SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of Fisica G. Occhialini

Dottorato di Ricerca in / PhD program: Fisica ed Astronomia
Ciclo / Cycle: XXXVI
Curriculum in: Fisica Subnucleare

Observation of the electroweak production of W^+W^- bosons with two jets with the CMS detector at the LHC and searches for new physics with dimension-6 Effective Field Theory

Cognome / Surname: **Pinolini**

Nome / Name: **Bianca Sofia**

Matricola / Registration number: **800348**

Tutore / Tutor: Prof. Marco Paganoni

Supervisor: Prof. Pietro Govoni

Cotutore / Co-tutor: Dr. Raffaele Gerosa

Coordinatore / Coordinator: Prof. Stefano Ragazzi

ANNO ACCADEMICO / ACADEMIC YEAR 2022/2023

Contents

Introduction

The Large Hadron Collider (LHC) is a circular accelerator capable of producing proton-proton collisions at a design centre-of-mass energy of $\sqrt{s} = 14$ TeV. These collisions occur at four distinct points along the accelerator ring. Two general-purpose detectors, namely ATLAS and CMS, are strategically positioned at two of these interaction points, located at opposite ends of the accelerator ring. This strategic placement is essential because it enables the machine to precisely adjust beam parameters to maximise collision intensity, thereby contributing to the highest operational efficiency in experimental operations.

The LHC operates in a cyclical manner, with alternating periods of data-taking (Run) and long shutdowns (LS) for detector upgrades and maintenance, as shown in Figure 1.1.

The LHC's initial data-taking phase, known as Run 1 (2010-2012), consisted in proton-proton collisions at $\sqrt{s} =$ 7-8 TeV, accumulating an integrated luminosity of approximately 30 fb^{-1}. During this phase, data collected by the ATLAS and CMS detectors led to the landmark discovery of the Higgs boson in 2012. After a long shutdown (LS1, from 2013 to 2015), proton collisions resumed at a higher $\sqrt{s}$ of 13 TeV in 2015, marking the commencement of another three-year data collection period known as Run 2. The LHC Run 2 yielded an integrated luminosity of around 138 fb^{-1}. Figure 1.2 displays the integrated luminosity delivered by the LHC during Run 1 and Run 2 data-taking years. The slope or derivative of the curve is proportional to the collision intensity, which was higher in 2018 compared to previous years. Comprehensive descriptions of the CERN accelerator complex and the CMS detector can be found in Chapter 2.

This thesis focuses mainly on the search for the electroweak production of two opposite sign W-bosons in association with two jets at $\sqrt{s} = 13$ TeV with the CMS detector. The analysed dataset, recorded during the Run 2 of the LHC, corresponds to an integrated luminosity of 138 fb^{-1}.

In this first chapter, a concise overview of the Standard Model (SM) of particle physics, with a specific focus on the electroweak sector, is provided in Section 1.1. Section 1.2 delves into Vector Boson Scattering (VBS) processes, highlighting their significance as they provide crucial insights into the electroweak sector of the SM and its role in elucidating the mechanism of spontaneous symmetry breaking. VBS processes underscore the importance of the Higgs boson within the SM as its presence ensures the theory's unitarity. However,

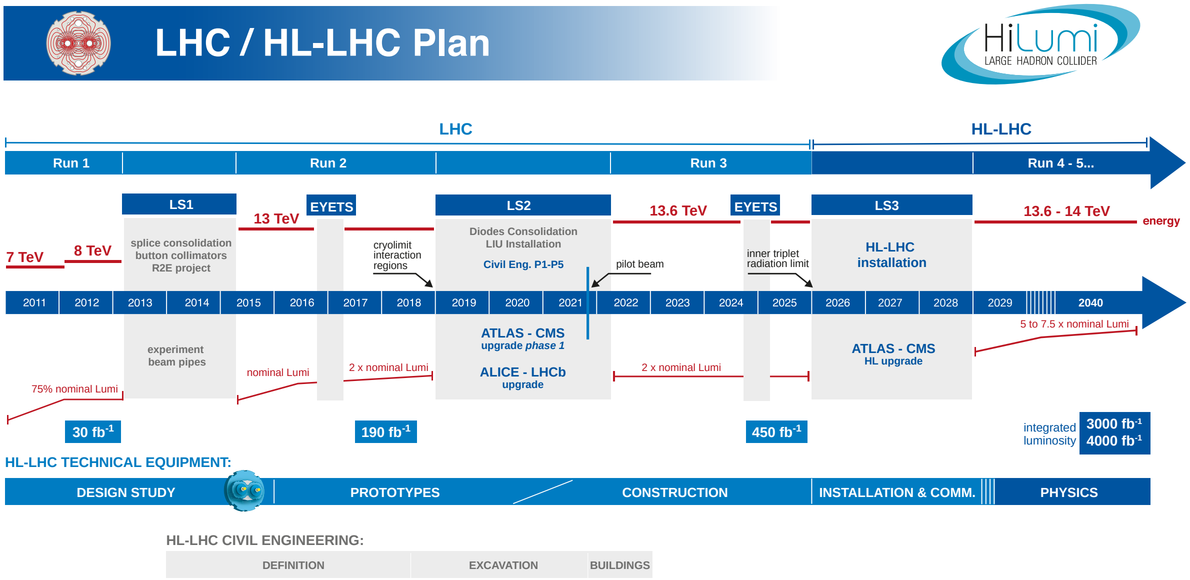


Figure 1.1: LHC baseline plan for long shutdowns (LS) and Run periods. The upper line displays the collision energy, while the lower line presents the luminosity. Due to COVID-19 restrictions, the start of Run 3 was rescheduled to February 2022. From Ref. [1].

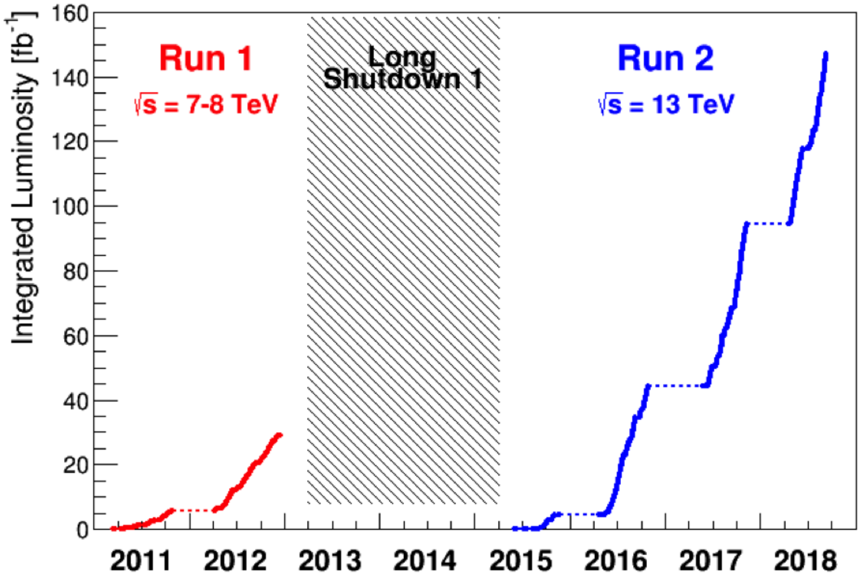


Figure 1.2: Integrated luminosities of Run 1 and Run 2. The Run 1 total is approximately 30 fb^{-1}, while the Run 2 total is around 138 fb^{-1}.

any deviations from the SM expectations could disrupt this delicate balance. Consequently, the study of VBS cross-sections can provide insights into new physics phenomena. Finally, in Section 1.3, the Standard Model Effective Field Theory (SMEFT) is explored as an expansion of the SM, serving as a framework for indirect searches of new physics at the LHC.

1.1 Standard Model of Particle Physics

The primary objective of particle physics is the investigation of the fundamental constituents of the universe (elementary particles) and their interactions (fundamental forces). At the core of this scientific pursuit lies the SM of particle physics, constructed within the framework of relativistic Quantum Field Theory (QFT). The SM comprehensively describes three of the four fundamental interactions in the universe: the strong force, the weak force, and the electromagnetic one. The gravitational force, despite being one of the fundamental forces of nature, cannot be easily described within the framework of QFT. Nevertheless, the SM retains its validity because on subatomic scales the gravitational force can be neglected due to its significantly weaker influence.
This section presents an overview of the foundational features of the SM, highlighting its achievements and outlining its inherent limitations. For a more detailed description of the SM, see Refs. [2, 3].

1.1.1 Fermions and Bosons

According to the SM, the fundamental constituents of matter are particles with spin $\frac{1}{2}$, referred to as fermions. In contrast, the mediators of the three fundamental forces are particles characterised by integer spin values, denoted as bosons.

Fermions

The fermionic sector comprises twelve distinct particles, divided into two groups, quarks and leptons, based on their sensitivity to the strong interaction. Quarks represent the sole fermionic entities that interact via the strong nuclear force, characterised by the exchange of gluons, which is why they are said to carry a colour charge. In contrast, leptons are colour-neutral particles, making them insensitive to the strong force. Moreover, fermions exhibit left- and right-handed chirality with respect to the weak force, a topic that will be explored in detail towards the conclusion of this section.

Within each group, leptons and quarks are further organised into three generations (or families), as detailed in Table 1.1. The particles constituting the first generation represent the familiar constituents of ordinary matter. The counterparts within the second and third generations might be considered as replicas of the first generation, sharing identical quantum numbers but differing in their masses. Antiparticles exhibit identical mass and spin characteristics but opposite physical charges (like the electric or the colour charge).

		Particle	Mass [GeV]	Q
Leptons	1st generation	electron e	$0.5 \cdot 10^{-3}$	-1
		neutrino ν_e	0	0
	2nd generation	muon μ	$106 \cdot 10^{-3}$	-1
		neutrino ν_μ	0	0
	3rd generation	tau τ	1.8	-1
		neutrino ν_τ	0	0
Quarks	1st generation	up u	$2 \cdot 10^{-3}$	$\frac{2}{3}$
		down d	$4.7 \cdot 10^{-3}$	$-\frac{1}{3}$
	2nd generation	charm c	1.3	$\frac{2}{3}$
		strange s	$93 \cdot 10^{-3}$	$-\frac{1}{3}$
	3rd generation	top t	172.8	$\frac{2}{3}$
		beauty b	4.2	$-\frac{1}{3}$

Table 1.1: The twelve fermionic constituents of the SM: six leptons and six quarks. The table includes their masses and electric charges (Q). Leptons are organised into three generations (or families), with each generation corresponding to a different flavour, such as the electron, muon, and tau, along with their respective neutrinos. In the quark sector, each generation exhibits a variety of flavours, including up, down, charm, strange, top, and bottom quarks. Source: Ref. [5].

Neutrinos, the subset made of those leptons which are neutral, interact exclusively through the weak force due to their lack of electric charge. While the SM assumes exact zero mass for neutrinos, experimental observations have provided upper limits (UL) on their masses, as cited in the Particle Data Group (PDG [4]).

In addition to their classification into quarks and leptons, fermions are also distinguished by their handedness, which refers to whether they are left-handed or right-handed. Left-handed fermions interact with the weak nuclear force, while right-handed fermions do not. This is because the weak force is mediated by W and Z bosons, which are associated with left-handed fermions. The right-handed fermions, on the other hand, interact with the strong nuclear force, which is mediated by gluons. Another important aspect of this distinction is that left-handed fermions interact more strongly with the Higgs field compared to right-handed fermions, which explains why left-handed fermions have a lower mass than right-handed fermions. This difference in mass contributes to the difference in lifetimes between left-handed and right-handed particles.

Particle	Spin	Mass [GeV]	Q
gluon g	1	0	0
photon γ	1	0	0
$W^{\pm}$	1	80.4	±1
Z	1	91.2	0
H	0	125.3	0

Table 1.2: The bosonic constituents of the SM: five vector bosons and one scalar boson (H). The table includes details about their spins, masses, and electric charges (Q). Gluons are the only vector bosons that carry a colour charge (RGB).

Bosons

Table 1.2 includes the vector bosons responsible for carrying the fundamental forces, and it features also an indispensable scalar boson, the Higgs boson, which arises from the mechanism of spontaneous symmetry breaking of the electroweak (EW) interaction. Details on this last mechanism will be provided in Section 1.1.3.

The vector bosons are pivotal in mediating the fundamental forces within the SM. Gluons are associated with the strong nuclear force, while photons mediate the electromagnetic force, governing interactions between electrically charged particles. $W^{\pm}$ bosons and the Z boson mediate the weak nuclear force. In particular, $W^{\pm}$ bosons are involved in numerous radioactive decay processes, while the Z boson is associated with neutral current interactions in neutrino scattering experiments. The Higgs boson is unique in the SM because it imparts mass to other particles through the Higgs field, without directly mediating a fundamental force. The strength of the interaction between a particle and the Higgs field is directly proportional to the mass of that particle.

1.1.2 Local Gauge Invariance Principle

The SM relies is a Yang-Mills QFT based on the symmetry group $SU(3) \times SU(2)_L \times U(1)$:

- $SU(3)$ corresponds to the strong force, as described by quantum chromodynamics (QCD). This component of the symmetry group deals with the behaviour of quarks and gluons.
- $SU(2)_L$ is associated with weak isospin[1], and $U(1)$ is associated with hypercharge. Following the symmetry breaking via the Higgs mechanism (Section 1.1.3), these groups give rise to the weak bosons and to the photon.

---

[1] The subscript L indicates that this symmetry transformation only applies to left-handed particles.

The cornerstone of the SM theoretical integrity lies in the principle of local gauge invariance. This principle dictates that the laws of physics remain unchanged (invariant) under local transformations, or "gauge" transformations. Within the framework of the SM, local gauge invariance gives rise to the existence of gauge bosons, including the photon, weak bosons, and gluons.

In summary, the SM is built on the foundation of symmetry groups and local gauge invariance, which together provide a powerful and elegant framework for describing the fundamental particles and their interactions.

Quantum Chromodynamics (QCD)

Quantum chromodynamics originates from the necessity of imposing Lagrangian invariance under local $SU(3)$ phase transformations. This requirement ensures that the Lagrangian governing free matter fields (Dirac terms) remains invariant under $SU(3)$ transformations, forming the basis for QCD. This introduces the concept of colour charge (red, green, blue), the charge characterising QCD interactions, in the same way that electric charge is associated with electromagnetic interactions. The invariance under $SU(3)$ transformations, in accordance with Noether's theorem (a fundamental principle in physics that connects symmetries to conserved quantities), leads to the conservation of colour charge. To maintain this invariance, the Lagrangian incorporates eight massless gauge bosons known as gluons, one for each generator of the $SU(3)$ group. These concepts together provide a comprehensive understanding of how quarks and gluons interact within the framework of the SM.

The QCD Lagrangian can be expressed as:

$$\mathcal{L}_{\mathrm{QCD}} = \sum_q \bar{\psi}_{q,a} \left(i\gamma^\mu \partial_\mu \delta_{ab} - \sqrt{4\pi\alpha_s}\gamma^\mu t^C_{ab}\mathcal{A}^C_\mu - m_q\delta_{ab} \right) \psi_{q,b} - \frac{1}{4} F^A_{\mu\nu} F^{A\mu\nu}, \tag{1.1}$$

with

$$F^A_{\mu\nu} = \partial_\mu \mathcal{A}^A_\nu - \partial_\nu \mathcal{A}^A_\mu - \sqrt{4\pi\alpha_s} f_{ABC}\mathcal{A}^B_\mu \mathcal{A}^C_\nu, \qquad [t^A, t^B] = i f_{ABC} t^C. \tag{1.2}$$

Within this equation, the components are as follows:

γ^μ : Dirac γ-matrices;

$\psi_{q,a}$: spinors for quarks with flavour q, mass m_q, and colour a;

$\mathcal{A}^C_\mu$, $C = 1, 2, \ldots, 8$: gluon fields.

t^C_{ab} : 3×3 matrix generators of the $SU(3)$ group; these describe how colour charges transform under the group's symmetry.

α_s : QCD coupling constant, which characterizes the strength of the strong force interactions;

$F^A_{\mu\nu}$: field tensor, which describes the strength and direction of the gluon fields. It is derived from the gluon field $\mathcal{A}^A_\mu$;

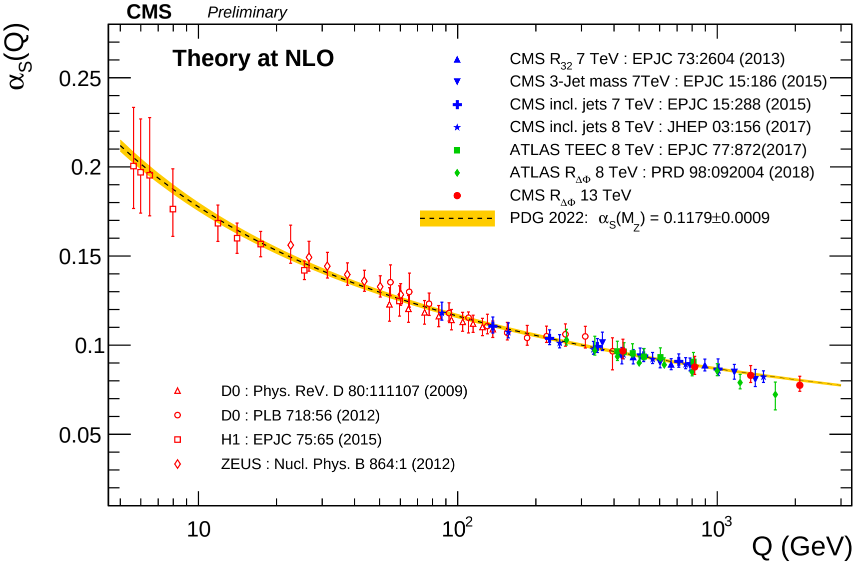


Figure 1.3: The QCD coupling constant (α_S) as a function of the transferred momentum (Q). From Ref. [6].

f_{ABC} : structure constants of the $SU(3)$ group. These provide the link between the commutator $[t^A, t^B]$ and the t^C generators, allowing to write any commutator of two t^A generators as a linear combination of the t^C generators, with the coefficients given by the structure constants f_{ABC}.

Experimentally, particles with colour charge such as gluons and quarks can't be observed isolated. Only colour-neutral particles (hadrons, i.e. bound states of quarks under the strong force) are observed. This phenomenon, termed "colour confinement", arises directly from the energy-dependent behaviour of the QCD coupling parameter α_s. The strong coupling constant α_s diminishes at high energies (corresponding to short-distance interactions) and increases at lower energies (associated with long-distance interactions), as illustrated in Figure 1.3. This is due to a phenomenon known as asymptotic freedom. According to Ref. [5], the current value of α_s at the mass scale of the Z boson is

$$\alpha_s(M_Z^2) = 0.1179 \pm 0.0009. \tag{1.3}$$

The behaviour of the strong coupling constant with energy implies that, when quarks are produced in high-energy collisions, they do not exist as free particles for long. In high-energy collisions, as the energy decreases, the tendency of quarks and gluons to separate also increases. This happens because the strong force, which holds these particles together, weakens with decreasing energy. However, this trend is countered by the fact that it becomes energetically favourable for them to disrupt the colour string binding them together and spontaneously create a quark-antiquark pair from the vacuum. This leads to the formation of bound states, primarily mesons ($q\bar{q}$ pairs) and baryons (qqq

combinations). This phenomenon is known as hadronisation, and it ultimately results in the creation of sprays of hadrons called jets.

Electroweak Sector

The unification of the electromagnetic and weak interactions into a single EW theory was a groundbreaking achievement by Glashow, Salam, and Weinberg [7, 8, 9]. This theory emerges from the requirement that the Lagrangian remains invariant under local gauge transformations of the form $SU(2)_L \times U(1)$.
In the EW sector of the SM, the Lagrangian includes terms that describe the behaviour of fermions across three generations:

$$\begin{aligned}\mathcal{L}_{\text{EW}} =& \overline{Q}_{Lj} i\gamma^\mu D_\mu Q_{Lj} + \overline{u}_{Rj} i\gamma^\mu D_\mu u_{Rj} + \overline{d}_{Rj} i\gamma^\mu D_\mu d_{Rj} + \overline{\ell}_{Rj} i\gamma^\mu D_\mu \ell_{Rj} + \\ &+ \overline{e}_{Rj} i\gamma^\mu D_\mu e_{Rj} - \frac{1}{4} W_a^{\mu\nu} W_{\mu\nu}^a - \frac{1}{4} B^{\mu\nu} B_{\mu\nu}.\end{aligned} \tag{1.4}$$

The subscripts j in the equations above indicate a summation over these generations. The relevant fields are as follows:

Q_L : left-handed doublet quark field;

u_R : right-handed singlet up-type quark field;

d_R : right-handed singlet down-type quark field;

ℓ_L : left-handed doublet lepton field;

e_R : right-handed singlet electron field;

$W_a^{\mu\nu}$ (with $a = 1, 2, 3$): field strength tensors for the weak isospin field;

$B^{\mu\nu}$: field strength tensors for the weak hypercharge field.

The EW gauge covariant derivative, denoted as D_μ, is defined as:

$$D_\mu \equiv \partial_\mu - ig' \frac{1}{2} Y_{\text{W}} B_\mu - ig \frac{1}{2} \vec{\tau}_{\text{L}} \cdot \vec{W}_\mu, \tag{1.5}$$

where:

B_μ : U(1) gauge field;

Y_{W} : weak hypercharge, which is the generator of the U(1) group;

$\vec{W}_\mu$: 3-component SU(2) gauge field;

$\vec{\tau}_{\text{L}}$: Pauli matrices, infinitesimal generators of the SU(2) group, with subscript L indicating their action on left-chiral fermions;

Lepton	T	T_3	Y	*Quark*	T	T_3	Y
ν_e	$\frac{1}{2}$	$\frac{1}{2}$	-1	u_L	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$
e_L	$\frac{1}{2}$	$-\frac{1}{2}$	-1	d_L	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{3}$
–	–	–	–	u_R	0	0	$\frac{4}{3}$
e_R	0	0	-2	d_R	0	0	$-\frac{2}{3}$

Table 1.3: Weak isospin (T), third component of weak isospin (T_3) and hypercharge (Y) for 1st generation leptons and quarks. Fermions of the left-handed variety form $SU(2)_L$ weak isospin doublets, whereas their right-handed counterparts exist as isospin singlets. An exception is made for neutrinos, which the theory considers to be massless, thereby lacking a right-handed component.

g' : coupling constants for the U(1) group;

g : coupling constants for the SU(2) group.

In order to preserve the $SU(2)_L$ gauge symmetry, three gauge bosons W_1, W_2, and W_3 need to be introduced. The strength of their interaction with matter-field is represented by g. The carriers of weak charged current interactions, namely the physical $W^{\pm}$ bosons, can be identified as a linear combination of the first two gauge bosons:

$$W^{\pm} = \sqrt{\frac{1}{2}}(W_1 \mp iW_2). \tag{1.6}$$

However, the $SU(2)_L$ symmetry alone cannot fully describe the weak interaction, as the third gauge boson W_3 cannot be directly identified as the physical Z boson.

