 23 particles in PYTHIA 8, and status code 2 in HERWIG++.



(a) multiplicity (b) major axis σ_1 (c) minor axis σ_2

(d) $p_T D$ (e) pull $|\vec{t}|$ (f) R

Figure 33: Normalized distributions of the considered variables for quark-gluon jet discrimination

$q\bar{q}$ respectively. Neglecting the splitting of the gluons, we can expect the ratio of multiplicities in quark and gluon jets approaching the color factor ratio [98]

$$\frac{< n >_g}{< n >_q} = \frac{C_A}{C_F} = \frac{9}{4} \tag{70}$$

for jets with high transverse momentum. At lower energies, this ratio will be lower due to non-perturbative effects and to gluon jets emitting more particles at larger angles, outside the cone of the jet algorithm.

6.2.1.2 *Jet shapes*

Because the gluon jets have higher multiplicities, the energy spectrum of the jet constituents is expected to be softer. The mean transverse energy of these soft particles relative to the jet axis is expected to be similar for both quark and gluon jets. This results in larger angles with respect to the jet axis for the constituents in the gluon jets. This means quark jets should produce, on average, narrower jets with respect to gluon jets of the same p_T. Jets have a conical structure that can be projected in the η-ϕ plane, and we can approximate its shape by an ellipse

which is characterized by the major and minor axis in the plane. These axes can be calculated using a 2x2 symmetric matrix

$$M = \begin{pmatrix} \sum_i p_{T,i}^2 \Delta\eta_i^2 & \sum_i p_{T,i}^2 \Delta\eta_i \Delta\phi_i \\ \sum_i p_{T,i}^2 \Delta\eta_i \Delta\phi_i & \sum_i p_{T,i}^2 \Delta\phi_i^2 \end{pmatrix} \tag{71}$$

where the sums extend over the jet constituents which have transverse momenta $p_{T,i}$. The $\Delta\eta_i$ and $\Delta\phi_i$ are the pseudorapidity and azimuthal distances with respect to the $p_{T,i}^2$ weighted direction of the constituents in η-ϕ space.

The major (σ_1) and minor (σ_2) axes are then computed using the eigenvalues $\lambda_{1,2}$ of the matrix M:

$$\sigma_1 = \sqrt{\frac{\lambda_1}{\sum_i p_{T,i}^2}} \qquad \sigma_2 = \sqrt{\frac{\lambda_2}{\sum_i p_{T,i}^2}} \tag{72}$$

One may also define an average width of the jet, by taking the quadratic mean of the two axes:

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} \tag{73}$$

It is also possible to construct similar variables using a different weight, i.e. one could replace the $p_{T,i}^2$ by another power of $p_{T,i}$ in the formulas above. In the CMS quark-gluon taggers it is common to use the squared transverse momentum as it has been found to exhibit the most optimal separation between quark and gluon jets.

6.2.1.3 $p_{\mathrm{T}}D$

The energy spectrum of the constituents can be expressed with the $p_{\mathrm{T}}D$ variable, defined as

$$p_{\mathrm{T}}D = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}} \tag{74}$$

where the sum runs over the jet constituents. A jet in which one particle carries most of the total momentum will have $p_{\mathrm{T}}D \to 1$, while jets composed out of a large number of soft particles will tend to $p_{\mathrm{T}}D$ values close to zero. Because gluon induced jets have a softer energy spectrum than quark induced jets, we can expect on average lower $p_{\mathrm{T}}D$ values for gluon induced jets.

6.2.1.4 *R*

The R variable is defined as the maximum transverse momentum fraction carried by a single constituent in the jet:

$$R = \frac{\max(p_{T,i})}{\sum_i p_{T,i}} \tag{75}$$

It has similar properties as the $p_T D$ variable, i.e. quark jets are more likely to have R values approaching unity.

6.2.1.5 *Jet pull*

The jet pull $\vec{t}$ expresses the asymmetry of the constituents within the jet and is defined as

$$\vec{t} = \frac{\sum_i p_{T,i}^2 |r_i| \vec{r}_i}{\sum_i p_{T,i}^2} \tag{76}$$

with $\vec{r}_i = (\Delta\eta_i, \Delta\phi_i)$ the difference between the jet axis (i.e. the sum of the four-momenta of the jet constituents) and the location of the particle with transverse momentum $p_{T,i}$. The pull variable was originally designed [99] to measure the color connection between neighbouring jets: if the jets are color-connected they tend to shower in the direction of the other jet. For the application in quark-gluon discrimination, only the absolute value $|\vec{t}|$ of the pull is considered. Due to the stronger color connection, gluon jets are expected to have larger pull values for a given p_T.

6.2.1.6 *Fractal fitting variables*

Recently, a new set of parameters based on the fractal nature of hadronic jets is proposed [100] for the Run II quark-gluon discriminator. The Branching Logarithmic Fit (BLF) parameters characterize the branching structure of a jet and it has the advantage of being very weakly correlated to the other variables. Even though these new variables seem promising at generator level, preliminary studies on reconstruction level show only very minimal improvement could be achieved. Due to their complexity, more studies on these BLF parameters are needed and they will not be included in the first versions of a Run II quark-gluon tagger.

6.2.2 *Choice of variables*

The performance of a classifier variable can be illustrated using a Receiver Operating Characteristic (ROC) curve. When selecting jets in a simulation sample by using

a cut on the classifier variable, one can define a quark jet efficiency, i.e. the fraction of quark jets passing the threshold, and a gluon jet rejection quantifying the fraction of gluon jets not passing that same threshold. Varying the threshold yields a ROC curve showing the quark jet efficiency as a function of the gluon jet rejection. As we want to reject as much gluon jets as possible while retaining most of the quark jets, large ROC integral values are associated with better performance.

Figure 34 shows the ROC curve for the studied classifier variables in $\sqrt{s} = 8$ TeV simulation. A few conclusions can be made:

- in the central region, where one can split the multiplicity into its charged and neutral components, the charged multiplicity significantly outperforms the neutral multiplicity; as expected, the multiplicity is the most discriminating variable for the higher p_T jets
- out of the jet shape variables, the minor axis performs better than the major axis, with the average width somewhere in between
- the pull variable does not provide much discrimination power
- out of the two variables accessing the energy spectrum information of the jet (R and $p_T D$), $p_T D$ has the best discrimination power

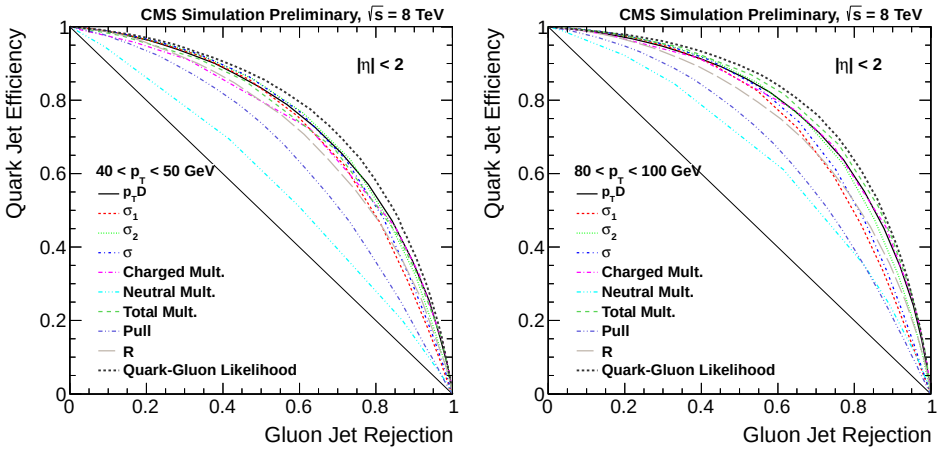
The 5-variable likelihood, developed before the $p_T D$ was studied, was build out of the charged ($|\eta| < 2$) or total ($|\eta| > 2$) multiplicity, σ_1, σ_2, R and the pull variable. Using the observations made on the single-variable performance, a new set of variables was defined for the 3-variable likelihood: the total multiplicity, σ_2 and $p_T D$. This is the minimal set of reasonably uncorrelated variables with the highest single-variable discriminator power.

A better discrimination power and stability to pile-up has been found by adding some restrictions to the PF candidates used as an input for the variables. Charged particles are only considered when they have a high-purity track associated to the main PV. In addition the track impact parameters need to satisfy $|d_z/\sigma(d_z)| < 5$ and $|d_0/\sigma(d_0)| < 3$. Neutral particles are required[2] to have a transverse momentum greater than 1 GeV.

6.3 CLASSIFIER ALGORITHMS

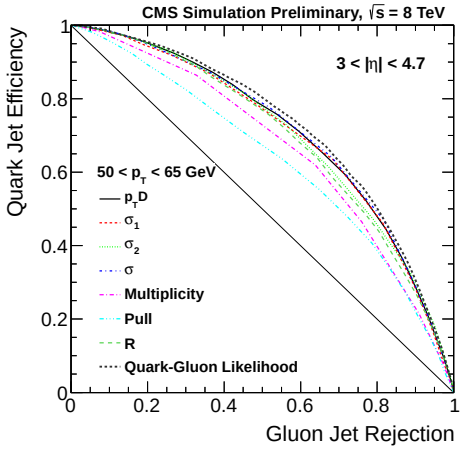
The information in the selected variables need to be combined in order to achieve the optimal discrimination. In other words, we want to map the input variables

2 Although this requirement was originally only applied for the multiplicity calculation during Run I, the studies with $\sqrt{s} = 13$ TeV simulation have shown it also improves the performance of the other variables. For this reason, it will be applied in the calculation of all variables in Run II.



(a) central, $40 < p_{\mathrm{T}} < 50$ GeV

(b) central, $80 < p_{\mathrm{T}} < 100$ GeV



(c) forward, $50 < p_{\mathrm{T}} < 65$ GeV

Figure 34: Single-variable performance comparison for quark-gluon discriminating variables. The ROC curves are shown for the above discussed variables, as well as for the 3-variable quark-gluon likelihood discriminant. The curves are shown for two p_{T} bins in the central region, and one p_{T} bin in the forward region

to one classifier variable, yielding a ROC integral which is larger than those from each of the individual input variables. This can be achieved by multi-variate techniques which belong to the family of supervised learning algorithms: these algorithms take a training sample[3], for which the desired output (i.e. quark or gluon) is known, as input and uses its information to built the mapping algorithm. After the training phase, the algorithm can be used to classify jets in data and simulation. The performance and advantages of different algorithms were studied and are described in this section. The categorization of the discriminant in bins of p_T, η and ρ is described in the next section.

6.3.1 *Likelihood discriminant*

The likelihood discriminant, more commonly known as the *naive-Bayes classifier*, is one of simplest and most transparent techniques available. It uses Probability Density Functions (PDF's), which are constructed for each of the input variables and for separately for light-quark and gluon jets. The PDF's are derived from the training sample in the form of histograms. and normalized to unity. For each jet, a set of input variables $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ is defined. The global PDF's for gluons (G) and light-quarks (Q) are then simply defined by the product of each variable's quark or gluon PDF (f_Q^i and f_G^i) computed at the given value x_i:

$$G(\mathbf{x}) = \prod_i f_G^i(x_i) \qquad\qquad Q(\mathbf{x}) = \prod_i f_Q^i(x_i) \tag{77}$$

These are used to construct the likelihood estimator

$$L(\mathbf{x}) = \frac{Q(\mathbf{x})}{Q(\mathbf{x}) + G(\mathbf{x})} \tag{78}$$

which can be interpreted as the probability of a jet to be originated from a quark parton. The likelihood algorithm has the advantage of being simple, transparent and fast. It results in an output value between 0 (gluon jets) and 1 (quark jets), where events naturally cluster towards these extreme values. The 5-variable likelihood was trained within the TMVA framework [101], while the 3-variable likelihood algorithm was implemented independently and optimized for direct use in CMSSW.

The performance of the likelihood algorithm relies on the accuracy of the underlying PDF's. If a large number of training events are available, one can allow small bin widths for the PDF, resulting in the ideal behaviour of a smooth PDF. Unfortunately, the statistics in the training sample are limited, which is even more the case when dividing the sample in categories (see section 6.4), as some categories (e.g. in the forward region or high p_T) are less populated than others. Statistical

3 In this study, the discriminators were trained on a simulated QCD sample for which the flavor was assigned as described above

fluctuations could therefore appear in the PDF's constructed from the training sample, resulting in overtraining. When measured on the training sample, overtraining leads to a seeming increase of classification performance, but will actually result in a decrease of performance when measured on an independent test sample. Two approaches can be followed to mitigate the effect of overtraining:

- *Smoothing* of the PDF's, which is applied within the TMVA framework for the 5-variable likelihood. In our case, a Kernel Density Estimator (KDE) smoothing with Gaussian kernel was used to smear the bins in the PDF. Smoothing algorithms allow to keep the same fine binning in the low-statistics categories as the one used in high-statistics categories. The smoothing was applied on all categories, even though there is a small risk of oversmoothing narrow structures in the PDF for the high-statistics categories.

- *Rebinning* of the low-statistics PDF's, applied in the 3-variable likelihood. Originally, only a global rebinning factor was chosen based on the number of entries in the histogram. This procedure is optimized for the Run II tagger, where the global rebinning factor is now chosen based on the number of empty bins and the scale of fluctuations in the histogram. In addition, bins are locally merged in the tails of the distribution in order to avoid left-over empty bins[4].

The biggest drawback of the likelihood discriminant is that it does not take into account the correlations between the input variables: as can be seen in equation 77 the contributions from each variable are simply multiplied. If two variables are highly correlated, their mutual information will enter twice in the product of equation 77. As a result, more importance is given to this mutual information with respect to other more independent information. This especially leads to problematic behaviour when the former is less discriminating than the latter. In such cases the likelihood output gets biased towards the mutual information, and does not reach its optimal performance. For this reason, the number of input variables was low for the developed likelihood algorithms: adding lower-performing variables would increase the chance of deteriorating the global likelihood output, instead of improving the performance. However, even with only three variables, this deteriorating effect could not be completely avoided: at high p_{T} the multiplicity variable significantly outperforms the other variables, and other the other variables can easily deteriorate its performance. Hence, in the high p_{T} categories the performance of the standard likelihood algorithm is found to be worse than the multiplicity alone.

4 Empty bins need to be avoided as $f_Q^i = 0$ ($f_G^i = 0$) results in an extreme value of 0 (1) in the likelihood output

6.3.2 *Variants on the standard likelihood discriminant*

During the studies for the Run II tagger, a couple of variants of the likelihood algorithm have been studied and compared to the original:

- A greater stability with respect to overtraining can be achieved by constructing the likelihood algorithm using Cumulative Distribution Functions (CDF's) instead of PDF. This variant starts by defining both the light-quark and gluon CDF and their complements by integrating the f^i distributions:

$$F_G^i(x_i) = \int_{-\infty}^{x_i} f_G^i(x_i)dx \qquad F_Q^i(x_i) = \int_{-\infty}^{x_i} f_Q^i(x_i)dx$$
$$\bar{F}_G^i(x_i) = \int_{x_i}^{+\infty} f_G^i(x_i)dx \qquad \bar{F}_Q^i(x_i) = \int_{x_i}^{+\infty} f_Q^i(x_i)dx \tag{79}$$

  A single-variable CDF-likelihood, for which the variable has typically lower values for quark jets than for gluon jets (e.g. the multiplicity), is then constructed as

$$G^i(x_i) = \frac{\bar{F}_G^i(x_i) - \bar{F}_Q^i(x_i)}{\bar{F}_G^i(x_i) + \bar{F}_Q^i(x_i)}$$
$$Q^i(x_i) = \frac{F_Q^i(x_i) - F_G^i(x_i)}{F_G^i(x_i) + F_Q^i(x_i)}$$
$$L_{CDF}^i(x_i) = \frac{Q^i(x_i)}{Q^i(x_i) + G^i(x_i)} \tag{80}$$

  If the variable has typically higher values for quark jets than for gluon jets (e.g. $p_T D$), then the CDF's and their complements should be interchanged in equation 80. Finally, the global CDF-likelihood is retrieved as

$$L_{CDF}(\mathbf{x}) = \frac{\prod_i Q^i(x_i)}{\prod_i Q^i(x_i) + \prod_i G^i(x_i)} \tag{81}$$

  Because the CDF-likelihood has more stable for fluctuations in the PDF, it allows for a much finer binning. It therefore performs better in few categories with very low statistics. On the other hand, in the high p_T region the CDF-likelihood is performing worse than the standard likelihood, revealing a higher susceptibility to the deteriorating effect of correlations. For this reason, and because of its more complex implementation, the CDF-likelihood is not further studied.

- One approach to counter the deteriorating effect of correlations, it to transform the input variables into a new set of less correlated variables in which linear correlations are removed. The linear correlation coefficients between two variables X and Y are defined as the ratio between their covariance and the product of their standard deviations:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\langle (X - \mu_X)(Y - \mu_Y) \rangle}{\sigma_X \sigma_Y} \tag{82}$$

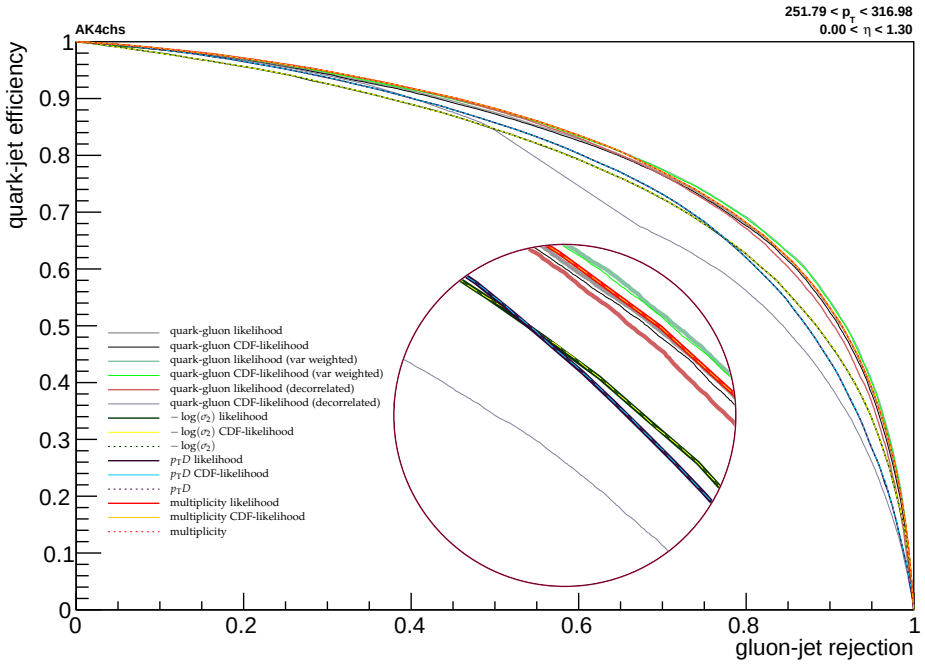
The goal is now to find a transformation, such that the new set of variables has no linear correlations and its correlation matrix becomes diagonal. For this reason, the covariance matrix $C_{ij} = \mathrm{cov}(X_i, X_j)$ is introduced. The linear correlations in a tuple $\mathbf{x}$ could then be removed by the transformation $\mathbf{x} \mapsto C^{-1/2}\mathbf{x}$. As expected, a modest gain in performance with respect to the standard likelihood is seen in the high p_T region, while there is no gain in the low p_T categories. In fact, in this low p_T region the performance is even worse than the standard likelihood, probably due to neglecting the more complicated non-linear correlations.

- Another approach to tackle the correlations problem, is to shift the power balance between the input variables by introducing weights in equation 77

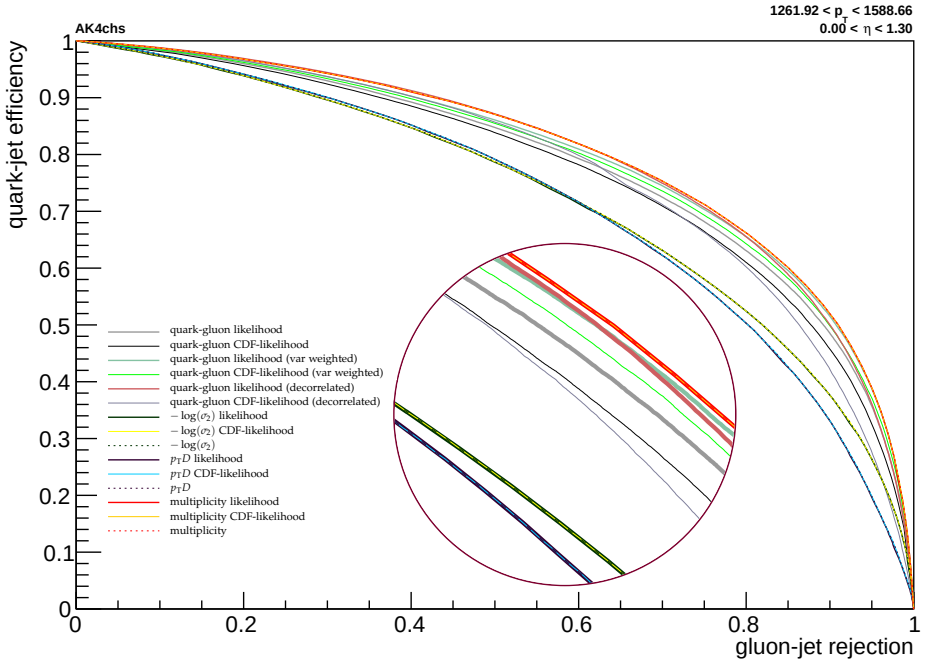
$$G(\mathbf{x}) = \prod_i (f_G^i(x_i))^{w_i} \qquad Q(\mathbf{x}) = \prod_i (f_Q^i(x_i))^{w_i} \tag{83}$$

This easy modification allows to improve on the ROC integral in the high p_T categories, by increasing the weight of the multiplicity variable while lowering the weights of the other variables, keeping $\sum_i w_i = N$. This modification has the biggest impact on jets for which the single-variable likelihood of the variables strongly disagree with each other. On the other hand, if the single-variable likelihood values are more or less in agreement, the global likelihood will not much differ of the one given by the standard algorithm (in which $w_i = 1$). This approach seems to be the best choice for improving the performance in the high p_T region: its easy implementation allows to gradually shift from the standard algorithm in the lower p_T categories towards a more asymmetric power balance in the high p_T categories. There is however some tuning needed to find the optimal weights for each p_T category.

Figure 35 shows the ROC curves for the different modifications of the likelihood algorithm in two categories with central jets ($\eta < 1.3$) with their p_T around 250-316 GeV and 1262-1589 GeV respectively. When focusing on the single-variable curves, no difference is seen between the standard likelihood, the CDF-likelihood and the ROC calculated on the variable itself. Hence we can conclude statistics is sufficient in these categories, and the algorithms are stable for overtraining. Above 250 GeV, the performance of the multiplicity starts to overtake the standard likelihood algorithm, which becomes even more pronounced for the very high p_T jets. By shifting the power balance to the multiplicity using the modification with weights, a clear gain with respect to the standard algorithm is seen in both categories. The approach with decorrelated variables has also a significant performance increase with respect to the standard algorithm, but its performance is worse in the lower p_T category.



(a)



(b)

Figure 35: Performance comparison of the various likelihood modifications in the high p_T region. The preliminary study is done using a simulated QCD sample at $\sqrt{s} = 13$ TeV (with preliminary Run II reconstruction), independent from the training sample. The ROC curves are shown for both the 3-variable likelihood and CDF-likelihood, and their variations with weighted or decorrelated variables. The ROC curves for each of the input variables are also shown, which are the same as the ROC curves for their respective single-variable (CDF-)likelihood.

6.3.3 *Other multi-variate techniques*

In addition to the likelihood algorithm, more advanced machine learning techniques were considered during the development of the 3-variable likelihood. In particular, the Multilayer Perceptron (MLP) method, an artificial neural network provided within the TMVA framework, was used to construct a MLP discriminant. Because the MLP method is able to treat correlations, the MLP performed better for high p_T jets compared with the 3-variable likelihood. A drawback with the MLP approach are the different shapes and ranges of the response distributions in the various categories, as opposed to the simple $[0, 1]$ output in the likelihood algorithm. Hence an additional transformation to a classification probability is needed, in order to have a discriminating value which can be unambiguously interpreted in the same way for all categories. Due to this more complex and slower implementation, the MLP discriminant was dropped in favour of the more transparent likelihood discriminant.

6.4 CATEGORIZATION

There are important reasons to train the likelihood discriminant in separate categories, all of them related to *confounding variables*: variables which are correlated to the input variables, but at the same time have different distributions for quark and gluon jets. One such example of a confounding variable is the jet transverse momentum which is very much correlated to our input variables. If we would group jets of a too large p_T range within one category, the quark and gluon PDF distributions would be smeared out more compared to a categorization with a narrow p_T range. The higher overlap between the PDF's would reduce the discrimination power of the variables. Even worse, in QCD samples, quarks jets have typically a higher p_T spectrum, while gluon jets are more represented in the low p_T region. Because of this, also the gap between the average quark and gluon multiplicity will diminish, as the quark PDF would be more dominated by high p_T jets than the gluon PDF. In other words, categorization helps us to narrow and separate the PDF distributions at the same time, increasing the performance significantly.

One could naively suggest to include these confounding variables in the set of output variables, and use another machine learning technique which is able to deal better with correlations. This would be the right thing to do in most high energy physics analyses, where the MC samples used for training and testing simulate the same physics process as the MC sample used in analysis. The quark-gluon jet discrimination case is different in this respect, as we typically train on a QCD sample while the tool is applied in a wide range of physics analyses, using MC samples simulating very different processes. These different processes could have very different distributions for the confounding variables. For this reason, we should avoid

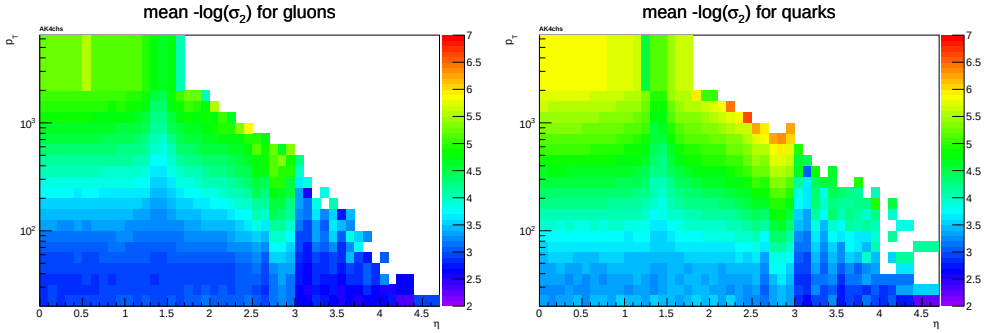
every bias from the confounding variables in the QCD sample, something which is best solved by categorization.

6.4.1 *Jet transverse momentum and pseudorapidity*

The two most obvious confounding variables are the jet transverse momentum and the pseudorapidity. The dependencies of the input variables on them are presented in Figure 36, which show the importance of categorization on these variables. The multiplicity variable has the strongest dependence on the jet p_T, the $p_T D$ variable is more stable. Narrowing the p_T categories has therefore the most significant impact on the multiplicity variable. The dependencies along the η-axis are more complicated: while the variables are reasonably stable in the barrel region of the detector, they show heavy dependencies in the endcap region. In addition, the transitions between HB and HE, and between HE and HF are easily seen: jets in these transition regions have lower multiplicities, smaller width and higher $p_T D$ values.

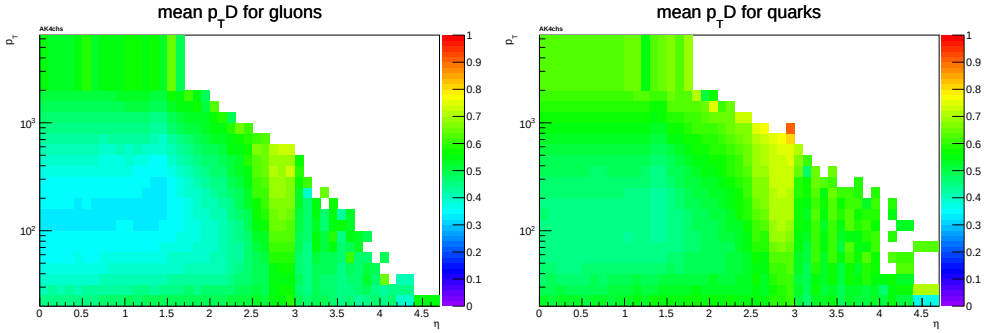
The 5-variable likelihood used 7 p_T categories between 30 and 600 GeV with divisions at 50, 80, 120, 170, 300 and 470 GeV. The likelihood output for higher p_T jets was assigned using the PDF's in the highest category. To further control for the p_T dependence, the likelihood output is retrieved from the categories which have their average p_T just above and below the tagging jet p_T. The final likelihood discriminant value is then the linear interpolation of these two outputs. The 3-variable likelihood dropped the interpolation in favor of a narrower categorization: 20 categories which span in logarithmic spacing between 20 and 2000 GeV, and an additional category for jets above 2000 GeV.

For the η categorization, different scenarios have been in use for the various tools. The 5-variable likelihood uses three categories: a central ($|\eta| < 2$), transition ($2 < |\eta| < 3$) and forward ($3 < |\eta| < 4.7$) region. For the 3-variable likelihood, the PDF's were only retrieved in the central ($|\eta| < 2$) and forward ($3 < |\eta| < 4.7$) region; the PDF's from the central region were applied up to $|\eta| < 2.5$, jets between $2.5 < |\eta| < 3$ used the PDF's from the forward region. As can be seen in Figure 37, the broad η categories were sufficient for the HB region were our input variables are very stable against η, resulting in a homogeneous response for the likelihood output. Also the HF region can be treated as one category, although there could be room for improvement if MC samples with higher statistics in the forward region would become available. The HE region, and its transitions with the HB and HF, are not treated well by the two categories scenario: this region is too different from the central and forward regions and need a separate PDF training on its own. In this respect, the 5-variable likelihood, having a third category, was a much better choice for VBF-type analyses, in which jets are typically found in the HF region. For the Run II tool, we propose an even finer categorization in η, resulting in a more homogeneous response of the likelihood output in the whole p_T-η grid.



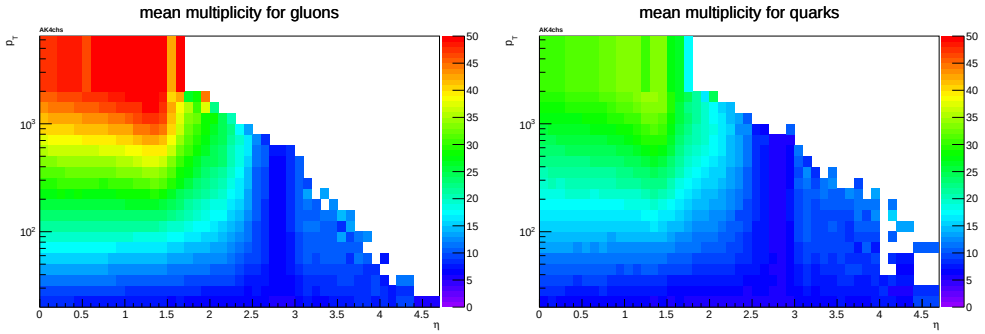
(a) mean $-\log(\sigma_2)$ for gluons

(b) mean $-\log(\sigma_2)$ for quarks



(c) mean $p_{\mathrm{T}}D$ for gluons

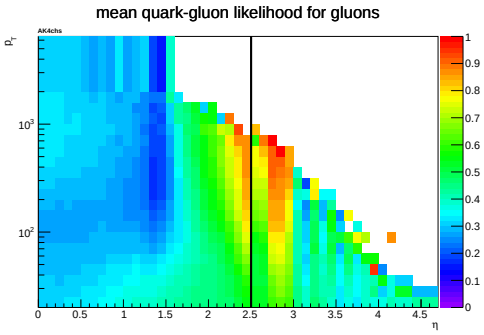
(d) mean $p_{\mathrm{T}}D$ for quarks



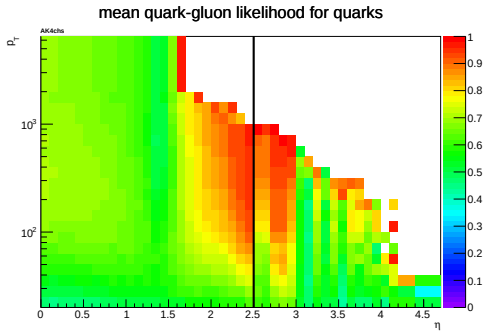
(e) mean multiplicity for gluons

(f) mean multiplicity for quarks

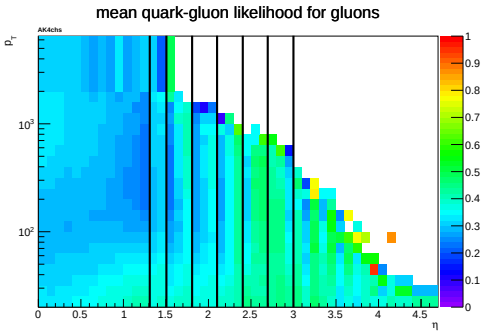
Figure 36: Dependencies of discriminating variables on jet transverse momentum and pseudorapidity, shown for a $\sqrt{s} = 13$ TeV QCD simulation, with preliminary Run II reconstruction.



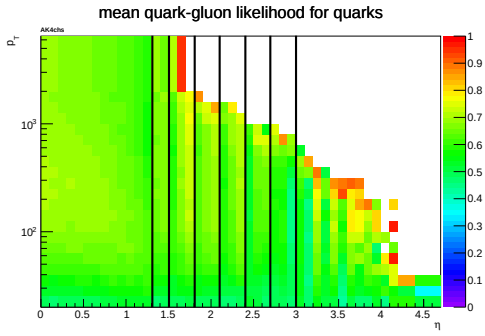
(a) mean likelihood output value for gluons, using two η categories



(b) mean likelihood output value for quarks, using two η categories

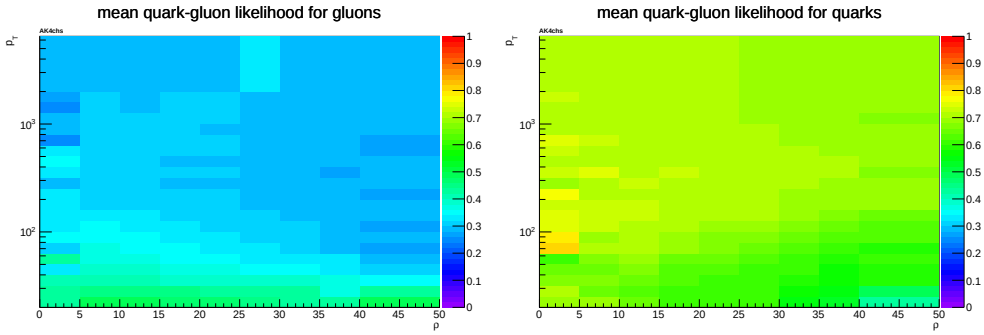


(c) mean likelihood output value for gluons, using eight η categories



(d) mean likelihood output value for quarks, using eight η categories

Figure 37: Effect of η categorisation on the quark-gluon likelihood. The mean likelihood output values as a function of p_T and η are shown for gluons (left) and quarks (right) using two different η categorisations: two categories as used in the 3-variable likelihood during Run I (above) and eight categories as a proposal for the Run II tool (below).



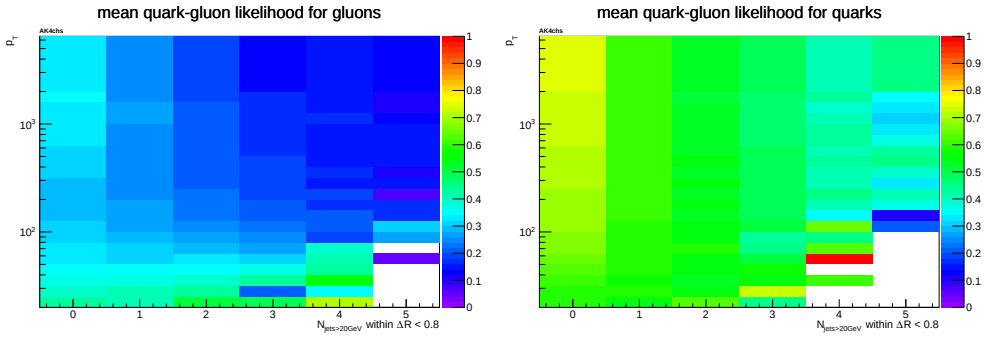
(a) mean likelihood for gluons (b) mean likelihood for quarks

Figure 38: Dependency of the quark-gluon likelihood on pile-up: the mean quark-gluon likelihood output, calculated without categorization or correction on ρ_{FG}, is shown as a function of p_T and ρ_{FG} for gluons (left) and quarks (right).

6.4.2 *Pile-up*

Also the amount of pile-up activity has an effect on the likelihood output: spurious pile-up particles could enter the multiplicity and change the energy sharing among the constituents in the jet. Even though this effect is partially mitigated by requiring the charged PF candidates to originate from the main PV, the neutral component is able to shift the variables to more gluon-like values in case of high pile-up activity. In order to control for pile-up we can use the same variable as used in the JES offset correction: the diffuse noise parameter ρ. The 3-variable likelihood used 45 categories in ρ_{k_T6} linear spaced between 0 and 45 GeV, plus an additional final category for higher values of ρ_{k_T6}. The 5-variable likelihood used another approach to correct for the pile-up dependence: for each variable the dependence on ρ_{k_T6} was measured by a linear fit, from which correction factors were extracted to apply on the variables. By applying the correction factors both in training and application, categorization was avoided.

The preliminary Run II studies have, however, shown the variables are very stable with respect to ρ_{k_T6} and ρ_{FG}, especially for high p_T jets in the central part of the detector. This is shown in Figure 38, where even in the absence of categorizations or corrections on ρ_{FG} a very homogeneous response of the mean likelihood output is seen. The slight dependency is small compared to separation between quarks and gluons in the variable distributions, and are therefore negligible. A light categorization might still be useful, especially for lower p_T jets and in the forward region, but a few categories would already be sufficient. In this way, more training statistics are available within each category, allowing to apply the finer η categorization proposed for Run II.



(a) mean likelihood for gluons (b) mean likelihood for quarks

Figure 39: Dependency of the quark-gluon likelihood on associated jets: the mean quark-gluon likelihood output (without categorization on associated jets) is shown as a function of p_T and the number of associated jets for both gluons (left) and quarks (right).

6.4.3 *Associated jets*

Recently, it was proposed [102] to include information of softer reconstructed jets constructed around the hard jet. We therefore construct the associated jet rate, the number of jets with $p_T > 20$ GeV, found within a cone of $\Delta R < 0.8$. During the fragmentation process, partons emitted under large angles could be reconstructed as an independent jet by the jet algorithm. As gluon jets are wider, they are more likely to be surrounded by one or more associated jets. In addition, the number of associated jets could also be influenced by the physics process, for example in processes with a boosted H, Z, W,... which decay into two jets ending up close to each other in $\eta - \phi$ space.

The dependence of the mean quark-gluon likelihood as a function of the number of associated jets n is shown in Figure 39: jets with one or more associated jets are more likely to have larger multiplicities, larger width and lower $p_T D$, resulting in more gluon-like values for these jets. The dependency on associated jets is more strongly compared to what we have see for the dependency on the pile-up contribution. If a categorization would be introduced for the Run II tool, it would need to restrict to only two categories ($n =$ and $n \geq 1$) as jets with higher number of associated jets are rare, and $n \geq 2$ categories would therefore not be enough populated to build a decent PDF. Outside the central region, also the $n \geq 1$ would not have sufficient statistics, restricting a possible additional jet categorization to the most central η category.

	5-variable likelihood	3-variable likelihood
first documented in	FSQ12019 [47]	JME13002 [97]
algorithm	likelihood (TMVA)	likelihood
input variables	charged multiplicity ($\|\eta\| < 2$)	multiplicity
	multiplicity ($\|\eta\| > 2$)	σ_2
	σ_1	$p_T D$
	σ_2	
	R	
	$\|\vec{t}\|$	
p_T categories	8 + interpolation	20
η categories	3	2
ρ correction	by linear fit	46 categories

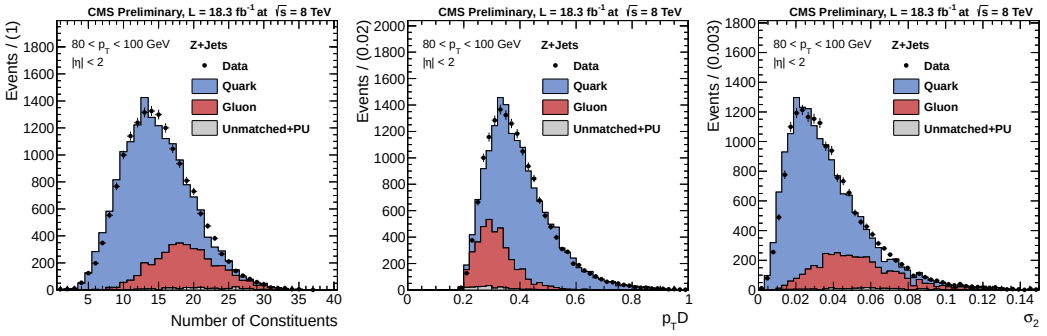
Table 4: Comparison of Run I quark-gluon jet discriminator tools

6.5 VALIDATION IN DATA

The studies on the input variables, multi-variate algorithms and categorization have all been performed on MC simulation. In order to check for possible MC-data discrepancies, the performance of the 3-variable likelihood has been validated on 8 TeV collision data. This is achieved by identifying two control samples, a Z+jets data sample which is expected to offer a relatively pure sample of quark jets, and a dijet data sample which is more gluon-enriched.

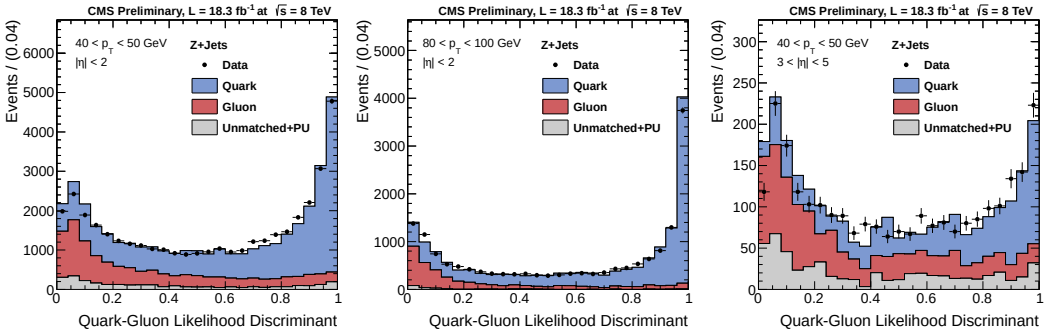
6.5.1 *Validation on Z+jet events*

For this validation study, Z+jet events in the dimuon channel were selected in the 2012 dataset, for a total integrated luminosity of 18fb^{-1}. The reconstructed dataset contained events passing the dimuon HLT path with respective thresholds of 17 GeV and 8 GeV on the p_T of each of the two muons. The muons were further required to fulfill the tight identification criteria and to have p_T greater than 20 and 10 GeV respectively. Z-bosons were identified by two oppositely charged muons with dimuon invariant mass in the 70-110 GeV range, i.e. within about 20 GeV of the nominal Z-boson mass. The leading jet needs to fulfill the tight jet identification criteria, and should fail for loose b-tagging [103] and pile-up jet identification [104] criteria. In order to have a pure Z + 1 jet sample, we require the subleading jet (i.e. second highest p_T-jet in the event) to have a p_T smaller than 30% of the p_T. In pp collisions, the Z + 1 jet events could be the result of Compton-like scattering events ($qg \to qZ$) or annihilation events ($q\bar{q} \to Z$) in which one gluon is radiated from the incoming quarks. Since protons are more likely to contain a gluon than an anti-quark, the Compton mechanism dominates the Z+jet production. As this process is more likely to result in a system where the dimuon system and the jet are balanced in the transverse plane, we can further purify the sample by requiring the azimuthal difference to be greater than 2.5 rad.



(a) multiplicity (b) $p_{\mathrm{T}}D$ (c) σ_2

Figure 40: Data-MC comparisons, for jets with $80 < p_{\mathrm{T}} < 100$ and $|\eta| < 2$ in Z+jet events, of the three input variables of the 3-variable likelihood. The data (black markers) are compared to a MADGRAPH + PYTHIA simulation, for which the different components are shown: quarks (blue), gluons (red) and unmatched/pile-up (grey).



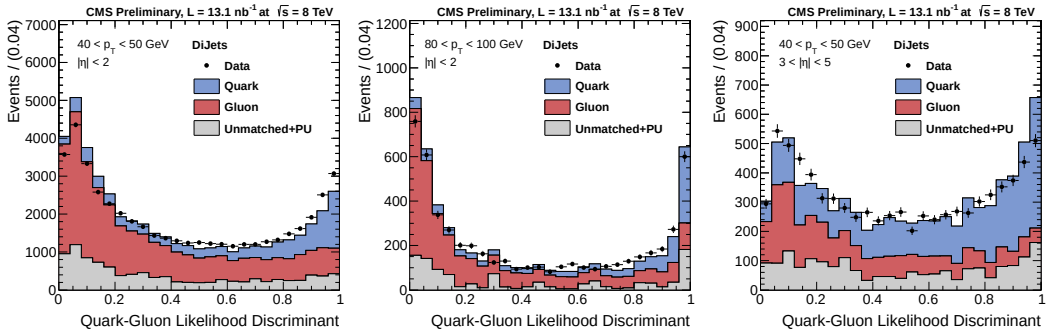
(a) central, $40 < p_{\mathrm{T}} < 50$ GeV (b) central, $80 < p_{\mathrm{T}} < 100$ GeV (c) forward, $40 < p_{\mathrm{T}} < 50$ GeV

Figure 41: Data-MC comparisons of the quark-gluon likelihood output in Z+jet events.

In figure 40, the input variables are compared between data and MADGRAPH + PYTHIA simulation events, on which the same selection is imposed. Similar level of agreement between data and simulation is observed in the other kinematic regions. Figure 41 shows the data-MC agreement for the likelihood output in three kinematic regions.

6.5.2 *Validation on dijet events*

Dijet production is mostly dominated by gluon jets, especially for lower p_{T} jets in the central region. In the forward region, quark jets dominate, hence it offers a cross check for Z+jet events in the forward region, as it has higher statistics. The sample has been collected using a prescaled zero bias trigger, corresponding to an integrated luminosity of 13.1 fb^{-1} taking into account the very large prescale factors deployed. The zero-bias trigger was chosen instead of a jet-based trigger,



(a) central, $40 < p_T < 50$ GeV (b) central, $80 < p_T < 100$ GeV (c) forward, $40 < p_T < 50$ GeV

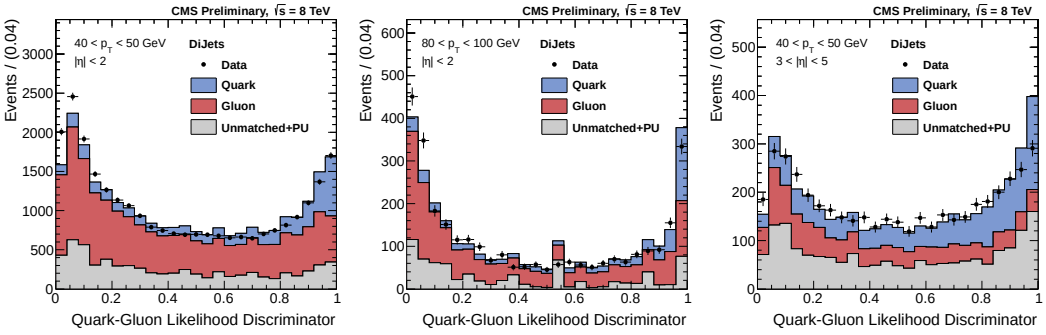
Figure 42: Comparison of the quark-gluon likelihood output in dijet events in data versus PYTHIA simulation

as it allows to reach the low p_T regime with full efficiency. Events where the two leading jets exceed $p_T > 30$ GeV are selected. Similar to the Z+jet case, the jets are forced to be back-to-back in the transverse plane by requiring their azimuthal difference to be greater than 2.5 rad, and events where the third jet exceeds 30% of the average p_T of the two leading jets are vetoed. Both leading jets are considered, fulfilling the usual jet identification criteria, and failing for the loose b-tag and pile-up criteria.

The likelihood output for data is compared to the PYTHIA simulation in Figure 42. The agreement between data and MC is worse than seen in the Z+jet validation, especially for lower p_T jets in the central region where PYTHIA predicts lower likelihood values compared to data. As the dijet sample is more dominated by the gluon component than the Z+jet sample, this effect could primarily be attributed to a mismodelling of the gluon fragmentation function in the PYTHIA simulation: gluon jets in data seem to be more similar to quark jets, compared to what is seen in PYTHIA. We further investigated this effect by performing a second validation with another parton shower and hadronization model. Figure 43 shows the same data as in Figure 42, in which the PYTHIA simulation is replaced by HERWIG++ simulation. Here we see the opposite effect: gluon jets in HERWIG++ are attributed with higher likelihood values compared to what is seen in data. Hence, we can conclude the performance of the quark-gluon jet discriminator in data is worse than expected from simulation PYTHIA simulation, but better than predicted by HERWIG++.

6.6 SYSTEMATIC UNCERTAINTIES

Because of the discriminator shape variations between data and MC (due to the hadronization model, the detector model, pile-up, etc ...) a generally applicable recipe is needed to estimate the shape uncertainty on the likelihood discriminant



(a) central, $40 < p_T < 50$ GeV (b) central, $80 < p_T < 100$ GeV (c) forward, $40 < p_T < 50$ GeV

Figure 43: Comparison of the quark-gluon likelihood output in dijet events in data versus HERWIG++ simulation

output. These shape differences are defined on the dijet data and MC samples, separately for quark and gluon jets. The shape variations established on the dijet events which has a more balanced quark-gluon composition than Z+jet events, therefore making it more sensitive to mismodelling of either distribution. The derived shape corrections are then applied to the Z+jet events to validate the method with an independent sample.

In order to vary the shape of the simulated sample, we need a function which maps the discriminator output $[0, 1]$ interval onto itself, allows to shift the population through the center, and towards the center or the extremes. Such a function is given by

$$g(x, a, b) = \frac{1}{2}\tanh(a \cdot \operatorname{arctanh}(2x - 1) + b) + \frac{1}{2} \tag{84}$$

in which (a, b) are the parameters to shift the population. These values are obtained by a minimization of the χ^2 obtained from a comparison between data and the simulated dijet events. The same functional form is applied independently on the quark and gluon distributions, so that both are modified independently to match the data. The parameters are retrieved in categories of η and p_T, and separately for PYTHIA and HERWIG++. The effect of this smearing function can be seen in Figure 44, which compares the likelihood output distributions for quark and gluon jets in the PYTHIA Z+jet simulation before and after applying the functional form. As expected, gluon jets in PYTHIA are smeared towards higher likelihood output values. In addition, a small shift of quark jets towards lower values is seen. Figure 45 compares the data to the simulation before and after the application of the smearing. The smearing parameters derived on the dijet case give also good closure on the Z+jet case.