To achieve a complete description, an additional $U(1)$ local gauge invariance is added, introducing a fourth gauge boson denoted as B with its coupling constant represented by g'. The weak hypercharge Y is related to both the electromagnetic charge Q (a conserved quantity in electromagnetic interactions) and the third component of weak isospin T_3 (conserved in weak interactions):

$$Q = T_3 + \frac{Y}{2}. \tag{1.7}$$

The values of Y for each fermion are provided in Table 1.3. In a groundbreaking development, Glashow, Weinberg, and Salam [7, 8, 9] demonstrated the successful unification of the weak and electromagnetic interactions within the framework of the SM. This unification is achieved by expressing the gauge bosons responsible for both the weak and electromagnetic neutral currents as combinations of the W_3 and B fields, leading to the identification of the electromagnetic (A) and the weak neutral (Z) field:

$$A = \cos\theta_W B + \sin\theta_W W_3, \tag{1.8}$$
$$\mathrm{Z} = -\sin\theta_W B + \cos\theta_W W_3. \tag{1.9}$$

Here, θ_W represents the weak mixing angle, often referred to as the Weinberg angle. This angle is directly linked to the coupling constants g and g' as follows:

$$\tan\theta_W = \frac{g'}{g}. \tag{1.10}$$

The Weinberg angle θ_W is a fundamental parameter in the SM. Experimental measurements [10] have provided a precise value for this parameter:

$$\sin^2\theta_W = 0.23101 \pm 0.00053. \tag{1.11}$$

This value represents a critical component in our understanding of the SM, as it quantifies the EW unification within particle physics.

The Weinberg transformations allow establishing a direct connection between the EW fields and the weak neutral field. The unified framework elegantly combines the weak and electromagnetic forces, providing a deeper understanding of the fundamental interactions that govern the behaviour of elemetary particles.

1.1.3 Spontaneous Symmetry Breaking Mechanism

In the previous section, it was established that both QCD and EW interactions stem from the local gauge invariance principle. However, to maintain the symmetries, the gauge bosons introduced should be massless. This holds true for the photon and gluons but not for the Z and $\mathrm{W}^{\pm}$ bosons, as confirmed by experimental evidence. Furthermore, this model fails to account for the masses of fermions.

The simplest approach to introducing mass terms for the W and Z bosons is through the Brout-Englert-Higgs (BEH) mechanism [11, 12]. This mechanism involves the introduction of a scalar doublet into the theory, allowing for the generation of mass terms for these bosons without breaking the underlying symmetry. The mass-generation process occurs in a subtle way that doesn't immediately appear in the equations. To implement this mechanism, a common choice is to introduce an isospin doublet consisting of two complex scalar fields with weak hypercharge $Y = 1$:

$$\varphi = \begin{pmatrix} \varphi^+ \\ \varphi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \varphi_1 + i\varphi_2 \\ \varphi_3 + i\varphi_4 \end{pmatrix}. \tag{1.12}$$

These fields are incorporated into the Lagrangian with a potential term:

$$V(\varphi) = \mu^2 \varphi^\dagger \varphi + \lambda(\varphi^\dagger\varphi)^2. \tag{1.13}$$

The shape of this potential depends on the values of the parameters μ^2 and λ. When μ^2 is negative and λ is positive, the potential has infinite minima. These minima are degenerate, meaning that they have the same energy.

In order to spontaneously break the symmetry and provide mass to elementary particles, a finite and non-zero value for the vacuum expectation value (VEV) of the scalar field, denoted as φ_0, has been selected. This value:

$$\varphi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad v = 246.22\,\mathrm{GeV}, \tag{1.14}$$

is derived from the Fermi theory of weak interactions [5]. This choice determines the ground state of the system, and results in the breaking of the $SU(2)_L \times U(1)$ symmetry, leading to the acquisition of masses by the associated gauge bosons. Within this framework, the Dirac mass terms for fermions are introduced through interactions with the Higgs field. The coupling of the Higgs field to fermions gives rise to the Dirac mass terms, ensuring that particles such as electrons, quarks, and neutrinos obtain mass. The Dirac mass terms are linked to the EW sector and the Higgs field, a key component of the SM that imparts mass to particles while preserving gauge invariance. The VEV is chosen in such a way that, after symmetry breaking, one linear combination of the $SU(2) \times U(1)$ generators remains massless, which is identified as the photon. This choice includes setting the first component of φ_0 to 0, and it is made to ensure that the photon remains massless $m_A = 0$, preserving the electromagnetic gauge symmetry.
φ_0 sets the scale for mass generation in the EW sector. Using the value depicted in Equation (1.14), the masses of various bosons, including the Higgs boson (m_H), the W$^\pm$ bosons (m_W), and the Z boson (m_Z), and the masses of the fermions (m_f with $f = e, \mu, \tau$), can be expressed as:

$$\begin{aligned} m_\mathrm{H} &= \sqrt{2v^2\lambda} \sim 125.35\,\mathrm{GeV}, \\ m_\mathrm{W} &= \frac{1}{2} v g \sim 80.39\,\mathrm{GeV}, \\ m_\mathrm{Z} &= \frac{1}{2} v \sqrt{g^2 + g'^2} \sim 91.19\,\mathrm{GeV}, \\ m_e &= \sqrt{2v^2\lambda_e} \sim 0.511\,\mathrm{MeV}, \\ m_\mu &= \sqrt{2v^2\lambda_\mu} \sim 105.66\,\mathrm{MeV}, \\ m_\tau &= \sqrt{2v^2\lambda_\tau} \sim 1'776.86\,\mathrm{MeV}. \end{aligned} \tag{1.15}$$

In the SM, the masses of particles, including bosons and fermions, are not predicted from theory alone but are instead considered as parameters that must be determined through experimental observations. Experimental measurements are fundamental for establishing the precise values of these parameters within the model, and this has been a critical and ongoing challenge in the world of particle physics. To date, the only undetermined masses are those of the neutrinos, which, in themselves, represent an extension beyond the SM (BSM).

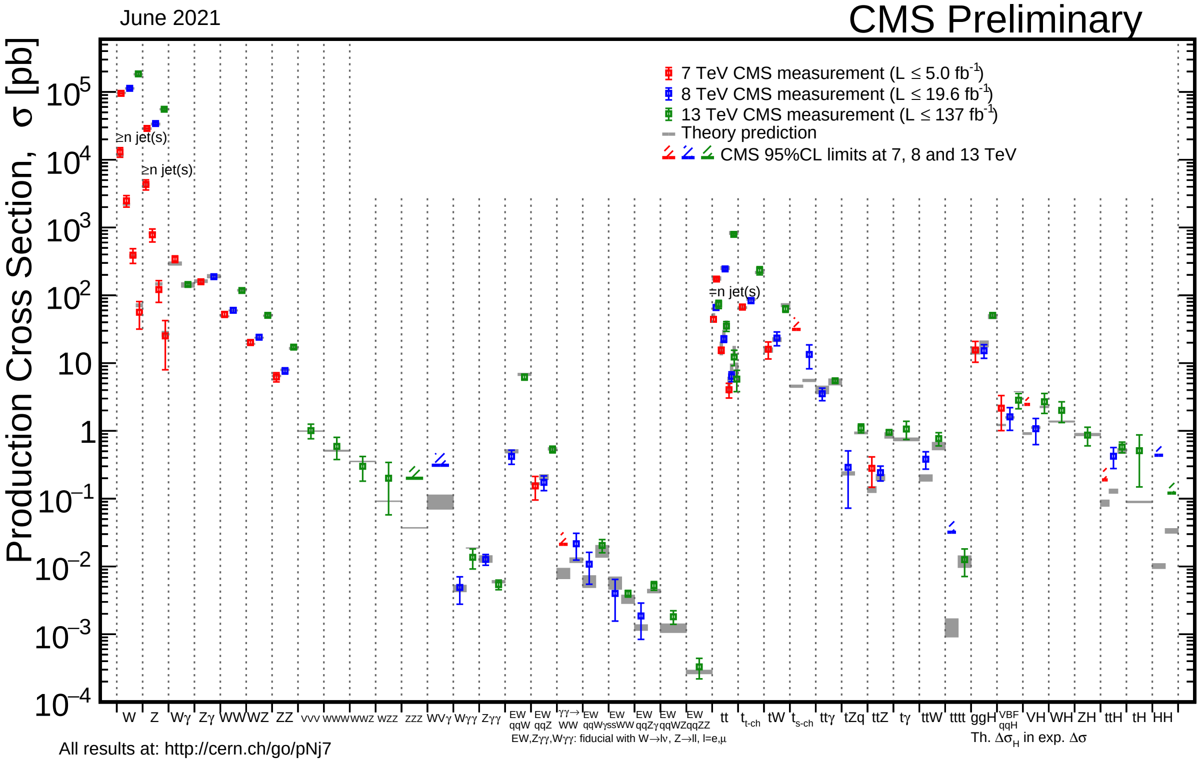


Figure 1.4: Summary of the cross-section measurements of SM processes. From Ref. [13].

1.1.4 Beyond the Standard Model

The SM has stood the test of time, with its validity confirmed by a vast set of experimental measurements at interaction energies up to the EW scale. These measurements encompass various aspects of EW physics, including the discovery of the Higgs boson. Recently, LHC experiments have further confirmed the validity of the SM by exploring rare EW processes never observed before as well as interactions at the TeV scale. Figure 1.4 serves as a summary of all these measurements in EW physics, including those related to the Higgs boson. The wealth of experimental data, combined with the SM predictions, forms a coherent and consistent framework for understanding the behaviour of fundamental particles and their interactions.

Despite the remarkable success of the SM in explaining experimental observations in particle physics, significant theoretical limitations and unexplained experimental evidences remain. In particular, the following challenges persist:

1. Hierarchy problem – In the context of the SM, there are two significant energy scales to consider: the EW scale, coinciding with the VEV of the Higgs field, $v = 246\,\text{GeV}$ (1.14), marks the typical scale of weak interactions. In contrast, the Planck scale, at approximately 10^{19} GeV, marks the energy level where quantum effects of gravity become significant. This scale stands as a point of theoretical pursuit for the unification of EW and gravitational interactions. As one moves to higher and

higher energies, the loop corrections to the Higgs boson mass progressively larger. These loop corrections can be on the order of 10^{17} GeV when approaching the Planck scale. The Higgs boson mass mesured close to the EW scale is about 125 GeV. The hierarchy problem arises because there is a huge difference between the EW scale and the Planck scale, and the loop corrections threaten to push the Higgs mass to values far beyond the EW scale. To avoid this problem, a fine-tuning of the higher order loop-induced contributions to the Higgs mass becomes necessary. In other words, the values of various parameters in the theory must be carefully adjusted to cancel out these large corrections and keep the Higgs mass close to the EW scale. This fine-tuning is considered unnatural and raises questions about the fundamental nature of the theory. Instead of relying on fine-tuning the theory, one alternative perspective is to contemplate the existence of new particles circulating within quantum loops. These particles could potentially mitigate the corrections to the Higgs boson mass, offering a more elegant solution.

2. Particle mass pattern – There are substantial disparities in mass among the particles of the SM. For example, the electron is much lighter than the top quark: $\frac{m_t}{m_e} \sim 10^{15}$. These mass differences are observed experimentally but are not explained within the SM itself.

3. Matter-antimatter asymmetry – Another experimental finding that underscores the limitations of the SM is the observed asymmetry between matter and antimatter in the universe. The SM does incorporate a phenomenon where weak interactions violate CP symmetry. This phenomenon, known as "CP violation", has been observed in the hadronic sector through flavour oscillations in meson systems (such as K and B mesons), and in D-mesons [14]. However, the observed CP violation within the SM is insufficient to account for the excess of matter over antimatter in the universe. This inconsistency necessitates a deeper exploration of physics BSM to provide a more complete explanation.

4. Neutrino mass – Experiments designed to study neutrino oscillations and a direct neutrino mass measurement from beta decays established that their masses, althoug not zero, are much lighter that the electron one ($m_\nu < 1\,\text{eV}$). This represents a challenge to the SM, as it does not naturally accommodate massive neutrinos.

5. Dark matter – The existence of dark matter, which was theorised based on observations in astrophysical and cosmological contexts, is another phenomenon that the SM cannot explain. Its presence is inferred from its gravitational effects on galaxies and the large-scale structure of the universe, but does not consist of any known SM particles.

6. Gravitational interaction – The SM does not incorporate the theory of general relativity, which describes the force of gravity. General relativity and the SM are separate theories that have not been unified within the SM framework. This represents a gap in our understanding of fundamental forces.

7. Unification of forces – The SM successfully describes the EW and strong nuclear forces as separate entities. However, it does not provide a comprehensive framework for unifying all the forces into a single, elegant grand unified theory. This leaves open the possibility of a deeper and more fundamental theory that could unify all known interactions.

Various theoretical models have been proposed to address these issues and extend our understanding BSM. Some of these models, such as supersymmetry (SUSY) or theories involving extra dimensions, introduced new particles and symmetries, and expanded the dimensionality of the space-time. However, as of today, no experimental evidence has been found to support these extensions, and the energy range explored by the LHC, the most powerful particle accelerator to date, has severely constrained the existence of new particles predicted by SUSY, placing significant limits on supersymmetric particles. Consequently, the focus of particle physics experiments has shifted from direct searches for new physics to indirect approaches. This entails the observation of deviations from SM predictions in the differential distributions of known processes. In this context, the Effective Field Theory (EFT) approach [15] serves as a model-independent method to investigate and characterise such deviations. In Section 1.3, this concept will be explored in more detail.

1.2 Vector Boson Scattering

The Higgs boson and the EW sector of the SM have not yet been subject to high-precision investigations, given that the Higgs boson's discovery only took place in 2012 [16, 17]. VBS processes [18] stand as valuable investigative tools for these sectors, offering insights into the mechanisms underlying the spontaneous breaking of EW symmetry. Furthermore, VBS processes possess sensitivity to potential anomalies that may unveil the nature of phenomena BSM. In the following sections, the details of VBS kinematic topology are explored (Section 1.2.1), its connection to the Higgs boson is examined (Section 1.2.2), and a comprehensive overview of ongoing VBS analyses conducted within the CMS experiment is presented (Section 1.2.3).

1.2.1 Signal Topology

VBS processes are exclusively governed by the EW interaction at the leading order (LO), characterised by an order of $\mathcal{O}(\alpha_{EW}^6 \alpha_S^0)$. This calculation takes also into account the final-state objects resulting from the decay of massive vector bosons. In Figure 1.4, a comprehensive range of cross-sections measured at the LHC by the CMS Collaboration is presented, encompassing various processes, including VBS, which stand out as some of the rarest production modes observed. In fact, their cross-sections typically fall within the range of a few femtobarns. Comparable measurements have been reported by the ATLAS Collaboration, as documented in Ref. [19].

When two protons collide at the LHC, their constituent partons may emit vector bosons. This process is depicted in the Feynman diagrams shown in Figure 1.5, specifically illustrating W^+W^- production. These emitted vector bosons subsequently interact, giving rise to the production of two additional vector bosons. Ultimately, these newly formed vector bosons can undergo decay, resulting in the emergence of six fermions within the final state. Figure 1.6 presents diagrams that does not involve an actual scattering between vector bosons, but belong to the order of $\mathcal{O}(\alpha_{EW}^6)$, and hence their inclusion in the signal is crucial to maintain gauge invariance.

When the produced vector bosons VV' are massive, they lead to three distinct final states depending on their decay modes:

- fully leptonic: $VV' \to \ell_1\nu_1\ell_2\nu_2$;
- semileptonic: $VV' \to \ell\nu qq'$;
- fully hadronic: $VV' \to q_1q_1'q_2q_2'$.

While fully leptonic decays offer high purity by minimising contamination from QCD-induced processes, they suffer from a low branching ratio, resulting in substantial statistical uncertainties. Advances in machine learning techniques, proficient at suppressing QCD-induced backgrounds, have created opportunities to explore final states beyond the leptonic modes. Specifically, semileptonic decays have emerged as a promising choice due to their balanced combination of statistical significance and signal purity.

Processes such as those depicted in Figure 1.7 are characterised by being $\mathcal{O}(\alpha_{EW}^4\alpha_S^2)$. In these processes, initial state quarks interact by exchanging a gluon, followed by the emission of W bosons from final state quarks without any interaction between them. These processes are referred to as QCD-induced production of W^+W^- bosons, owing to the presence of two QCD vertices in the Feynman diagrams at LO. They represent an irreducible background for the EW signal, as they share the same final state. The interference between the EW and QCD components occurs at an order of $\mathcal{O}(\alpha_{EW}^5\alpha_S)$, and its contribution typically amounts to just a few percent relative to the signal yield.

The two initial-state quarks, following their interaction in VBS processes, give rise to two jets in the final state, often referred to as "tagging jets" or "VBS jets". These jets are notable for their substantial invariant mass m_{jj} and a significant gap in pseudorapidity $\Delta\eta_{jj}$. The presence of such a di-jets system represents a central characteristic of VBS processes. Indeed, the two initial-state quarks undergo weak interactions, involving the exchange of a fraction of their energy and experiencing minor deviations from their original trajectories.

Kinematic selections applied to these variables play a crucial role in distinguishing the VBS signal from other processes that may mimic the VBS final state. These selections, in particular, prove indispensable in separating the signal from the inherent QCD-induced background. The tagging jets associated with the EW signal exhibit a high invariant mass and substantial separation in pseudorapidity. In contrast, jets linked to W^+W^- QCD-induced background (or the interference) tend to be less energetic and more centrally located. This discrepancy arises due to the strong interaction between quarks, leading to a significant energy exchange and substantial deflection of the interacting particles.

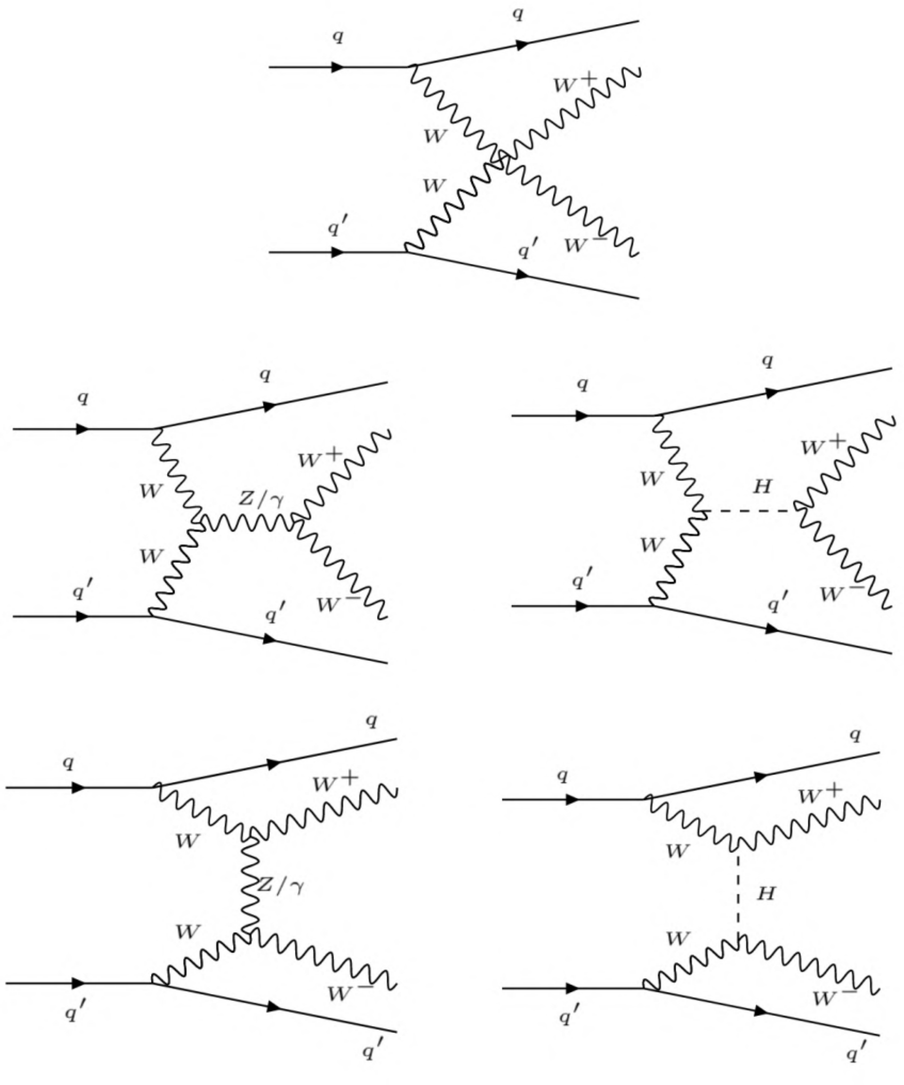


Figure 1.5: Illustrative VBS Feynman diagrams at $\mathcal{O}(\alpha_{EW}^6)$ contributing to $\mathrm{W}^+\mathrm{W}^-$ EW production.

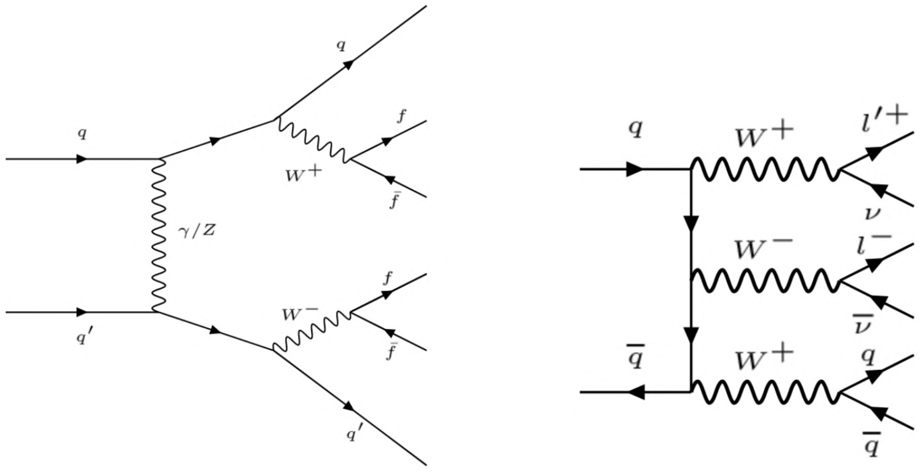


Figure 1.6: Additional diagrams at order $\mathcal{O}(\alpha_{EW})^6$. Their importance lies in preserving gauge invariance.

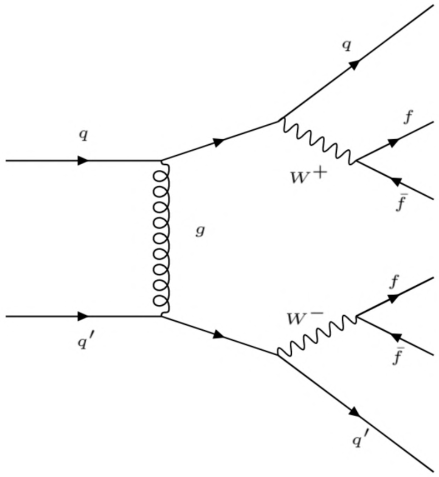


Figure 1.7: Example of Feynman diagrams depicting the QCD-induced production of W^+W^- bosons at an order of $\mathcal{O}(\alpha_{EW}^4\alpha_S^2)$.

Consequently, the imposition of topological selections on m_{jj} and $\Delta\eta_{jj}$ stands as the most effective approach for distinguishing between EW and QCD production and defining a VBS-like phase space characterised by high m_{jj} and $\Delta\eta_{jj}$ values.

1.2.2 Higgs Boson and Unitarity of VBS Cross-Sections

When calculating the amplitude for the $\mathrm{W^+W^-} \to \mathrm{W^+W^-}$ scattering process, the matrix element diverges as $\frac{s^2}{M_W^4}$, when solely taking into account the diagram with the quartic gauge coupling depicted in Figure 1.5.

Incorporating the diagrams with vector boson exchange in the s- and t-channels (the left column of Figure 1.5) does not resolve this divergence issue, as the cross-section continues to diverge with the centre of mass energy (s) [20]:

$$-iM(\mathrm{W^+W^-} \to \mathrm{W^+W^-}) \sim \frac{s}{M_W^2} \quad \text{as} \quad s \to \infty. \tag{1.16}$$

The divergence is only rectified by introducing diagrams in which a scalar boson is exchanged, as illustrated in the right column of Figure 1.5. Remarkably, this scalar particle happens to be the Higgs boson, with a mass of 125 GeV.