The change in discriminating performance in the PYTHIA dijet simulation after applying the smearing function is shown in Figure 46: The efficiency on quark and gluon jets for a fixed cut (likelihood discriminant greater than 0.5) is shown before and after the smearing, in both the central and forward region. As expected, the

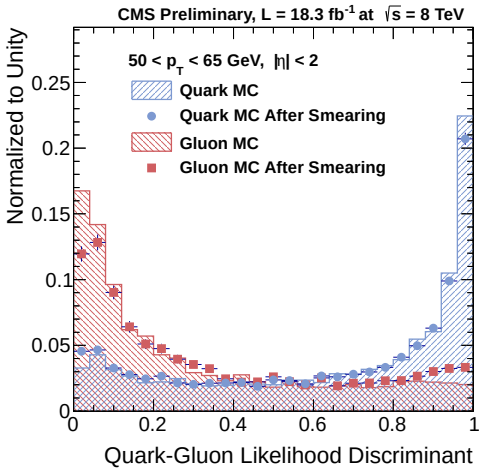
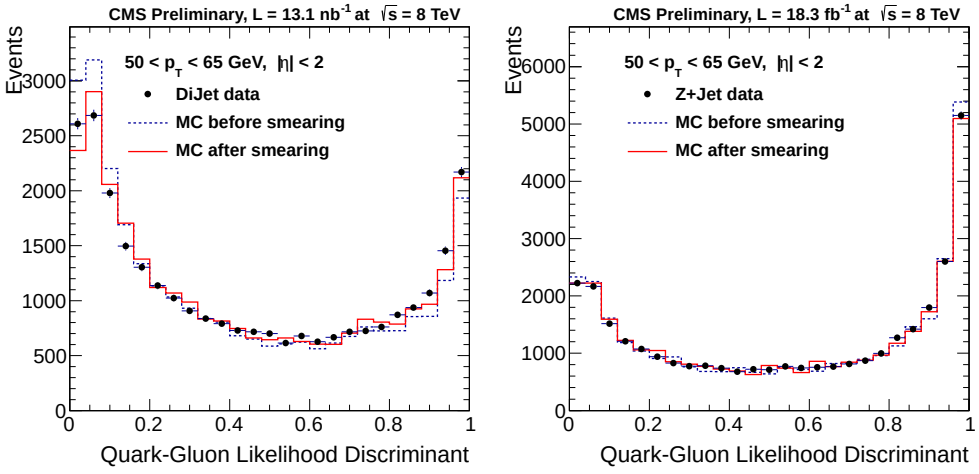
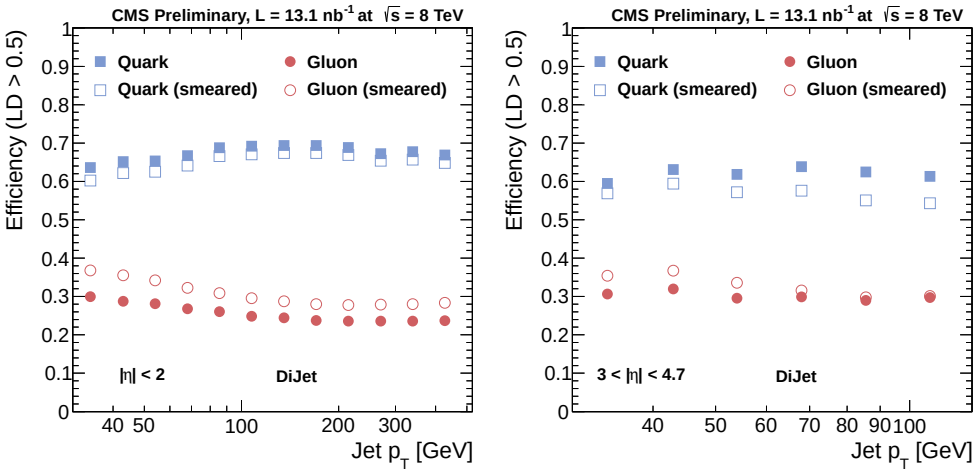


Figure 44: Effect of the data-driven smearing, derived on dijet events, on Z+jet simulated events (PYTHIA), separately for quark (blue) and gluon (red) jets. The smeared distributions (markers) are compared to the simulation before the application of the smearing (filled histograms).



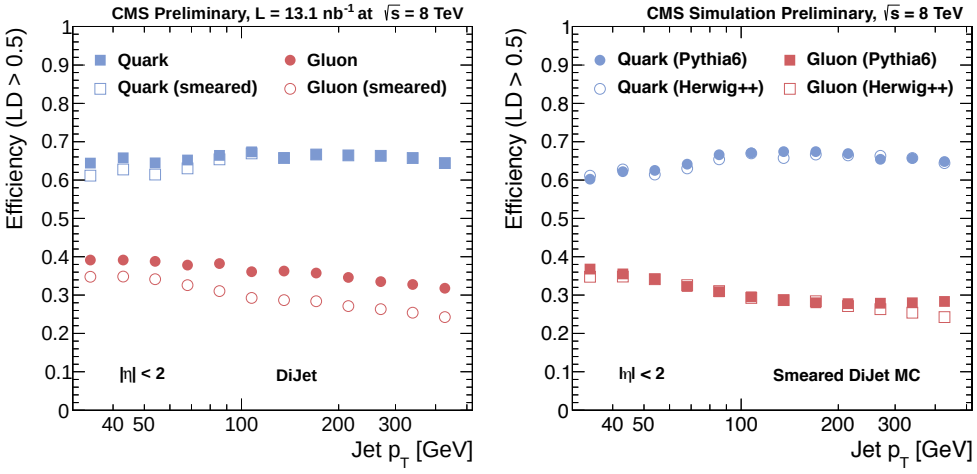
(a) dijet events (b) Z+jet events

Figure 45: Validation of the smearing function method in dijet and Z+jet data events compared to PYTHIA simulation. The data are compared to the simulation before and after the application of the smearing.



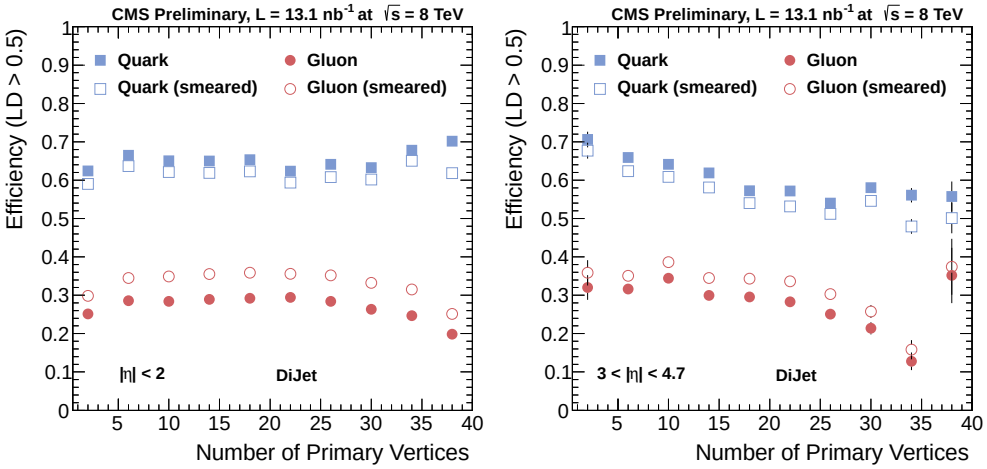
(a) central ($|\eta| < 2$) (b) forward ($3 < |\eta| < 4.7$)

Figure 46: Change in discriminating performance on PYTHIA simulation, by comparing the fraction of quark and gluon jets which have a quark-gluon likelihood discriminant greater that 0.5, as a function of the jet transverse momentum, before and after smearing.



(a) (b)

Figure 47: Change in discriminating performance on HERWIG++ simulation, by comparing the fraction of quark and gluon jets which have a quark-gluon likelihood discriminant greater than 0.5, as a function of the jet transverse momentum, before and after smearing. (a) The performance on HERWIG++ is increased after applying the smearing, and (b) is in very good agreement with the performance on PYTHIA.



(a) central ($|\eta| < 2$) (b) forward ($3 < |\eta| < 4.7$)

Figure 48: Pileup robustness in the central and forward region. The change in discriminating performance, evaluated as the efficiency of quark and gluons to pass a fixed cut (likelihood discriminant greater than 0.5), before and after the smearing is shown as a function of the number of reconstructed primary vertices in the event.

change in quark efficiency is rather small, while the discrimination performance of gluons seem to be up to 15% worse in data, compared to the expectations in PYTHIA simulation. Figure 47 shows a similar plot for HERWIG++ dijet events, which is underestimating the performance on gluons in data. After applying the smearing functions, PYTHIA and HERWIG++ are both in agreement with the data, and predict the same discriminator performances.

The smearing functions were only derived in bins of η and p_T. To confirm their robustness against pile-up, figure 48 shows the change in discriminating performance as a function of the number of primary vertices in the event. As can be seen, the performance is very robust against pile-up, both before and after the smearing, especially in the central region of the detector, where tracking is available.

6.7 BOOSTED AND HEAVY-FLAVOUR JETS

The quark-gluon discrimination tools are optimized for discriminating jets initiated by light-quarks (u, d and s) from jets initiated by gluon jets. There are, however, other fundamental particles in the hard jet which could initiate a jet, which could be misidentified either as a quark or a gluon jet:

- Jets initiated by b quarks have larger multiplicities and a larger angular spread compared to light quarks, especially at low p_T where effects due to the longer decay chain of the b quark dominate. As a result b-initiated jets

have gluon-like discriminator values in the low p_T region. Towards higher jet transverse momenta, the parton shower produces more particles and the fragmentation will be more similar to light-quark jets. In the high p_T regime, b-jets are identified as light quarks with the same efficiency as light-quark jets. On the other hand, dedicated b-tagging algorithms, such as the Combined Secondary Vertex tagger [103], can be used to efficiently select b-jets, both in the low and high p_T regime. Analyses focusing on b-jets are therefore better served by these dedicated algorithms.

- The same effect is seen for c quarks, although less strong: jets initiated by c quarks have discriminator outputs which are halfway in between the light-quark jets and the b jets.

- In the higher p_T region ($p_T > 300$ GeV), jets could be associated to a boosted W-, Z-, H-boson or top quark for which the opening angle between their decay products becomes so small that the boosted object ends up as a single jet. These jets have a similar multiplicity, energy sharing and width as gluon jets and will be identified as gluon jets. Analyses which try to identify the quark jets from these decays can therefore only use the quark-gluon discrimination tool on jets up to about 300 GeV, and should switch to dedicated tagging algorithms for boosted topologies [105] for higher p_T jets.

7

MEASUREMENT OF THE ELECTROWEAK PRODUCTION OF A Z BOSON IN ASSOCIATION WITH TWO JETS AT 7 TEV

The analysis strategy starts with the selection of $lljj$ events in both data and MC samples, together with some basic requirements on the dilepton invariant mass and the jet kinematics to further purify the samples for our signal process. Reweighting factors are introduced to mitigate data-MC discrepancies due to different pile-up distributions, trigger and lepton inefficiencies, and the LO description of the generated MC events. The background and signal are further separated by means of a BDT, of which the output shapes are used to fit the amount of signal and background contributions to data. Finally, the theoretical and experimental uncertainties are carefully taken into account to estimate the error on the cross section measurement. This chapter describes the measurement on the $\sqrt{s} = 7$ TeV data, while the improved analysis on the larger $\sqrt{s} = 8$ TeV data is discussed in the next chapter.

7.1 SAMPLES AND TRIGGERS

Candidate events were selected by various HLT triggers which require the presence of two muons or two electrons. Such events are reconstructed and stored in the *doubleElectron* and *doubleMuon* primary datasets, and are centrally produced and shared by multiple analysis groups within the CMS collaboration. The HLT trigger paths used correspond to the lowest-threshold unprescaled trigger available in a given data taking period. The *doubleElectron* HLT paths requires events with the presence of two electrons in the event, exceeding a transverse momentum of respectively 17 GeV and 8 GeV. In addition, a set of loose calorimeter and/or tracker-based identification and isolation criteria were applied in order to avoid these paths to be fired by QCD events, keeping the rate to acceptable levels. For muons, a *doubleMuon* HLT path requiring events with both muons exceeding $p_{\mathrm{T}} > 6$ GeV was originally used for the early data acquisition period. Due to the rising instantaneous luminosity, this original trigger became prescaled, and new unprescaled triggers with higher thresholds were introduced. In this way, the muon triggers gradually evolved towards the same asymmetrical cuts as used for the electron triggers: $p_{\mathrm{T}} > 17$ GeV for the leading lepton and $p_{\mathrm{T}} > 8$ GeV for the second lepton.

Only events which are certified for data analysis are further processed, i.e. events collected when all subdetectors were fully operationally and no issues were reported during data quality monitoring [106]. The total integrated luminosity for the analyzed data amounts to $\mathcal{L} = 5.0 \pm 0.1$ fb^{-1}, which is estimated using the pixel cluster counting method [107]. This method evaluates the luminosity based on the number of pixel clusters occurring on average in a zero-bias event, which is proportional to the number of collisions per bunch crossing. The expected number of pixel clusters per inelastic collision is calibrated during Van Der Meer scans, which scan the LHC beams through each other to determine the size of the beams at their collision point.

The data is compared to the signal and background MC samples which are centrally produced within the CMS collaboration. An overview of these samples is given in Table 5, together with their equivalent luminosity ($\mathcal{L} = N_{\text{events}}/\sigma$).

	generator	σ (pb)	$\mathcal{L}$ (fb^{-1})
EW $Z(ll)jj$	MADGRAPH + PYTHIA	0.754	2475
DY ll + jets	MADGRAPH + PYTHIA	3048	11.88
$t\bar{t}$ + jets	MADGRAPH + PYTHIA	165	361.3
WW	PYTHIA	43	98
WZ	PYTHIA	18.2	708
ZZ	PYTHIA	5.9	1344
QCD ($100 < H_T < 250$ GeV)	MADGRAPH + PYTHIA	$4.194 \cdot 10^6$	$5.66 \cdot 10^{-3}$
QCD ($250 < H_T < 500$ GeV)	MADGRAPH + PYTHIA	$1.985 \cdot 10^5$	0.104
QCD ($500 < H_T < 1000$ GeV)	MADGRAPH + PYTHIA	5856	2.5
QCD ($H_T > 1000$ GeV)	MADGRAPH + PYTHIA	122.6	51.6

Table 5: Signal and background MC samples for the 7 TeV analysis. The EW Zjj includes all lepton flavors and is generated in a slightly larger phase space as in Table 1, explaining the difference in cross section.

7.2 EVENT SELECTION

Events which have passed the trigger selection are scanned for Z-boson candidates. The leptons are required to exceed $p_{\text{T}} > 20$ GeV and to be constructed withing the fiducial region $|\eta| < 2.4$ of the tracking system. Additionally, they have to fulfill the identification and isolation criteria as described in Chapter 5. The two highest p_{T} leptons of the same flavor that are oppositely-charged are identified as possible Z-boson candidates. Their four-momenta are combined to reconstruct the dilepton invariant mass M_{ll}, and the event is selected if the dilepton invariant mass is within 20 GeV of the nominal Z boson mass, i.e. $|M_{ll} - m_Z| < 20$ GeV.

The two PF candidates associated with the selected lepton pair are removed from the PF candidates collections. The remaining PF candidates are clustered to PF jets,

using the anti-k_T algorithm with distance parameter $R = 0.5$. In this way we avoid the leptons to be constructed as part of the jets, even though the isolation requirements already removes most of these cases. In order to select candidate EW Zjj events, events are required to have at least two jets in the event within $\mid \eta \mid < 3.6$. At the time of carrying out this measurement, the JES corrections were not yet fully optimized to describe the most forward pseudorapidities correctly, and it was therefore chosen to limit the η-range of the jets to 3.6. The two highest p_T jets were selected as the tagging jets, and were required to exceed $p_T > 65$ GeV and $p_T > 40$ GeV respectively, exploiting the energetic jets as a key feature of EW Zjj production. This selection was chosen by maximizing the significance[1] on simulated events.

Figure 49 shows the kinematic distributions for the tagging jets for events which passed our selection. Good agreement between data and simulation is observed for most of these variables, though the DY simulation slightly overestimates the number of events with very high p_T jets. As expected, the signal contribution with respect to the DY and other background increases for larger values of the dijet invariant mass. A similar behaviour is seen for the pseudorapidity difference $\Delta\eta_{jj}$, which also encodes the forward-backward jet topology of the signal events. The tagging jets are expected to be quark-like, while DY events have a mix of quark and gluon jets, which can be seen in the quark-gluon likelihood values shown in Figure 50.

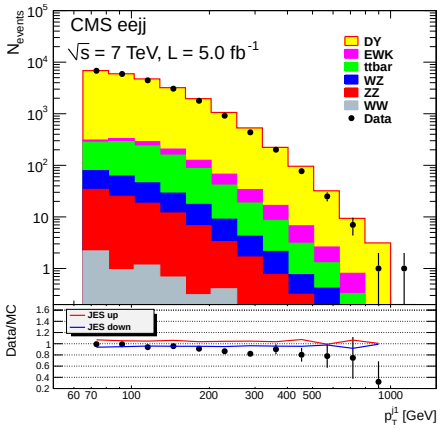
In what follows below, we will measure the cross section both with and without requiring the events to have $|y^*_Z| < 1.2$, in which y^*_Z is the Zeppenfeld variable for the Z-boson, as defined in equation 49 and shown in Figure 51. This requirement is again chosen to maximize the significance, but also avoids events with higher values for $|y^*_Z|$. The high $|y^*_Z|$ region is troubled with statistical fluctuations, high interference effects (see Figure 13) and imperfect modeling by the LO description of MADGRAPH. All of these effects complicate the cross section measurement, and it is therefore interesting to study the effect of restricting the analysis to the lower $|y^*_Z|$ values.

7.3 SCALE FACTORS

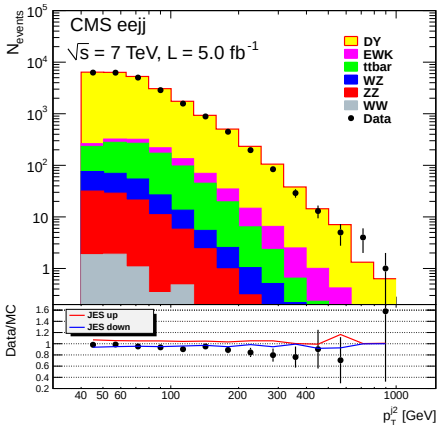
7.3.1 *Pile-up reweighting*

Because the true pile-up distribution in data was not known at the time of generating the MC simulations, MC events were produced according to an assumed pile-up distribution. The true pile-up distribution in data is calculated using the bunch-by-bunch luminosity, assuming a total pp inelastic cross section of 68 mb [108]. For

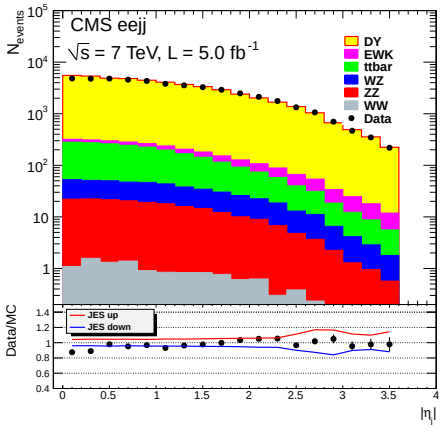
1 The significance is defined as $\frac{N_S}{\sqrt{N_B}}$, i.e. the number of signal events with respect to the standard deviation on the number of background events.



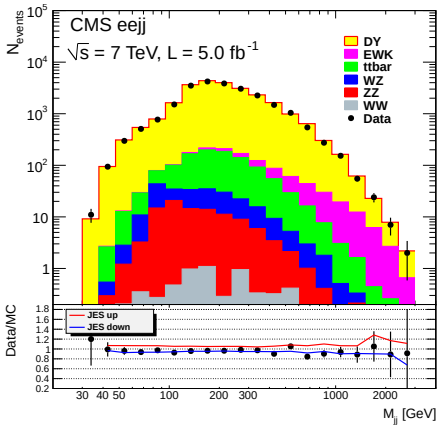
(a) p_{T} of the leading jet



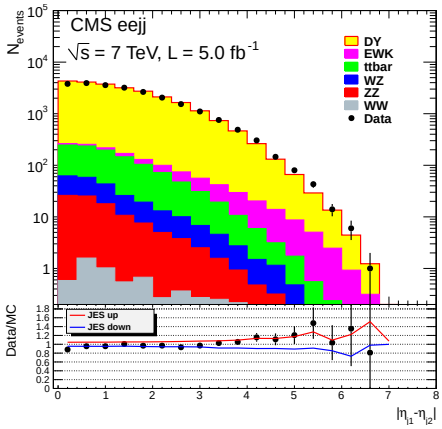
(b) p_{T} of the second jet



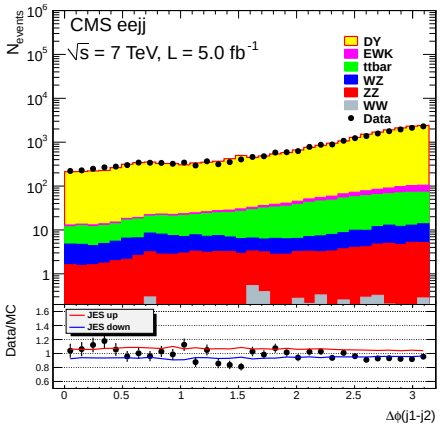
(c) η of the tagging jets



(d) dijet invariant mass

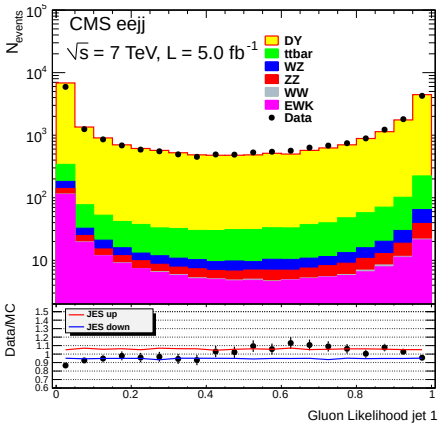


(e) $\Delta\eta$ between the tagging jets

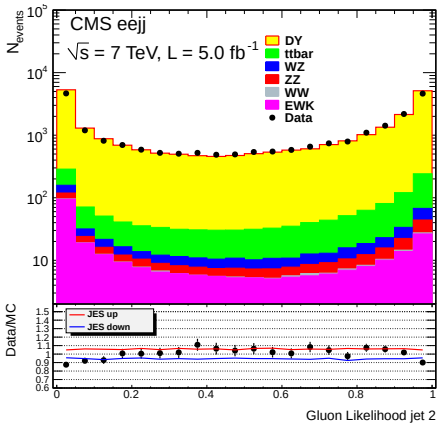


(f) $\Delta\phi$ between the tagging jets

Figure 49: Kinematic distributions for the tagging jets in events passing our selection, shown for data and compared to the expected contributions from signal and background processes evaluated from simulation. The bottom panel shows the ratio of data over the expected contributions of signal plus backgrounds along with the statistical uncertainties. The red and blue lines show how the simulation prediction would be affected when using the lower and upper uncertainty bound of the JES corrections.

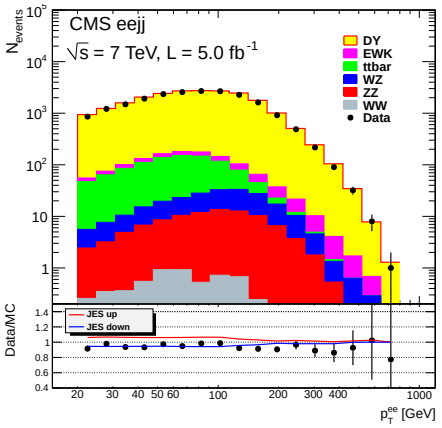


(a) Quark-gluon likelihood for the leading jet

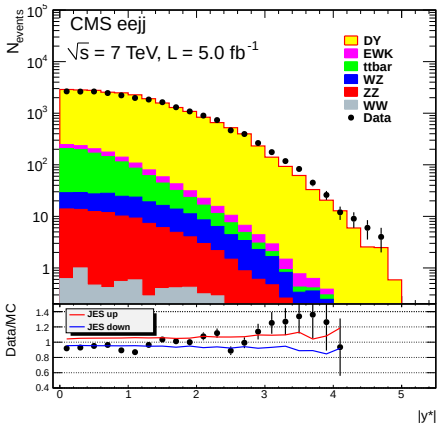


(b) Quark-gluon likelihood for the second jet

Figure 50: Quark-gluon likelihood distributions for the tagging jets in events passing our selection, shown for data and compared to the expected contributions from signal and background processes evaluated from simulation. The distribution for the 5-variable likelihood peaks at 0 for quark-like jets and at 1 for gluon-like jets.



(a) p_{T} of the Z-boson



(b) $|y^*_Z|$ distribution

Figure 51: Distributions of $p_T(Z)$ and y^*_Z in events passing our selection, shown for data and compared to the expected contributions from signal and background processes evaluated from simulation. The bottom panel shows the ratio of data over the expected contributions of signal plus backgrounds along with the statistical uncertainties. The red and blue lines show how the simulation prediction would be affected when using the lower and upper uncertainty bound of the JES corrections.

each MC sample, the pile-up distribution is saved before any selections are done, and events are reweighted in order to match the true distribution in data.