Figure 1.8 illustrates the dependence of cross-sections on the centre-of-mass energy $\sqrt{s}$ for various VBS processes in different scenarios. The divergence is mitigated, and unitarity is restored when a particle like the Higgs boson is considered.

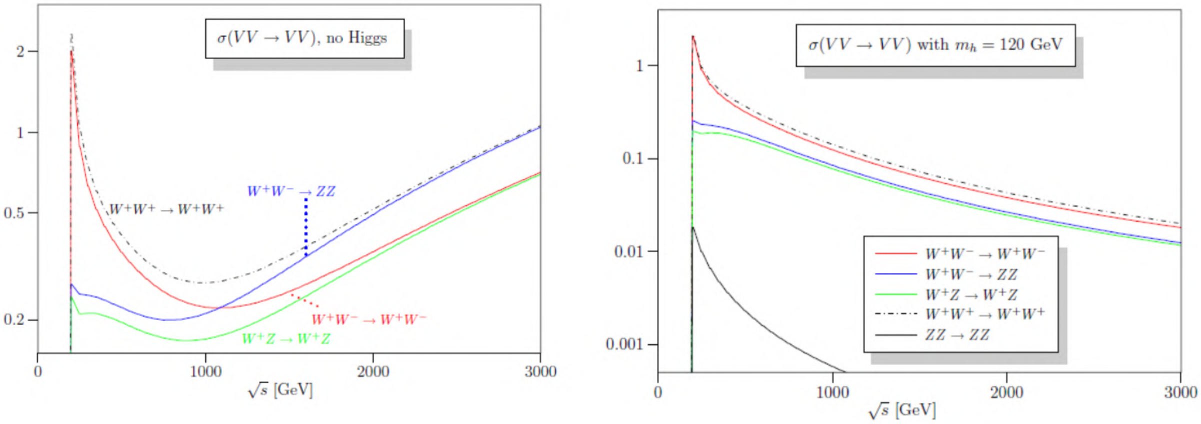


Figure 1.8: Cross-sections (in nb) for five distinct scattering processes involving longitudinal weak gauge bosons. The left panel illustrates the scenario without a Higgs boson, while the right panel represents the SM with a Higgs boson mass of 120 GeV. Adapted from Ref. [21].

VBS processes underscore the importance of this scalar boson within the SM as its presence ensures the theory's unitarity. However, any deviations from the SM expectations could disrupt this delicate balance. Alterations in the triple or quartic gauge couplings can impact the interaction vertices, resulting in cross-section behaviours that deviate from the expected values. Consequently, the study of VBS cross-sections serves as a potent tool for identifying potential anomalies in these couplings, which in turn can provide insights into new physics phenomena.

1.2.3 VBS Measurements in CMS

The study of EW production involving two vector bosons in association with two jets has been ongoing at the LHC since its inception during Run 1, with an integrate luminosity of approximately 30 fb^{-1}. However, the limited data volume and the relatively small cross-sections of these processes have thus far precluded any definitive claims of observation. It was only with the advent of the Run 2 dataset, which amounted to 138 fb^{-1}, that the first confirmed observation of a VBS process was made [22]. This achievement was largely facilitated by the increase in the centre-of-mass energy to $\sqrt{s} = 13$ TeV in Run 2 and the significant expansion of data collected. These enhancements substantially boosted the sensitivity of these searches, which had been previously obstructed by their very low cross-sections, as illustrated in Figure 1.4.

Figure 1.9 presents a comprehensive summary of CMS results in this domain, encompassing measurements from both Run 1 and Run 2 (as of May 2021). A corresponding picture consistent with CMS results can be drawn for ATLAS findings [19]. In general, these results align well with predictions from the SM, with statistical uncertainty being the primary source of error for most of the measurements.

In the following, some of the most recent results published by the CMS Collaboration are briefly discussed to provide an overview of the current state of knowledge regarding VBS processes. This context will help elucidate the significance of the work presented in this thesis.

The Golden Channel

The initial observation of a VBS process occurred in the final state with massive W bosons decaying leptonically. Specifically, the CMS Collaboration reported this discovery with the 2016 dataset, amounting to 35.9 fb^{-1}, in the context of the VBS production of two same-sign W bosons [22]. This particular channel, often referred to as the “golden channel”, was chosen due to its favourable signal-to-background ratio in terms of EW and QCD contributions.

Subsequently, a full Run 2 analysis by CMS [23] presented simultaneous cross-section measurements for the production of same-sign WW and WZ boson pairs in association with two jets in the fully leptonic final state. This final state comprises either two charged leptons ($\ell^{\pm}\ell'^{\pm}$, with $\ell, \ell' = e, \mu$) from the W pair decays, or three charged leptons ($\ell^{\pm}\ell'^{\pm}\ell'^{\mp}$) originating from WZ boson decays. The EW production of WZ bosons in association with two jets was observed with a statistical significance of 6.8 standard deviations, while an expectation of 5.3 standard deviations had been anticipated. In-depth studies of inclusive and differential cross-sections were conducted for both same-sign WW and WZ production,

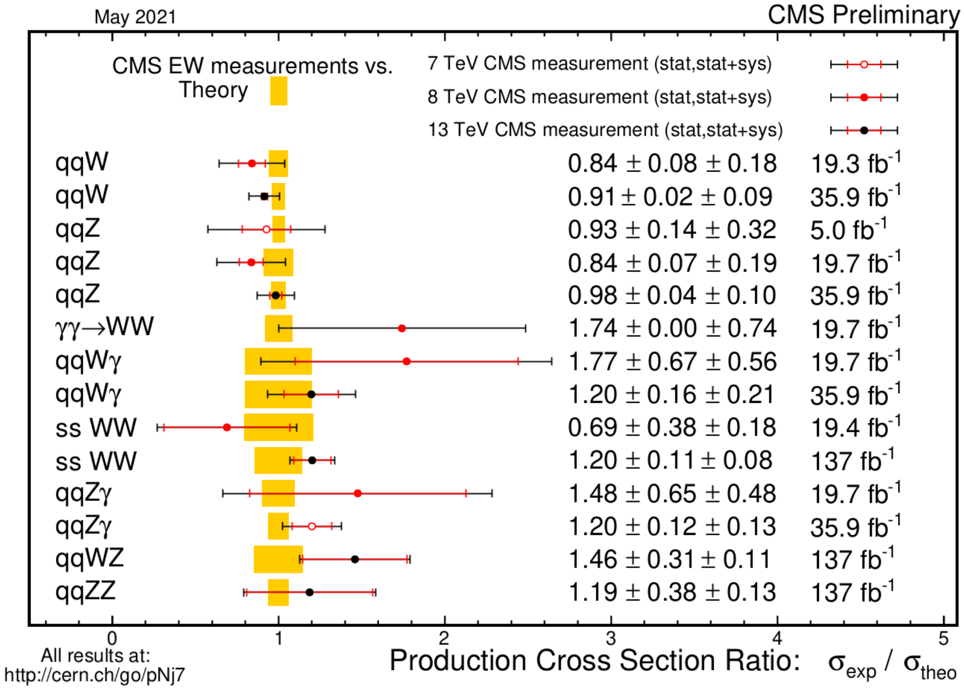


Figure 1.9: Overview of the cross-section ratios for pure EW interactions involving gauge bosons, with data updated as of May 2021.

demonstrating good agreement with predictions from the SM.

A recent analysis by CMS exploited the golden channel to obtain the first measurements of polarised cross-sections. This analysis used 137 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV and was conducted in the fully leptonic final state, which ensures optimal signal purity [24]. Machine learning techniques were employed to discriminate between various polarised components, ultimately facilitating the separation of the inclusive signal from the SM background. In particular, the key technique employed in this analysis involved training boosted decision trees (BDT) using discriminating kinematic and angular observables. These observables played a crucial role in distinguishing various polarised components within the final states (as illustrated in Figure 1.10). Furthermore, this technique was instrumental in disentangling the inclusive signal from the SM background.
The results, derived for helicity eigenstates in both the parton-parton and WW centre-of-mass reference frames, set a 95% confidence level upper limit of 1.17 fb (0.88 expected) on the cross-section of longitudinally polarised same-sign WW bosons in the bosons' centre-of-mass reference frame. Additionally, the observed (expected) statistical significance for the same-sign WW EW signal, featuring at least one longitudinally polarised W boson, in the WW centre-of-mass reference frame, was found to be 2.3 (3.1) standard deviations. All measured cross-sections are aligned with the expectations of the SM.

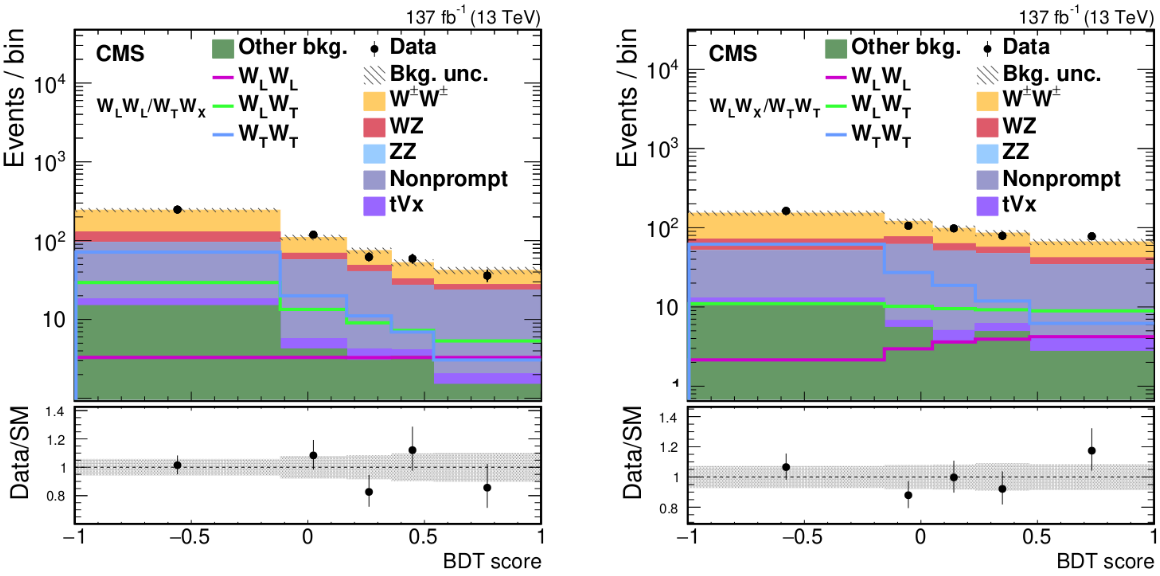


Figure 1.10: Post-fit distributions showcasing the output scores of the signal Boosted Decision Tree (BDT) used for the $W_L^{\pm}W_L^{\pm}/W_X^{\pm}W_T^{\pm}$ (left) and $W_L^{\pm}W_X^{\pm}/W_T^{\pm}W_T^{\pm}$ (right) cross-section measurements. Here, 'L' and 'T' signify the longitudinal and transverse states of polarisation, while 'X' may denote either 'L' or 'T'. In the lower panel of each figure, the ratio between the observed data events and the total SM prediction is presented. The shaded gray regions denote the uncertainties associated with the predicted yields, while vertical bars represent the statistical uncertainties in the data. These plots are sourced from Ref. [24].

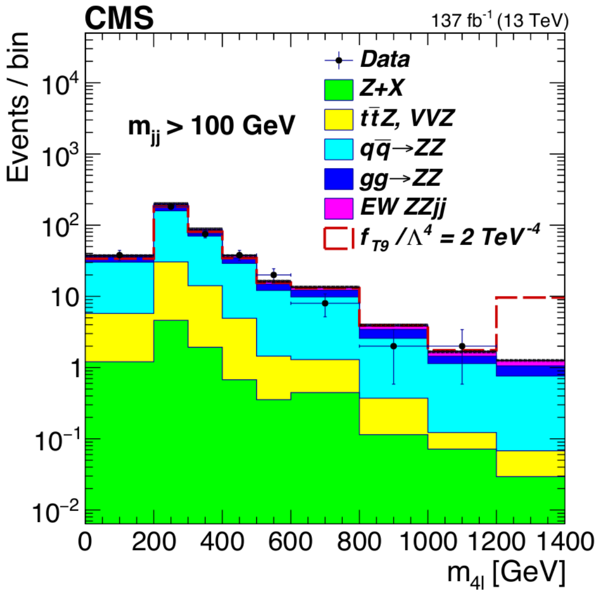


Figure 1.11: Post-fit distributions of the four-lepton invariant mass for $\frac{f_{T9}}{\Lambda^4}$. Data points with error bars represent the observed data, while the filled histograms depict the fitted signal and background contributions. Source: Ref. [25].

ZZ and Zγ VBS

The production of VBS ZZ bosons in the leptonic decay mode, while offering high purity, has yet to be observed due to its lower branching ratio. Nevertheless, CMS published evidence for the EW signal in this channel [25] with an observed (expected) statistical significance of 4.0 (3.5) standard deviations in the final state with four leptons, utilising the complete Run 2 dataset. The analysis reported a measured fiducial cross-section of $\sigma_{\text{fid}} = 0.33^{+0.11}_{-0.10}$ (stat) $^{+0.04}_{-0.33}$ (syst) fb, consistent with the expectations from the SM. This channel offers an excellent platform for probing anomalies and has led to the establishment of limits within an EFT framework, employing the invariant mass of the four final-state leptons $m_{4\ell}$ as a discriminating variable (see Figure 1.11). The ZZ channel currently sets the most stringent limit at a 95% confidence level on the neutral current operator T_8, bounding it within $-0.43 < \frac{f_{T8}}{\Lambda^4} < 0.43$.

A recent publication from the CMS Collaboration marked the observation of VBS production with γ−Z with a statistical significance exceeding five standard deviations [26]. The corresponding paper presents inclusive cross-sections for both the EW signal in isolation (5.21 ± 0.52 (stat) ± 0.56 (syst) fb) and the combined EW+QCD production (14.7 ± 0.80 (stat) ± 1.26 (syst) fb). All measured cross-sections align closely with predictions from the SM at LO. Additionally, the Zγ analysis places stringent constraints on a set of anomalous quartic gauge couplings defined within the EFT framework.

Semileptonic Channels

CMS has recently delved into final states distinct from pure leptonic ones [27], leading to the first-ever observation of the EW production of a WV pair in the semileptonic channel, with an observed (expected) significance of 4.4 (5.1) standard deviations.

The appeal of the semileptonic channel lies in its potential to augment the number of signal events, thanks to a higher branching ratio compared to the leptonic channel. However, this enhancement comes at the cost of an increased QCD background, particularly due to events involving W+jets.

Advanced machine learning methods were instrumental in disentangling the signal from the background.

The measured fiducial cross-section for the EW WV signal stands at $1.90^{+0.53}_{-0.46}$ pb, in good agreement with the theoretical prediction of $2.23^{+0.08}_{-0.11}$ (scale) $^{+0.05}_{-0.05}$ (pdf) pb. The results are in agreement with SM expectations.

WWjj VBS

Finally, a part of the CMS Collaboration's latest research efforts, undertaken during the course of this thesis, delved into the purely EW production of a pair of W bosons featuring opposite sign charges in association with two jets ($\mathrm{W}^+\mathrm{W}^-jj$). This channel had previously evaded detection, because of the significant $\mathrm{t\bar{t}}$ background, making it challenging to identify. Moreover, the final state was affected by contamination from a stronger QCD-induced background compared to the same-sign WW channel. The contamination from Drell-Yan (DY) production becomes significant in final states with same flavour leptons (e.g., ee and $\mu\mu$).

The development of sophisticated machine learning techniques played a pivotal role in achieving optimal discrimination between the signal and the two primary background processes, $\mathrm{t\bar{t}}$ and QCD-induced $\mathrm{W}^+\mathrm{W}^-$.

These efforts culminated in the discovery of this process, achieving a statistical significance of 5.6 standard deviations. As a result, the first observation of purely EW $\mathrm{W}^+\mathrm{W}^-jj$ production was confirmed, and a measurement of its cross-section was presented. The details of this groundbreaking result, as well as the methodologies employed, are thoroughly explained in Chapter 4.

1.3 Standard Model Effective Field Theory

Direct searches for new physics, both model-dependent, such as SUSY, and more model-independent, have, as of now, revealed no significant deviations from the predictions of the SM. With the LHC now entering a regime of fixed centre-of-mass energy precision physics, it is becoming increasingly relevant to focus on indirect searches for new physics that are as model-independent as possible. In situations where it is postulated that the energy scale of this new physics is well beyond the reach of the LHC, the adoption of an EFT framework becomes a powerful tool for capturing its potential effects.

The SMEFT is a theoretical framework that expands upon the SM. It accomplishes this by introducing higher-dimensional operators derived from SM fields and symmetries. SMEFT is a model-independent framework specifically designed to accommodate potential extensions of the SM that involve not only the presence of heavy particles but also new interactions. For a more detailed theoretical description of EFTs and SMEFT, see Refs. [28, 29].

In the SMEFT framework, the Lagrangian can be expressed as follows:

$$\mathcal{L}_{\mathrm{SMEFT}} = \mathcal{L}_{\mathrm{SM}} + \mathcal{L}^{(5)} + \mathcal{L}^{(6)} + \mathcal{L}^{(7)} + \ldots, \tag{1.17}$$

where $\mathcal{L}^{(d)}$ for $d > 4$ is given by:

$$\mathcal{L}^{(d)} = \sum_{\alpha} \frac{C_{\alpha}^{(d)}}{\Lambda^{d-4}} Q_{\alpha}^{(d)}. \tag{1.18}$$

$\mathcal{L}_{\mathrm{SM}}$ represents the SM Lagrangian. $\mathcal{L}^{(5)}$, $\mathcal{L}^{(6)}$, $\mathcal{L}^{(7)}$, and so forth correspond to higher-dimensional operators $Q_{\alpha}^{(d)}$ with a dimension $d > 4$.

Now, a closer examination of Equation (1.18) will be undertaken. The operators $Q_{\alpha}^{(d)}$ are suppressed by powers of $(d-4)$ relative to the new physics scale Λ. This implies that, with the consideration of operators at higher energy scales, their contributions become relatively weaker compared to the cutoff scale Λ. The $(d-4)$ factor reflects how rapidly this influence diminishes as one move to higher energy scales. The term $C_{\alpha}^{(d)}$ represents the Wilson coefficients, determining the strength or magnitude of each operator. The index α runs over a complete and non-redundant set of operators that are invariant under the SM gauge symmetries. In alignment with the guidelines established by the LHC EFT Working Group [30], the so-called Warsaw basis [31] is employed in this project. The Lagrangian is defined according to the SMEFTsim package implementation [32, 33].

The SMEFT framework primarily focuses on even-dimensional operators, such as dimension-6, dimension-8, and so on. This is because odd-dimensional operators, such as dimension-5, are typically associated with physical processes that lead to neutrino masses, which are beyond the scope of the SMEFT at colliders.

Concerning the operators $Q_{\alpha}^{(d)}$, to obtain the Warsaw basis some constraint are considered. First of all, this basis takes into account only dimension-6 operators. These operators are the lowest-dimensional operators that can have a significant impact on low-energy observables while still providing a relatively simple and manageable framework for extending the SM. The dimension-6 operators are particularly important for describing deviations from the SM predictions at energy scales accessible by current experiments, such as those at the LHC. Including higher-dimensional operators, while possible, would introduce a much larger number of terms and complicate the analysis without necessarily significantly improving the precision of predictions for current experiments.

Others constraints to consider when forming the Warsaw basis are the CP conservation and the flavour symmetry. The Warsaw basis usually assumes CP conservation, meaning that the Lagrangian is invariant under CP transformations. This simplifies the basis by reducing the number of independent operators. Moreover, the fermionic operators are required to be invariant under a $U(3)^5$ flavour symmetry[2]. This constraint reduces the number of independent flavour-violating operators and simplifies the analysis, particularly in the context of quark and lepton flavour physics. With these constraints are applied, the Warsaw basis includes a set of 59 dimension-6 operators.

Focusing on the influence of dimension-6 operators operating at the order of Λ^{-2}, the scattering matrix amplitude $\mathcal{A}$ for a specific process undergoes a correction:

$$\mathcal{A} = \mathcal{A}_{\mathrm{SM}} + \sum_{\alpha} \frac{c_\alpha^{(6)}}{\Lambda^2} \mathcal{A}_{Q_\alpha}. \tag{1.19}$$

Here, $\mathcal{A}_{\mathrm{SM}}$ stands for the SM amplitude, while $\mathcal{A}_{Q_\alpha}$ corresponds to the overall amplitude resulting from the insertion of the operator Q_α, with α running over a the chosen set of operators. The latter amplitude exhibits a linear relationship with respect to $\frac{c_\alpha^{(6)}}{\Lambda^2}$, where $c_\alpha^{(6)}$ represents the Wilson coefficient associated with the dimension-6 operator Q_α.

Consequently, the expected number of events N within a specific phase-space region scales with the Wilson coefficients as described in the following equations:

$$N \propto |\mathcal{A}|^2 = |\mathcal{A}_{SM}|^2 + \sum_{\alpha} \frac{c_\alpha}{\Lambda^2} 2\,\mathrm{Re}\left(\mathcal{A}_{SM}\mathcal{A}^\dagger_{Q_\alpha}\right) + \sum_{\alpha,\beta} \frac{c_\alpha c_\beta}{\Lambda^4}\left(\mathcal{A}_{Q_\alpha}\mathcal{A}^\dagger_{Q_\beta}\right), \tag{1.20}$$

$$N = N_{\mathrm{SM}} + \sum_{\alpha}\left(\frac{c_\alpha}{\Lambda^2} N_\alpha^{\mathrm{int}} + \frac{c_\alpha^2}{\Lambda^4} N_\alpha^{\mathrm{quad}}\right) + \sum_{\alpha\neq\beta}\left(\frac{c_\alpha c_\beta}{\Lambda^4} N_{\alpha,\beta}^{\mathrm{mix}}\right). \tag{1.21}$$

In Equation (1.20), when $\alpha = \beta$, the final term simplifies to $\frac{c_\alpha^2}{\Lambda^4}|\mathcal{A}_{Q_\alpha}|^2$. The overall outcome combines the SM prediction, a term with linear dependence on the Wilson coefficients arising from the interference between SM and BSM amplitudes, and a pure BSM contribution. The latter exhibits quadratic dependence on the Wilson coefficients and is conventionally referred to as the "quadratic term". Equation (1.21) further divides the last term into individual and mixed quadratic contributions.

The outcome presented in Equation (1.21) is applicable to both integrated observables and bin-by-bin representations of differential observables. For a specific observable, the quantities N_{SM}, N_α^{int}, N_α^{quad}, $N_{\alpha,\beta}^{\mathrm{mix}}$ can be estimated numerically. With a total of n operators contributing, there are n independent linear terms, n individual quadratic terms, and $\frac{n(n-1)}{2}$ mixed terms. Consequently, the entirety of EFT contributions can be determined by evaluating N at a total of $\frac{n(n+3)}{2}$ (20 for $n = 5$) independent points across the parameter space $\{c_\alpha\}$. In this work, these evaluations are conducted through Monte Carlo (MC) event simulations, as outlined in Chapter 5.

---

[2] The $U(3)^5$ flavour symmetry is used to describe how different types of quarks (up, down, strange, charm, and beauty) transform under flavour-changing processes, such as weak interactions. There are five different $U(3)$ subgroups within the larger $U(3)$ flavour symmetry, each representing transformations specific to an individual quark flavour.

Theoretical predictions suggest that SMEFT operators can affect both vertex interactions and particle propagators, primarily through correction of masses and decay widths. However, considering the limitations of presently available tools [29], generating comprehensive simulations at quadratic order for the dimension-6 Wilson coefficients that affect propagators is currently unfeasible. As a practical workaround, the exclusion of all propagator corrections from fundamental analyses is chosen, with the focus solely on the vertex EFT corrections.

1.3.1 SMEFT in Vector Boson Scattering

VBS signatures hold a prominent place in the quest for new physics. In fact, as example, VBS processes offer tree-level sensitivity to effective operators responsible for modifications of triple (TGC) and quartic gauge couplings (QGC). While dedicated dimension-6 phenomenological studies have been conducted for specific VBS signatures, such as ZZ [34] and same-sign WW [35], a first attempt to combine VBS and diboson results on SMEFT operators was recently presented in [36]. This projection suggests that although VBS currently has a visible but relatively small impact on global EFT fits, its significance is expected to grow in the future.

In Chapter 5, the sensitivity of the W^+W^- VBS channel to a set of five dimension-6 bosonic operators within the Warsaw basis will be explored. These operators include $Q_W, Q_{HW}, Q_{HWB}, Q_{HD}, Q_{H\Box}$ as defined in 5.1. This choice was due to the fact that the VBS, being a multiboson process, would be more sensible to bosonic operators. This study aims to evaluate the projected sensitivities to these effective operators of the W^+W^- VBS channel. This approach involves working at the reconstructed level, commonly referred to as “reco-level”. This entails the analysis of experimental data post-reconstruction and calibration, rendering it ready for detailed physics analysis. Consideration is given to all sources of background and source of uncertainties to replicate a fully realistic analysis. The primary goal is to conduct a comprehensive analysis using CMS data in one of the channels most sensitive to VBS processes, the W^+W^- fully leptonic. This approach allows studying the data in a form that closely resembles the final output, ensuring that the analysis is as close to real experimental conditions as possible. Limits on the Wilson coefficients are then extracted, to be regarded as indicative of the maximum sensitivity of these processes to EFT effects.

The CMS Experiment at the LHC

The LHC [37] is a circular accelerator designed to explore the frontiers of high-energy physics, probing energy scales of several TeV. Its primary mission was to investigate the EW sector and discover the Higgs boson, a task successfully accomplished in 2012 [16, 17]. Subsequently, after this landmark discovery, research at the LHC has focused – among the other things – on the properties of the Higgs boson, contributing significantly to our understanding of the symmetry-breaking mechanism.

The LHC continues to explore various research paths, including verifying the SM's consistency and exploring potential extensions to it. In this chapter, a comprehensive overview of the LHC machine is provided (Section 2.1) and briefly describe the CMS detector (Section 2.2), which collected the data used in this study. Following this, in Section 2.3, a description of the algorithm employed to reconstruct the final state particles, commonly referred to as physics objects, is provided. Lastly, in Section 2.4, the High-Luminosity LHC (HL-LHC) upgrade is introduced, addressing both the prospects and the challenges that come with it. Additionally, the upgrade planned for the Electromagnetic Calorimeter (ECAL) will be explored.

2.1 The Large Hadron Collider

The LHC is a circular accelerator located in the vicinity of the city of Geneva, spanning the border between France and Switzerland. It is situated underground at a depth of approximately 100 meters, and boasts a circumference of approximately 27 kilometres, making it the most extensive and most energetic particle accelerator ever constructed.

The LHC is capable of producing collisions involving protons (lead ions), with a nominal centre-of-mass energy reaching up to 14 TeV (5.5 TeV) and achieves luminosities of up to $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ ($10^{27}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$). Around the LHC ring, four points of interaction are strategically positioned, each equipped with a different detector designed for specific purposes. ATLAS and CMS are general-purpose detectors used for a wide range of SM and BSM physics research. In contrast, the LHCb detector focuses on processes involving the bottom quark and seeks evidence of CP violation. Meanwhile, ALICE serves as a dedicated tool for the examination of lead ion collisions.

To bend the paths of the two counter-rotating proton beams, the LHC employs two opposite dipole magnetic fields. Consequently, two separate pipes run along the ring, each dedicated to one of the beam directions. The accelerator structure consists of eight straight sections, each 528 meters in length, and eight long arc sections, each extending 2.9 kilometers. Collisions occur in four of these eight straight sections, each of which is equipped with one of the four previously mentioned detectors. The LHC relies on powerful 8 Tesla magnetic dipole fields in the arc sections to bend the proton trajectories. These magnets are constructed using Niobium-Titanium wires covered with Copper and are maintained in a superconducting state at a temperature of 1.9 Kelvin using a cryogenic Helium-4 system. Magnetic quadrupoles are strategically placed to focus the proton beams in the centre of the pipes along the LHC ring. A system of three quadrupoles precedes each interaction point to precisely adjust the beam's focalisation, enhancing the rate of hard interactions.

The LHC typically distributes protons into up to 2800 bunches, each containing approximately $1.15 \cdot 10^{11}$ protons. These bunches travel along the ring with a nominal time separation of 25 ns. Proton bunches are generated from ionised hydrogen atoms and are initially accelerated to 450 GeV by a sequence of machines, including Linac2, Proton Synchrotron Booster (PSB), Proton Synchrotron (PS), and Super Proton Synchrotron (SPS), as shown in Figure 2.1. Once the proton bunches are prepared, they are injected into the LHC, where their energy is further increased to 7 TeV using a 400 MHz superconducting cavity system located in one of the straight sections of the accelerator. Stable beam conditions allow for collisions to occur at a rate of 40 MHz.

The instantaneous luminosity ($\mathcal{L}$) delivered by a colliding machine is defined as the number of particle collisions per unit cross-sectional area per unit time:

$$\mathcal{L} \cdot \sigma = \frac{dN}{dt}. \tag{2.1}$$

Here, dN is the number of particle collisions observed in a given time frame dt, and σ is the cross-sectional area over which the collisions occur. $\mathcal{L}$ solely depends on the characteristics of the circulating beams and is a crucial measure of a collider's performance. Integrating the instantaneous luminosity over time yields the integrated luminosity (L):

$$L = \int_{t_0}^{t_1} \mathcal{L} dt = \frac{N}{\sigma}, \tag{2.2}$$

where L is directly associated with the number of events collected N for a particular process with cross-section σ during the LHC's operational period. The instantaneous and integrated luminosity delivered by the LHC in the 2010-2018 period are reported in Figure 2.2.

In each collision between two proton bunches (bunch crossing), multiple interactions occur. The number of these interactions per crossing, referred to as pile-up, is a significant parameter characterising the LHC environment. Figure 2.3 illustrates the average number of interactions per bunch crossing, as measured by the CMS experiment during LHC data-taking period. During Run 2, an average of 20-40 simultaneous collisions occurred in a bunch crossing.

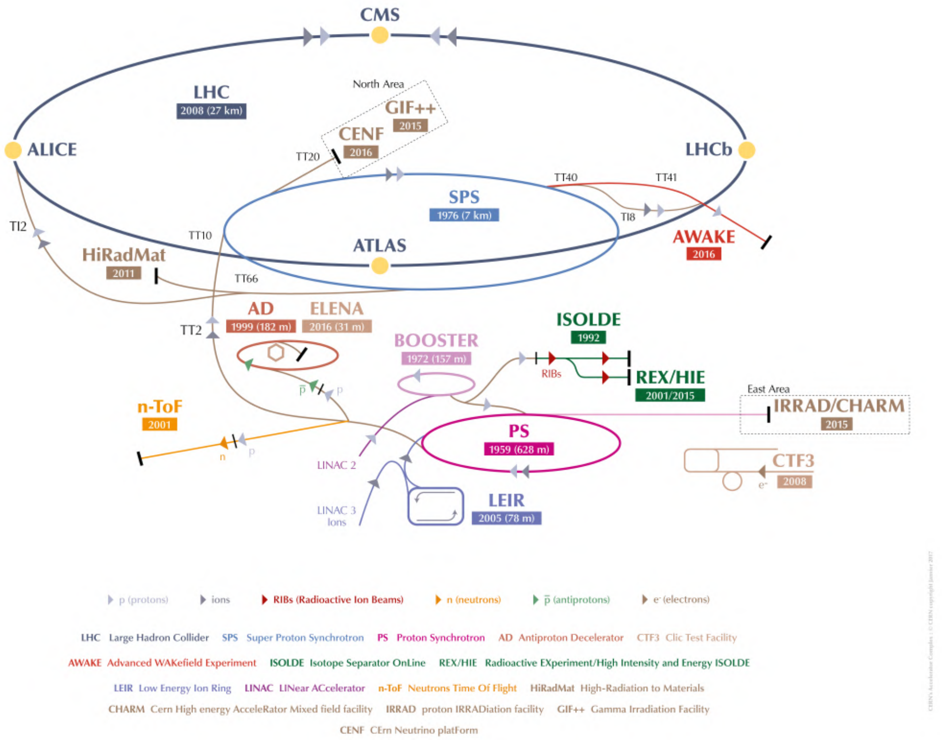


Figure 2.1: The LHC complex. From Ref. [38].

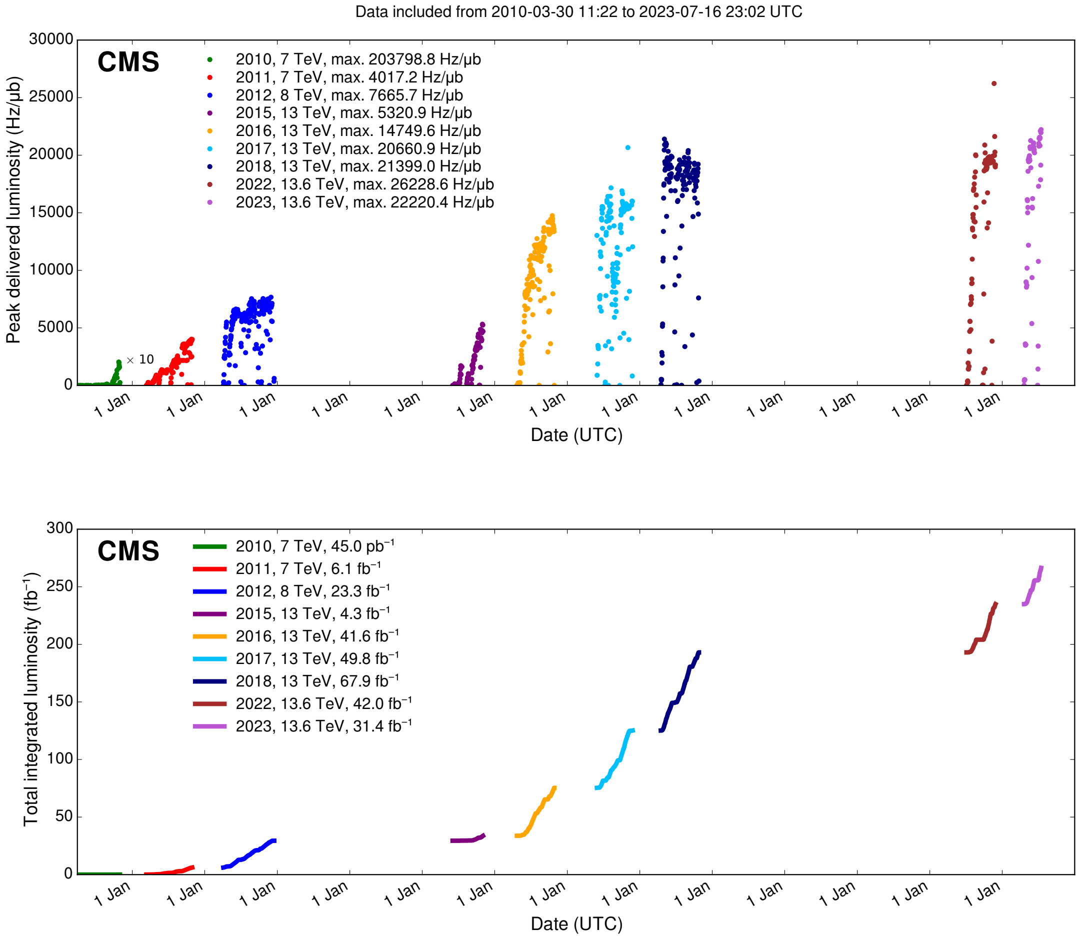


Figure 2.2: Top: Instantaneous luminosity versus day delivered by the LHC to CMS during proton-proton collisions throughout the period from 2010 to 2023. Bottom: Integrated luminosity versus day delivered by the LHC to CMS during proton-proton collisions over the same period, with each year starting at the endpoint of the previous year. From Ref. [39].

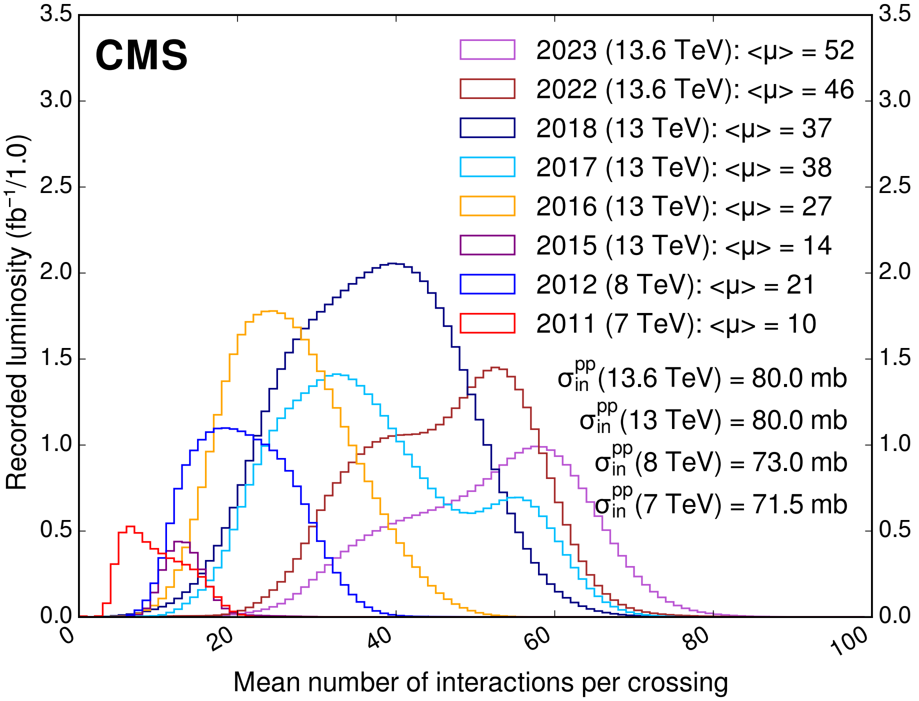


Figure 2.3: Distribution of the average number of interactions per crossing (pile-up) for proton-proton collisions throughout the period from 2011 to 2023. The overall mean values and the minimum bias cross-sections are also shown. From Ref. [39].

The data used for this thesis corresponds to the period from 2016 to 2018, known as Run 2, with an integrated luminosity of approximately 138 fb^{-1}. After the ongoing Run 3, a third LS phase will follow, during which extensive upgrades to the ATLAS and CMS detectors will prepare them for the high-luminosity phase (HL-LHC) expected to start in 2026. The HL-LHC aims to accumulate an integrated luminosity approximately ten times greater than that achieved by the LHC, as will be discussed in Section 2.4.

2.2 The Compact Muon Solenoid

The Compact Muon Solenoid (CMS) detector [40] is a versatile multi-purpose apparatus situated around one of the four interaction points of the LHC, as can be seen in Figure 2.1. It features a cylindrical shape, measuring 21.6 meters in length, 14.6 meters in diameter, and weighing a total of 14'000 tons. The CMS detector comprises several subdetectors, each serving a specific purpose. Information collected from these various components is combined to reconstruct the complete kinematics of final-state particles.

Closest to the collision point, the tracker system, composed of silicon pixels and strips, records the tracks of charged particles (Section 2.2.1). Surrounding the tracker are the electromagnetic calorimeter and the hadronic calorimeter. The former measures the energy of electrons and photons (Section 2.2.2), while the latter quantifies observables related to hadrons (Section 2.2.3). All three detectors are located within a superconducting solenoid measuring 13 meters in length and 6 meters in diameter, providing a magnetic field of 3.8 Tesla, ensuring precise momentum measurements for even high-energy charged particles. Beyond the solenoid, only the muon chambers are situated (Section 2.2.4), where the return magnetic field saturates the 1.5 meters of iron within the holding structure, housing the muon stations. In Figure 2.4, a schematic representation of the CMS detector with all its subdetectors is presented. For a comprehensive discussion of the CMS detector, see Ref. [40].

The CMS coordinate system has its origin at the nominal interaction point within the experiment. It employs a right-handed coordinate system: the x-axis points radially toward the LHC's center, the y-axis extends vertically upward, and the z-axis aligns with the beam direction. A spherical coordinate system (r, φ, ϑ) is used, well-suited to the detector's geometry. The azimuthal angle φ is measured from the x-axis in the (x, y) plane, and the radial coordinate in this plane is denoted by r. The polar angle ϑ is measured from the z-axis.

Pseudorapidity is defined as

$$\eta = -\ln\left(\tan\frac{\vartheta}{2}\right), \tag{2.3}$$

and is often preferred over ϑ. In fact, QCD-mediated scattering processes, the most common interactions at the LHC, exhibit a nearly uniform distribution in η. Additionally, for boosts along the z-axis, the η difference for massless particles is a Lorentz invariant quantity.

In the (ϑ, φ) plane, to expresses the angular separation between two particles, this following

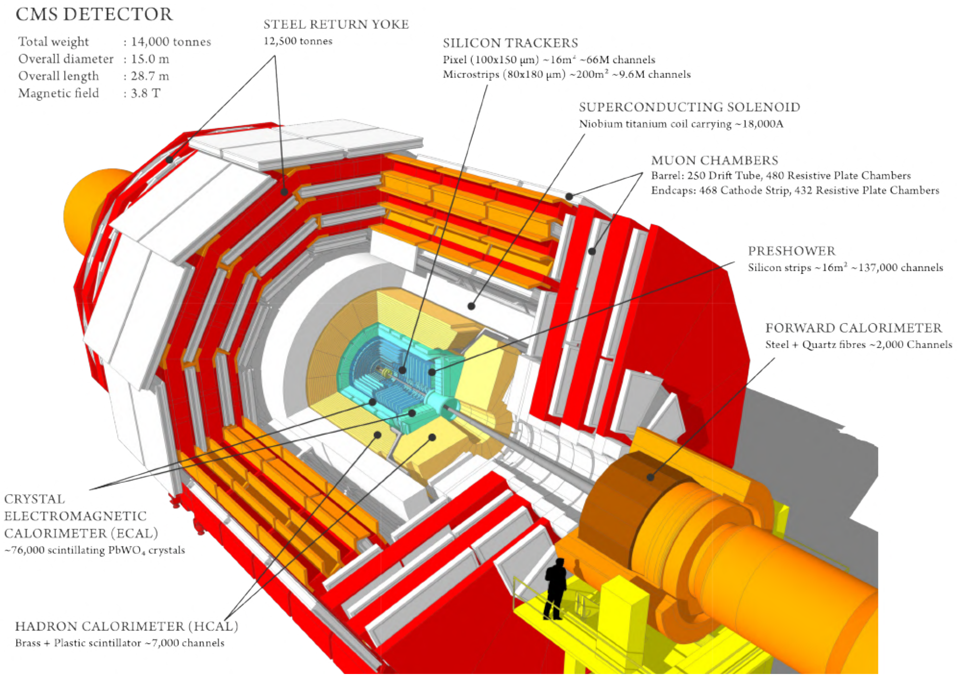


Figure 2.4: Sectional view of the CMS detector. The LHC beams circulate in opposite directions along the central axis of the CMS cylinder, colliding at the centre of the CMS detector. From Ref. [41].

quantity can be used:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2}. \tag{2.4}$$

For boosts along the z-axis, the ΔR calculated for massless particles is a Lorentz invariant observable.

In the transverse plane, transverse momentum (p_T) and transverse energy (E_T) represent momentum projections and magnitudes within this plane, respectively, while p_T^{miss} indicates the momentum imbalance measured in this plane. The focus on transverse quantities is primarily due to the nature of the collider experiments. When protons collide, the momentum along the beamline direction is often dominated by the initial states of the protons, leaving the physics of interest in the transverse plane. Moreover, focusing on transverse quantities simplifies calculations and analysis. For example, the transverse mass, which is a combination of the transverse energy and the magnitude of the transverse momentum, is an invariant under Lorentz transformations, making it a convenient quantity to work with.

2.2.1 Tracker

The inner tracking system of the CMS detector spans 5.8 meters in length and 2.5 meters in diameter. It requires high spatial granularity and rapid response to function effectively in the LHC environment. The tracker's responsibilities include precisely reconstructing the interaction vertex and secondary vertices, crucial for identifying the presence of b-hadrons. Furthermore, it must accurately separate the trajectories of charged particles and correlate them with the corresponding bunch crossings in a challenging environment.

As can be seen in Figure 2.2, LHC operations may involve an average of approximately 50 simultaneous interactions per bunch crossing, and the detector is exposed to a high particle flux, necessitating radiation hardness. The silicon material chosen for the tracker possesses all three required attributes. To minimise radiation damage, the system operates at -20 °C and maintains a constant temperature through an efficient cooling system. In fact, by operating the system at a lower temperature, the kinetic energy of particles is reduced, which can help mitigate the effects of radiation damage on the silicon material.

The CMS tracker, depicted in Figure 2.5, consists of a silicon pixel detector and a silicon strip detector. The entire tracker incorporates 66 million pixels, covering an area of approximately 1 square meter, and 9.3 million strips, with a total active silicon area of 198 square meters. The tracker provides coverage in pseudorapidity up to 2.5.

In the closest proximity to the interaction point, the tracker features three cylindrical layers of silicon pixel detectors situated at radii of 4.4, 7.3, and 10.2 cm. Two pixel modules are placed on each endcap adjacent to the cylinder. During the year-end technical stop of 2016/17, the pixel system underwent an upgrade, transitioning to four layers/three disks of low-mass silicon pixels [42]. In the upgraded detector, the innermost layer within the barrel is positioned closer to the origin at $r = 3$ cm. Additionally, a fourth layer has been added to the outermost layer in the barrel system, and an extra pixel disk is introduced at each endcap. These enhancements enhance track and secondary vertex reconstruction precision while reducing material content in the tracking region.

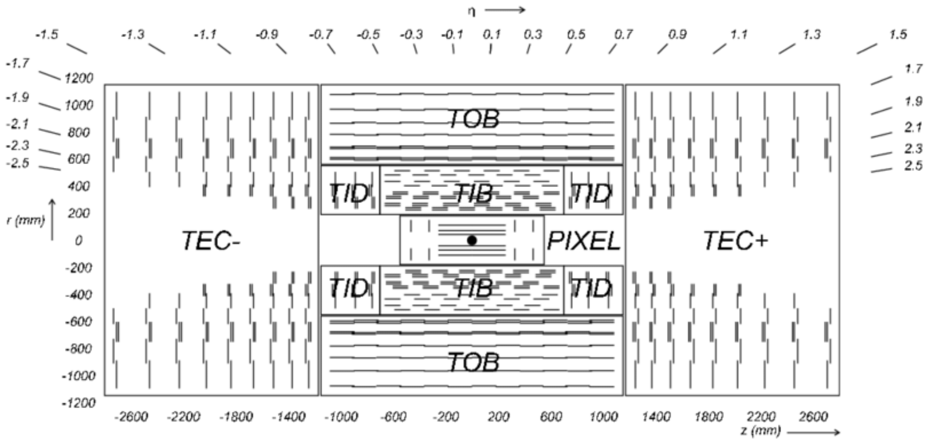


Figure 2.5: Schematic overview of the CMS tracking system before the pixel upgrade. The figure displays the different substructures: Tracker Inner Barrel and Disks (TIB/TID); the Tracker Outer Barrel (TOB); the Tracker EndCaps (TEC+ and TEC-). Source: Ref. [40].