7.3.2 *Lepton selection efficiencies*

Both the trigger and analysis lepton selections efficiencies are not necessarily the same in data and MC due to imperfections in the event and detector simulation. A tag and probe technique [109] is therefore used to measure these efficiencies independently in data and MC for $Z/\gamma \rightarrow ll$ events. In this technique, a set of *tag* lepton candidate are selected by very tight trigger, identification and isolation criteria, therefore having a very high efficiency to pass the less strong criteria under study. Another set of lepton candidates, known as the *probes*, are selected with very loose selection criteria. Those are paired to the tag leptons in order to find pairs for which their combined invariant mass is consistent with the mass of the Z resonance. A fit to the invariant mass distribution is applied to subtract the contribution of non-resonant dilepton pairs to the mass window of the resonance. The probes are then classified as whether they pass or fail the selection criteria under study. The efficiency of a single lepton to pass the criteria is then defined as the number of passed probes over the total number of probes. Because the tag leptons are a subset of the probe electrons, the efficiency is calculated as

$$\epsilon = \frac{2N_{TT} + N_{TP}}{2N_{TT} + N_{TP} + N_{TF}} \tag{85}$$

in which N_{TT} is the number of pairs in which both leptons are tagged, N_{TP} contains pairs in which the probe passed the selection but is not tagged, and N_{TF} represents the pairs with a failed probe. By applying this procedure in both data and MC, scale factors $\rho = \epsilon_{\text{data}}/\epsilon_{\text{MC}}$ are constructed in bins of p_{T} and η, which are applied to reweight the MC events. The data/MC scale factors are provided centrally within CMS. In our 7 TeV analysis, the scale factors were only applied in the dimuon channel, only taking into account the η dependency while neglecting the p_{T} dependency which was found to be negligible for muons with $p_{\text{T}} > 20$ GeV. No scale factors have been applied for electrons, though no significant data/MC differences were observed for the efficiencies at the $WP90$ working point [110].

7.4 SIGNAL EXTRACTION

7.4.1 *Multi-variate analysis*

Even though our event selection improves the signal to background ratio, the signal process is still covered below a huge background of DY events. One way to extract the signal is to introduce even tighter cuts on the variables to select regions with an

enriched signal contribution, for example by requiring very high M_{jj} or $\Delta\eta_{jj}$ values. A more elegant way to extract the signal is by use of a multi-variate analysis which maps a set of input variables onto one output variable, with maximal separation between signal and background. The signal and background shapes for the output variable distribution could then be fitted to the output distribution in data.

Multiple multi-variate analysis methods, provided by the TMVA framework [101], have been tried in order to find the best performing and most robust method for our analysis. We have settled with the BDT method, which implements the adaptive boost algorithm [111], and is trained using MC events of the signal and DY background sample. An independent simulation of both samples was used to verify and validate the performance. The set of input variables, decorrelated before feeding it to the BDT, is constructed out of following observables:

- the transverse momenta of the leading, second jet and Z-boson ($p_T^{j_1}$, $p_T^{j_2}$ and p_T^Z)
- the dijet invariant mass (M_{jj})
- the pseudorapidity difference between the tagging jets ($\Delta\eta_{jj}$)
- the pseudorapidity of the Z-boson ($|\eta_Z|$)
- the sum of the pseudorapidities of both tagging jets ($|\eta_{j_1} + \eta_{j_2}|$)
- the differences in azimuthal angles between the tagging jets ($\Delta\phi_{jj}$), between the leading jet and the Z-boson ($\Delta\phi_{Zj_1}$), and between the second jet and the Z-boson ($\Delta\phi_{Zj_2}$)
- the 5-variable quark-gluon likelihood for each of the tagging jets

Figure 52 shows the normalized signal and background distributions for each of these input variables. Some of the most separating variables are caused by the energetic forward-backward dijet system in signal events, resulting in larger values for $p_T^{j_1}$, $p_T^{j_2}$, M_{jj} and $\Delta\eta_{jj}$ compared to DY Zjj events. Because the jets are found in the opposite forward-backward halves of the detector, $|\eta_{j_1} + \eta_{j_2}|$ has on average lower values for the signal with respect to the DY background where both jets could end up on the same side. The Z-boson is slightly more boosted in the signal, leading to larger p_T^{ll} values, while its central production results in lower $|\eta_{ll}|$ values. In signal events, the high p_{T} values of the jets require them to be balanced in the azimuthal plane, resulting in $\Delta\phi_{jj}$ values close to π. The Z-boson is typically found on the side of the second jet, resulting in small $\Delta\phi_{Zj_2}$ values while $\Delta\phi_{Zj_1}$ also peaks close to π. On the other hand, the DY Zjj events are more likely to have both jets balancing the Z-boson, resulting in larger $\Delta\phi_{Zj_2}$ and smaller $\Delta\phi_{jj}$ values.

The output distribution of the BDT is shown in Figure 53 for data and all simulation components. A high BDT output value corresponds with a high probability to find a signal event. Even though statistics are low at the high end of the BDT output, one can clearly observe how the signal contributes to the data.

7.4.2 *Fit results*

The expected signal and background shapes are fitted to the shape of the data distribution using the `TFractionFitter` method provided by the `ROOT` data analysis framework [112]. It employs a maximum likelihood fit using Poisson statistics such that data statistical uncertainties are taken into account [113]. The fit optimized for two free parameters s and b, which are the ratios of the fitted to the expected event yields of the signal and DY background respectively. The contributions form other backgrounds were fixed to their expected event yields from simulation.

The best fit yields the following parameters for s and b in the muon and electron channels:

$$\begin{aligned} s = 1.17 \pm 0.27 \text{ (stat)} \quad b = 0.957 \pm 0.010 \text{ (stat)} \qquad & (eejj \text{ events}) \\ s = 0.85 \pm 0.18 \text{ (stat)} \quad b = 0.937 \pm 0.007 \text{ (stat)} \qquad & (\mu\mu jj \text{ events}) \end{aligned}$$

in which we find the measured signal to be in agreement with the prediction by MADGRAPH within statistical uncertainties. The DY background has a very low statistical uncertainty and is found below unity, though this can be explained by the systematical uncertainties described in the next section. The fit of the muon channel is in good agreement with the results of a parallel analysis, which was carried out in the muon channel, using JPT jets instead of PF jets. The JPT analysis implemented the same event selection criteria and its fit resulted in $s = 0.90 \pm 0.19$ (stat) and $b = 0.905 \pm 0.006$ (stat).

The right plot in Figure 53 shows the BDT output shape with its EW Zjj and DY Zjj contributions evaluated form the fit while retaining the normalization of the other backgrounds to their simulation estimates. The bottom panel in this plot shows also the observed significance in each bin i, given by

$$S_i^{\text{observed}} = \frac{N_i^{\text{data}} - N_i^{\text{bkg}}}{\sqrt{N_i^{\text{bkg}} + \left(\Delta B_i^{\text{JES}}\right)^2}} \tag{86}$$

where N_i^{data} and N_i^{bkg} are the number of observed events and the number of simulated background events respectively. The background events include the DY contribution as evaluated by the fit as well as all other backgrounds estimated from

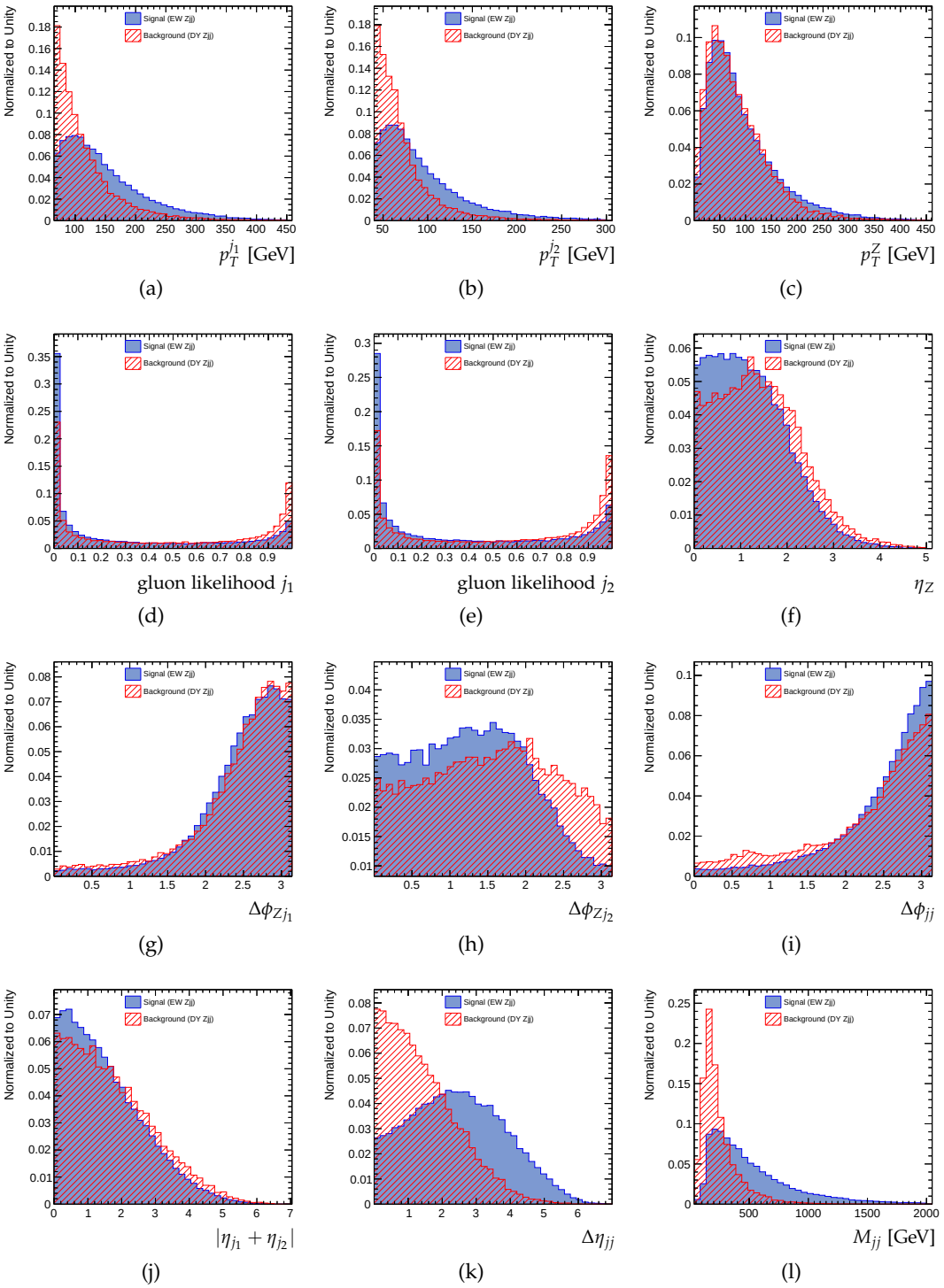


Figure 52: Normalized distributions for signal and background for each of the BDT input variables

simulation. In order to account for the dominant experimental uncertainty, due to the JES, ΔB_i^{JES} is introduced calculated as

$$\left(\Delta B_i^{\text{JES}}\right)^2 = \frac{1}{2}\left[\left(N_i^{\text{bkg}} - N_{i,\text{JES+}}^{\text{bkg}}\right)^2 + \left(N_i^{\text{bkg}} - N_{i,\text{JES-}}^{\text{bkg}}\right)^2\right] \tag{87}$$

in which $N_{i,\text{JES+}}^{\text{bkg}}$ and $N_{i,\text{JES-}}^{\text{bkg}}$ are the number of simulated background events varied by the positive and negative JES systematic uncertainty. The observed significance is compared with the expected significance, given by

$$S_i^{\text{expected}} = \frac{N_i^{\text{EW } Zjj}}{\sqrt{N_i^{\text{bkg}} + \left(\Delta B_i^{\text{JES}}\right)^2}} \tag{88}$$

where $N_i^{\text{EW } Zjj}$ is the number of EW Zjj events, evaluated by the fit. The expected signal significance is 0 at low BDT values but raises to 2 for the higher BDT bins. A similar behaviour is seen for the observed significance.

In the muon channel, the fit procedure has been repeated with events required to pass the same selection plus the $|y^*| < 1.2$ requirement. The BDT has been retrained for the $|y_Z^*| < 1.2$ selection, and the best fit yields

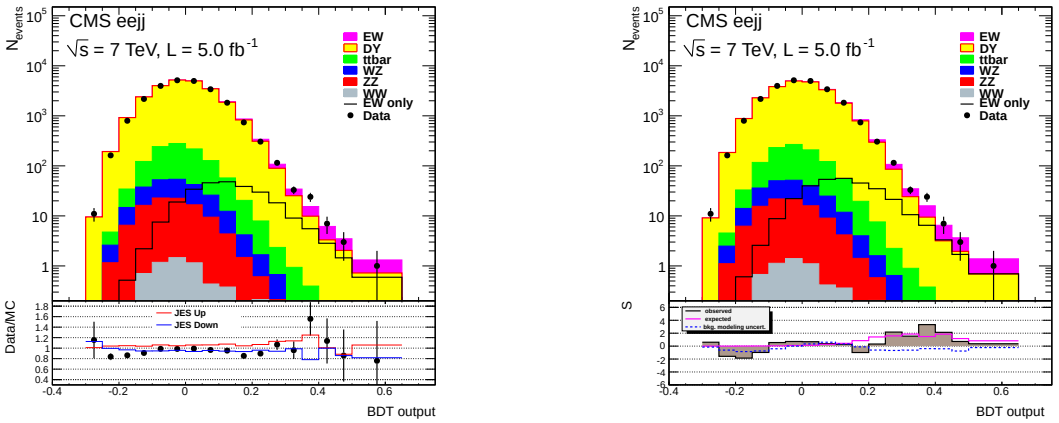
$$s = 1.37 \pm 0.25 \text{ (stat)} \quad b = 0.862 \pm 0.007 \text{ (stat)} \qquad (|y_Z^*| < 1.2)$$

which is in good agreement with the JPT analysis which measured $s = 1.50 \pm 0.26$ (stat) and $b = 0.863 \pm 0.007$ (stat). Because the y_Z^* variable is not very well modelled by the MADGRAPH simulation, which predicts slightly higher y_Z^* values for the DY simulation compared to the observed data, a lower background contribution is measured in the fit. Through its correlations with some of the BDT input variables, the y_Z^* requirement also affects the shape of the signal, background and data output distributions and the statistical fluctuations on them. Hence it is possible to find a different result for the fitted signal yield, though it is still compatible with unity within statistical and systematical errors.

The fitting procedure was not only applied on the BDT shape, but also on the dijet invariant mass distribution, which is the most discriminating input variable of the BDT. This fit was only extracted in the muon channel and yields

$$s = 1.14 \pm 0.30 \text{ (stat)} \quad b = 0.897 \pm 0.008 \text{ (stat)} \qquad (M_{jj} \text{ fit}, |y_Z^*| < 1.2)$$

which is in good agreement with the JPT analysis which measured $s = 1.14 \pm 0.28$ (stat) and $b = 0.869 \pm 0.008$ (stat). The s and b values found using the M_{jj} fit lay between those achieved with the BDT fit using the whole y_Z^* range, and the BDT fit with $|y_Z^*| < 1.2$.



(a) BDT output before fit

(b) BDT output after fit

Figure 53: BDT output distribution, before and after a fit of the EW Zjj signal and DY background contribution to the data. The bottom panel on the right plot shows the significance observed in data (black line) and expected from simulation (purple line) in each bin. The dashed blue line shows the background modeling uncertainty.

	$eejj$ events	$\mu\mu jj$ events
Background modeling	0.16	0.14
Signal acceptance	0.06	0.05
$t\bar{t}$ cross section	0.03	0.03
diboson cross sections	0.02	0.02
total theoretical uncertainty	0.17	0.16
JES and JER	0.29	0.21
Pile-up modeling	0.03	0.03
Statistics of simulation	0.19	0.12
Quark-gluon discriminator tool	0.02	0.02
Lepton selection	0.02	0.02
total experimental uncertainty	0.35	0.25
luminosity	0.03	0.02

Table 6: Absolute values of the theoretical and experimental uncertainties on the expected Zjj yield, as measured by a fit of the BDT output shapes in the full y^* range using PF jets.

7.4.3 *Systematic uncertainties*

7.4.3.1 *Theoretical uncertainties*

- **Background modeling**
  The LO description by MADGRAPH is not sufficient to describe the DY background accurately. The dijet invariant mass spectrum is therefore calculated at NLO using MCFM, and compared to the M_{jj} shape in MADGRAPH at parton level. The DY background is then reweighted in M_{jj} in order to match the NLO description, and the fit is repeated. The systematic uncertainty is then assigned as the difference between the original fit and the new fit. The background modeling uncertainty is also shown in Figure 53, where it is calculated for each bin i as

$$\frac{N_i^{\mathrm{MCFM}} - N_i^{\mathrm{bkg}}}{\sqrt{N_i^{\mathrm{bkg}} + \left(\Delta B_i^{\mathrm{JES}}\right)^2}} \tag{89}$$

  with N_i^{MCFM} the number of background events obtained from the new fit. The background modeling uncertainty seems to explain the deficit in events at low BDT values. For the high BDT values, the background modeling uncertainty also lowers the number of DY events. Hence, the fit with the NLO shape results in a higher signal yield, which accounts to an uncertainty of 0.16 in the electron channel and 0.14 in the muon channel.

- **Signal acceptance**
  An uncertainty has to be taken into account for the acceptance of signal events through our event selection cuts. The calculation of the cross sections using listed in Table 1, were done with similar requirements as those used in the analysis. The good agreement between the MADGRAPH and VBFNLO cross sections indicate this uncertainty is not larger than 5%.

- **Normalization of residual backgrounds**
  The diboson and $t\bar{t}$ backgrounds are fixed in the fit, and their contributions correspond directly to their theoretical cross section. The uncertainty on their cross sections is therefore propagated to an uncertainty on the fit, and the effect of the $t\bar{t}$ and diboson cross sections on the fit are estimated to be about 2% and 1% respectively.

7.4.3.2 *Experimental uncertainties*

- **Jet energy scale and resolution**
  For each event in the simulations, the jet energy is varied up and down with

the JES uncertainty, after which the BDT and the fit of its shapes are again evaluated. The systematic uncertainty is taken as the difference of the fit value s from the original fit to the measured fit value from the fit with the adjusted energy.

The fit is also repeated with an additional smearing of the p_{T} resolution in simulation, using the correction factors shown in Figure 32. The systematic uncertainty for the JER corresponds with the variations in the fitted values of s, with and without the smearing.

- **Pile-up modeling**
  The pile-up distribution is not expected to affect the identification and isolation of the leptons, or the corrected energy of individual jets. Pile-up activity can however influence how the jet clustering algorithm is run, in which it can slightly distort the parameters quantifying the reconstructed dijet system. It is therefore important the pile-up distribution in MC matches the one in data, and we need to assign an uncertainty for it. The pile-up distribution in data is based on the assumption of a total pp inelastic cross section of 68 mb. The pile-up distribution is recalculated by varying the pp inelastic cross section by $\pm 5\%$, and its effect is propagated through the analysis up to the fit, where its effect can be deduced.

- **Quark-gluon jet discrimination**
  The systematic uncertainty on the quark-gluon likelihood performance has been calculated by smearing the quark-gluon likelihood distributions in MC for both tagging jets, as described in section 6.6, and measuring the deviation of the fit due to this smearing.

- **Lepton selection efficiencies**
  A conservative 1% uncertainty on the data-to-simulation scale factor for the efficiency on lepton reconstruction, identification, isolation and trigger is assigned for each lepton, resulting in a 2% overall uncertainty for the Z boson selection.

- **Statistics of simulation**
  The fit with the `TFractionFitter` tool was performed without taking the finite statistics of the MC samples into account. In order to estimate an uncertainty due to the MC statistics, an envelope is created around the signal and background MC distributions by shifting all bin contents simultaneously up or down by its statistical uncertainty. This generates two alternatives to the signal and background BDT shapes, on which the fit is repeated and the maximum deviation is taken as the uncertainty.

- **Luminosity**
  In addition to the above uncertainties, we account for an uncertainty on the integrated luminosity, which is estimated at 2.2% [107].

7.5 SUMMARY

The fit parameter s returned the ratio of the measured EW Zjj yield over the expected yield from the MADGRAPH simulation. We can therefore calculate the cross section as $\sigma_{\text{meas}} = s \times \sigma_{MG}(\text{EW } lljj)$, in which the MADGRAPH cross section $\sigma_{MG}(\text{EW } lljj) = 162$ fb per lepton flavor is obtained with the parton-level requirements listed in table 1. The cross section in the electron and muon channel are then given by

$$\sigma_{ee}^{\text{EW}} = 190 \pm 44 \text{ (stat.)} \pm 57 \text{ (exp. syst.)} \pm 27 \text{ (th. syst.)} \pm 4 \text{ (lum.) fb}$$
$$\sigma_{\mu\mu}^{\text{EW}} = 138 \pm 29 \text{ (stat.)} \pm 40 \text{ (exp. syst.)} \pm 25 \text{ (th. syst.)} \pm 3 \text{ (lum.) fb}$$

and are compatible with the JPT analysis which measured $\sigma_{\mu\mu}^{\text{EW}} = 146 \pm 31$ (stat.) $\pm$ 42 (exp. syst.) $\pm$ 26 (th. syst.) $\pm$ 3 (lum.) fb. The measurements in the electron and muon channel using PF jets were combined, taking into account their experimental and systematical uncertainties are fully correlated. The combined cross section is

$$\sigma_{ll}^{\text{EW}} = 154 \pm 24 \text{ (stat.)} \pm 46 \text{ (exp. syst.)} \pm 27 \text{ (th. syst.)} \pm 3 \text{ (lum.) fb}$$

This measurement is in good agreement with the NLO cross section of 166 fb, as calculated by VBFNLO. The significance of this measurement is 2.6 standard deviations, and a further reduction of statistical and systematical uncertainties was needed to fully claim observation of the EW Zjj process. Fortunately, this was achieved on the larger dataset in 2012, using $\sqrt{s} = 8$ TeV pp-collisions, which is described in the next chapter.

8

MEASUREMENT OF THE ELECTROWEAK PRODUCTION OF A Z BOSON IN ASSOCIATION WITH TWO JETS AT 8 TEV

Using $p\,p$-collisions at $\sqrt{s} = 8$ TeV, the analysis from the former chapter has been repeated. The 2012 dataset provides us a larger integrated luminosity, such that the statistical error can be reduced. The systematic uncertainties are studied in more detail and incorporated in a new fit procedure which treats the systematic uncertainties as nuisances. The PF jet analysis is again performed in both electron and muon channels, while being cross checked by a parallel JPT analysis in the muon channel. An additional cross check has been performed by a third analysis, also performed using PF jets and in both electron and muon channels, but introducing a data-driven method to model the DY background.

8.1 SAMPLES AND TRIGGERS

The analysis is again carried out on the *doubleElectron* and *doubleMuon* primary datasets. The same HLT dilepton triggers which were in use at the end of the 2011 run, were also used for the 2012 data: one lepton is required to have $p_T > 17$ GeV, while the other has to pass $p_T > 8$ TeV, and additional identification and isolation requirements are required for the electrons. The total integrated luminosity for the certified 2012 data run amounts to $19.7\ \text{fb}^{-1}$, as measured by the pixel cluster counting method.

The signal and background MC samples are shown in Table 7. Because the inclusive DY + jets sample is mainly dominated by events which contain 0 additional jets at parton level, it suffers from low statistics after the requirement of two jets. Events with higher parton multiplicities are therefore replaced by separately generated samples to increase statistics.

8.2 EVENT SELECTION

After passing the trigger, the Z-bosons candidates are reconstructed in the same way as for the 7 TeV analysis, but with the updated lepton isolation and identification criteria as mentioned in Chapter 5. Events were again required to have at least two PF jets, constructed with the anti-k_T algorithm with distance parameter

	generator	σ (pb)	$\mathcal{L}$ (fb^{-1})
EW $Z(ll)jj$	MADGRAPH + PYTHIA	0.8938	3332.6
DY ll + jets	MADGRAPH + PYTHIA	3531.9	8.62
DY ll + 1 jet	MADGRAPH + PYTHIA	671.660	35.8
DY ll + 2 jets	MADGRAPH + PYTHIA	216.703	100.9
DY ll + 3 jets	MADGRAPH + PYTHIA	61.180	180.1
DY ll + 4 jets	MADGRAPH + PYTHIA	27.585	232.1
$t\bar{t}$ + jets (full leptonic decays)	MADGRAPH + PYTHIA	25.80	465.6
$t\bar{t}$ + jets (semi leptonic decays)	MADGRAPH + PYTHIA	107.67	231.8
$t\bar{t}$ + jets (full hadronic decays)	MADGRAPH + PYTHIA	112.33	371.8
single t (tW-channel)	POWHEG + PYTHIA	11.1	44.8
single $\bar{t}$ ($\bar{t}W$-channel)	POWHEG + PYTHIA	11.1	44.5
single t (s-channel)	POWHEG + PYTHIA	3.79	68.6
single $\bar{t}$ (s-channel)	POWHEG + PYTHIA	1.76	79.5
single t (t-channel)	POWHEG + PYTHIA	56.4	66.6
single $\bar{t}$ (t-channel)	POWHEG + PYTHIA	30.7	63.0
WW	PYTHIA	54.838	182
WZ	PYTHIA	22	454
ZZ	PYTHIA	7.6	1289
W + jets	MADGRAPH + PYTHIA	36703.2	1572.3
QCD ($100 < H_T < 250$ GeV)	MADGRAPH + PYTHIA	$1.036 \cdot 10^7$	$4.839 \cdot 10^{-3}$
QCD ($250 < H_T < 500$ GeV)	MADGRAPH + PYTHIA	$2.76 \cdot 10^5$	$9.805 \cdot 10^{-2}$
QCD ($500 < H_T < 1000$ GeV)	MADGRAPH + PYTHIA	8426	3.632
QCD ($H_T > 1000$ GeV)	MADGRAPH + PYTHIA	204	67.9

Table 7: Signal and background MC samples for the 8 TeV analysis. The EW Zjj includes all lepton flavors and is generated in a slightly larger phase space as in Table 1, explaining the difference in cross section. The DY background is also simulated in parton multiplicity bins, increasing the available statistics for this process.

$R = 0.5$ and excluding the selected lepton pair. Due to a better description of the JES in the forward region, it was possible to extend the selection of these jets up to pseudorapidity $|\eta| < 4.7$. Furthermore, the p_T thresholds were reduced to $p_T > 50$ GeV for the leading jet and $p_T > 30$ GeV for the second jet.

The cross section measurement is again carried out using events with $y_Z^* < 1.2$, as higher y_Z^* variables are still badly described by simulation. In addition, we require the tagging jets to have a dijet invariant mass exceeding $M_{jj} > 200$ GeV, to be well above the minimum dijet invariant mass in the EW Zjj MC sample. It also purifies the search region, as the backgrounds are strongly dominating over the signal for M_{jj} values below 200 GeV.

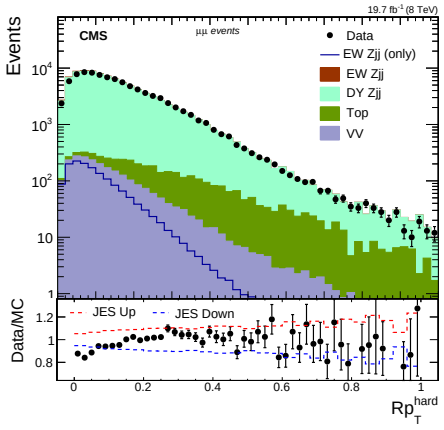


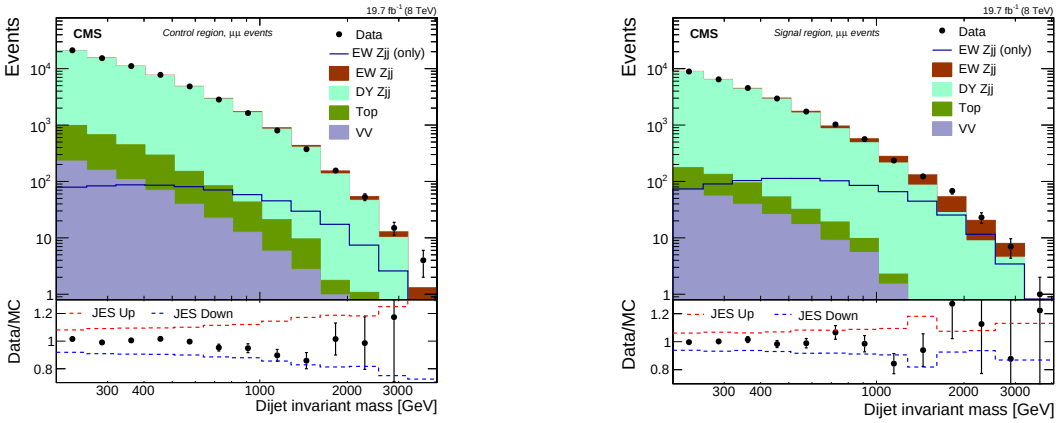
Figure 54: Distribution of p_T^{hard} in the dimuon channel for Zjj events with dijet invariant mass $M_{jj} > 200$ GeV. The contributions of the signal and different background sources are shown stacked in which the background sources are grouped in DY Zjj (up to 4 additional partons), top ($t\bar{t}$ and single top contributions), VV (including WW, WZ, ZZ and W+jets). The events from QCD simulation did not pass initial event selection criteria. The bottom panel shows the ratio between data and background expectation, as well as the uncertainty envelope due the JES.

In order to verify the agreement between data and the background estimates by MC, we have introduced a control and signal region based on an event balance variable, Rp_T^{hard}, defined as

$$Rp_T^{\text{hard}} = \frac{|\vec{p}_T^{\,j_1} + \vec{p}_T^{\,j_2} + \vec{p}_T^{\,Z}|}{|\vec{p}_T^{\,j_1}| + |\vec{p}_T^{\,j_2}| + |\vec{p}_T^{\,Z}|} = \frac{|\vec{p}_T^{\,\text{hard}}|}{|\vec{p}_T^{\,j_1}| + |\vec{p}_T^{\,j_2}| + |\vec{p}_T^{\,Z}|} \tag{90}$$

The Rp_T^{hard} is therefore an estimator for the p_{T} of the hard process relative to the sum of the transverse momenta of the two jets and the Z-boson. This variable, shown in Figure 54, is expected to peak at 0, especially for the signal were these three objects (the tagging jets and the Z-boson) are balanced with respect to each other. For background processes, extra jets or missing objects are expected to spoil the balance, resulting in a larger deviation from 0. The signal region is defined by the events which pass $Rp_T^{\text{hard}} < 0.14$, in which the cut value was chosen by optimizing $\frac{N_S}{\sqrt{N_B}}$, whereas events failing this requirement are used as the control region. The purpose of the control and signal region is illustrated in Figure 55 which shows the dijet invariant mass M_{jj} in the control region, where a good agreement between data and MC is observed, and the signal region, where the background processes are strongly reduced and the signal has a higher relative contribution.

Figure 56 shows the pseudorapidity separation $\Delta\eta_{jj}$ between the two tagging jets and the z_Z^* variable in the control region. As for the 7 TeV analysis, the $\Delta\eta_{jj}$ distribution shows a slight excess of data events with respect to MC towards higher $\Delta\eta_{jj}$



(a) control region

(b) signal region

Figure 55: Distribution of the dijet invariant mass M_{jj} of the tagging jets for $\mu\mu$ events in the signal ($Rp_T^{\text{hard}} < 0.14$) and control ($Rp_T^{\text{hard}} > 0.14$) region.

values, albeit still within the JES uncertainty. The z_Z^* distribution, which follows equation 50 for the Z-boson:

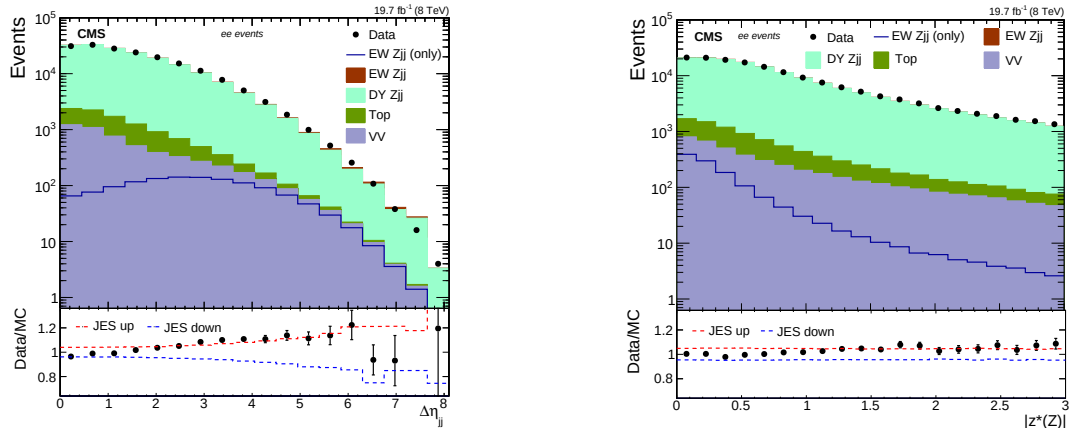
$$z_Z^* = \frac{y_Z^*}{\Delta y_{jj}} \tag{91}$$

shows a very good agreement between data and MC. This variable is closely related to the y_Z^* variable, used in the 7 TeV analysis and showed an upward trend for higher y_Z^* values, but the division by Δy_{jj} allows to cancel out the data/MC disagreements such that the z_Z^* distribution is reasonably flat. The z_Z^* distribution also confirms the rapidity of the Z-boson lays in between the rapidities of the tagging jets ($|z_Z^*| < 0.5$) for the vast majority of signal events. This behaviour is less pronounced for the background processes.

8.3 SCALE FACTORS

8.3.1 *Pile-up reweighting*

The MC events are reweighted in order to match the true pile-up distribution in data, which was calculated using the bunch-by-bunch luminosity, assuming a total pp inelastic cross section of 69.4 mb.



(a) $\Delta\eta_{jj}$ (b) $|z_Z^*|$

Figure 56: Distributions of the pseudorapidity difference $\Delta\eta_{jj}$ between the tagging jets and the z_Z^* variable for *ee* events in the control region.

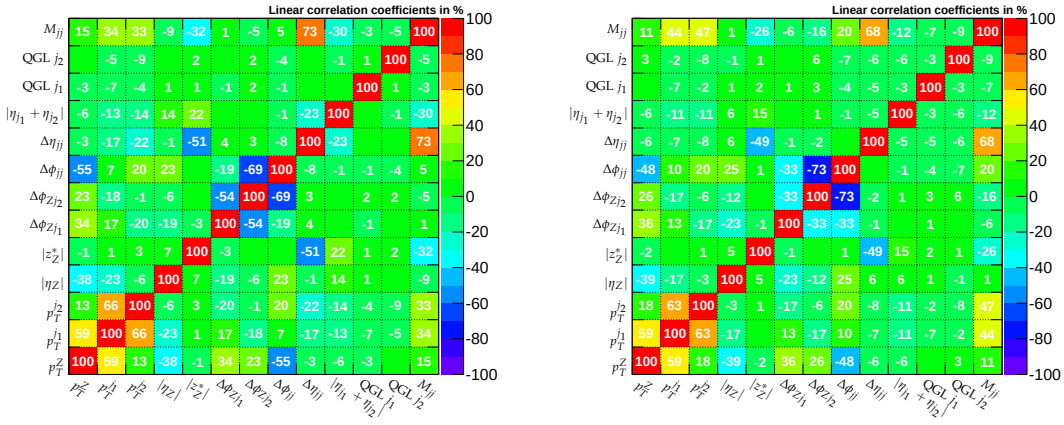
8.3.2 *Lepton selection efficiencies*

The data/MC scale factors for the trigger and analysis lepton selections were implemented in both the electron and muon channel. The scale factors were derived and implemented as a function of the p_T and η of the lepton [114].

8.4 SIGNAL EXTRACTION

8.4.1 *Multi-variate analysis*

The EW Zjj and DY Zjj components of inclusive $lljj$ spectrum are again optimally separated by use of a BDT discriminator, provided by the TMVA framework. The choice of variables was, however, re-evaluated by studying the correlations among them as well as the stability of the input variables with respect to the agreement between data and MC. Figure 57 shows the linear correlation coefficients of the studied discrimination variables, which include the input variables from the 7 TeV analysis and the z_Z^* variable. It is easily observed that the quark-gluon likelihood values for both of the tagging jets have no correlations with the other variables, hence they add independent discrimination power to the analysis and are again selected for the BDT. Among three of the most powerful discriminating variables, M_{jj}, $\Delta\eta_{jj}$ and z_Z^*, strong correlations are observed. We have therefore chosen to drop the $\Delta\eta_{jj}$ variable from the BDT, because its agreement between data and MC is slightly worse compared to the other two. Likewise, the p_T of each of the tagging jets was dropped, as the MC overestimates the DY contribution at higher values of $p_T^{j_1}$ and $p_T^{j_2}$, a feature which was also observed in the 7 TeV analysis (Figure 49).



(a) Correlation matrix for EW Zjj

(b) Correlation matrix for DY Zjj

Figure 57: Correlation matrix of the studied input variables for both signal and background samples

Leaving out the transverse momenta also reduces the influence of the JES on the BDT, resulting in a smaller uncertainties Finally, the variables describing the azimutal angles between the jets and the Z-boson, even though reasonably described by MC, were not used for the 8 TeV analysis due to their smaller separation power. This leaves a BDT based on 7 strong and independent variables, shown in Figure 58, for which their data distributions are very well described by the simulation. The BDT distributions are shown for the control and signal region in Figure 59. The control region shows fair agreement between data and MC within statistical and JES uncertainties. In the signal region, the contribution of the EW Zjj component is clearly noticeable, and dominates over the DY background for high BDT values.

8.4.2 *Fit procedure*

A new procedure, which is similar to the methodology used for the CMS Higgs analysis [115] using asymptotic formulas [116], has been adopted to extract the signal cross section from the BDT signal, background and data shapes. The goal of the fit procedure is to determine the strength factors μ and ν for the EW Zjj and DY Zjj processes respectively as

$$\mu \equiv \sigma(\text{EW } Zjj)/\sigma_{\text{LO}}(\text{EW } Zjj)$$
$$\nu \equiv \sigma(\text{DY } Zjj)/\sigma_{\text{th}}(\text{DY } Zjj)$$

in which $\sigma_{\text{LO}}(\text{EW } Zjj)$ is the MADGRAPH cross section for the signal sample and $\sigma_{\text{th}}(\text{DY } Zjj)$ the theoretical NNLO cross section for the DY process, both given in Table 7. Naively, one could estimate the number of events expected in each bin of the BDT data histogram to be the sum of the signal and background contributions

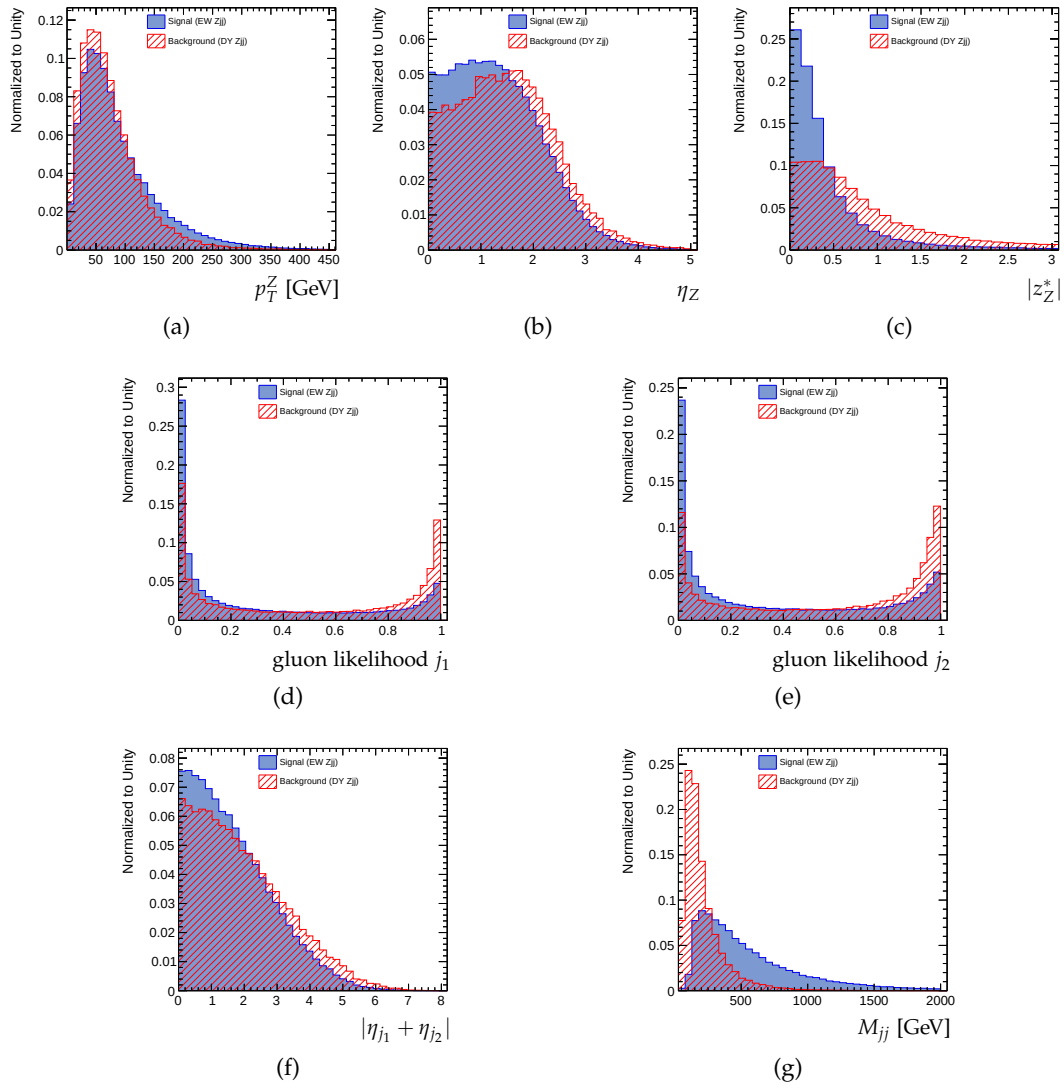
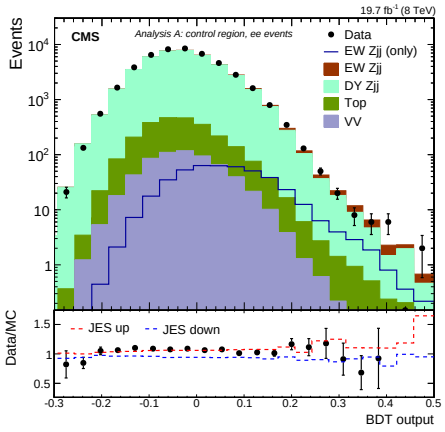
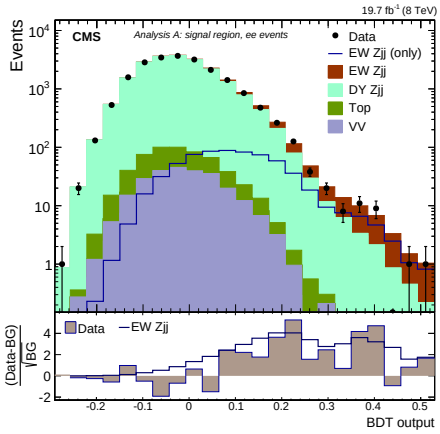


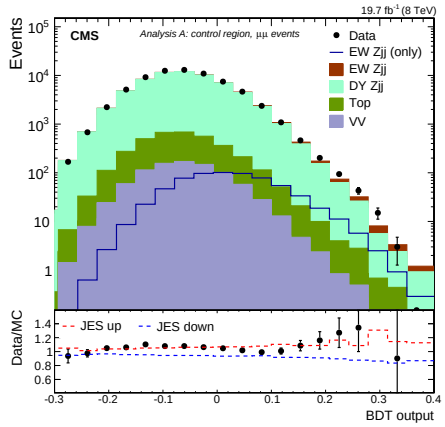
Figure 58: Normalized distributions for signal and background for each of the BDT input variables, for Zjj events with $M_{jj} > 200$ GeV.



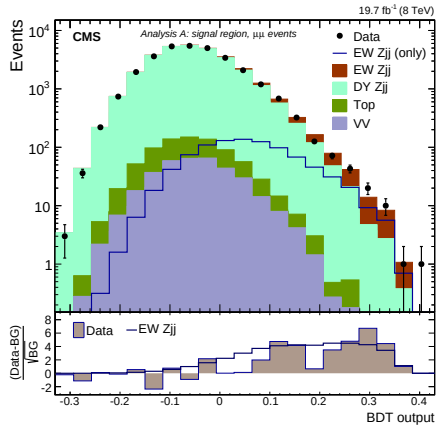
(a) control region, *ee* events



(b) signal region, *ee* events



(c) control region, $\mu\mu$ events



(d) signal region, $\mu\mu$ events

Figure 59: Output distributions for the BDT discriminants for the control and signal region in *ee* and $\mu\mu$ events. For the control region, the bottom panels show the ratio of data to MC and the impact on the MC shape by shifting the JES up and down by one standard deviation. In the signal region, the bottom panels show the significance observed in data compared to those expected for the signal.

in which the two main components, EW Zjj and DY Zjj, are weighted with μ and ν respectively. However, we need to take into account the interference between those two contributions, and we use therefore a more complicated parametric model:

$$\hat{N}^{lljj}(\mu,\nu) = \mu N_{\text{EW } Zjj} + \sqrt{\mu\nu} N_I + \nu N_{\text{DY } Zjj} + N_{\text{res}} \tag{92}$$

in which $N_{\text{EW } Zjj}$, $N_{\text{DY } Zjj}$ and $N_{\text{res}}$are the yields for the signal, DY and residual backgrounds respectively, and N_I is the expected contribution of the interference to the total yield, estimated from equation 51.

The parameters of the model, μ and ν, are determined by a binned maximum likelihood fit, in which the likelihood $\mathcal{L}$ is built assuming a Poisson distribution in each bin. The effect of the systematic uncertainties on the expected rates and shapes of the BDT distribution can be taken into account by treating them as a set of nuisance parameters $\boldsymbol{\theta}$. This is achieved by scanning the profile likelihood ratio test statistic $\lambda(\mu,\nu)$, defined as

$$\lambda(\mu,\nu) = \frac{\mathcal{L}(\mu,\hat{\hat{\nu}},\hat{\hat{\boldsymbol{\theta}}})}{\mathcal{L}(\hat{\mu},\hat{\nu},\hat{\boldsymbol{\theta}})} \tag{93}$$

The estimators $\hat{\mu}$,$\hat{\nu}$ and $\hat{\boldsymbol{\theta}}$ in the denominator denote the unconditional maximum likelihood estimates of these parameters. The numerator contains $\hat{\hat{\nu}}$ and $\hat{\hat{\boldsymbol{\theta}}}$ which denote the conditional maximum likelihood estimate for a given signal strength μ. The optimal values for μ, ν are found by scanning for the largest value of $\lambda(\mu,\nu)$, and immediately yield the strength of the nuisance parameters as given by the maximization of $\mathcal{L}(\mu,\nu,\hat{\hat{\boldsymbol{\theta}}})$.

8.4.3 *Fit results*

The fitted signal strengths in the electron and muon channel are given by

$$\mu = 0.82 \pm 0.11\ (\text{stat}) \pm 0.19\ (\text{syst}) \qquad (eejj \text{ events})$$
$$\mu = 0.86 \pm 0.10\ (\text{stat}) \pm 0.18\ (\text{syst}) \qquad (\mu\mu jj \text{ events})$$

The maximum likelihood fit approach also allows to combine both channels assuming lepton universality, obtaining the signal strength

$$\mu = 0.84 \pm 0.07\ (\text{stat}) \pm 0.19\ (\text{syst})$$

and is therefore compatible with the SM prediction at LO by MADGRAPH.

8.4.4 *Systematic uncertainties*

Most of the systematic uncertainties are treated as nuisance parameters, and their post-fit parameters are listed in Table 8. In order to estimate the systematic uncertainty associated with a given nuisance parameter, the fit was repeated while

keeping the other nuisance parameters fixed such that the nuisance parameter of interest was the only one contributing to the retrieved systematic error. In this section, we detail the theoretical and experimental uncertainties and discuss their effect on the fit.

8.4.4.1 *Theoretical uncertainties*

- **Background modelling**
  The effect of virtual corrections to the MADGRAPH-based description of the DY Zjj process is again studied using MCFM. The DY Zjj process is calculated in MCFM at LO and NLO, such that their distributions could be compared at parton-level. In this analysis, the scale factors were derived as a function of both M_{jj} and y_Z^*. The scale factors are found to increase steeply with $|y_Z^*|$, which again confirms the need to avoid events with $|y_Z^*| > 1.2$ entering the cross section measurement. Unlike the other systematics, this uncertainty is not treated as nuisance parameter. Instead, the analysis is repeated with the MCFM-reweighted BDT histogram to assign a value for this uncertainty. The MCFM-reweighted fit yields $\mu = 0.88$ in the electron channel and $\mu = 0.89$ in the muon channel, which results in an uncertainty of 0.06 and 0.03 in the electron and muon channel respectively. The DY Zjj strength parameter ν is constrained between 0.9 and 1.1. The post-fit values for the DY normalization, listed in Table 8, result in an estimated DY strength of 0.962 ± 0.070 and 0.972 ± 0.072 in the electron and muon channel respectively. The DY normalization yields an uncertainty on the signal fit of 0.04 in both channels. Combining the uncertainties on the DY shape and its normalization, we find 0.07 and 0.05 for the electron and muon channel respectively.

- **Signal acceptance**
  Studies performed using different MC generators indicate that the uncertainty on the signal acceptance is not larger than 6%. When performing the fit without this nuisance, the signal strength is affected with 0.03 in the electron channel and 0.04 in the muon channel. The uncertainty on the signal acceptance directly works on the signal strength μ and does not affect the backgrounds. The central value and error of the associated nuisance parameter is therefore not shifted after the fit.

- **Normalization of residual backgrounds**
  The cross sections of the top and diboson backgrounds have an uncertainty arising from the parton density functions and factorization/renormalization scales. The cross section uncertainties are based on references [58, 117, 33] and yield a negligible uncertainty on the signal strength of less than 1%.