The silicon strip tracker, positioned outside the pixel region between $20\,\mathrm{cm} < r < 116\,\mathrm{cm}$, is divided into three primary components: the Tracker Inner Barrel and Disks (TIB/TID), the Tracker Outer Barrel (TOB), and the Tracker EndCaps (TEC+ and TEC-, with the sign indicating the location along the z-axis). The TIB and TID cover the region up to $r = 55$ cm. The TIB comprises four barrel layers, while the TID incorporates three disks situated at each end of the barrel. The TOB consists of six layers, with the outermost layer at $r = 116$ cm. Finally, each TEC is composed of nine disks, expanding coverage in the z-direction up to 280 cm.

All barrel layers employ silicon strips, although with differing thicknesses and lengths corresponding to the distance from the interaction points. Inner layers feature finer strips with smaller dimensions, gradually reducing granularity in the outer layers as particle track densities decrease. This design maintains an average occupancy of approximately 2-3% per bunch crossing.

2.2.2 Electromagnetic Calorimeter (ECAL)

The Electromagnetic Calorimeter (ECAL) of CMS is designed to be both hermetic and homogeneous, ensuring excellent energy resolution. It consists of 61'200 lead tungstate ($PbWO_4$) crystals in the central barrel section (EB) and 7'324 crystals in each of the two endcaps (EE). Additionally, a preshower detector (ES) is placed in front of each EE. The crystals are read out using avalanche photodiodes (APDs) in the barrel and vacuum phototriodes (VPTs) in the endcaps. Similar to the tracker, the ECAL was engineered to meet the demands of fast response, fine spatial granularity, and radiation hardness.

$PbWO_4$ was chosen for its high density ($8.28\,\mathrm{g/cm^3}$), short radiation length[1] ($X_0 = 0.89$

---

[1] The radiation length (X_0) is defined as the mean distance over which the energy of an electron or

cm), and small Molière radius[2] (2.2 cm). These characteristics enable the construction of a compact calorimeter with fine granularity. Moreover, approximately 80% of the scintillation light is emitted within the first 25 ns, matching the time intervals between LHC bunch crossings. The light output is temperature-dependent, with relatively low emission (equivalent to 4.5 photoelectrons per MeV) at the operating temperature of 18°C. The scintillation light falls within the blue-green spectrum, peaking at 420–430 nm.

Figure 2.6 provides a schematic representation of the ECAL's various components. The EB covers the pseudorapidity region up to $|\eta| = 1.479$. It consists of 170 rings centred in the z-direction, with each ring fixed at a specific η position and housing 360 crystals that span the φ angle. The total volume occupied by these crystals is 8.14 m^3, with a combined weight of 67.4 tons. These crystals measure 230 mm in length (equivalent to 25.8 X_0) and have a tapered shape with front (rear) surfaces measuring $22 \times 22\,\mathrm{mm}^2$ ($26 \times 26\,\mathrm{mm}^2$). The crystals are arranged in a quasi-projective geometry, with their axes forming a 3° angle relative to the nominal interaction vertex. This configuration eliminates gaps between crystals aligned with particle trajectories. Crystals are grouped into alveolar structures known as submodules, which, in turn, form modules. Four modules are housed within each supermodule, resulting in a total of 36 supermodules, each covering a 20° segment in φ and containing 1'700 crystals. The EE extends from $|\eta| = 1.479$ to $|\eta| = 3$, with each endcap divided into two Dees where crystals are arranged in an $x - y$ grid. These crystals are 220 mm in length (equivalent to 24.7 X_0) and have front (rear) surfaces measuring $28.62 \times 28.62\,\mathrm{mm}^2$ ($30 \times 30\,\mathrm{mm}^2$). The preshower detector comprises two layers of silicon strip sensors (active elements) preceded by two layers of lead (absorbers). This sampling calorimeter is positioned before each endcap, covering the pseudorapidity range of $1.653 < |\eta| < 2.6$ and possessing a thickness of 20 cm. It serves to identify photons resulting from neutral pion decay in the endcaps and aids in the position measurements of electrons and photons.

The energy resolution of a calorimeter can be expressed as the quadratic sum of three components, namely the stochastic term (S), the noise term (N), and the constant term (C):

$$\frac{\sigma_E}{E} = \sqrt{\left(\frac{N}{E}\right)^2 + \left(\frac{S}{\sqrt{E}}\right)^2 + C^2}. \tag{2.5}$$

Values for these three terms were determined in a test beam with electrons prior to data collection [43], yielding $S = 2.8\%$, $N = 12\%$, and $C = 0.3\%$. This formula holds for energies below 500 GeV; beyond this threshold, the electromagnetic shower is not fully contained within the calorimeter. The stochastic term S accounts for fluctuations in the lateral containment of electromagnetic showers, variations in the number of photoelectrons emitted per GeV in APD or VPT, and the randomness of energy deposit in the preshower absorber. The noise term N encompasses all electronic, digitisation, and pileup noise contributions, including photodetector noise from leakage current. Lastly, the constant

---

photon decreases to about 1/e (approximately 36.8%) of its initial value.

[2] The Molière radius (R_M) is a parameter describing the lateral spread of a particle shower in a material with density ρ and radiation lenght X_0: $R_M \approx \frac{0.885}{\rho X_0}$.

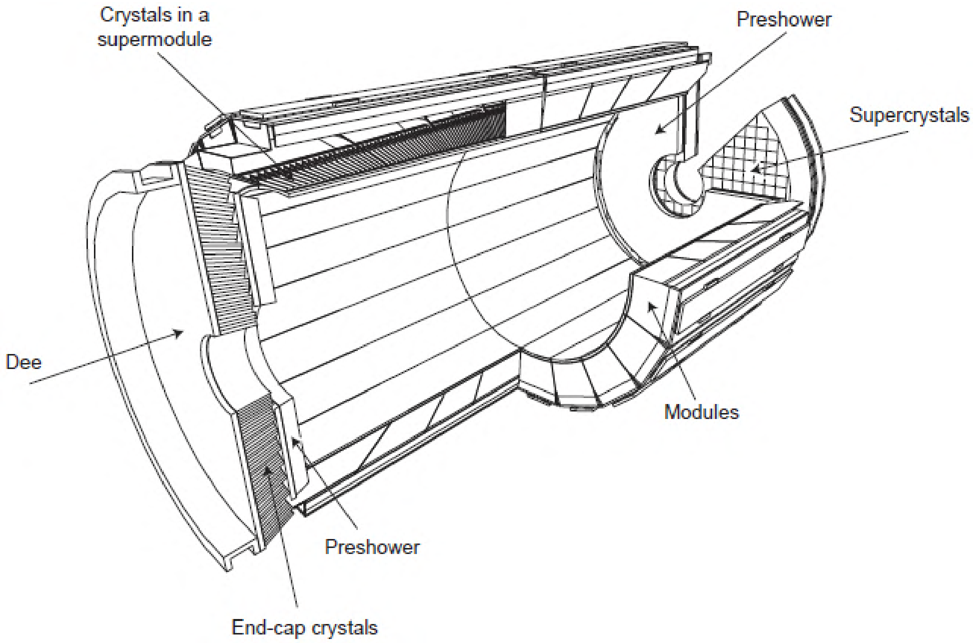


Figure 2.6: Layout of the ECAL, including the barrel, two endcaps, and the preshower. Source: Ref. [40].

term C results from non-uniform light collection along the longitudinal direction and imperfections in crystal intercalibration. A minor contribution comes from energy leakage from the back of the crystal. These effects are energy-independent, making the constant term the primary contributor to energy resolution at higher energies. Consequently, precise calibration to standardise crystal responses is crucial for achieving optimal ECAL performance.

Response Monitoring and Calibration of the ECAL

During LHC operation, the $PbWO_4$ crystals are exposed to radiation damage caused by both electromagnetic and hadronic interactions. Electromagnetic interactions lead to the formation of colour centres in the crystals, reducing crystal transparency. Fortunately, this electromagnetic damage is largely self-healed at the operating temperature of the ECAL, thanks to the spontaneous annealing of these centres. However, hadron-induced damage poses a more serious threat to the crystals than electromagnetic interactions. It results in a shift of the transmission band, causing significant losses in light transmission. Unlike electromagnetic damage, this type of damage is not recoverable over time and accumulates during data-taking.

To monitor changes in crystal response during LHC operation and prevent degradation of the ECAL resolution, a high-precision light monitoring system (LM) is employed. This system uses laser light with a wavelength of 447 nm, which is injected every 40 minutes at a fixed position on the front (or rear) face of each crystal and collected by the APD (VPT) on the crystal's rear face in the EB (EE). Laser pulses are controlled by a PN photodiode, typically associated with a set of 100-200 crystals known as a harness. This setup ensures that the measured response remains stable despite fluctuations in the injected light. The relative response of a crystal to laser light is calculated as the amplitude ($\mathcal{A}$) of the photodetector signal normalised to the signal measured by the PN diode, as follows:

$$R_{\mathrm{ch}} = \frac{\mathcal{A}_{\mathrm{APD}}}{\mathcal{A}_{\mathrm{PN}}}. \tag{2.6}$$

These changes are most pronounced for crystals near the beam line that receive higher radiation doses. In particular, response variations of up to 10% are observed in the barrel, increasing to 62% at the tracker's limit ($|\eta| < 2.5$) and reaching 96% in the region closest to the beam pipe.

The LM system continuously monitors crystal response variations every 40 minutes to correct for transparency losses. Laser corrections are determined for each crystal over time using the formula:

$$LC_{\mathrm{ch}}(t) = \Big(\frac{R_{\mathrm{ch}}(0)}{R_{\mathrm{ch}}(t)}\Big)^{\alpha}, \tag{2.7}$$

where $R_{\mathrm{ch}}(t)$ and $R_{\mathrm{ch}}(0)$ represent the crystal's response to laser light at time t and the start of each year's data-taking ($t = 0$), respectively. The parameter α accounts for different paths of laser and scintillation light and has an average value of about 1.5 in EB and 1 in EE. These values, determined in a test beam before data-taking and refined during it to consider transparency losses, ensure precise calibration of crystal responses. A residual drift in the energy scale over time, observed in the barrel, is attributed to radiation-induced damage in PN photodiodes. An effective correction for Run 2 data was derived using the E/p method to address this residual drift.

Following the application of time-dependent calibration, intercalibration is performed to equalise crystal responses at the same pseudorapidity coordinate. Intercalibration methods include the π^0, E/p, Zee, and φ-symmetry methods, each offering complementary benefits based on different data samples and techniques. These methods are combined as a weighted sum, with weights determined by the relative precision of each method.

Finally, the absolute energy scale is derived and applied as a function of η to equalise crystal responses in different η regions. This scale is determined using events from Z boson decays into electrons, ensuring agreement between observed invariant masses for electrons in data and MC simulations across various η-rings.

2.2.3 Hadronic Calorimeter (HCAL)

The CMS Hadron Calorimeter (HCAL) was designed with high granularity, allowing it to not only measure the energy of jets but also determine their direction. Similar to the ECAL, the HCAL is a hermetic calorimeter designed to contain hadronic showers effectively. This containment capability is vital for estimating missing transverse momentum

p_T^{miss}, associated with weakly interacting particles like neutrinos or potential exotic particles.

The HCAL system consists of four subdetectors:

- the Hadron Calorimeter Barrel (HB) and Endcaps (HE) are placed inside the solenoid magnet, covering pseudorapidity regions of $|\eta| < 1.3$ and $1.3 < |\eta| < 3$, respectively;
- the Outer Hadron Calorimeter (HO) extends to $|\eta| < 1.3$, covering the same region as HB;
- the Forward Hadron Calorimeter (HF) provides rapidity coverage up to $|\eta| = 5.2$.

The HB is a sampling calorimeter made of brass absorber and plastic scintillator, read out using wavelength shifters. Due to space limitations between the ECAL and the solenoid magnet coil, the HB has a limited amount of absorber material. Therefore, an additional subdetector, the HO, is added outside the magnet coil to ensure adequate shower containment. The HO also uses plastic scintillator as the active material.

The HE is a sampling calorimeter placed in the endcap iron yokes, consisting of alternating layers of brass absorber and plastic scintillator, similar to the HB.

The HF, covering the high pseudorapidity region ($3 < |\eta| < 5.2$), is a Cherenkov calorimeter with quartz fibers as the active material and steel as the absorber material. Its radiation-hard design allows for the identification of forward jets, a crucial signature for processes such as VBS and Vector Boson Fusion (VBF). Additionally, the HF contributes to more reliable p_T^{miss} measurements.

During Run 2, the HCAL readout system, including photodetectors and electronics, underwent upgrades to enhance detector performance [44].

2.2.4 Muon Chambers

The CMS muon system serves three main purposes: muon identification, momentum measurement, and triggering. It is designed to deliver precise and robust muon measurements across a wide range of LHC energies and angles. The system takes advantage of the high magnetic field and uses magnet return yokes as additional hadron absorbers.

Muon momentum resolution achieved solely by the muon system is approximately 9% for low values of η and transverse momenta up to 200 GeV [45]. However, multiple scattering of muons before reaching the muon system due to budget material present before the muon station impacts resolution negatively. For 1 TeV muons, the resolution ranges from 15% to 40%, depending on η. Combining inner tracker information significantly improves performance, resulting in an order of magnitude improvement at low p_T and reaching 5% resolution at 1 TeV in the barrel region. The muon momentum resolution is further optimized through an alignment system that measures muon subsystem positions relative to other CMS detector components.

The muon system consists of three types of gaseous particle detectors arranged cylindrically in the barrel and endcaps. The detectors include:

- Drift chambers (DT) with rectangular drift cells, covering the pseudorapidity region $|\eta| < 1.2$ with a uniform magnetic field, low muon rates, and minimal neutron background.

- Cathode strip chambers (CSC) grouped in four stations within each endcap, covering the region $0.9 < |\eta| < 2.4$ with a non-uniform magnetic field. CSCs offer a fast response time, fine granularity, and radiation hardness, crucial for high muon rates and radiation doses.

- Resistive plate chambers (RPC) located up to $|\eta| < 1.6$ in both the barrel and endcaps. RPCs provide an independent trigger system with fast response and good time resolution. Redundancy is employed to reduce background contamination, enhancing time and p_T resolution.

2.2.5 Trigger

The CMS trigger system serves to reduce the extremely high event rate produced by LHC bunch crossings to a manageable rate for data storage. This reduction occurs in two stages, ultimately retaining only the most interesting events. The first stage, known as the Level-1 (L1) [46] trigger, reduces the rate to 100 kHz, and the subsequent stage, the High-Level Trigger (HLT) [47], further narrows it down to 1 kHz. A software component, the Trigger Supervisor, controls the configuration and operation of these trigger components.

The L1 trigger is a hardware system that analyses every bunch crossing with a latency of approximately 4 μs. During this time, it assesses raw data from muon and calorimeter subsystems to determine whether an event should be passed on to the HLT trigger. The L1 trigger employs a hierarchical structure, starting with Local Triggers (or Trigger Primitive Generators) that analyse energy deposits in calorimeter trigger towers, track segments, and hit patterns in muon chambers. Data from the calorimeters are aggregated by Regional Triggers (RTC) to create trigger objects such as electrons, photons, or jets, which are then passed to the global calorimeter trigger (GCT). Similarly, muon chamber hits are processed through pattern comparators, segment finders, and track finders to identify muon candidates, which are transferred to the global muon trigger (GMT). The GMT and GCT combine information to make a final decision about event rejection, which is communicated to the subdetectors through the Timing, Trigger, and Control system.

Accepted events proceed to the HLT trigger, where further selection criteria are applied to decide whether an event should be permanently stored. The HLT consists of a large farm of thousands of computers performing calculations to reconstruct and identify objects from the full dataset acquired by all subdetectors. Specific HLT paths are defined based on data analysis requirements, and they filter and select interesting events. Special HLT paths are also defined for data calibration and monitoring purposes. Events accepted by an HLT path are archived for further analysis.

2.3 Particle Flow Reconstruction Algorithm

In this section, an overview of the Particle Flow reconstruction algorithm (PF) [48] is provided. This algorithm is employed for the precise identification and reconstruction of final state particles within the CMS environment. The CMS detector is equipped with subdetectors (described in Section 2.2) optimised for the PF algorithm's performance. The PF algorithm uses data from these subdetectors to reconstruct various physical objects, including jets, muons, electrons, and photons. Additionally, it calculates missing transverse momentum p_T^{miss} and identifies jets containing b-quarks. This algorithm remains effective even in scenarios with multiple simultaneous interactions, as it can identify and subtract particles from pileup vertices.

The identification and reconstruction of physical objects begin with PF elements, which consist of tracks in the tracker and muon station, as well as clusters of energy in the calorimeters. Then, the PF algorithm proceeds by connecting these fundamental elements to obtain PF blocks. The identification and reconstruction of different objects start from these PF blocks following a specific order.

Muons are identified and reconstructed by connecting their tracks in the muon station with those in the inner tracker, as detailed in Section 2.3.1. Subsequently, the particles corresponding to these muon candidates are removed from the PF blocks. For electron reconstruction, energy deposits in the ECAL are matched with tracks in the silicon detector, as explained in Section 2.3.2. Any associated tracks and energy clusters linked to electrons are then filtered out from the event. In the same processing step, isolated photons are also identified. Additionally, in accordance with Section 2.3.3, neutral and charged hadrons are reconstructed by combining information from HCAL, ECAL, and the tracker, which are later merged to create jets. Finally, the total energy imbalance is calculated as described in Section 2.3.3. To illustrate the concept of the PF algorithm, refer to Figure 2.7.

2.3.1 Muons

Muons are classified into three distinct categories based on the methodology employed in their reconstruction [49]: standalone muons, tracker muons, and global muons. The standalone muons are reconstructed solely from hits in the muon stations, and their reconstruction involves a track finder that uses the Kalman filter technique [50]. However, standalone muons often suffer from contamination by cosmic muons and typically exhibit poorer momentum resolution compared to global or tracker muons. Consequently, they are generally not used at the analysis level.

The tracker muons category encompasses candidates with a silicon tracker track that matches at least one segment in the DT or CSC muon stations. Tracker muons are particularly efficient at reconstructing muons with low transverse momentum (p_T) because such muons often scatter from material before reaching the muon station. However, their purity can be compromised by the presence of hadron background, as some hadron showers may penetrate the innermost layer of the muon station.

Finally, global muons represent a combination of both tracker and muon station infor-

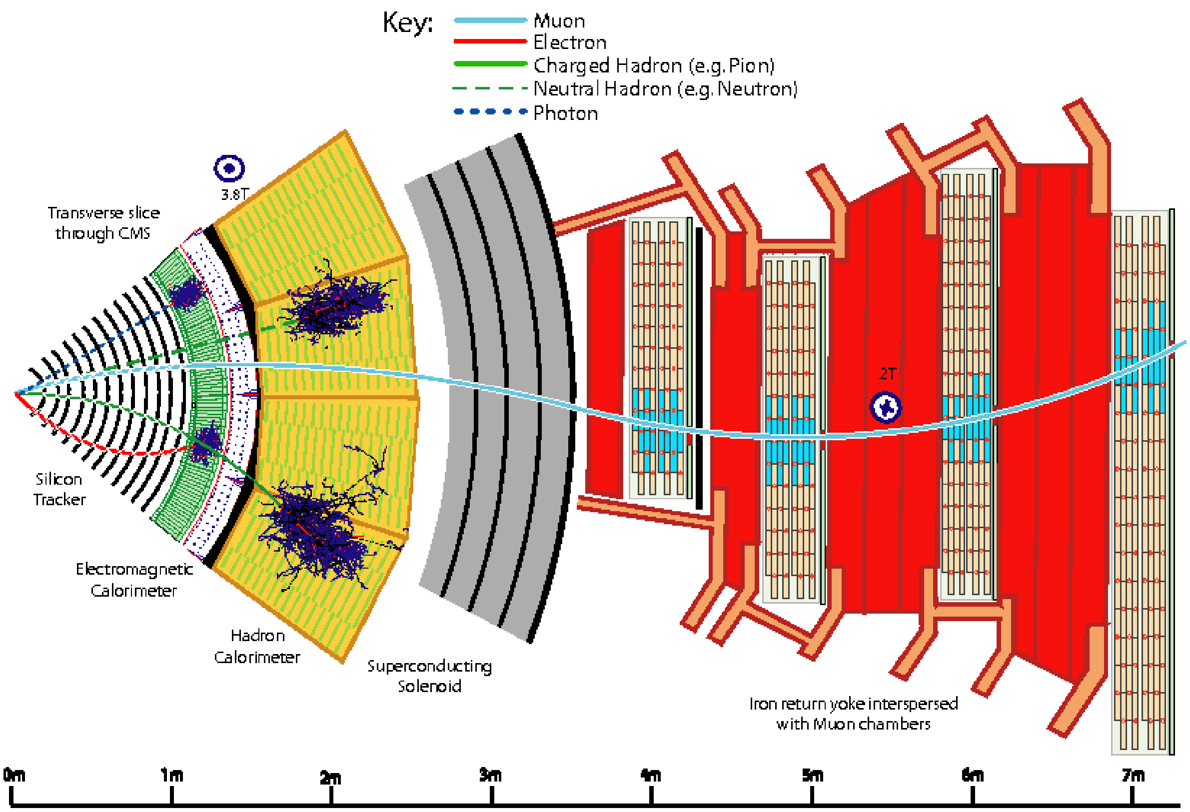


Figure 2.7: A schematic representation of particle interactions within a transverse slice of the CMS detector, extending from the beam interaction region to the muon detector. From Ref. [48].

mation. These muons are typically associated with high p_T values and exhibit a low misidentification rate. The integration of information from both subsystems also yields more accurate momentum measurements. In particular, global muons offer a more precise measurement of p_T than tracker muons for p_T values exceeding 200 GeV.

In summary, the overall efficiency in reconstructing muons reaches an impressive 99% within the geometrical acceptance of the CMS detector.

2.3.2 Electrons

Electron reconstruction begins by identifying clusters of energy in the ECAL that are associated with tracks detected in the silicon tracker. However, as electrons pass through the tracker's material, they can emit photons due to bremsstrahlung. These emitted photons tend to be concentrated in the φ direction, with minimal dispersion in η.

To reconstruct electrons accurately, the process starts with identifying individual crystals in the ECAL that record energy above a certain threshold. Nearby crystals that meet specific criteria are then included in the reconstruction, expanding both in the φ and η directions for the EB or in the (x, y) plane for the EE. These initially identified clusters are eventually merged to form what are known as superclusters. For more detailed information on this process, see Ref. [51].

The tracker employs two different approaches depending on the level of bremsstrahlung emission associated with the electron. For cases with low bremsstrahlung, a standard Kalman filter is employed, which is the standard approach for all charged particles. However, in situations where bremsstrahlung leads to substantial energy loss and trajectory alterations, the Gaussian Sum Filter (GSF) [52] proves more effective. The GSF has the ability to recover hits that have deviated due to changes in the electron's path.

Electrons can be reconstructed using two distinct approaches: one starting from an ECAL supercluster and the other from a GSF track. In the ECAL-based approach, the position of the ECAL supercluster is extrapolated back to the tracker to check for a matching track. A critical energy-based selection criterion is applied to differentiate electrons from hadrons. Specifically, the ratio of the supercluster's energy to the corresponding energy deposited in the HCAL must be less than 0.15. This helps avoid confusion with hadrons, which can deposit energy in the ECAL. The HCAL energy calculation includes contributions from all the towers within a cone of $\Delta R = 0.15$ around the electron's direction. This approach is particularly suited for high-energy electrons. In the tracker-based approach, the GSF track is extrapolated to match an ECAL supercluster. This method is chosen to ensure high efficiency and accuracy in reconstructing both low p_T electrons and non-isolated electrons. The momentum resolution for reconstructed electrons can vary from 1.7% to 4.5%. Meanwhile, the efficiency of the reconstruction process varies depending on the electron's position in η and the extent of bremsstrahlung emission, ranging from 88% to 98% in the barrel and from 90% to 96% in the endcaps for electrons with p_T in the range of 10 to 100 GeV [51].