- **Interference between EW Zjj and DY Zjj**
  The difference observed when repeating the fit without its interference term relative to the nominal result is used to estimate the uncertainty on the in-

name	b-only fit	$s+b$ fit	$\rho(\theta,\mu)$
DY norm	-0.78 ± 0.66	-0.38 ± 0.70	+0.21
Top norm	-0.07 ± 0.99	-0.02 ± 0.99	+0.01
VV norm	$+0.02 \pm 0.99$	-0.01 ± 0.99	-0.01
JER	$+0.70 \pm 0.74$	-0.11 ± 0.79	-0.03
JES	$+0.79 \pm 0.60$	-0.29 ± 0.77	-0.22
JES norm	-0.35 ± 0.94	-0.18 ± 0.94	-0.00
pile-up	$+0.35 \pm 0.95$	$+0.22 \pm 0.94$	+0.01
quark-gluon tagger	$+0.01 \pm 0.84$	$+0.19 \pm 0.77$	+0.03
signal acceptance	$+0.00 \pm 1.00$	$+0.00 \pm 1.00$	-0.30
lepton efficiencies	-0.27 ± 0.96	-0.13 ± 0.96	-0.11
luminosity	-0.23 ± 0.97	-0.12 ± 0.97	-0.09
bin 1	$+0.00 \pm 0.98$	$+0.00 \pm 0.98$	-0.00
bin 2	$+0.00 \pm 0.98$	$+0.00 \pm 0.98$	-0.00
bin 3	$+0.00 \pm 0.98$	$+0.00 \pm 0.98$	-0.00
bin 4	$+0.03 \pm 0.84$	$+0.08 \pm 0.84$	+0.01
bin 5	$+0.03 \pm 0.84$	-0.07 ± 0.86	-0.01
bin 6	-0.29 ± 0.79	-0.08 ± 0.80	+0.01
bin 7	-0.27 ± 0.80	-0.14 ± 0.82	+0.06
bin 8	$+0.31 \pm 0.79$	$+0.73 \pm 0.83$	+0.12
bin 9	-0.53 ± 0.80	$+0.07 \pm 0.81$	+0.10
bin 10	-1.46 ± 0.82	-0.90 ± 0.83	+0.11
bin 11	-0.41 ± 0.84	$+0.00 \pm 0.85$	+0.07
bin 12	$+0.30 \pm 0.83$	$+0.51 \pm 0.84$	+0.04
bin 13	-0.63 ± 0.81	-0.78 ± 0.81	-0.03
bin 14	$+1.29 \pm 0.79$	$+0.74 \pm 0.81$	-0.11
bin 15	$+0.89 \pm 0.83$	$+0.24 \pm 0.85$	-0.17
bin 16	$+0.63 \pm 0.79$	-0.31 ± 0.83	-0.21
bin 17	$+1.49 \pm 0.81$	$+0.18 \pm 0.87$	-0.20
bin 18	$+1.73 \pm 0.82$	$+0.63 \pm 0.88$	-0.21
bin 19	$+0.40 \pm 0.81$	-0.52 ± 0.88	-0.14
bin 20	$+0.82 \pm 0.83$	-0.07 ± 0.91	-0.12
bin 21	$+0.09 \pm 0.85$	-0.54 ± 0.93	-0.05
bin 23	$+1.29 \pm 0.81$	$+0.32 \pm 0.92$	-0.09
bin 24	-0.66 ± 0.98	-0.64 ± 0.98	+0.00
bin 25	$+0.17 \pm 0.87$	-0.09 ± 0.95	-0.02
bin 26	$+0.59 \pm 0.76$	$+0.01 \pm 0.95$	-0.03
bin 27	-0.12 ± 0.98	-0.12 ± 0.98	+0.00
μ		$+0.82 \pm 0.16$	+1.00

(a) *ee* channel

name	b-only fit	$s+b$ fit	$\rho(\theta,\mu)$
DY norm	-0.58 ± 0.65	-0.28 ± 0.72	+0.28
Top norm	-0.08 ± 0.99	-0.00 ± 0.99	+0.01
VV norm	$+0.02 \pm 0.99$	$+0.01 \pm 0.99$	-0.00
JER	$+0.15 \pm 1.00$	-0.10 ± 0.78	-0.01
JES	$+0.52 \pm 0.61$	-0.38 ± 0.81	-0.28
JES norm	-0.27 ± 0.94	-0.12 ± 0.95	+0.01
pile-up	$+0.66 \pm 0.91$	$+0.29 \pm 0.88$	-0.02
quark-gluon tagger	-0.19 ± 0.58	-0.20 ± 0.54	-0.02
signal acceptance	$+0.00 \pm 1.00$	-0.00 ± 1.00	-0.34
lepton efficiencies	-0.21 ± 0.96	-0.08 ± 0.97	-0.11
luminosity	-0.18 ± 0.97	-0.07 ± 0.97	-0.09
bin 1	$+0.00 \pm 0.98$	$+0.00 \pm 0.98$	-0.00
bin 2	$+0.00 \pm 0.98$	$+0.00 \pm 0.98$	-0.00
bin 3	$+0.06 \pm 0.85$	-0.04 ± 0.87	+0.07
bin 4	-0.74 ± 0.73	-0.57 ± 0.73	-0.04
bin 5	-0.08 ± 0.78	$+0.17 \pm 0.76$	+0.02
bin 6	-0.21 ± 0.75	$+0.05 \pm 0.75$	+0.04
bin 7	-0.01 ± 0.79	$+0.35 \pm 0.80$	+0.02
bin 8	-1.69 ± 0.81	-0.98 ± 0.82	+0.11
bin 9	-0.15 ± 0.83	$+0.59 \pm 0.84$	+0.15
bin 10	-0.83 ± 0.83	-0.28 ± 0.83	+0.10
bin 11	$+0.64 \pm 0.85$	$+0.92 \pm 0.85$	+0.00
bin 12	-0.08 ± 0.80	-0.28 ± 0.81	-0.02
bin 13	$+0.10 \pm 0.82$	-0.61 ± 0.83	-0.10
bin 14	$+0.85 \pm 0.80$	-0.19 ± 0.83	-0.14
bin 15	$+2.00 \pm 0.78$	$+0.67 \pm 0.83$	-0.22
bin 16	$+1.54 \pm 0.82$	$+0.16 \pm 0.87$	-0.21
bin 17	-0.03 ± 0.81	-1.28 ± 0.87	-0.17
bin 18	$+1.12 \pm 0.79$	-0.25 ± 0.88	-0.17
bin 19	$+1.63 \pm 0.78$	$+0.18 \pm 0.88$	-0.15
bin 20	$+1.77 \pm 0.83$	$+0.59 \pm 0.91$	-0.12
bin 21	$+1.18 \pm 0.80$	$+0.22 \pm 0.92$	-0.08
μ		$+0.86 \pm 0.15$	+1.00

(b) $\mu\mu$ channel

Table 8: Post-fit nuisance parameters for the background-only fit (in which the signal strength μ is fixed to zero) and for the fit with floating signal strength ($s+b$ fit). Each of those columns shows the shift in value and its uncertainty, relative to one standard deviation of the uncertainty. The last column shows $\rho(\theta,\nu)$ which is the correlation of the nuisance parameter to the strength parameter μ, and can be used to assess the importance of the systematic uncertainty for the measurement.

terference between signal and background. The difference in fitted signal strength results in an uncertainty of 0.14 for both the electron and muon channel.

8.4.4.2 *Experimental uncertainties*

- **Jet energy scale and resolution**
  As was done for the 7 TeV analysis, the transverse momenta of the jets are shifted up and down according to their JES uncertainty on an event-by-event basis, and its effect is propagated to construct an up and down variant of the BDT templates for each of the signal and background samples. In this way, an envelope is defined around the nominal BDT distribution, for which the up and down variations correspond to the -1 and $+1$ values of the nuisance parameters. The JES does not only affect the shape of the BDT distribution, but has also an effect on the expected event yields through the selection criteria. A separate nuisance parameter is therefore added to quantify the JES normalization. The total systematic uncertainty due to the JES shape and normalization is estimated at 0.06 in the *ee* channel and 0.05 in $\mu\mu$ channel.

  A similar approach is used for the JER: for each event, the p_{T} distribution of the jets is altered using the smearing factors given by Figure 32, and its effect is used to create a new BDT template. The systematic uncertainty for the JER is 0.02 in both channels.

- **Pile-up modelling**
  Assuming an uncertainty of 5% on the total inelastic cross section, an up and down variation on the data pile-up distribution was calculated. Repeating the BDT analysis using this modified pile-up reweighting yields another up and down variation for the BDT templates. The effect of the associated nuisance parameter on the fit is 0.01 in the electron channel, and 0.02 in the muon channel.

- **Quark-gluon discrimination**
  Following the prescription in Section 6.6, the quark-gluon likelihood distribution has been smeared for both of the tagging jets, and propagated to a new BDT template. The impact of this smearing on the fitted signal strength is rather low, making this uncertainty negligible (< 0.01) with respect to the other uncertainties.

- **Lepton selection efficiencies**
  A total 3% uncertainty is assigned for the trigger and selection efficiencies of the leptons. This yields an uncertainty of 0.04 on the fitted signal strength in both channels.

- **Statistics of simulation**
  The DY sample has few statistics in the high BDT region, hence we have to take into account the effect of its statistical fluctuations on the likelihood fit. Because statistical fluctuations behave independently in each bin of the BDT template, a separate nuisance parameter was added for each of the bins. Most of these nuisance parameters have shown little impact on the fitted signal strength, with only a few bins yielding an uncertainty close to 0.01.

- **Luminosity**
  A 2.6% uncertainty is assigned to the value of the luminosity [118], and its associated uncertainty on the fitted signal strength is 0.03.

8.5 COMPARISON WITH OTHER ANALYSIS METHODS

In addition to the main 8 TeV analysis described in this thesis (*analysis A*), two other parallel analyses were independently developed within the CMS collaboration. Although these analysis use the same or similar baseline selection and also use a BDT method to discriminate the signal from the background, their are subtle differences in the final implementation. In general, they show very good agreement with analysis A and together they make a strong and robust case for the cross section measurement.

8.5.1 *Analysis B: using JPT jets*

Analysis B repeats the analysis using JPT jets which was also carried out at 7 TeV, again only making use of the dimuon channel. They uses the same baseline selection as analysis A, but their BDT is built on a larger and different set of input variables, listed in Table 9. Also the fit procedure and nuisance parameters are handled in exactly the same way as analysis A. The resulting fit for the signal strength is $\mu = 0.89 \pm 0.09\,(\text{stat}) \pm 0.17\,(\text{syst})$, confirming the results of the muon channel of analysis A. Also each of the systematic uncertainties, listed in Table 10, is found in excellent agreement as those obtained in analysis A.

8.5.2 *Analysis C: data-driven approach*

A third analysis introduces a data-driven background model for the DY Zjj background. The production of $\gamma + 2$ jets is expected to closely resemble the production of DY Zjj and could therefore provide a possible data-based model to describe the kinematics of the tagging jets. This approach has the advantage not being sensitive to imperfections and mismodelling in the LO MC description used for analysis A

and B. On the other hand, it requires to correct for the the differences between the photon and Z samples, which are mostly mitigated by reweighting the p_{T} of the photon to the transverse momentum of the Z boson. Because the low p_{T} region in the photon sample has to deal with backgrounds from QCD multijet production, and also is affected by prescaled triggers, the photon or Z boson were required to have $p_{\mathrm{T}} > 50$ GeV. Furthermore, in this analysis the rapidity of the photon and Z boson were limited to $|y| < 1.4442$, the physical boundary of the barrel region of the CMS ECAL. From simulation, it was deduced there are no significant differences in dijet kinematic between the photon and Z samples if one requires $M_{jj} > 2M_Z$.

The data-driven approach needs also to deal with the contamination of EW γjj, which of course yields signal-like distributions for the tagging jets. Fortunately, from simulation it is estimated that the ratio of EW γjj to the total number of photon events is a factor ~ 5 times smaller than the ratio expected between EW Zjj and DY Zjj yields. Nevertheless, this contamination is subtracted from the data-driven background distributions for DY Zjj.

Similar to analysis A, the data-driven analysis makes use of PF jets and uses a BDT for the signal extraction. Due to the main background being derived from the photon control sample, it needs to avoid variables related to the Z boson or leptons, but instead is relying solely on input variables related to the tagging jets listed in Table 9. The events were split up in four categories for M_{jj} values in the intervals $450 - 550$ GeV, $550 - 750$ GeV, $750 - 1000$ GeV, and above 1000 GeV, which have been chosen to have similar numbers of signal events.

The fit of the BDT templates, shown in Figure 61, yields a signal strength of 0.88 ± 0.16 (stat) ± 0.18 (syst) for the combination of the ee and $\mu\mu$ channels. The fit results of the individual ee and $\mu\mu$ channels, are listed in 10, together with the systematic uncertainties. The data-driven analysis shows a smaller uncertainty for the interference between EW Zjj and DY Zjj, which can be explained by its selection of events with higher dijet invariant masses, where the interference effects become less strong (see Figure 12). On the other hand, this analysis has a higher statistical uncertainty as well as a higher uncertainty on the DY Zjj shape.

8.6 SUMMARY

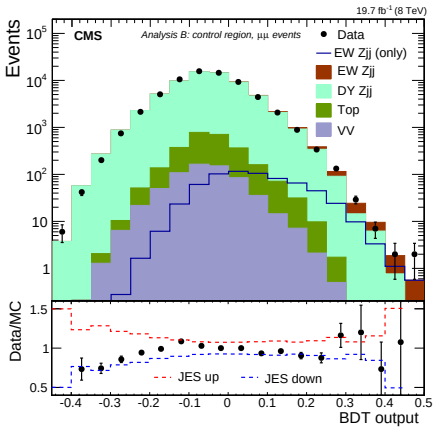
Using the signal strength obtained by analysis A, $\mu = 0.84 \pm 0.07$ (stat) ± 0.19 (syst), we measured a cross section of

$$\sigma_{ll}^{\mathrm{EW}} = 174 \pm 15 \text{ (stat.)} \pm 40 \text{ (syst.) fb} = 174 \pm 42 \text{ (total) fb}$$

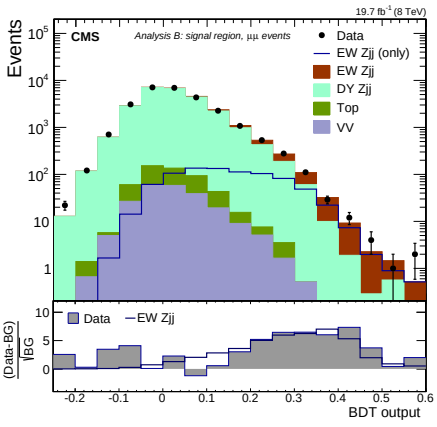
in the kinematic region defined by $M_{ll} > 50$ GeV, $M_{jj} > 120$ GeV, $p_T^j > 25$ GeV and $|\eta_j| < 5$. The background-only hypothesis is excluded with a significance greater than 5σ.

Table 9: Comparison of the selections and variables used in three different analyses.

	Analysis A	*Analysis B*	*Analysis C*
Channels	ee, $\mu\mu$	$\mu\mu$	ee, $\mu\mu$ binned in M_{jj}
Selection	$p_T^{j_1,j_2} > 50, 30$ GeV $Rp_T^{\text{hard}} < 0.14$ $\lvert y^* \rvert < 1.2$ $M_{jj} > 200$ GeV		$p_T^Z > 50$ GeV $\lvert y_Z \rvert < 1.4442$ $M_{jj} > 450$ GeV
Jets	PF	JPT	PF
Variables used			
M_{jj}	•	•	•
$p_T^{j_1}, p_T^{j_2}$		•	•
η_{j_1}, η_{j_2}			•
$\Delta_{\text{rel}}(jj) = \frac{\lvert \vec{p}_T^{\,j_1} + \vec{p}_T^{\,j_2} \rvert}{p_T^{j_1} + p_T^{j_2}}$			•
$\Delta\eta_{jj}$		•	
$\lvert\eta_{j_1}\rvert + \lvert\eta_{j_2}\rvert$	•	•	•
$\Delta\phi_{jj}$		•	•
$\Delta\phi_{Zj_1}$		•	
y_Z		•	
η_Z	•		
z_Z^*	•		
p_T^Z	•	•	
Rp_T^{hard}		•	
q/g discriminator	•		•
DY Zjj model	MC-based	MC-based	data-based

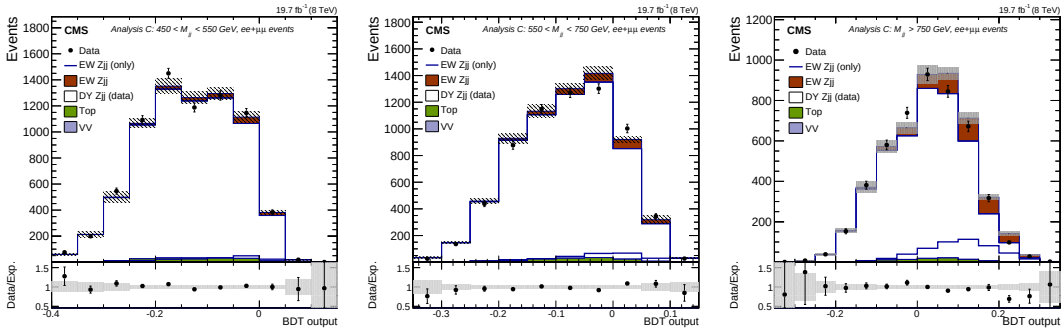


(a) control region, $\mu\mu$ events



(b) signal region, $\mu\mu$ events

Figure 60: Output distributions for the BDT discriminants in the JPT analysis.



(a) $450 < M_{jj} < 550$ GeV (b) $550 < M_{jj} < 750$ GeV (c) $750 < M_{jj}$ GeV

Figure 61: Output distributions for the BDT discriminants in the data-driven analysis, for $ee + \mu\mu$ events in different M_{jj} categories.

Table 10: Comparison of the fitted signal strengths in the different analyses and channels including the statistical and systematic uncertainties.

	Analysis A			Analysis B	Analysis C		
	ee	$\mu\mu$	$ee + \mu\mu$	$\mu\mu$	ee	$\mu\mu$	$ee + \mu\mu$
Luminosity	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Trigger/lepton selection	0.04	0.04	0.04	0.04	0.04	0.04	0.04
JES+residual response	0.06	0.05	0.05	0.04	0.06	0.05	0.05
JER	0.02	0.02	0.02	0.02	0.04	0.04	0.03
Pileup	0.01	0.02	0.02	0.01	0.01	0.01	0.01
DY Zjj	0.07	0.05	0.07	0.08	0.14	0.12	0.13
q/g discriminator	<0.01	<0.01	<0.01	-	<0.01	<0.01	<0.01
Top, dibosons	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Signal acceptance	0.03	0.04	0.04	0.04	0.06	0.06	0.06
DY/EW Zjj interference	0.14	0.14	0.14	0.13	0.06	0.08	0.08
Systematic uncertainty	0.19	0.18	0.19	0.17	0.17	0.17	0.18
Statistical uncertainty	0.11	0.10	0.07	0.09	0.24	0.21	0.16
$\mu = \sigma/\sigma_{\text{th}}$	0.82	0.86	0.84	0.89	0.91	0.85	0.88

9

HADRONIC ACTIVITY IN Zjj EVENTS

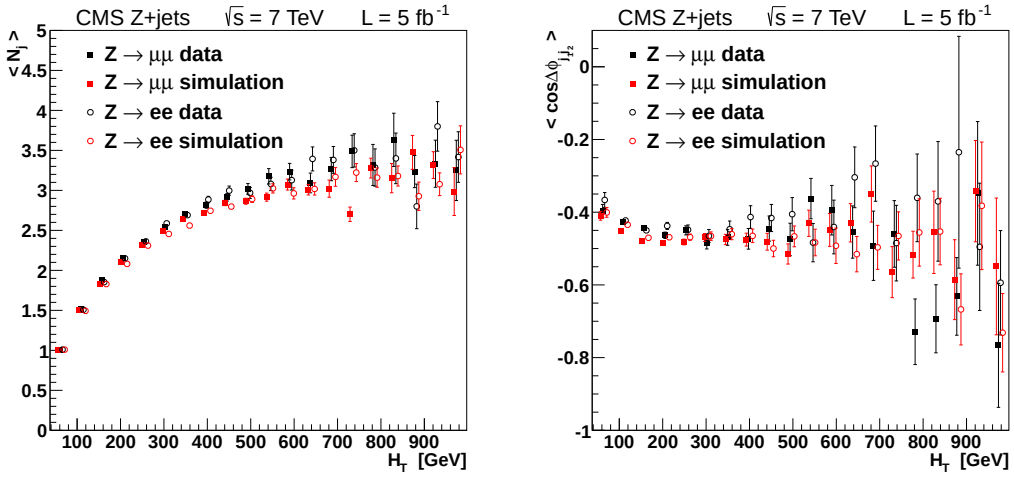
Because the hadronization process in EW Zjj events develops in the forward region, between each of the tagging jets and their proton remnants, little hadronic activity is expected in the central pseudorapidity ranges of the detector. In order to test the properties of the hadronic activity and how they are predicted by simulation, a few studies were carried which are presented in this chapter. Section 9.1 describes the measurement of jet radiation patterns in Z+jet events, while Section 9.2 discusses the hadronic activity in the pseudorapidity interval of the tagging jets in $Z + 2$ jet events, measured using soft track-jets. For completeness, the results from jet activity studies in a high purity EW Zjj signal region, carried out in the framework of the data-driven analysis are given in the last section.

9.1 RADIATION PATTERNS IN Z+JETS EVENTS

The measurements in this section were inspired by Ref. [119], where it was suggested to investigate the additional jet radiation in Vjj events, and a comparison was made between different MC models. A first measurement studied the average number of jets as a function of the total H_T in the event, defined as the total scalar sum of all jets within $|\eta| < 4.7$. This was studied in Z plus at least one jet events, in which the Z boson was selected using the same criteria of the analyses described in Chapters 7 and 8. The jets were required to have a transverse momentum exceeding 40 GeV. Figures 62 and 63 show this measurement in 7 and 8 TeV data respectively. Good agreement is observed between the measurements in electron and muon channels, and both 7 and 8 TeV measurements indicate the jet radiation for high H_T values is slightly underestimated by the MADGRAPH + PYTHIA simulation of DY Zjj. Figure 63 also shows the prediction[1] for the EW Zjj, even though its contribution to data is estimated at less than 1%. As expected, due to the EW Zjj signature of highly energetic tagging jets and low central hadronic activity, even the high values of H_T are dominated by these tagging jets, whereas DY Zjj events have a higher chance of containing a third jet ($p_{\mathrm{T}} > 40$ GeV) contribution to the H_T.

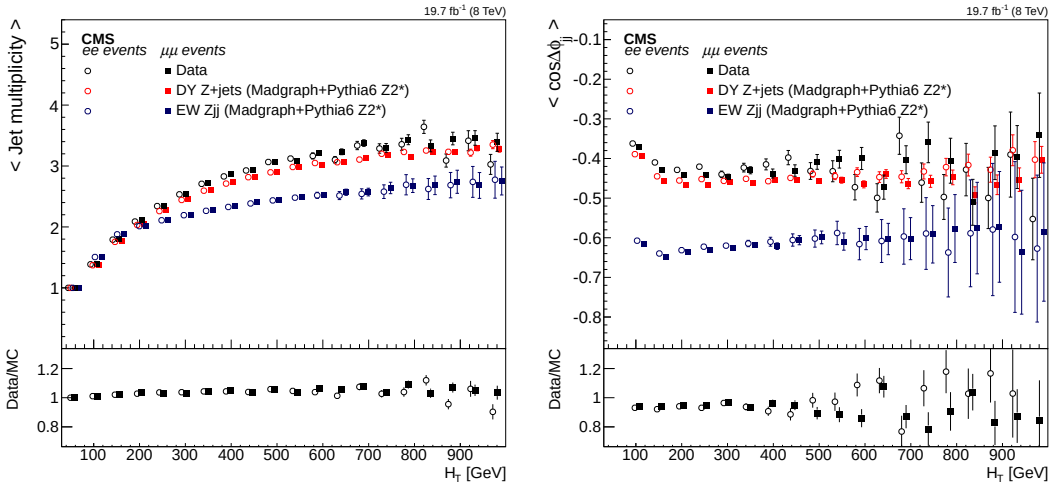
Figures 62 and 63 also show the dependency of $\cos\Delta\phi_{jj}$ on the H_T variable, in which $\cos\Delta\phi_{jj}$ is defined as the cosine of the azimuthal angle difference between

1 Note that it is also possible for EW Zjj to have an average of less than two jets, if one of the two tagging jets does not exceed the $p_{\mathrm{T}} > 40$ GeV requirement.



(a) (b)

Figure 62: Average number of jets with $p_{\mathrm{T}} > 40$ GeV as a function of their total H_T in Z plus at least one jet events and average $\cos\Delta\phi jj$ as a function of the total H_T in events containing a Z boson and at least two jets. The $\sqrt{s} = 7$ TeV data is compared to DY Zjj simulation. Both data and MC points, slightly separated at each ordinate for clarity, are shown with their statistical uncertainties.



(a) (b)

Figure 63: Average number of jets with $p_{\mathrm{T}} > 40$ GeV as a function of their total H_T in Z plus at least one jet events and average $\cos\Delta\phi jj$ as a function of the total H_T in events containing a Z boson and at least two jets. The $\sqrt{s} = 8$ TeV data is compared to simulations for DY Zjj and EW Zjj. The bottom panel shows the ratio of data to the DY Zjj expectation. Both data and MC are shown with their statistical uncertainties.

the two jets with $p_{\mathrm{T}} > 40$ GeV which span the largest pseudorapidity gap in the event[2]. In DY Zjj events, these two jets have an average $\cos\Delta\phi_{jj}$ of about -0.5, independently of H_T, indicating the jets are roughly separated by 120°, and are balanced in the azimuthal plane by the Z-boson or by additional jets in the event. In EW Zjj events, the $\cos\Delta\phi_{jj}$ are found to be lower, and the two jets with highest pseudorapidity span are more likely to balance each other in the azimuthal plane.