2.3.3 Jets and Missing Transverse Momentum

Following the reconstruction of muons, electrons, and photons, their associated tracks and clusters are removed from the event data. What remains, in terms of tracks and clusters, is attributed to two fundamental components: hadrons (both charged and neutral) and non-isolated photons, forming the building blocks for the formation of jets.

Photons and hadrons are reconstructed based on information from the ECAL and the HCAL. Within the tracker's coverage region ($|\eta| < 2.5$), HCAL clusters that can be linked to a track are identified as charged hadrons. ECAL clusters without corresponding tracks in the tracker are classified as photons. HCAL clusters with no association to any track are attributed to neutral hadrons. In regions beyond the tracker's coverage ($|\eta| > 2.5$), ECAL clusters that can be matched with an HCAL cluster are associated with hadrons (either neutral or charged), while ECAL clusters alone (without a link to HCAL clusters) are identified as photons.

Jets are then constructed from these PF candidates using the anti-k_T algorithm [53, 54] with a specific distance parameter ($R = 0.4$, often referred to as AK4 jets). The p_T of an AK4 jet is calculated as the vectorial sum of the p_T of all the PF candidates contained within it.

To account for various effects, such as pileup and differences in the response of the detector, jet energy scale corrections (JEC) are applied. These corrections are determined as functions of both the pseudorapidity and the transverse momentum of the jets. They play a critical role in ensuring the accuracy of jet energy measurements. The achieved resolution in measuring jet energy depends on the jet's energy and its position in pseudorapidity. In the central region of the CMS barrel, for instance, jet energy resolutions are typically around 15-20% at 30 GeV, approximately 10% at 100 GeV, and as low as 5% at 1 TeV [55].

Once all the PF candidates have been reconstructed, the negative sum of their transverse momenta (p_T) is computed. This quantity is referred to as the missing transverse momentum (p_T^{miss}) [56] and is used to account for particles that are undetectable by the detector, such as neutrinos. Several factors influence the measurement of p_T^{miss}, including experimental resolutions, misreconstructions of particles, non-hermeticity of the detector, and the presence of additional non-primary vertices. Extensive studies, conducted on both real data and simulated samples, have been undertaken [56] to enhance the efficiency and precision of p_T^{miss} measurements. This precise measurement of transverse momentum is particularly vital for the analysis conducted in this thesis, as it involves the leptonic final state of W bosons.

2.4 The High Luminosity Large Hadron Collider

To unlock the full discovery potential of the LHC, the High-Luminosity LHC (HL-LHC) [57] upgrade is dedicated to enhancing the original design's performance, quantified through integrated luminosity, by a factor of ten. The upgrade is currently underway, and physics

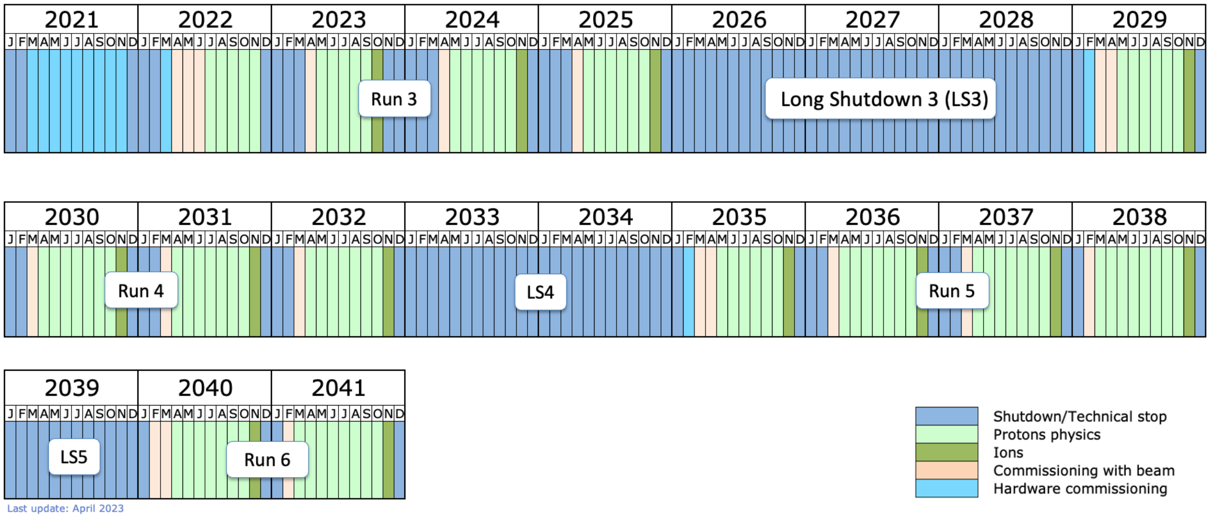


Figure 2.8: Long-term operational timeline of the LHC, with Run 4, Run 5, and Run 6 representing the HL-LHC phase.

	Peak lumi. ($\times 10^{34}$ cm^{-2} s^{-1})	Peak pile-up	Lumi/year (fb^{-1})
LHC 2017	1.7	50	40
HL-LHC	7.2	140	250

Table 2.1: Projected HL-LHC operating parameters, with 2017 reference values. This table includes values on the peak instantaneous luminosity, peak pile-up (indicating the number of concurrent interactions per LHC bunch crossing), and the total integrated luminosity expected per year. From Ref. [58].

experiments are expected to start data collection no earlier than 2029. The operational timeline for both the LHC and the HL-LHC spanning two decades, from 2021 to 2041, is visually represented in Figure 2.8.

The HL-LHC project aims to provide proton-proton collisions at 14 TeV, allowing an integrated luminosity of 250 fb^{-1} per year, with the ultimate goal of reaching 3000 fb^{-1} approximately twelve years after the upgrade [58]. The HL-LHC projected operating parameters are listed alongside the parameters of the LHC in 2017 in Table 2.1.

2.4.1 Present Luminosity Limitations and Hardware Constraints

The instantaneous luminosity $\mathcal{L}$, already defined in Equation (2.1), can be expressed as [57]:

$$\mathcal{L} = \gamma \frac{n_\mathrm{b} N^2 f_\mathrm{rev}}{4\pi\beta^* \epsilon_\mathrm{n}} R, \tag{2.8}$$

where R represents the luminosity geometrical reduction factor. This factor accounts for geometric and focusing effects on the collision rate and is defined as:

$$R = \left(\sqrt{1 + \frac{\theta_c \sigma_z}{2\sigma}}\right)^{-1}. \tag{2.9}$$

Here is a breakdown of the parameters and some of their nominal values for both the LHC and HL-LHC:

γ: Proton beam energy in units of rest mass.

n_b: Number of bunches per beam.

N: Bunch population.

f_rev: Revolution frequency (11.2 kHz).

β^*: Beam beta function at the collision point. Describes beam focusing.

ϵ_n: Transverse normalised emittance. Measures beam compactness.

θ_c: Full crossing angle between beams. Controls collision angle.

σ: Transverse r.m.s. size (16.7 μm).

σ_z: Longitudinal r.m.s. size (7.55 cm).

It should be noted that the nominal values enclosed in brackets remain the same for both the LHC and HL-LHC. For differences in other parameters between the LHC and HL-LHC, refer to Table 2.2, which provides a comprehensive overview of the primary parameters essential for high luminosity operations at the HL-LHC.

Before discussing the new changes for HL-LHC, it is crucial to identify the systems that may need improvements due to potential breakdowns and accelerated aging or if they could become bottlenecks in higher radiation environments:

- Inner Triplet Magnets: Some components of the inner triplet quadrupoles and corrector magnets could experience radiation damage, possibly leading to sudden electrical breakdowns.

- Cryogenics: A new cryogenic facility is being built to cool the superconducting radio freqeuncy and magnets separately. This separation will avoid the need to warm up the entire circular section of the accelerator when maintenance is required in the triplet region, making the accelerator more flexible and available for ongoing operations and upgrades.

- Collimation: The current collimation system requires an upgrade to accommodate higher beam intensities and protect the new triplets in specific areas.

Parameter	Nominal LHC	HL-LHC 25 ns	HL-LHC 50 ns
Beam energy in collision [TeV]	7	7	7
Bunch population N	1.5×10^{11}	2.2×10^{11}	3.5×10^{11}
Number of bunches per beam n_b	2808	2748	1404
Number of collisions in ATLAS (IP1) and CMS (IP5)	2808	2736	1404
Crossing angle [μrad]	285	590	590
β^* [m]	0.55	0.15	0.15
ϵ_n [μm]	3.75	2.50	3.00
r.m.s. longitudinal size	7.55×10^{-2}	7.55×10^{-2}	7.55×10^{-2}
Peak luminosity [cm^{-2} s^{-2}]	1.70×10^{34}	7.18×10^{34}	8.44×10^{34}
Levelled luminosity [cm^{-2} s^{-2}]	-	5.00×10^{34}	2.50×10^{34}

Table 2.2: High Luminosity LHC Parameters. The baseline operation mode for the HL-LHC uses a 25 ns bunch spacing. However, a contingency plan includes the option of a 50 ns bunch spacing in case unexpected effects impact the performance at 25 ns. From Ref. [57].

- Dispersion Suppressor (DS) Regions: These areas need attention due to the potential issue of off-momentum particle leakage into the main superconducting dipoles, which can significantly limit LHC performance. "Leakage of off-momentum particles" refers to the unintended presence of particles that don't have the desired or optimal momentum or energy, that have deviated from the ideal trajectory as they circulate through the accelerator. The most promising approach involves replacing an LHC main dipole with with shorter, higher-field dipoles (11 T, 11 m) of equal bending strength (120 T·m). This change enables the installation of specialised collimators, mitigating the issue.

- Radiation to Electronics (R2E) and Superconducting Links: Efforts are underway to replace radiation-sensitive electronics boards with radiation-hard cards, designed and manufactured to be more resilient to the damaging effects of ionising radiation. Additionally, some power converters may be relocated to the surface to improve LHC availability using superconducting links.

- Quench Protection System (QPS), Machine Protection, and Remote Manipulation: Systems like QPS for superconducting magnets, machine protection, and remote manipulation need upgrades to ensure safety and efficiency, especially with the increased performance levels.

These improvements are essential for maintaining the reliability and efficiency of the accelerator, especially in the context of higher radiation environments.

2.4.2 Luminosity Levelling

The operation of the HL-LHC encounters limitations due to two primary factors. Firstly, there is the issue of energy deposit resulting from collision debris within the magnet region. Secondly, there is the imperative to control peak pile-up events in the detector. To tackle these challenges, the HL-LHC employs a strategic approach called "luminosity levelling" [57]. Instead of maintaining a constant peak luminosity, the collider operates at a consistent luminosity level, intentionally kept below its maximum potential. This approach effectively mitigates "luminosity burn", a phenomenon where protons are consumed during collisions.

To maximise integrated luminosity under the levelled luminosity constraint, the key is to maximise fill length, namely the period when the accelerator is actively used for experiments. Achieving this involves maximising the injected beam current. Additionally, several other factors play crucial roles in reaching the required 3 fb^{-1}/day:

- Minimising the average machine turnaround time, namely the time it takes to complete a cycle of operations, from the end of one experimental run to the beginning of the next.

- Ensuring the average operational fill length exceeds the luminosity levelling time, namely the time required to adjust and stabilise the luminosity.

- Maintaining good overall machine efficiency, which accounts for the available physics time (portion of time when the accelerator is operational and can be used for conducting experiments and collecting data) after accounting for downtime for fault recovery.

Closely related to machine efficiency is physics efficiency, which is a critical factor for maximising integrated luminosity. Physics efficiency represents the fraction of time per year dedicated to delivering actual collisions for experiments. In the context of achieving the desired 3 fb^{-1}/day luminosity, efficiency is just as crucial as optimising the accelerator's virtual peak performance. It ensures that the available operational time is effectively used for conducting experiments and collecting data, aligning with the goal of luminosity levelling to enhance the overall performance of the HL-LHC.

2.4.3 HL-LHC Parameters and Key Upgrade Systems

A common approach for achieving a luminosity upgrade involves reducing β^* through the implementation of stronger and larger aperture low-β^* triplet quadrupoles. However, as β^* is reduced to achieve tighter beam focusing, it has a cascading effect on the accelerator design. The reduction in β^* necessitates an increase in the crossing angle. This is because, with smaller beam sizes at the interaction point (IP), there is a risk of beam collisions outside the beam pipe, which could lead to beam loss and potential damage to accelerator components. To mitigate this risk, it becomes necessary to increase the crossing angle, ensuring stable and safe beam operation while maximising luminosity. To accommodate the increased crossing angle, even larger aperture triplet magnets, an expanded aperture for the first separation dipole[3], and additional adjustments to the matching section are required. Additionally, the increased crossing angle reduces the size of the luminous region, which is the area where particle collisions occur, and, as a result, may reduce the potential gain in peak luminosity.

In the standard LHC setup, the practical limit for β^* is around 30-40 cm, compared to the nominal value of 55 cm. However, a novel approach called the Achromatic Telescopic Squeeze (ATS) scheme greatly improves beam focusing and control, allowing for a remarkably low β^* value of 15 cm. This innovation significantly boosts luminosity while maintaining beam stability. To achieve this reduced β^*, the triplet quadrupoles must double their aperture, requiring a 50% stronger magnetic field than the current LHC. This necessitates the use of advanced superconducting technology based on Nb_3Sn.

One drawback of significantly reducing β^* is the requirement for a larger crossing angle, which leads to a reduction in the geometrical luminosity reduction factor R. When the beam is more strongly focused with a β^* of 0.55 m, R is 0.85, but as the focusing changes to a β^* of 0.25 m, R decreases to 0.5. To counteract this effect, various methods can be employed. The most efficient and elegant solution for compensating the lower R involves the use of special superconducting radio-frequency crab cavities. These cavities

---

[3] The first magnetic dipole is encountered by particles after they exit the IP. Its role is to bend the paths of particles in such a way that they are separated into different beamlines.

can generate transverse electric fields that rotate each bunch longitudinally, aligning them in such a way that they effectively collide head-on and overlap perfectly at the collision points. Crab cavities unlock the full potential of the small β^* values offered by the ATS scheme and larger triplet quadrupole magnets.

2.4.4 ECAL Upgrade

Upgrades are underway for the ECAL (Section 2.2.2) to ensure its performance aligns with the luminosity objectives of the HL-LHC. These enhancements comprise two crucial aspects: the replacement of front-end and off-detector electronics within the EB, and the complete replacement of the EE. The primary motivation for the EB upgrade in Phase II lies in stringent trigger requirements. These necessitate achieving a latency of 12.5 μs, a significant increase from the current 4 μs, and accommodating a L1 trigger rate of approximately 750 kHz, compared to the current 100 kHz. Meeting these criteria entails replacing front-end cards in the barrel and off-detector electronics. For a more detailed description of the ECAL upgrade refer to Ref. [59].

In Phase II, CMS plans to leverage per-particle time information with a remarkable precision of 30 ps. This innovation promises to enhance event reconstruction and reinforce resilience against pile-up. Simulation studies have shown that incorporating timing information for electrons, photons, and charged hadrons significantly broadens the potential for precise measurements within the Higgs sector and extends the search for new physics BSM.

Crystals and Photoreadout

Studies have been conducted to assess the impact of radiation on lead tungstate crystals.The enduring reduction in light transmission due to hadron irradiation is the primary concern for utilising the calorimeter at the HL-LHC. Previous test beam campaigns have generated data for characterising and forecasting how the crystals will evolve under the anticipated aging conditions at the HL-LHC. An increase in the constant term of energy resolution, reaching around 1.5% at high η values in the EB, is anticipated by the end of LHC Phase II. This aligns with the requirements of physics and, consequently, the EB crystals will be kept.

Throughout LHC Phase II, APDs will remain in operation. However, a significant increase in their dark current due to silicon damage from hadron irradiation is expected. To mitigate this effect, the operational temperature of the EB will be lowered from the Phase I nominal temperature of 18°C to 9°C. Projections suggest that at 9°C, the dark current for APDs will reach approximately 50 μA (equivalent to 250 MeV energy-equivalent noise) for regions with $\eta = 0$ and about 100 μA at $\eta = 1.45$ after 3000 fb^{-1} of operation. Figure 2.9 illustrates the anticipated evolution of the dark current at these two operating temperatures. Additionally, there is consideration of further lowering the temperature to 6°C after approximately 1600 fb^{-1} of operation, although this scenario is still under evaluation.

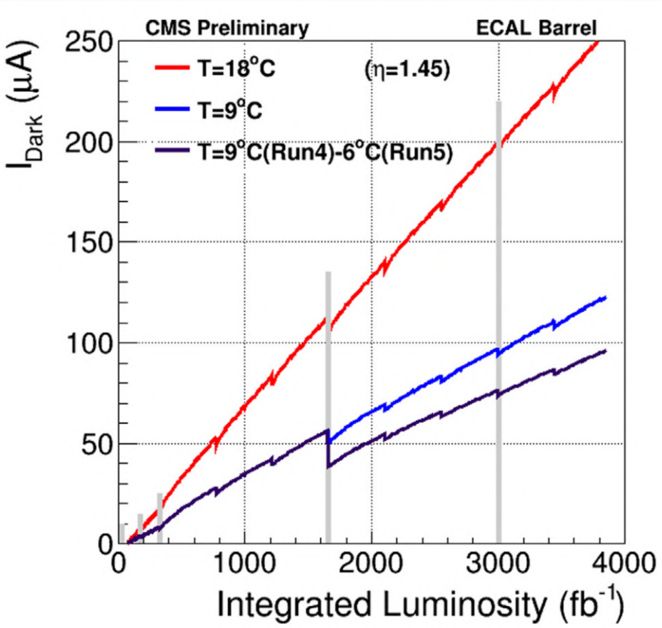


Figure 2.9: Anticipated evolution of APD dark current for APDs at $|\eta| = 1.45$ operated at 18°C (red line), at 9°C (blue line), and in a mixed scenario including a further step to 6°C after the LHC Long Shutdown 4. Source: Ref. [59].

Electronics Upgrade

In the electronics upgrade, modifications will be made to the front-end components and cards while preserving the mechanical structure of the cooling system and the motherboards. These motherboards are positioned beneath the cooling structure and serve to distribute low voltage to the electronics, provide bias voltage to the photodetectors, and route APD signals to the very front-end cards.

The very front-end card has been redesigned to ensure low noise, a shorter pulse shaping duration, and a higher sampling rate, as deeply explained in Section 3.3. This is accomplished using trans-impedance amplifiers (TIA), enabling the generation of a voltage representation of the photocurrent at the APD output. Two gain settings, × 1 and × 10, will be employed to maintain signal sensitivity up to 2 TeV. These TIA signals will be digitised at 160 MHz and will be transmitted to the front-end cards through a LiTE-DTU unit. This transmission will incorporate a lossless data compression algorithm to reduce bandwidth requirements.

The front-end card will use new data transmission technologies, including lpGBT and Versatile Link plus, to transmit single-crystal data sampled at 160 MHz (a significant improvement from the current 40 MHz) to the back-end electronics system.

This data will be ready for use at the Level-1 trigger stage, eliminating the need for on-detector data selection and buffering. This upgrade will substantially enhance the Level-1 trigger's granularity by a factor of 25, offering improved precision and aiding in the rejection of unwanted signals, including those resulting from direct ionisation in the APDs (referred to as “spikes”).

ECAL Signal Reconstruction

In this chapter, a closer look is taken at how signals are reconstructed in the ECAL. When particles interact with the ECAL, they trigger a chain of events leading to the measurement of their energy. This energy undergoes a sophisticated reconstruction process, which will be elaborated on in Section 3.1. A crucial component of the ECAL signal reconstruction is the Multifit algorithm [60], an iterative algorithm that reconstructs the amplitude using templates directly derived from experimental data. Section 3.1.1 provides an in-depth explanation of the Multifit algorithm. Alternatively, the Weights method [61] offers a distinct approach to estimate signal amplitudes by calculating them through weighted combinations of time samples. Section 3.1.2 will provide a detailed explanation of the Weights method. The choice between the Multifit and the Weights method depends on specific experimental needs, as will be detailed in Section 3.1.3.

In this project, the primary focus was on developing and refining the Weights method for Phase II, addressing both amplitude and timing jitter reconstruction. The timing jitter is defined as the nominal time of maximum and the actual time of maximum. This method involved the derivation of optimised weights, designed to accommodate the Phase II pulse shapes. An in-depth exploration of the mathematical procedures used to extract optimised weights, a critical aspect of this method's success, and a comprehensive understanding of the algorithm's inner workings will be provided in Section 3.2 for amplitude reconstruction and Section 3.3 for timing jitter reconstruction. To validate the method's reliability and effectiveness, closure tests were conducted, and the results and insights thereof are also covered in the respective sections.

The modules developed for this project have been successfully integrated into the official CMS software. This achievement marks a significant milestone in the practical application of the Weights method within the CMS framework for the high-luminosity phase.

3.1 ECAL Energy Reconstruction

The ECAL, thoroughly described in Section 2.2.2, assumes a pivotal role in the reconstruction of photons, electrons, and jets. Beyond its crucial role in measuring the electromagnetic component of particle showers, precision in these measurements is essential.

Hence, ensuring precise energy reconstruction in the ECAL is of utmost importance to maintain the reliability of analyses conducted within the CMS Collaboration.

When particles, such as photons or electrons, interact with the lead tungstate ECAL crystals, different physical processes take place. Initially, the incident particles may undergo interactions like pair production and bremsstrahlung within the crystal's structure. As these interactions unfold, they give rise to what is known as an electromagnetic shower. An electromagnetic shower is a cascade of secondary particles, including electrons and positrons, which are generated as a result of the initial interactions. These secondary particles continue to scatter and interact within the crystal, further producing additional photons and other secondary particles in the process. The energy from these secondary particles leads to the excitation of the lead tungstate crystal lattice. This excited state is temporary, and as the crystal returns to its ground state, it emits scintillation photons. These scintillation photons carry a portion of the energy deposited by the incident particles and can be collected by photodetectors for further analysis. The scintillation photons are subsequently transformed into an analog electrical signal, a crucial step in the overall detection process. This signal is then subject to amplification, shaping, and digitisation through a multigain preamplifier (MGPA) with three different ADC gains (×1, ×6, and ×12). The choice of the gain is performed dynamically; the highest gain, in which the pulse is not saturated, is automatically chosen. In this way, the energy range covered for a single channel spans from 35 MeV up to 1.7 (2.8) TeV in the EB (EE). The reconstructions algorithm then comes into play to extract pulses information like the amplitude or the timing jitter. In the following paragraphs, two reconstructions algorithms will be explained in details, along with their pros and cons, and a final operational comparison between them.

3.1.1 Multifit Algorithm

The Multifit algorithm [60] plays a crucial role in accurately reconstructing the amplitudes of signals received by ECAL crystals. Its primary goal is to handle the issue of signal contamination arising from different bunch crossings within the data acquisition window, causing an overlap of Out-Of-Time (OOT) signals with the In-Time (IT) signals. IT signals are the ones originating from the current bunch crossing and are the ones of primary interest. OOT signals, on the other hand, come from previous or subsequent bunch crossings and can interfere with the accurate measurement of the IT signal, thus degrading the energy resolution of ECAL.