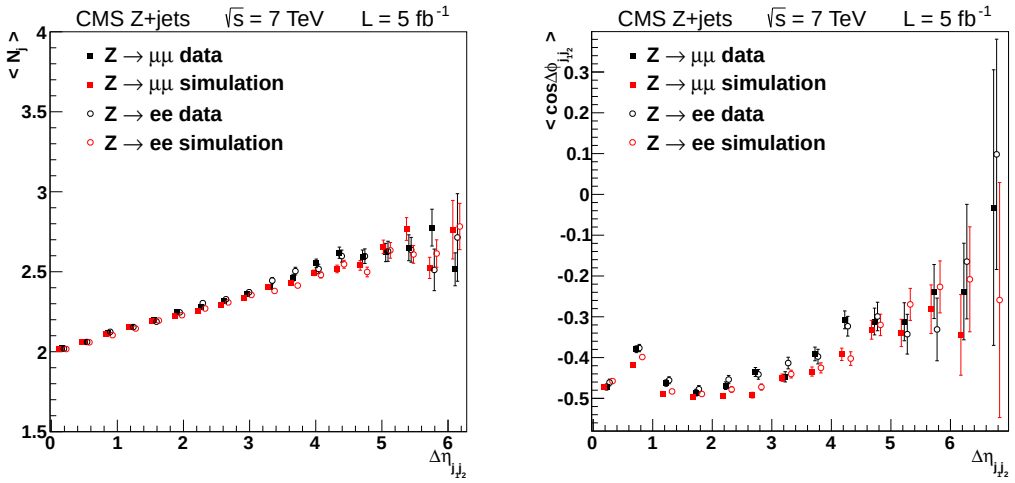
The average number of additional jets and average $\cos\Delta\phi_{jj}$ is also studied as a function of the pseudorapidity separation $\Delta\eta_{jj}$ between the two jets with $p_{\mathrm{T}} > 40$ GeV which span the largest pseudorapidity separation in the event. These measurements are shown in Figure 64 and 65 for the 7 and 8 TeV analyses respectively. As expected, the data follows the DY Zjj prediction which shows a higher ability to radiate additional jets as the pseudorapidity span increases. On the other hand, events in the EW Zjj simulation are less likely to have additional jets, and the average number of jets is independent of $\Delta\eta_{jj}$. The measurement of $\cos\Delta\phi_{jj}$ as a function of $\Delta\eta_{jj}$ shows an interesting feature for data and DY Zjj: the value of $\cos\Delta\phi_{jj}$ in the $0.5 < \Delta\eta_{jj} < 1$ bin is significantly above those of its neighbouring bins. This can be explained by DY Zjj events in which both jets are the result of a gluon splitting $g \to gg$ (see Figure 11c) or $q \to qg$, and as a result yield small values for $\Delta\eta_{jj}$ and $\Delta\phi_{jj}$, such that they pull up the average value for $\cos\Delta\phi_{jj}$. The lowest bin is less affected, as the jet distance parameter ($R = 0.5$) avoids two jets could be reconstructed separately as the splitting partons are too close to each other in η and ϕ.

9.2 CHARGED HADRONIC ACTIVITY USING SOFT TRACK-JETS

In this section we try to focus on the rapidity interval between the tagging jets in Zjj events, as selected by the same requirements as used in the 7 and 8 TeV analyses. Because the hadronic activity is expected to be suppressed in the case of pure EW Zjj events, we need a way to quantify soft (small) hadronic activity without being affected by contributions of pile-up interaction. This can be achieved by a collection of soft track-jets, in which only charged tracks originating from the main PV are considered, and provide a clean method to probe energies down to a few GeV. The soft track-jets are built from tracks passing the following criteria:

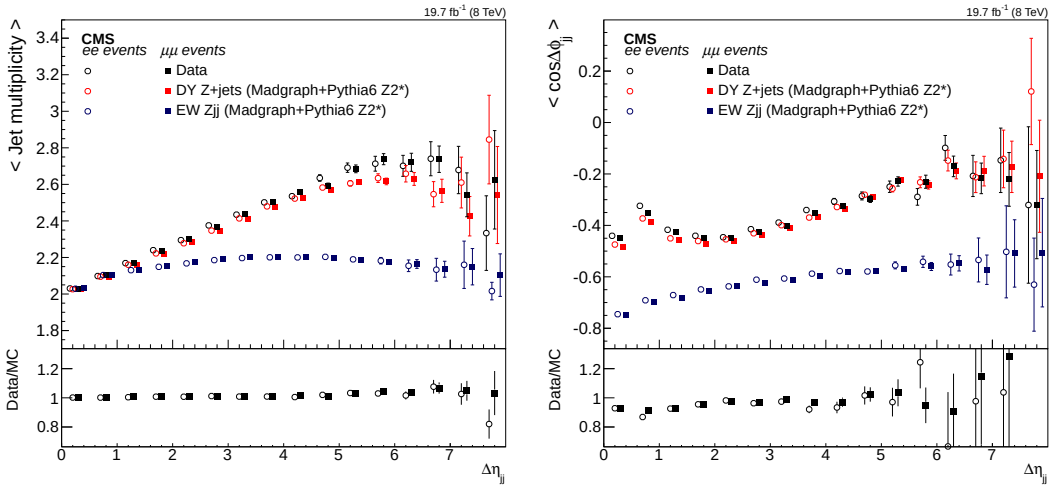
- it has a high-purity flag and has a transverse momentum exceeding 300 MeV
- the uncertainty on its momentum is less than 20%
- it is not associated to one of the two main leptons in the $Z + 2$ jet event

2 The choice for the two jets with largest pseudorapidity separation is based on Ref. [119]. Note that this is not necessarily the pseudorapidity difference between the two hardest (i.e. tagging) jets in the event



(a) (b)

Figure 64: Average number of jets with $p_T > 40$ GeV and $\cos\Delta\phi_{jj}$ as a function of the pseudorapidity span between the two most separated jets ($p_T > 40$ GeV) in η. The $\sqrt{s} = 7$ TeV data is compared to DY Zjj simulation. Both data and MC points, slightly separated at each ordinate for clarity, are shown with their statistical uncertainties.



(a) (b)

Figure 65: Average number of jets with $p_T > 40$ GeV and $\cos\Delta\phi_{jj}$ as a function of the pseudorapidity span between the two most separated jets ($p_T > 40$ GeV) in η. The $\sqrt{s} = 8$ TeV data is compared to simulations for DY Zjj and EW Zjj. The bottom panel shows the ratio of data to the DY Zjj expectation. Both data and MC are shown with their statistical uncertainties.

- it is not associated to PF candidates belonging to the two leading PF jets in the event
- the closest PV to the track along the z-axis is the main PV
- the transverse distance d_z to the main PV is smaller than 2 mm

The selected tracks are clustered together using the anti-k_T algorithm with distance parameter $R = 0.5$. Only soft track-jets found in the pseudorapidity span between the two tagging jets,

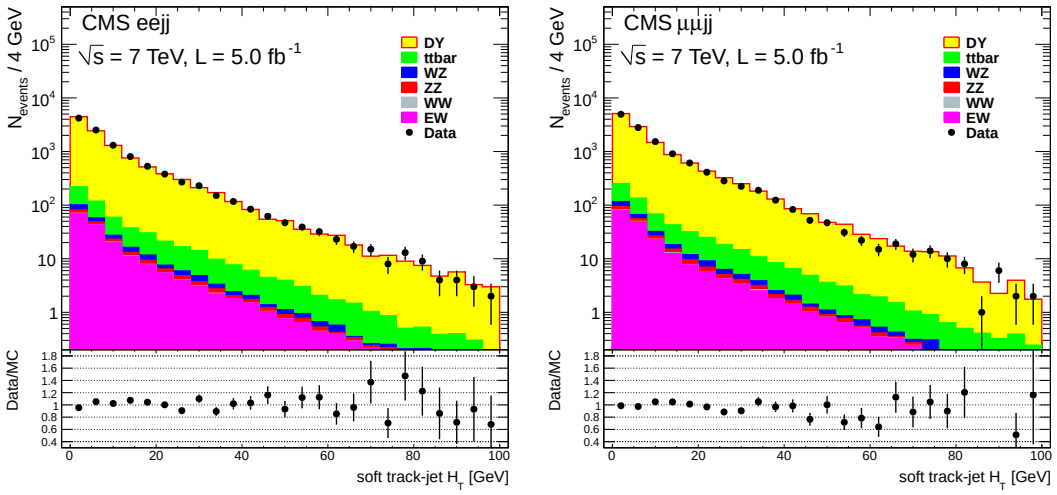
$$\eta_{\text{min}}^{\text{tag jet}} + 0.5 < \eta_{\text{stj}} < \eta_{\text{max}}^{\text{tag jet}} - 0.5 \tag{94}$$

are selected for the study of the central hadronic activity. The three leading soft-track jets ($p_{\text{T}} > 1$ GeV) in this region are combined using the scalar sum of their transverse momenta into the soft track-jet H_T variable. While in this thesis, we only verify the properties of this variable, and its good agreement between simulation and data, note that has also been used as a discriminating variable in the VBF $H \to bb$ analysis [43, 44].

The soft track-jet H_T distribution is shown in Figure 66 and show excellent agreement between data and simulation. It is interesting to note that the soft track-jet H_T distribution is also strongly peaking at 0 for the DY Zjj and other backgrounds, and the soft track-jet variable does not provide much separation power on its own. This is explained by the smaller pseudorapidity spans which occur between the tagging jets, which leave little room for additional hadronic activity. Instead, the strength of the soft track-jet H_T variable appears only when we study it as a function of M_{jj} or $\Delta\eta_{jj}$, shown in Figure 68. For a given M_{jj} or $\Delta\eta jj$ value, the soft track-jet H_T is expected to be significantly lower for EW Zjj signal events compared to DY Zjj. Furthermore, very good agreement between the data and (dominating) DY Zjj simulation is observed. In these plots, the contribution from the EW Zjj signal is estimated to be fairly low, at the level of 1%. However, it is expected to evolve as a function of the different values, and it reaches up to 20% for the highest M_{jj} values ($M_{jj} > 1$ TeV). It is also worth to note how the soft track-jet H_T scales linearly with $\log M_{jj}$ and saturates for $\Delta\eta_{jj} > 5$. The latter is easily explained by the limited acceptance of the CMS tracker.

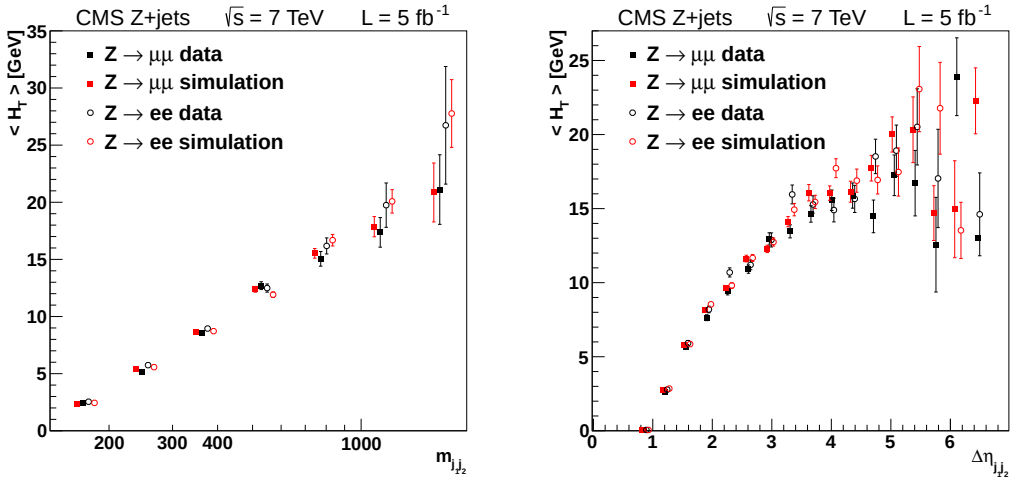
9.3 CENTRAL JET ACTIVITY STUDIES IN A HIGH-PURITY SIGNAL REGION

In addition to the above measurements, carried out in the framework of this thesis, the additional jet activity within pseudorapidity span of the tagging jets was also studied by the data-driven analysis. They focused on Zjj events with a dijet invariant mass exceeding 1250 GeV, allowing to probe a region with high signal purity. Figure 69 shows the number of jets with $p_{\text{T}} > 15$ GeV found in the pseudorapidity interval between the tagging jets, as well as the scalar sum H_T of their transverse



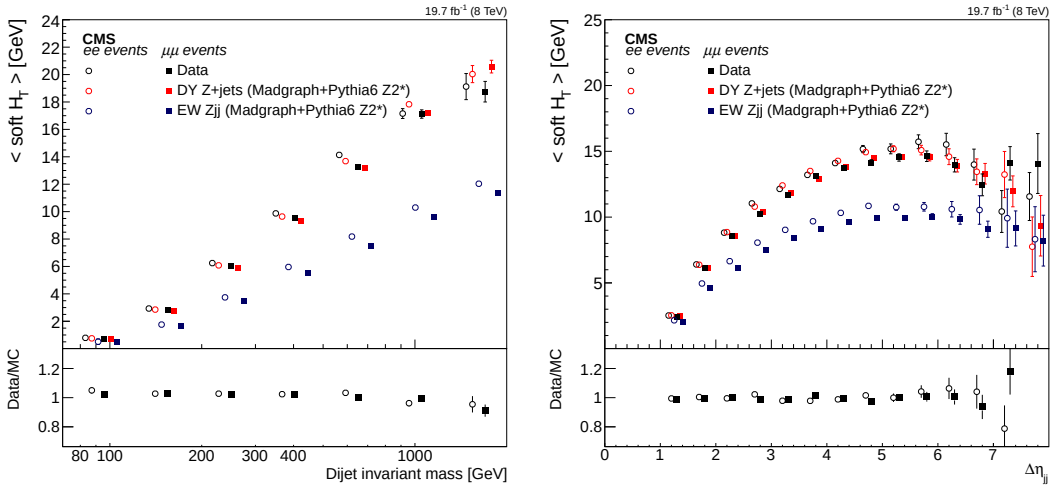
(a) *ee* events (b) $\mu\mu$ events

Figure 66: Soft track-jet H_T distribution, calculated as the scalar sum of the transverse momenta of the tree leading soft track-jets ($p_T > 1$ in the pseudorapidity interval between the tagging jets. The data, using Zjj events with $p_T^{j_1,j_2} > 65,40$ GeV at $\sqrt{s} = 7$ TeV, are compared to the stacked contributions of signal and backgrounds.



(a) soft-track jet H_T vs. M_{jj} (b) soft-track jet H_T vs. $\Delta\eta_{jj}$

Figure 67: Soft track-jet H_T in the pseudorapidity interval between the tagging jets, as a function of M_{jj} and $\Delta\eta_{jj}$. The $\sqrt{s} = 7$ TeV data is compared to the DY Zjj simulation.

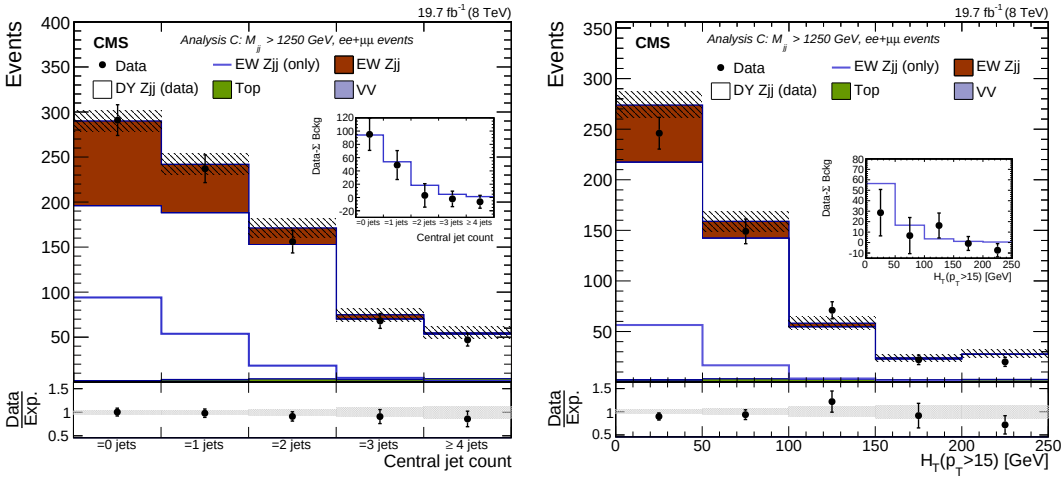


(a) soft-track jet H_T vs. M_{jj}

(b) soft-track jet H_T vs. $\Delta\eta_{jj}$

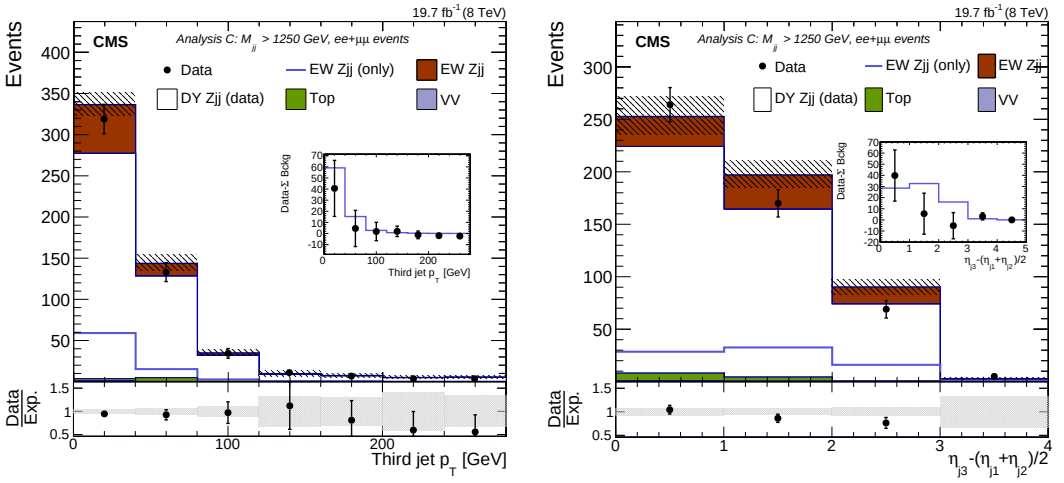
Figure 68: Soft track-jet H_T in the pseudorapidity interval between the tagging jets as a function of M_{jj} and $\Delta\eta_{jj}$. The $\sqrt{s} = 8$ TeV data is compared to the DY Zjj and EW Zjj simulations. The bottom panel shows the ratio of data to the DY Zjj expectation.

momenta. The data, in excellent agreement with the simulation, indicates the presence of the EW Zjj signal, which has the expected feature of suppressed central jet multiplicity and H_T compared to the DY background. The transverse momentum $p_T^{j_3}$ and Zeppenfeld variable $\eta_{j_3}^*$ of the leading additional jet in the pseudorapidity interval are shown in Figure 70. The leading jet seems to be slightly more central in data compared to simulation, though the deviations are still acceptable given the high statistical and systematic uncertainties. The $p_T^{j_3}$ and H_T variables could also be used to compute the efficiency of a central jet veto, as shown in Figure 71. The gap fraction is defined as the fraction of events for which the tested variable does not exceed a given threshold. Good agreement is observed between the data and the signal plus background expectation, for both the data-driven and simulation-based background models for DY Zjj.



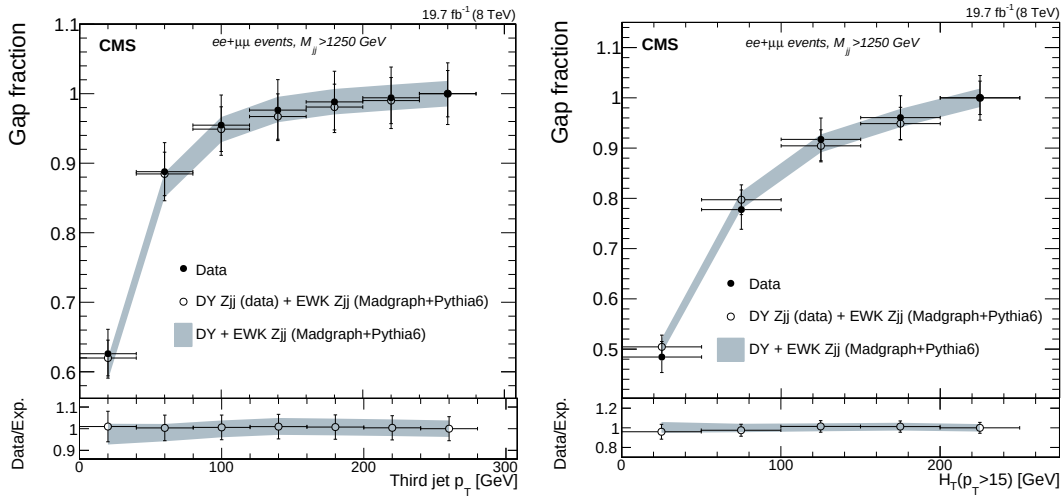
(a) central jet multiplicity (b) H_T of central jets

Figure 69: Central jet ($p_{\mathrm{T}} > 15$ GeV) multiplicity and their corresponding H_T within the pseudorapidity interval between the tagging jets, in events with $M_{jj} > 1250$ GeV. The expected contributions of signal and backgrounds are shown stacked in the main panel and compared to the observed data. The inset shows the signal-only contribution which is compared to the residual data after subtraction of the backgrounds. The bottom panels show the ratio of data to MC, including the systematic uncertainties represented by the shaded bands.



(a) $p_T^{j_3}$ of the third jet (b) $\eta^*_{j_3}$ of the third jet

Figure 70: The transverse momentum $p_T^{j_3}$ and $\eta^*_{j_3}$ of the leading additional jet within the pseudorapidity interval of the tagging jets, in events with $M_{jj} > 1250$ GeV.



(a) gap fraction using $p_T^{j_3}$

(b) gap fraction using H_T

Figure 71: Gap fraction for the $p_T^{j_3}$ and H_T variables, shown for data as well for two different signal plus background predictions where DY Zjj is modelled either from γjj data (white circles) or from simulation (shaded bands).

10

SUMMARY AND OUTLOOK

In this thesis we presented the measurement of the EW production of a Z boson in association with two jets in pp collisions, using data collected at $\sqrt{s} = 7$ and 8 TeV by the CMS experiment. The search for this signal exploits the typical VBF features: a dijet system in which the jets are produced in the forward and backward regions of the detector, while the VBF product, in our case the Z-boson, decays in the central region of the detector. Furthermore, we introduced a quark-gluon jet discrimination tool, which helps to separate signal events, dominated by quark induced jets, from backgrounds where jets can also originate from gluons. The performance of the quark-gluon jet discrimination tool has been documented in CMS-PAS-JME-13-002.

After a first indication of the signal by the 7 TeV analysis, the improved 8 TeV analysis fully established the presence of this signal and rejected the background-only hypothesis. The analyses were carried out in both the $Z \to ee$ and $Z \to \mu\mu$ channels, in which the leptons and jets were reconstructed using the CMS PF algorithm and the signal and background components were evaluated using MC simulations. The measured cross sections at $\sqrt{s} = 7$ and $\sqrt{s} = 8$ TeV are in agreement with those predicted by the SM within statistical and systematic uncertainties, and have been confirmed by two other analyses which have been performed in parallel with slightly different methodology. Additionally, the central hadronic activity in Zjj events and jet radiation patterns in Z plus at least one jet events have been studied in data and show good agreement with the simulation. The results of the cross section measurements and hadronic activity studies at $\sqrt{s} = 7$ and 8 TeV have been published in references [47] and [48] respectively.

There are many opportunities for a follow-up analysis, and more in general for VBF-type analyses, during the LHC Run II, when data is being collected at $\sqrt{s} = 13$ TeV. The higher centre-of-mass energy will cause the tagging jets to be produced at more forward pseudorapidities, resulting in a higher pseudorapidity separation between them and a higher dijet invariant mass for VBF events. Furthermore, the LHC will gradually increase its instantaneous luminosity, collecting much more statistics than it did during Run I. Hence, it will be a lot easier to collect events at high M_{jj}, where the EW Zjj contribution to the total event yield is getting stronger. This will reduce the statistical error on the EW Zjj cross section measurement, as well as providing a better populated high-purity signal region where VBF features like the low central hadronic activity can be studied with greater precision. As we have seen in Chapter 2, the chromo-electroweak interference between EW Zjj and

DY Zjj is expected to decrease in the high M_{jj} region. Using the higher luminosity in Run II, it might be possible to study this interference as a function of M_{jj}. If the statistical and interference uncertainty are reduced, the modelling of the DY Zjj would become the dominating uncertainty. Fortunately, NLO generators are becoming more widespread, and it will be possible to simulate NLO events at detector level. Using a NLO description for the background and signal simulations could result in a better description of some important VBF variables like y^*, $\Delta\eta_{jj}$ and the transverse momenta of the leading jets. This would help to built a more stable analysis with a smaller uncertainty on the DY Zjj description or signal acceptance.