To address this challenge, the Multifit algorithm employs a χ^2 minimisation technique. This technique is used to simultaneously calculate the amplitude of up to ten consecutive signal templates, that represent the expected signal shapes, each shifted in time by 25 ns – the typical time interval between bunch crossings. The templates used in the Multifit algorithm are directly measured from the experimental data. For each ECAL crystal, templates are obtained using events that involve the collisions of isolated proton bunches, characterised by a larger time spacing than the typical 25 ns between bunch crossings. In collisions involving two isolated bunches, there is no OOT pile-up contribution, making

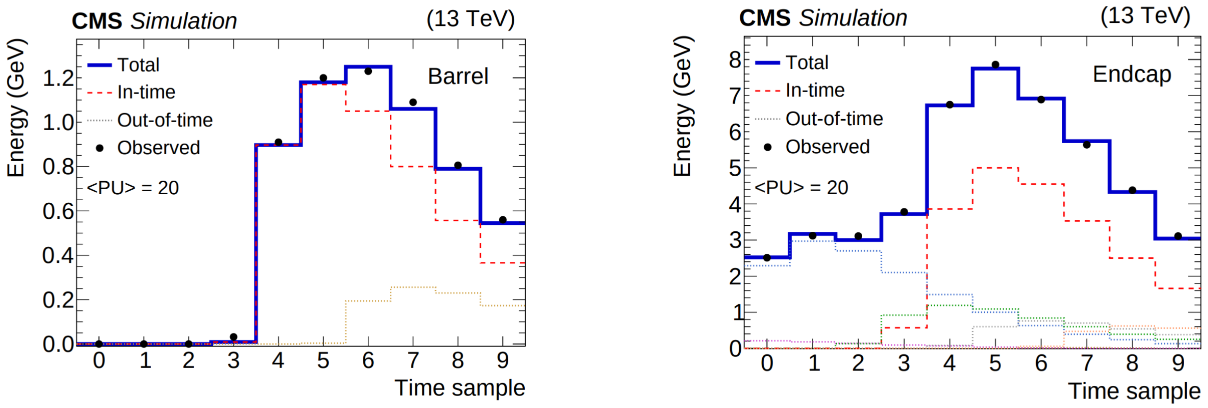


Figure 3.1: Shown here are two examples of pulse fits for simulated events with 20 average pile-up interactions and a 25 ns bunch spacing. These signals originate from individual crystals and result from a 10 GeV photon shower in the barrel (left) and in an endcap (right). In the left panel, one out-of-time (OOT) pulse is fitted in addition to the in-time (IT) pulse. In the right panel, six OOT pulses, along with the IT pulse, are fitted. The filled circles with error bars represent the 10 digitised samples. The red dashed distributions correspond to the fitted in-time pulses, while the dotted multicoloured distributions represent the OOT pulses with positive amplitudes. The solid dark blue histograms depict the sum of all fitted contributions. Different colours within the dotted distributions denote OOT pulses with varying BX, with a generic gray dotted line representing them in the legend. From Ref. [60].

these events ideal for template generation. The weighted average pulse shape from these events, adjusted to have the same amplitude, serves as the IT signal template. OOT templates maintain the same shape as the IT signal template but are shifted in time to represent signals originating from different bunch crossings.

Figure 3.1 illustrates an application of the Multifit procedure, showcasing its implementation for both the barrel (on the left) and the endcap (on the right). The figure presents a comprehensive display of all contributing templates, represented by dashed and dotted lines, which collectively influence the measured pulse (depicted as black circles).

In essence, the Multifit algorithm is a sophisticated tool used to disentangle the contributions of IT and OOT signals within the ECAL data and to reconstruct the amplitude of IT signals by fitting templates to the data.

Sensitivity of the Amplitude Reconstruction to Pedestal Drifts

In the Multifit method, the pedestal mean is used to calculate the pedestal-subtracted template amplitudes, used to compute the χ^2. And so, the Multifit amplitude reconstruction relies on pedestal baseline values and signal pulse templates obtained from periodic measurements, making it sensitive to potential variations over time.

The average pedestal value and electronic noise are separately measured for the differ-

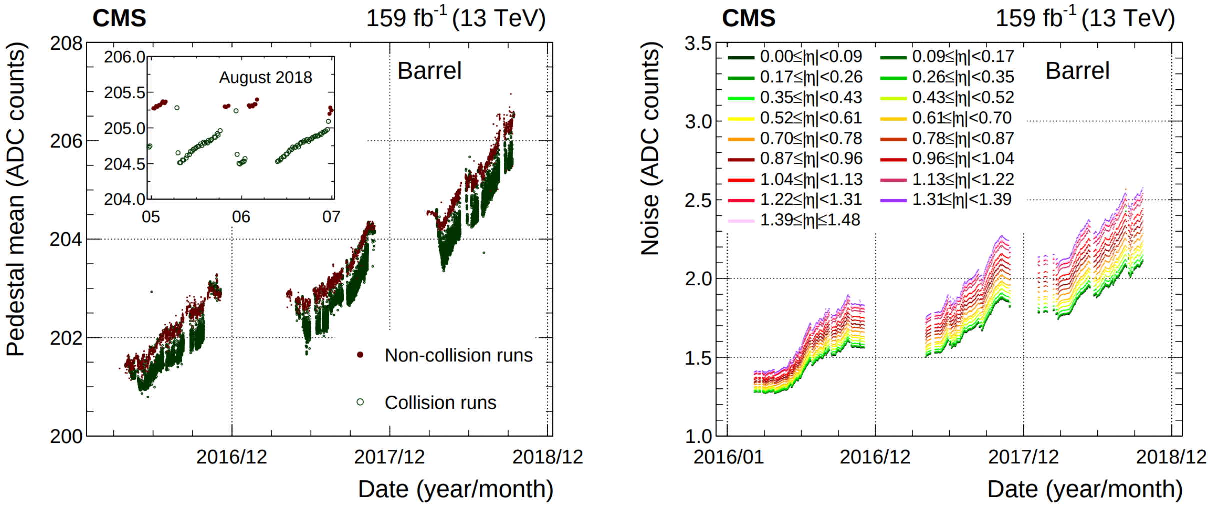


Figure 3.2: History of the pedestal mean value for the ECAL barrel (left) and its noise (right), measured during the 2016-2018 data taking period. The inset in the left panel shows an enlargement of two days in August 2018, to show in more detail the variation of the pedestal mean during LHC fills. From Ref. [60].

ent gain values from empty LHC bunches. The procedure applied for Run 2 in explained in Ref. [60]. The time evolution of the pedestal mean in the EB during Run 2 is shown in Figure 3.2 (left). It exhibits a long-term upward drift, that depends on integrated luminosity, while short-term effects depend on instantaneous luminosity, associated with variations in the readout electronics. The variation in pedestal value with time is similar at any crystal's pseudorapidity (η), but the magnitude increases with pseudorapidity due to higher irradiation. The evolution of electronic noise in the barrel is displayed in Figure 3.2 (right) and shows a steady increase over time, primarily due to increased APD dark current from higher radiation exposure. For the EB, where 1 ADC count is approximately equivalent to 40 MeV, the noise translates to around 65 MeV at the beginning of 2017 and 80 MeV at the end of proton-proton running in the same year. This observed increase in electronic noise can be attributed to transparency loss, namely the gradual reduction in the transparency of the crystal material over time due to its exposure to radiation. A minor decrease in noise induced by APD dark current is noticeable after long periods without irradiation, such as after year-end LHC stops. For the EE, the single-channel noise related to the VPT signal remains constant with time, at approximately 2 ADC counts. However, the energy-equivalent noise increases over time and with absolute pseudorapidity due to the crystal's transparency loss, which is strongly dependent on $|\eta|$ and time, caused by higher irradiation. Therefore, the average noise at the end of 2017 in the EE ranges from approximately 150 MeV up to $|\eta| \approx 2$, increasing to as much as 500 MeV at the limit of the CMS tracker acceptance ($|\eta| \approx 2.5$).

Any bias in measuring the pedestal mean would result in a nearly linear bias in the fitted amplitude. Figure 3.3 illustrates the absolute amplitude bias for 50 GeV energy

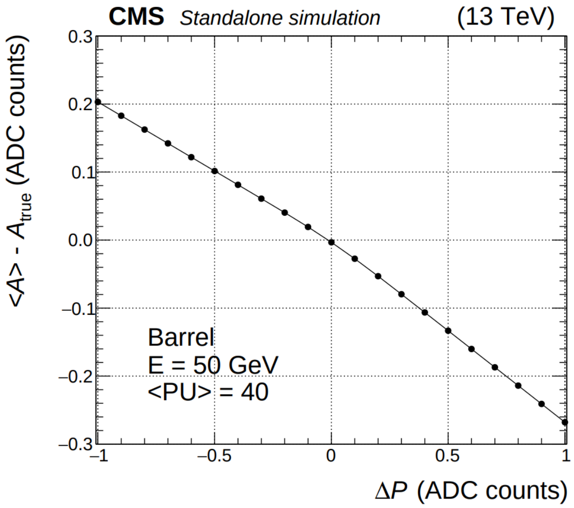


Figure 3.3: Reconstructed amplitude bias for the IT amplitude $\langle A\rangle - A_{\text{true}}$, as a function of pedestal shifts ΔP, for a single-crystal pulse of E = 50 GeV in the EB. From Ref. [60].

deposits in a single crystal in the barrel as a function of pedestal baseline shift. The behaviour in the endcaps is similar. A shift of ±1 ADC count results in an amplitude bias of up to 0.3 ADC counts in a single crystal, equivalent to a shift of around 300 MeV in a 5×5 crystal matrix in the barrel. Given that pedestal baseline can drift by as much as 2 ADC counts in a year, as seen in Figure 3.2 (left), and this drift affects all crystals coherently, it introduces a significant bias in the range of approximately (0.5–1%), even in the typical energy range of W, Z, and Higgs boson decay products. Hence, it is crucial to monitor and periodically correct the pedestals used in the reconstruction inputs.

3.1.2 Weights Method

The Weights method [61] will now be addressed. An estimate of the signal amplitude, denoted as $\hat{\mathcal{A}}$, is computed using a linear combination of discrete time samples, as shown in the equation below:

$$\hat{\mathcal{A}} = \sum_{i=1}^{N} w_i \cdot S_i. \tag{3.1}$$

Here, the weights are denoted as w_i, while S_i represents the time sample values in ADC counts, and N indicates the number of samples in a pulse, with i denoting the sample index. In summary, the estimation of signal amplitude is achieved through a weighted sum of the time samples.

Ensuring that the estimator $\hat{\mathcal{A}}$ accurately represents the true amplitude $\mathcal{A}$ is crucial. This requirement implies that:

$$\sum_{i=1}^{N} w_i \cdot f_i = 1. \tag{3.2}$$

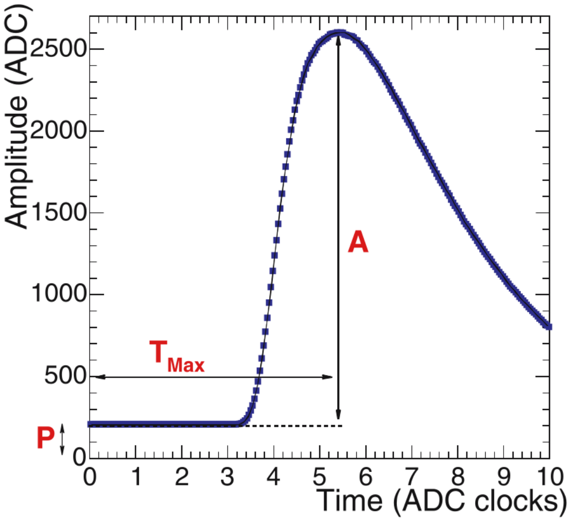


Figure 3.4: Profile of the signal Run 1 pulse using an electron beam of 120 GeV. The peaking time T_{Max}, the pedestal P and the amplitude of the signal A are shown. From Ref. [61].

In this equation, f_i signifies the value of the function $f(t)$, describing the temporal evolution of the signal pulse at time t_i for sample i. This function is normalised to have an amplitude of 1. An example of $f(t)$, obtained using an electron beam of 120 GeV, is shown in Figure 3.4.

The signal amplitude is estimated using a least squares method to minimise the variance of the estimator. The optimal weights are derived through the minimisation of the χ^2 statistic, defined by the equation:

$$\chi^2 = \sum_{i,j} (S_i - G_i) \cdot \mathbf{C}_{ij}^{-1} \cdot (S_j - G_j). \tag{3.3}$$

Breaking down the components of this equation:

S_i represents the magnitude of the sample i in ADC counts recorded at time t_i;

G_i $(\mathcal{A}, P, T_{\text{Max}})$ describes the characteristics of the signal pulse. Here, $\mathcal{A}$ is the true amplitude, P is the pedestal, and T_{Max} is the peaking time;

$\mathbf{C}$ is the covariant matrix. It represents the noise correlations between time samples i and j, and it is derived from data collected in the absence of a signal. In the following paragraph, an in-depth description of this covariance matrix will be provided.

When considering two free parameters, $\mathcal{A}$ and P, and neglecting noise correlation, the most accurate estimation of the amplitude can be obtained through the Equation (3.1) using the following weights:

$$w_i = \frac{f_i - \sum_j^N f_j}{\sum_j^N f_j^2 - (\sum_j^N f_j)^2/N}. \tag{3.4}$$

An additional set of weights can be derived to assess the peaking time parameter T_{Max} since the samples also carry information about it.

$$w_i = \frac{\left(\sum_j^N \frac{df_j^2}{dt}\right) f_i - \left(\sum_j^N f_j \frac{df_j}{dt}\right) \frac{df_i}{dt}}{\left(\sum_j^N f_j^2\right)\left(\sum_j^N \frac{df_j^2}{dt}\right) - \left(\sum_j^N f_j \frac{df_j}{dt}\right)^2}, \tag{3.5}$$

where the terms $\frac{df_j}{dt}$ and $\frac{df_j^2}{dt}$ represents the first and second derivative of the function f_j with respect to t. These weights are also called "pedestal-subtracting weights", since they perform a subtraction of the pedestal on an event-by-event basis. The detailed derivation of these formulas will be provided in the following paragraph.

Deriving the Weights

The optimal weights for amplitude and timing jitter reconstruction are obtained through the use of a least squares method. The χ^2 expression defined in Equation (3.3) can be represented in matrix notation as:

$$\chi^2 = (\mathbf{S} - \mathbf{G}(\mathcal{A}, \delta t, P))^\top \mathbf{C}^{-1} (\mathbf{S} - \mathbf{G}(\mathcal{A}, \delta t, P)). \tag{3.6}$$

Here, $\mathbf{S}$ is a vector composed of the time samples S_i, with N elements, and $\mathbf{G}(\mathcal{A}, \delta t, P)$ is modeled by:

$$G(t_i; \mathcal{A}, \delta t, P) = \mathcal{A} f(t_i + \delta t) + P. \tag{3.7}$$

$\mathbf{G}$ gives the expected measurement value at time t_i, taking into account the signal's amplitude $\mathcal{A}$, timing jitter δt, and pedestal P. Here, $f(t)$ describes the time evolution of the signal pulse.
When δt is sufficiently small, Equation (3.7) can be linearised, and G can be represented as a linear function of the free parameters:

$$G(t_i; \mathcal{A}, \delta t, P) \simeq \mathcal{A} f(t_i) + (\mathcal{A}\delta t) \frac{df}{dt}(t_i) + P. \tag{3.8}$$

In this way, Equation (3.6) becomes:

$$\chi^2 = (\mathbf{S} - \mathcal{A}\mathbf{F} - (\mathcal{A}\delta t)\mathbf{F}' - P\mathbf{1})^\top \mathbf{C}^{-1} (\mathbf{S} - \mathcal{A}\mathbf{F} - (\mathcal{A}\delta t)\mathbf{F}' - P\mathbf{1}), \tag{3.9}$$

where $\mathbf{F}$ is a vector containing $f(t_i)$, $\mathbf{F}'$ is a vector containing $\frac{df}{dt}(t_i)$, and $\mathbf{1}$ has all its vector elements equal to 1.

Minimising χ^2 of Equation 3.9 with respect to $\mathcal{A}$, $\mathcal{A}\delta t$, and P, a linear system of three equations is obtained:

$$\begin{pmatrix} \mathbf{F}^\top \mathbf{C}^{-1} \mathbf{F} & \mathbf{F}^\top \mathbf{C}^{-1} \mathbf{F}' & \mathbf{F}^\top \mathbf{C}^{-1} \mathbf{1} \\ \mathbf{F}'^\top \mathbf{C}^{-1} \mathbf{F} & \mathbf{F}'^\top \mathbf{C}^{-1} \mathbf{F}' & \mathbf{F}'^\top \mathbf{C}^{-1} \mathbf{1} \\ \mathbf{1}^\top \mathbf{C}^{-1} \mathbf{F} & \mathbf{1}^\top \mathbf{C}^{-1} \mathbf{F}' & \mathbf{1}^\top \mathbf{C}^{-1} \mathbf{1} \end{pmatrix} \begin{pmatrix} \hat{\mathcal{A}} \\ \hat{\mathcal{A}}\hat{\delta t} \\ \hat{P} \end{pmatrix} = \begin{pmatrix} \mathbf{F}^\top \\ \mathbf{F}'^\top \\ \mathbf{1}^\top \end{pmatrix} \mathbf{C}^{-1} \mathbf{S}. \tag{3.10}$$

Let's denote the matrix on the left-hand side of Equation (3.10) as $\mathbf{M}$. The solution to this system can be expressed as:

$$\begin{pmatrix} \hat{\mathcal{A}} \\ \hat{\mathcal{A}}\delta\hat{t} \\ \hat{P} \end{pmatrix} = \mathbf{M}^{-1} \begin{pmatrix} \mathbf{F}^{\top} \\ \mathbf{F}'^{\top} \\ \mathbf{1}^{\top} \end{pmatrix} \mathbf{C}^{-1}\mathbf{S} = \mathbf{W}\mathbf{S}. \tag{3.11}$$

In this equation, $\mathbf{W}$ represents the matrix of weights.

The minimal χ^2 value can be computed as follows by substituting the parameters of Equation (3.6) with the solutions of Equation (3.11):

$$\chi^2_{\text{Min}} = \mathbf{S}^{\top}(\mathbb{1} - (\mathbf{F}\mathbf{F}'\mathbb{1})\mathbf{W})^{\top}\mathbf{C}^{-1}(\mathbb{1} - (\mathbf{F}\mathbf{F}'\mathbf{1})\mathbf{W})\mathbf{S} \tag{3.12}$$

$$= \mathbf{S}^{\top}\mathbf{M}_{\chi^2}\mathbf{S}, \tag{3.13}$$

where $\mathbb{1}$ denotes the identity matrix. In this equation, $\mathbf{M}_{\chi^2}$ is a matrix that allows computing χ^2_{Min} event-by-event without having to minimise χ^2.

For the sake of simplicity, scenarios where the noise correlation between individual samples can be considered negligible will be the primary focus. The noise is the root mean square deviation of the reconstructed amplitude when no signal is input. Thus the ECAL noise is measured by examining the variation of the reconstructed amplitude when data taken with no signal (often called "pedestal runs"), are reconstructed. The concept of the noise correlation matrix is thus introduced in the absence of any signal, and it is mathematically defined as:

$$\mathbf{C}_{ij} = \langle n_i \cdot n_j \rangle, \tag{3.14}$$

where $n_i = S_i - P_i$ is the difference between the sample value S_i and the pedestal P_i. Consequently, the diagonal elements of this matrix (C_{ii}) represent the squared noise contribution from each individual sampling process (σ^2). In situations where there is a complete absence of noise correlations among the samples, the noise correlation matrix $\mathbf{C}$ simplifies to $\sigma^2\mathbb{1}$.

In this study, only the following two scenarios were considered:

1. If the parameter $\mathcal{A}$ and P are the only free parameters, simplifying Equation (3.10) gives two sets of weights used to compute the two estimators. In particular, optimal weights for the amplitude reconstruction can be computed using the following formula:
$$w_i = \frac{f_i - \sum_j^N f_j}{\sum_j^N f_j^2 - (\sum_j^N f_j)^2/N}. \tag{3.5}$$
2. Jitter-compensating weights can be obtained by considering $\mathcal{A}$ and δt as a free parameters. This leads to the following expression for the optimal set of weights:
$$w_i = \frac{\left(\sum_j^N \frac{df_j^2}{dt}\right) f_i - \left(\sum_j^N f_j \frac{df_j}{dt}\right)\frac{df_i}{dt}}{\left(\sum_j^N f_j^2\right)\left(\sum_j^N \frac{df_j^2}{dt}\right) - \left(\sum_j^N f_j \frac{df_j}{dt}\right)^2}. \tag{3.6}$$

Optimisation and Noise Reduction

The reconstruction of signal amplitude can rely on a single sample taken at the signal peak. In scenarios of synchronous running, adjustments to the pulse maximum time, represented by the parameter T_{Max}, can ensure the peak aligns with one of the samplings. By independently determining and subtracting the pedestal from the maximum, a single sample can yield the pulse amplitude. This approach is relatively robust against potential timing jitter. However, using a greater number of samples offers the advantage of reducing noise. In the following, a comprehensive overview of the underlying concept is provided. For a detailed exposition of this concept, refer to Ref. [61]. In the absence of noise correlation between time samples, and with knowledge of both the pedestal P and peaking time T_{Max} (enabling $G = G(A) = Af(t)$ from Equation (3.7)), the square root of the variance for the estimated amplitude is given by:

$$\sigma_{\mathcal{A}} = \frac{\sigma}{\sqrt{\sum f_i^2}}, \tag{3.15}$$

where σ represents the noise in a single digitisation, known as single-sampling noise. Consequently, increasing the number of samples decreases the noise in the reconstructed amplitude. Moreover, better noise reduction is achieved when samples close to the peak, containing more signal, are used. In this context, it was found that with the Run 1 pulse shape of Figure 3.4, using just five samples provides approximately 60% noise reduction, with marginal improvement when employing more samples. In the analysis for Phase II, all 16 samplings are used to reconstruct the amplitude and the timing jitter.

3.1.3 Comparative Analysis

The Multifit algorithm exhibits notable robustness when confronted with OOT pile-up, making it particularly suitable for scenarios with significant OOT pile-up. However, a key limitation of the Multifit approach lies in its sensitivity to variations in the pedestal level. In fact, the Multifit amplitude reconstruction relies on pedestal baseline values obtained from periodic measurements, making it sensitive to potential variations over time. The average pedestal value and electronic noise are separately measured from empty LHC bunches. Over the course of Run 2, the pedestal mean and electronic noise in the EB steadily drifted due to elevated APD dark current caused by higher radiation exposure, as depicted in Figure 3.2. Any bias in measuring the pedestal has a linear impact on the fitted amplitude, as demonstrated in Figure 3.3. Therefore, it is imperative to monitor and periodically correct the pedestals used in the reconstruction inputs to ensure accurate results.

In contrast, the Weights method streamlines the reconstruction process while displaying robustness against pedestal variations. In fact, this algorithm was developed to provide a dynamic subtraction of the pedestal, which is estimated on an event-by-event basis through the weights in Equations (3.4) and (3.5). Another noteworthy advantage of the Weights method lies in its analytical, robust, and rapid nature, as opposed to the recursive Multifit algorithm. Nevertheless, it may suffer accuracy compromises when confronted with substantial OOT pile-up contributions.

The choice between the Multifit and Weights methods depends on specific experimental requirements. In situations of significant OOT pile-up, the Multifit method may be the preferred choice, excelling in managing high OOT pile-up scenarios. On the contrary, when prioritising robustness and insensitivity to pedestal variations, the Weights method stands out as the best option, particularly when OOT pile-up is not the dominant factor.