The first step towards a Run II analysis is already taken with the update of the quark-gluon jet discrimination tool. The Run II quark-gluon likelihood is constructed using a narrower categorization in pseudorapidity in the HF region, improving its performance in this region where most of the VBF jets are found. Also a better performance for high p_T jets is expected due to the better handling of the correlations among the input variables. Furthermore, the tool is now fully implemented in the CMS software and highly optimized in speed such that it can be easily implemented by the many analyses which could benefit from quark-gluon jet discrimination.

ACKNOWLEDGEMENTS

This thesis brings an end to four interesting years in which I had the opportunity to meet and collaborate with many brilliant and nice people. Hereby I would like to express my gratitude to everybody who helped me to complete my PhD research.

First of all, I would like to thank Paolo Azzurri, which was involved in every aspect of my research and was always available to discuss results and new ideas. Without Paolo's guidance and support, this thesis would not have been written. Also special thanks to my promotor Pierre Van Mechelen, to give me the chance to start a PhD and for welcoming me into the friendly atmosphere of the University of Antwerp's particle physics group.

Many thanks to the people involved with the measurement of the EW Zjj cross section. A special mention to Alex Van Spilbeeck, which was not only an amazing office mate, but also helped to construct the 7 TeV analysis together with me and Paolo. Also thanks to Pedro Silva, the force behind the data-driven analysis, and Vladimir Gavrilov, Olga Kodolova, Alexander Nikitenko from the JPT analysis for the many interesting discussions, analysis ideas and cross checks between our results. The development of the quark-gluon discrimination tool was a challenging project, and was only possible by a collaborative effort including Andrea Marini, Francesco Pandolfi, Paolo Azzurri, Nick Van Remortel and Sunil Bansal. We had many interesting discussions about the physics and techniques behind this tool, and I am sure every member of this group gained a lot of new insight about the subject along the journey.

In addition to people already mentioned above, I would like to thank the other members of the particle physics group at the University of Antwerp for providing a very nice work environment: Benoît Roland, Hans Van Havermaet, Merijn Van de Klundert, Paolo Gunnellini, Albert Knutsson, Sarah Van Mierlo, Sara Alderweireldt, Sten Luyckx, Xavier Janssens, Frederik Van der Veken, and all others which were part of the group during these years. Also thanks to the many people in the wider CMS collaboration which helped me one way or another.

Finally, thanks to my family and friends for the support they gave me all these years. A special mention goes to Kelly, Lieselotte, Gregory, Mathias, Sien, Pieter, Karen en Nadja, with whom I studied together during my bachelor and master's years in Ghent.

SAMENVATTING

Het Standaard Model (SM) van de deeltjesfysica is een theorie die alle gekende fundamentele deeltjes en hun interacties, met uitzondering van de zwaartekracht, beschrijft. De voorspellingen van deze succesvolle theorie zijn in de laatste decennia bevestigd door verschillende deeltjesfysica experimenten. Toch zijn er ook enkele zaken die het SM niet kan verklaren, en vermoeden we dat het SM niet de ultieme theorie is. Dankzij de LHC, waarin men proton op elkaar laat botsen op een nooit eerder bereikt energieniveau, hebben we de mogelijkheid om het SM in een nieuwe situatie te bestuderen. Wanneer in een botsing tussen twee protonen, in elk van de protonen een quark is die een W of Z-boson uitgestuurd, waarbij deze vector bosons vervolgens fuseren spreken we van een vector boson fusie (VBF) proces. Dit proces bestaat dus uit enkel electrozwakke interacties. De quarks die de W/Z bosonen hebben uitgestuurd worden lichtjes afgebogen van de protonbundel en vormen jets in de CMS detector waarmee we dit onderzoek uitvoeren. Een VBF proces wordt dan ook gekenmerkt door twee energetische jets die worden geproduceerd onder een groot rapiditeitsinterval en met een hoge invariante massa van het dijet systeem, terwijl het product van de fusie zich eerder centraal in de detector zal bevinden. Er zijn verschillende deeltjes die het product kunnen zijn van een VBF proces, waaronder een ander vector boson of een Higgs boson. Ook bestaan er uitbreidingen op het SM die nieuwe exotische deeltjes voorspellen als het resultaat van een VBF proces. In deze thesis leggen we echter de focus op de VBF productie van een Z-boson, een deeltje dat al uitgebreid bestuurd is geweest in eerdere deeltjesfysica experimenten. Dit laat ons toe om de strategieën voor VBF analyses uit te testen zodat deze later kunnen toegepast worden op minder gekende VBF processen.

Het is echter niet mogelijk om een puur VBF-Z signaal, te isoleren: de Zjj signatuur kan immers ook door andere puur electrozwakke (EW) processen gecreëerd worden en de probabiliteitsamplitudes van deze processen interfereren sterk met elkaar. Gelukkig hebben deze andere processen vergelijkbare eigenschappen en is het nog steeds mogelijk om de analyse uit te voeren, waarbij we een meting uitvoeren van de totale electrozwakke productie van een Z boson in associatie met twee jets. De moeilijkheid van deze analyse ligt in de overweldigende achtergrond veroorzaakt door het Drell-Yan (DY) proces dat ook tot een Zjj signatuur kan leiden via een combinatie van electrozwakke en sterke interacties. Om een optimale separatie tussen signaal en achtergrond te bereiken hebben we gebruik gemaakt van een BDT methode om de verschillende discriminerende variabelen optimaal te combineren. De vorm van de BDT distributies voor data, signaal en achtergrond werd vervolgens gebruikt om het aandeel van signaal en achtergrond in de data te meten en te vergelijken met de SM voorspelling.

Aangezien de twee jets in het EW Zjj proces afkomstig zijn van quarks, terwijl de jets in het DY Zjj proces zowel door quarks als gluonen kunnen geïnitieerd worden, hebben we ook een quark-gluon jet discriminatie methode ontwikkeld. Hierbij werden enkele jet substructuur variabelen bestudeerd in MC simulaties en werden hun verschillende distributies voor quark en gluon jets gebruikt om een quark-jet probabiliteit aan de jet toe te kennen. Deze methode werd geverifieerd in data, en met de tweede data run van de LHC in het vooruitzicht werd de methode terug geoptimaliseerd om in de toekomst een nog betere performantie en precisie te bekomen.

In het kader van de EW Zjj analyse werden ook nog enkele metingen van de hadronische activiteit uitgevoerd. Aangezien de twee jets in een EW Zjj of VBF proces in de voor- en achterwaartse regios van de detector worden gevormd, zal de hadronische activiteit, veroorzaakt door de sterke kracht, zich tot deze regios beperken. In het centrale gedeelte van de detector verwachten we dus een lagere hadronische activiteit ten opzichte van de achtergrond processen. Hoewel de analyse deze aanname lijkt te bevestigen, is er echter nog meer data nodig om dit ten gronde te kunnen bestuderen.

BIBLIOGRAPHY

[1] M. Peskin and D. Schroeder, *An Introduction to Quantum Field Theory*. Westview Press, 1995.

[2] C. N. Yang and R. L. Mills, "Conservation of isotopic spin and isotopic gauge invariance," *Phys. Rev.*, vol. 96, pp. 191–195, Oct 1954.

[3] S. Glashow, "Partial symmetries of weak interactions," *Nucl. Phys.*, vol. 22, pp. 579–588, 1961.

[4] A. Salam and J. Ward, "Gauge theory of elementary interactions," *Phys. Rev.*, vol. 136, pp. B763–B768, Nov 1964.

[5] S. Weinberg, "A model of leptons," *Phys. Rev. Lett.*, vol. 19, pp. 1264–1266, Nov 1967.

[6] M. Gell-Mann, "The interpretation of the new particles as displaced charge multiplets," *Nuovo Cim.*, vol. 4, no. S2, pp. 848–866, 1956.

[7] K. Nishijima, "Charge Independence Theory of V Particles," *Progress of Theoretical Physics*, vol. 13, pp. 285–304, Mar. 1955.

[8] T. Nakano and K. Nishijima, "Charge Independence for V-particles," *Progress of Theoretical Physics*, vol. 10, pp. 581–582, Nov. 1953.

[9] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons," *Phys. Rev. Lett.*, vol. 13, pp. 321–323, Aug 1964.

[10] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," *Phys. Rev. Lett.*, vol. 13, pp. 508–509, Oct 1964.

[11] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global conservation laws and massless particles," *Phys. Rev. Lett.*, vol. 13, pp. 585–587, Nov 1964.

[12] F. Halzen and A. Martin, *Quarks and leptons: an Introductory Course in Modern Particle Physics*. Wiley, 1984.

[13] S. Chatrchyan *et al.*, "Measurement of the ratio of the inclusive 3-jet cross section to the inclusive 2-jet cross section in pp collisions at $\sqrt{s}$ = 7 TeV and first determination of the strong coupling constant in the TeV range," *Eur.Phys.J.*, vol. C73, no. 10, p. 2604, 2013.

[14] K. Olive *et al.*, "Review of Particle Physics," *Chin.Phys.*, vol. C38, p. 090001, 2014.

[15] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, *et al.*, "Event generation with SHERPA 1.1," *JHEP*, vol. 0902, p. 007, 2009.

[16] A. Buckley, J. Butterworth, S. Gieseke, D. Grellscheid, S. Hoche, *et al.*, "General-purpose event generators for LHC physics," *Phys.Rept.*, vol. 504, pp. 145–233, 2011.

[17] G. Altarelli and G. Parisi, "Asymptotic Freedom in Parton Language," *Nucl.Phys.*, vol. B126, p. 298, 1977.

[18] V. Gribov and L. Lipatov, "Deep inelastic e p scattering in perturbation theory," *Sov.J.Nucl.Phys.*, vol. 15, pp. 438–450, 1972.

[19] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.," *Sov.Phys.JETP*, vol. 46, pp. 641–653, 1977.

[20] V. Sudakov, "Vertex parts at very high-energies in quantum electrodynamics," *Sov.Phys.JETP*, vol. 3, pp. 65–71, 1956.

[21] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual," *JHEP*, vol. 0605, p. 026, 2006.

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

[23] M. Bahr, S. Gieseke, M. Gigg, D. Grellscheid, K. Hamilton, *et al.*, "Herwig++ Physics and Manual," *Eur.Phys.J.*, vol. C58, pp. 639–707, 2008.

[24] S. Gieseke, D. Grellscheid, K. Hamilton, A. Papaefstathiou, S. Platzer, *et al.*, "Herwig++ 2.5 Release Note," 2011.

[25] CMS Collaboration, "Measurement of the Underlying Event Activity at the LHC with $\sqrt{s} = 7$ TeV and Comparison with $\sqrt{s} = 0.9$ TeV," *JHEP*, vol. 1109, p. 109, 2011.

[26] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, "MadGraph 5: Going Beyond," *JHEP*, vol. 1106, p. 128, 2011.

[27] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, "Matching matrix elements and shower evolution for top-quark production in hadronic collisions," *JHEP*, vol. 0701, p. 013, 2007.

[28] J. Alwall, S. Hoche, F. Krauss, N. Lavesson, L. Lonnblad, *et al.*, "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions," *Eur.Phys.J.*, vol. C53, pp. 473–500, 2008.

[29] C. Oleari, "The POWHEG-BOX," *Nucl.Phys.Proc.Suppl.*, vol. 205-206, pp. 36–41, 2010.

[30] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX," *JHEP*, vol. 06, p. 043, 2010.

[31] P. Nason, "A New method for combining NLO QCD with shower Monte Carlo algorithms," *JHEP*, vol. 0411, p. 040, 2004.

[32] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method," *JHEP*, vol. 0711, p. 070, 2007.

[33] J. M. Campbell and R. Ellis, "MCFM for the Tevatron and the LHC," *Nucl.Phys.Proc.Suppl.*, vol. 205-206, pp. 10–15, 2010.

[34] K. Arnold, M. Bahr, G. Bozzi, F. Campanario, C. Englert, *et al.*, "VBFNLO: A Parton level Monte Carlo for processes with electroweak bosons," *Comput.Phys.Commun.*, vol. 180, pp. 1661–1670, 2009.

[35] K. Arnold, J. Bellm, G. Bozzi, M. Brieg, F. Campanario, *et al.*, "VBFNLO: A Parton Level Monte Carlo for Processes with Electroweak Bosons – Manual for Version 2.5.0," 2011.

[36] K. Arnold, J. Bellm, G. Bozzi, F. Campanario, C. Englert, *et al.*, "Release Note – Vbfnlo-2.6.0," 2012.

[37] J. Bjorken, "Rapidity gaps and jets as a new physics signature in very high-energy hadron hadron collisions," *Phys.Rev.*, vol. D47, pp. 101–113, 1993.

[38] D. L. Rainwater, R. Szalapski, and D. Zeppenfeld, "Probing color singlet exchange in Z + two jet events at the CERN LHC," *Phys.Rev.*, vol. D54, pp. 6680–6689, 1996.

[39] H. Chehime and D. Zeppenfeld, "Single w- and z-boson production as a probe for rapidity gaps at the superconducting super collider," *Phys. Rev. D*, vol. 47, pp. 3898–3905, May 1993.

[40] F. Schissler and D. Zeppenfeld, "Parton Shower Effects on W and Z Production via Vector Boson Fusion at NLO QCD," *JHEP*, vol. 1304, p. 057, 2013.

[41] U. Baur and D. Zeppenfeld, "Measuring three vector boson couplings in qq $\to$ qqW at the SSC," 1993.

[42] V. Khoze, M. Ryskin, W. Stirling, and P. Williams, "A Z monitor to calibrate Higgs production via vector boson fusion with rapidity gaps at the LHC," *Eur.Phys.J.*, vol. C26, pp. 429–440, 2003.

[43] CMS Collaboration, "Search for the Standard Model Higgs Boson produced through Vector Boson Fusion and decaying to $b\bar{b}$," 2015.

[44] S. Alderweireldt, *Search for the Standard Model Higgs Boson produced through Vector Boson Fusion and decaying to $b\bar{b}$*. PhD thesis, University of Antwerp, 2015.

[45] G.-C. Cho, K. Hagiwara, J. Kanzaki, T. Plehn, D. Rainwater, *et al.*, "Weak boson fusion production of supersymmetric particles at the CERN LHC," *Phys.Rev.*, vol. D73, p. 054002, 2006.

[46] B. Dutta, A. Gurrola, W. Johns, T. Kamon, P. Sheldon, *et al.*, "Vector Boson Fusion Processes as a Probe of Supersymmetric Electroweak Sectors at the LHC," *Phys.Rev.*, vol. D87, no. 3, p. 035029, 2013.

[47] CMS Collaboration, "Measurement of the hadronic activity in events with a Z and two jets and extraction of the cross section for the electroweak production of a Z with two jets in pp collisions at $\sqrt{s} = 7$ TeV," *JHEP*, vol. 1310, p. 062, 2013.

[48] CMS Collaboration, "Measurement of electroweak production of two jets in association with a Z boson in proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$," *Eur.Phys.J.*, vol. C75, no. 2, p. 66, 2015.

[49] ATLAS Collaboration, "Measurement of the electroweak production of dijets in association with a Z-boson and distributions sensitive to vector boson fusion in proton-proton collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector," *JHEP*, vol. 1404, p. 031, 2014.

[50] CMS Collaboration, "Measurement of the cross section of the electroweak production of a W boson with two jets in pp collisions at sqrt(s) = 8TeV," 2015.

[51] CMS Collaboration, "Study of vector boson scattering and search for new physics in events with two same-sign leptons and two jets," *Phys.Rev.Lett.*, vol. 114, no. 5, p. 051801, 2015.

[52] ATLAS Collaboration, "Evidence for Electroweak Production of $W^{\pm}W^{\pm}jj$ in pp Collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector," *Phys.Rev.Lett.*, vol. 113, no. 14, p. 141803, 2014.

[53] J. Pumplin, D. Stump, J. Huston, H. Lai, P. M. Nadolsky, *et al.*, "New generation of parton distributions with uncertainties from global QCD analysis," *JHEP*, vol. 0207, p. 012, 2002.

[54] C. Oleari and D. Zeppenfeld, "QCD corrections to electroweak $l\nu_l jj$ and l^+l^-jj production," *Phys.Rev.*, vol. D69, p. 093004, 2004.

[55] H.-L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, *et al.*, "New parton distributions for collider physics," *Phys.Rev.*, vol. D82, p. 074024, 2010.

[56] K. Melnikov and F. Petriello, "Electroweak gauge boson production at hadron colliders through $\mathcal{O}(\alpha_s^2)$," *Phys.Rev.*, vol. D74, p. 114017, 2006.

[57] R. J. Gonsalves and C. Wai, "Chromoelectroweak interference and parity violating asymmetries in the production of an electroweak boson + two jets in hadron collisions," *Phys.Rev.*, vol. D49, p. 190, 1994.

[58] M. Czakon, P. Fiedler, and A. Mitov, "Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $\mathcal{O}(\alpha_S^4)$," *Phys.Rev.Lett.*, vol. 110, p. 252004, 2013.

[59] L. Evans and P. Bryant, "LHC Machine," *JINST*, vol. 3, p. S08001, 2008.

[60] F. Marcastel, "CERN's Accelerator Complex." Oct 2013.

[61] S. Dailler, "Cross section of LHC dipole." Apr 1999.

[62] J.-L. Caron, "LHC Layout." Sep 1997.

[63] ALICE Collaboration, "The ALICE experiment at the CERN LHC," *JINST*, vol. 3, p. S08002, 2008.

[64] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *JINST*, vol. 3, p. S08003, 2008.

[65] CMS Collaboration, "The CMS experiment at the CERN LHC," *JINST*, vol. 3, p. S08004, 2008.

[66] LHCb Collaboration, "The LHCb Detector at the LHC," *JINST*, vol. 3, p. S08005, 2008.

[67] LHCf Collaboration, "The LHCf detector at the CERN Large Hadron Collider," *JINST*, vol. 3, p. S08006, 2008.

[68] MoEDAL Collaboration, "The Physics Programme Of The MoEDAL Experiment At The LHC," *Int.J.Mod.Phys.*, vol. A29, p. 1430050, 2014.

[69] TOTEM Collaboration, "The TOTEM experiment at the CERN Large Hadron Collider," *JINST*, vol. 3, p. S08007, 2008.

[70] CMS Collaboration, "Alignment of the CMS Silicon Tracker during Commissioning with Cosmic Rays," *JINST*, vol. 5, p. T03009, 2010.

[71] CMS Collaboration, "Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in pp Collisions at $\sqrt{s} = 7$ TeV," *JINST*, vol. 8, p. P09009, 2013.

[72] CMS Collaboration, "Performance of the CMS Drift Tube Chambers with Cosmic Rays," *JINST*, vol. 5, p. T03015, 2010.

[73] GEANT4 Collaboration, "GEANT4: A Simulation toolkit," *Nucl.Instrum.Meth.*, vol. A506, pp. 250–303, 2003.

[74] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET," CMS-PAS-PFT-09-001.

[75] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker," *JINST*, vol. 9, no. 10, p. P10009, 2014.

[76] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s}$ = 7 TeV," *Journal of Instrumentation*, vol. 7, p. 2P, Oct. 2012.

[77] CMS Collaboration, "Electron reconstruction and identification at $\sqrt{s}$ = 7 TeV," CMS-PAS-EGM-10-004.

[78] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV," *JINST*, vol. 10, no. 06, p. P06005, 2015.

[79] M. Pioppi, "A pre-identification for electron reconstruction in the cms particle-flow algorithm," *Journal of Physics: Conference Series*, vol. 119, no. 3, p. 032039, 2008.

[80] G. P. Salam, "Towards Jetography," *Eur.Phys.J.*, vol. C67, pp. 637–686, 2010.

[81] M. Cacciari, G. P. Salam, and G. Soyez, "The Anti-k_t jet clustering algorithm," *JHEP*, vol. 0804, p. 063, 2008.

[82] CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in CMS," *Journal of Instrumentation*, vol. 6, p. 11002, Nov. 2011.

[83] CMS Collaboration, "The Jet Plus Tracks Algorithm for Calorimeter Jet Energy Corrections in CMS," CMS-PAS-JME-09-002.

[84] CMS Collaboration, "Commissioning of Track-jets in pp Collisions at 7 TeV," CMS-PAS-JME-10-006.

[85] CMS Collaboration, "Jet Energy Scale and Resolution in the CMS Experiment," CMS-PAS-JME-13-004.

[86] M. Cacciari, G. P. Salam, and G. Soyez, "The Catchment Area of Jets," *JHEP*, vol. 0804, p. 005, 2008.

[87] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas," *Phys.Lett.*, vol. B659, pp. 119–126, 2008.

[88] S. Niki and T. Eirini, "Performance of the Particle-Flow jet identification criteria using proton-proton collisions at $\sqrt{s} = 8$ TeV," CMS-AN-2014/227.

[89] OPAL Collaboration, "A study of differences between quark and gluon jets using vertex tagging of quark jets," *Z.Physik C Particles and Fields*, vol. 58, no. 3, pp. 387–403, 1993.

[90] OPAL Collaboration, "A Model independent measurement of quark and gluon jet properties and differences," *Z.Phys.*, vol. C68, pp. 179–202, 1995.

[91] DELPHI Collaboration, "Energy dependence of the differences between the quark and gluon jet fragmentation," *Z.Phys.*, vol. C70, pp. 179–196, 1996.

[92] ALEPH Collaboration, "Quark and gluon jet properties in symmetric three jet events," *Physics Letters B*, vol. 384, no. 1-4, pp. 353–364, 1996.

[93] ALEPH Collaboration, "Studies of quantum chromodynamics with the ALEPH detector," *Phys. Rep.*, vol. 294, pp. 1–165. 166 p, Dec 1996.

[94] CDF Collaboration, "Fragmentation differences of quark and gluon jets at Tevatron," *Int.J.Mod.Phys.*, vol. A20, pp. 3723–3725, 2005.

[95] DØ Collaboration, "Subjet multiplicity in quark and gluon jets at DØ," *Nucl.Phys.Proc.Suppl.*, vol. 79, pp. 494–496, 1999.

[96] A. C. Marini, "Quark and Gluon Jet Separation and QCD studies," master thesis, 2011.

[97] CMS Collaboration, "Performance of quark/gluon discrimination in 8 TeV pp data," CMS-PAS-JME-13-002.

[98] J. Gaffney and A. Mueller, "$\alpha(Q^2)$ corrections to particle multiplicity ratios in gluon and quark jets," *Nuclear Physics B*, vol. 250, no. 1-4, pp. 109 – 142, 1985.

[99] J. Gallicchio and M. D. Schwartz, "Seeing in Color: Jet Superstructure," *Phys.Rev.Lett.*, vol. 105, p. 022001, 2010.

[100] J. Davighi, "Branching structure of QCD jets: new jet observables for quark-gluon discrimination," CERN-STUDENTS-Note-2014-215.

[101] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," *PoS*, vol. ACAT, p. 040, 2007.

[102] B. Bhattacherjee, S. Mukhopadhyay, M. M. Nojiri, Y. Sakaki, and B. R. Webber, "Associated jet and subjet rates in light-quark and gluon jet discrimination," *JHEP*, vol. 1504, p. 131, 2015.

[103] CMS Collaboration, "Identification of b-quark jets with the CMS experiment," *JINST*, vol. 8, p. P04013, 2013.

[104] CMS Collaboration, "Pileup Jet Identification," CMS-PAS-JME-13-005.

[105] CMS Collaboration, "Identification techniques for highly boosted W bosons that decay into hadrons," *JHEP*, vol. 1412, p. 017, 2014.

[106] L. Tuura, A. Meyer, I. Segoni, and G. Della Ricca, "CMS data quality monitoring: Systems and experiences," *Journal of Physics: Conference Series*, vol. 219, no. 7, p. 072020, 2010.

[107] CMS Collaboration, "Absolute Calibration of the Luminosity Measurement at CMS: Winter 2012 Update," CMS-PAS-SMP-12-008.

[108] CMS Collaboration, "Inelastic pp cross section at 7 TeV," CMS-PAS-FWD-11-001.

[109] N. Adam, J. Berryhill, V. Halyo, A. Hunt, and K. Mishra, "Generic Tag and Probe Tool for Measuring Efficiency at CMS with Early Data," CMS-AN-2009/111.

[110] V. Candelise, M. Casarsa, F. Cossutti, G. Della Ricca, B. Gobbo, C. La Licata, M. Marone, D. Montanino, D. Scaini, A. Schizzi, and T. Umer, "Study of the associated production of a Z boson and jets in pp collisions at $\sqrt{s} = 7$ TeV," CMS-AN-2012/376.

[111] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, Aug. 1997.

[112] R. Brun and F. Rademakers, "Root - an object oriented data analysis framework," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 389, no. 1 - 2, pp. 81 – 86, 1997.

[113] R. Barlow and C. Beeston, "Fitting using finite monte carlo samples," *Computer Physics Communications*, vol. 77, no. 2, pp. 219 – 228, 1993.

[114] L. Kaur Saini, I. Kravchenko, and Y. Maravin, "A study of efficiencies and scale factors for cut-based electron identification at CMS experiment using data from proton-proton collisions at $\sqrt{s} = 8$ TeV," CMS-AN-2014/055.

[115] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys. Lett.*, vol. B716, pp. 30–61, 2012.

[116] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *European Physical Journal C*, vol. 71, p. 1554, Feb. 2011.

[117] N. Kidonakis, "Differential and total cross sections for top pair and single top production," *Proceedings, 20th International Workshop on Deep-Inelastic Scattering and Related Subjects (DIS 2012)*, pp. 831–834, 2012.

[118] CMS Collaboration, "CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update," CMS-PAS-LUM-13-001.

[119] T. Binoth *et al.*, "The SM and NLO Multileg Working Group: Summary report," *Physics at TeV colliders. Proceedings, 6th Workshop, Les Houches, France, June 8-26, 2009*, pp. 21–189, 2010.