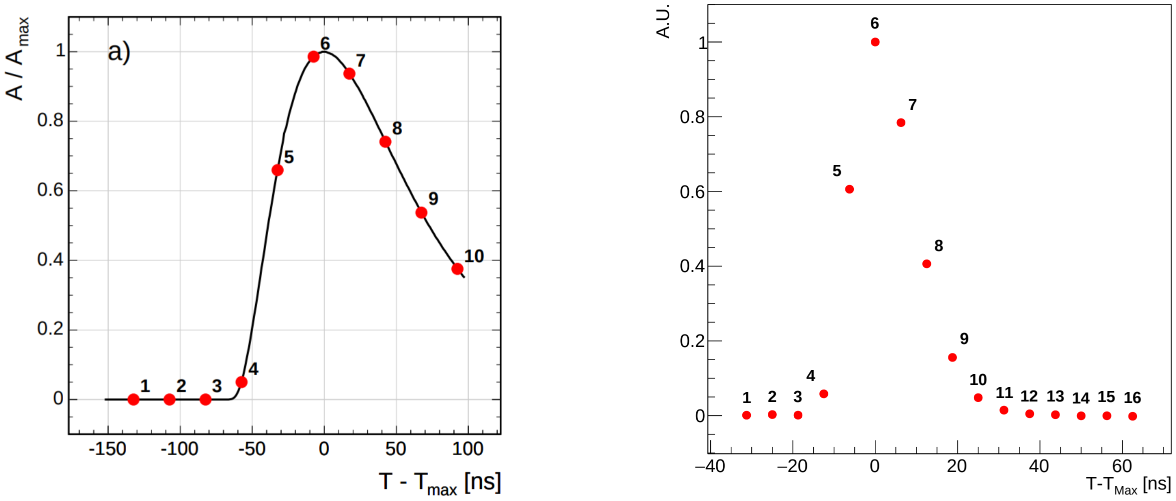
In Phase II, OOT pile-up is expected to be less problematic due to a reduced pulse shape width and a higher sampling frequency of 160 MHz, providing a sampling interval of 6.25 ns. This advantage becomes even more evident when comparing the typical pulse shapes of Phase I and Phase II, as illustrated in Figure 3.5. The Phase II pulse shape showcases the reduced width and increased frequency, which are instrumental in mitigating OOT pile-up. Additionally, the pedestal may vary with detector aging, potentially influenced by changes in noise levels. Consequently, the Weights method emerges as a viable candidate for ECAL reconstruction in Phase II. However, the approach of the Collaboration favours maintaining both algorithms accessible for selection, depending on evolving detector conditions. When OOT pile-up becomes predominant, the Multifit algorithm is the preferred choice, whereas the Weights method gains precedence when prioritising robustness and insensitivity to pedestal variations, not only in offline reconstruction but also in real-time processing at L1, contributing to the overall effectiveness of the experiment. In fact, the use of the Weights method extends beyond offline reconstruction. It is highly probable that the reconstruction at L1 will also rely on the Weights method, employing the same algorithm that can subsequently be used in offline and HLT processing. This seamless integration of the Weights method across different stages of data processing ensures consistency and maximises the adaptability of the reconstruction process.

For Phase II, the Weights method has been successfully implemented for this thesis project, while the Multifit algorithm is still under development. The modules developed for the Weights method have been integrated into the official CMS software. This achievement is a major step forward in applying the Weights method effectively for the high-luminosity phase.

3.2 Amplitude Reconstruction for Phase II

The objective of this section is to employ the Weights method for the precise extraction of reconstructed amplitude for Phase II. Initially, a set of optimal weights tailored specifically to the new Phase II pulse shapes was derived. Subsequently, a dedicated algorithm was developed to calculate the reconstructed amplitude using these custom weights.

Phase II introduces significant changes compared to Phase I pulses, primarily driven by two key factors: an increased number of samples, with 16 samples for each pulse replacing the previous 10, and the introduction of a new pulse shape, as can be seen from the comparison between Figure 3.5a and Figure 3.5b. These changes are aimed at improving the energy resolution and precision of the ECAL detector. The increase in the number of



(a) Typical Phase I ECAL pulse shape. The dots indicate 10 discrete samples of the pulse, from a single event, pedestal subtracted and normalised to the maximum amplitude. A detailed description of this plot can be found in Ref. [61].

(b) Averaged pulse, computed as described in Section 3.2.1, representing the typical Phase II ECAL pulse shape.

Figure 3.5: Typical Phase I (a) and Phase II (b) ECAL pulse shapes, as functions of the difference between the time (T) of the ADC sample and the peaking time (T_{max}). Note that the 6-th sample does not coincide with the pulse maximum time. For the plot (b), however, an approximation was made and was considered T_6 as T_{Max}.

samples allows for a more detailed characterisation of the pulse shape, while the narrower shape helps in reducing the impact of pile-up and improving the energy measurement accuracy.

3.2.1 Averaged Pulse

To accurately reconstruct amplitudes for a wide range of pulses with the Equation (3.1), a crucial step involves the computation of weights that can effectively account for the characteristics of these pulses. To achieve this objective, it is essential to provide the weights-computation algorithm with a representative input pulse. This representative pulse is known as the “averaged pulse”.
The process of computing the averaged pulse involves several steps:

1. Generation of simulated pulses: initially, a set of 100 simulated events resembling the response of ECAL to the production of a top-antitop quarks pair at 14 TeV, is generated. Figure 3.6 displays ten randomly selected simulated pulses as an example.

2. Selection of "interesting pulses": among the generated pulses, a subset called "interesting pulses" is selected. These are pulses with the 6-th sample exceeding a certain threshold, set at 100 ADC counts. The 100-events sample has 342 interesting pulses.

3. Pedestal subtraction: for each of the selected pulses, the pedestal, representing the baseline signal level, is subtracted. The pedestal for each pulse is determined by averaging the last four samples.

4. Normalisation of pulses: subsequently, each of the interesting pulses is normalised relative to the 6-th sample. This value is referred to as "maxADC" for the sake of clarity.

5. Weighted average: the final step involves computing the averaged pulse using a weighted mean. In this weighted mean, the maxADC values serve as the weights, ensuring that pulses with higher maxADC values contribute more significantly to the computation of the averaged pulse. The i-th sample of the averaged pulse a_i is computed as follows:

   $$a_i = \frac{\sum_{j=1}^{M}(\mathrm{S}_{i,j} \cdot \mathrm{maxADC}_j)}{\sum_{j=1}^{M} \mathrm{maxADC}_j}. \tag{3.16}$$

   In this formula, M represents the number of signal pulses, N represents the number of samples in each pulse, $S_{i,j}$ represents the i-th sample of the j-th pulse, and maxADC_j represents the 6-th sample for the j-th pulse (before normalisation).

Figure 3.7 presents a plot featuring 10 interesting pulses, chosen randomly, along with the plot of the averaged pulse. Minor differences in pulses are observable due to timing shifts caused by the absence of vertex correction. Vertex correction is a process that takes into account the precise location where the particle interaction occurred within the detector. This correction helps align the timing information of different pulses, particularly when particles originate from different points within the detector. As a result, minor differences in the timing of the observed pulses are observable. These differences are due to the variations in the production points of the particles within the detector. For a standalone representation of the averaged pulse, refer to Figure 3.5b.

The resulting set of weights, derived using Equation (3.4), is as follows:

$$\begin{aligned}&-0.121016, -0.119899, -0.120923, -0.0848959, 0.261041, \underline{0.509881}, 0.373591, 0.134899,\\ &-0.0233605, -0.0913195, -0.112452, -0.118596, -0.121737, -0.121737, -0.121737, -0.121737\end{aligned}$$

The most substantial weight corresponds to the 6-th one (underlined), aligning with expectations. Moreover, by ensuring that the sum of weights equals to zero, the estimator $\hat{\mathcal{A}}$ remains unbiased.

3.2.2 Closure tests

Finally, a closure test was executed on an independent 1000-events set, featuring 3205 interesting pulses (with maxADC > 100 ADC counts). Figure 3.8 displays the relation-

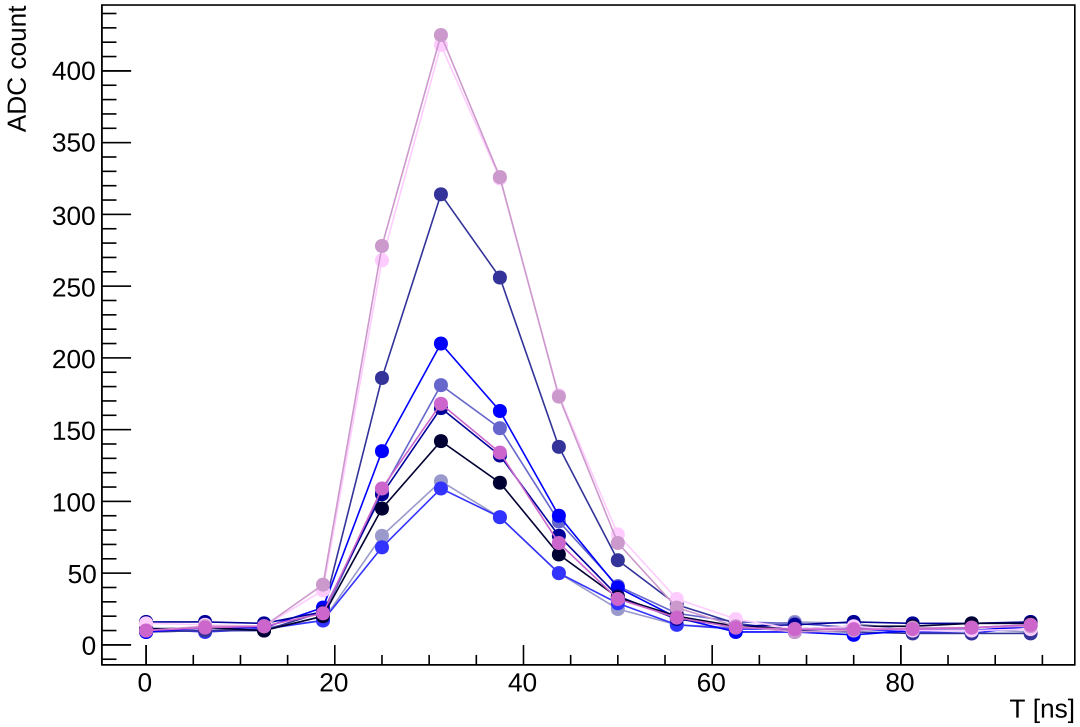


Figure 3.6: A selection of 10 privately simulated Phase II pulses corresponding to the production of a top-antitop quarks pair, chosen randomly. The dots indicate 16 discrete samples of the pulses.

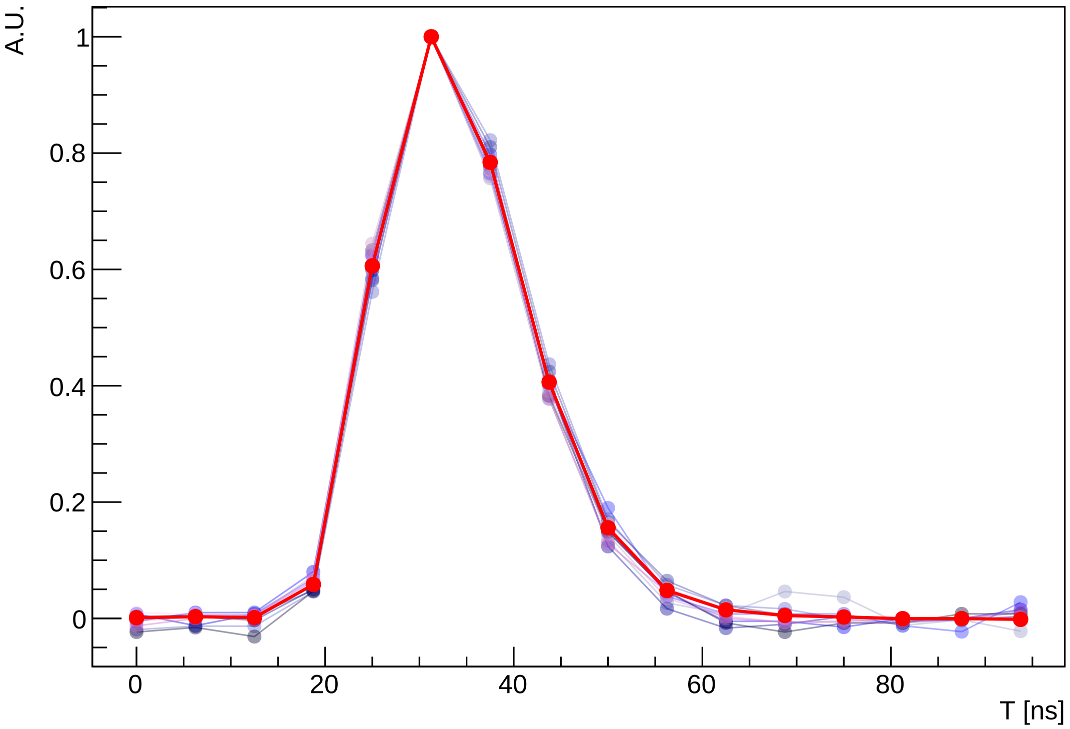


Figure 3.7: A plot depicting 10 randomly selected, pedestal-subtracted, and normalised pulses, along with the averaged pulse (in red). The pedestal has been computed averaging over the last four samples. Minor differences in pulses are observable due to timing shifts, owing to the absence of vertex correction.

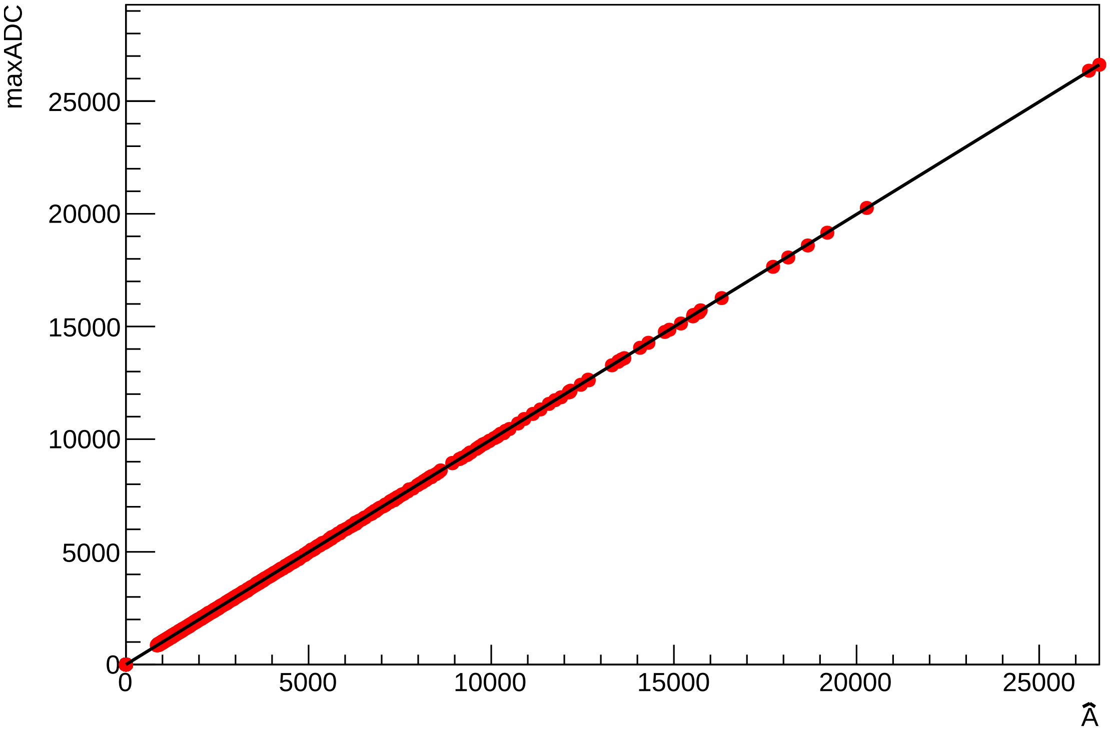


Figure 3.8: Relationship between maxADC pedestal subtracted (with the pedestal computed averaging over the last four sample), and the amplitude reconstructed through the Weights method ($\hat{\mathcal{A}}$). Only the pulses with maxADC > 100 ADC counts were considered. The black line represents the bisector, indicating when maxADC equals $\hat{\mathcal{A}}$.

ship between maxADC pedestal subtracted, and the amplitude reconstructed through the Weights method ($\hat{\mathcal{A}}$). The black line represents the bisector, indicating when maxADC equals $\hat{\mathcal{A}}$. Figure 3.9, on the other hand, portrays the ratio between the $\hat{\mathcal{A}}$ and maxADC (pedestal subtracted).

In both cases, the results are satisfactory, since the obtained average ratio between $\hat{\mathcal{A}}$ and maxADC stands at 1.0026. This ratio is marginally greater than 1, and this outcome aligns with expectations, considering that the 6-th sampling point is not perfectly synchronised with the actual maximum. Consequently, it tends to yield a slightly lower value. As a result, maxADC, calculated as the difference between the 6-th sampling point and the pedestal, will also be marginally smaller than the true amplitude. This outcome serves to validate the effectiveness of the Weights method for reconstructing the amplitude in cases where the pulses exhibit Phase II characteristics.

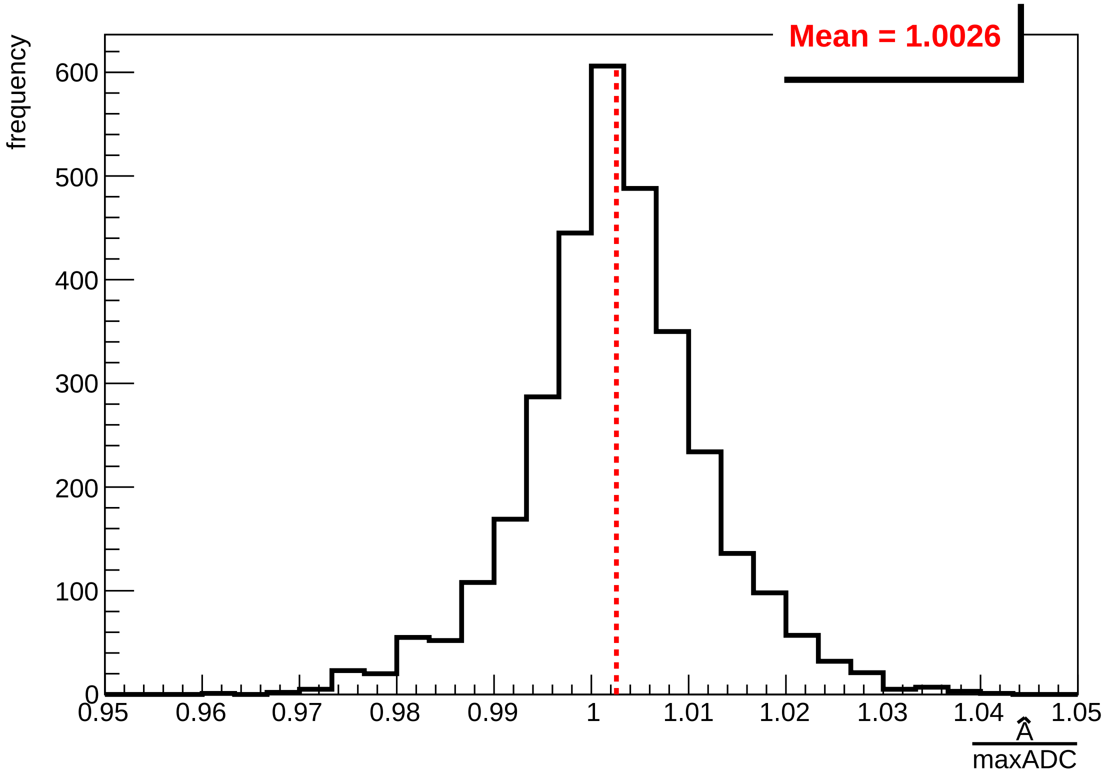


Figure 3.9: Ratio between the amplitude reconstructed through the Weights method ($\hat{\mathcal{A}}$) and maxADC, pedestal subtracted (with the pedestal computed averaging over the last four sample). Only the pulses with maxADC $>$ 100 ADC counts were considered. The red dashed line represents the mean value.

3.3 Timing Jitter Reconstruction for Phase II

This section is dedicated to timing jitter reconstruction for the Phase II ECAL pulses. The main goal is the computation of a set of optimised weights specifically tailored to accommodate the pulse shapes characteristic of Phase II ECAL (Figure 3.5b).

Timing jitter, denoted as δt, represents the temporal deviation in the arrival time of particle signals detected by the ECAL. A weighted sum of signal samples (in ADC counts) is relied upon, as expressed by the equation:

$$E[\delta t] = \sum_{i=1}^{16} w_i S_i. \tag{3.17}$$

Here, $E[\delta t]$ is the expected timing jitter, and the summation comprehends the contributions of 16 signal samples, each associated with a specific weight factor, w_i.

Details regarding the mathematical procedure for weight extraction can be found in Section 3.1.2. However, the weights can be computed using the formula in Equation (3.5). To compute the derivative component required in this methodology, $\frac{df}{dt}$, the derivative of the alpha-beta function is employed:

$$S(t) = \mathcal{A}\Big(1 + \frac{t - t_0}{\alpha\beta}\Big)^{\alpha} e^{-\frac{t-t_0}{\beta}}. \tag{3.18}$$

This function models the signal samples S as a function of time t with the following parameters:

$\mathcal{A}$ represents the amplitude of the signal

t_0 represents the time of the pulse peak

α,β determine the shape of the pulse: α controls the rise time, while β controls the decay time.

This function is used to fit the pulses and then to retrieve the derivative when needed for the weights computation. An example of fit can be seen in Figure 3.10.
To compute the timing jitter weights, the same averaged pulse used for the amplitude reconstruction in Section 3.2.1 was used.
Finally, the resulting set of weights, derived using Equation (3.5), is as follows:

$$\begin{aligned}&0.429452, 0.442762, 0.413327, 0.858327, 4.42324, 2.04369, -3.42426, -4.16258,\\&-2.36061, -0.725371, 0.0727267, 0.326005, 0.402035, 0.404287, 0.434207, 0.422775\end{aligned}$$

The sum of the weights in 0 to ensure an unbiased estimator.

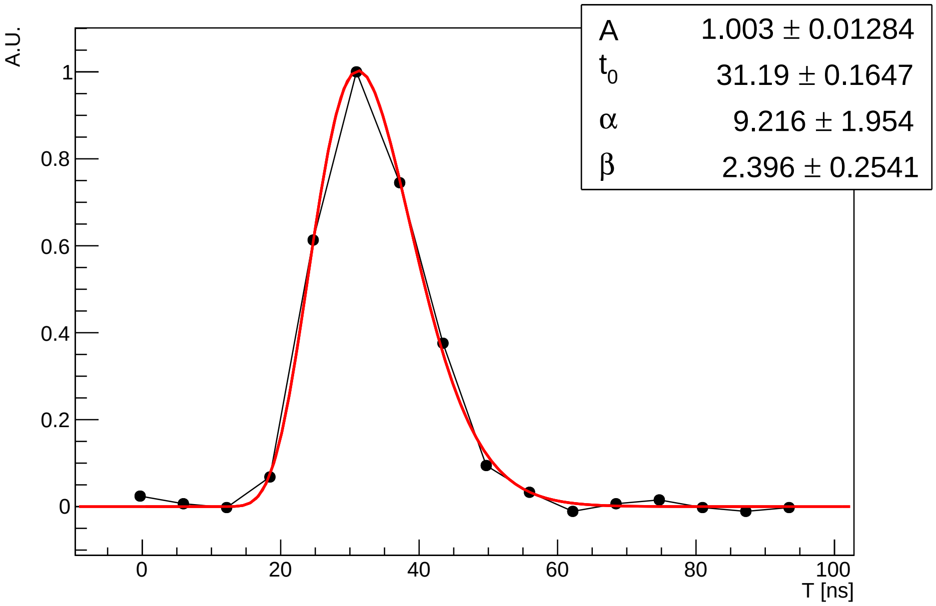


Figure 3.10: Randomly selected pulse fitted with the alpha-beta function defined in Equation (3.18). The pulse has been previously pedestal-subtracted and normalised to the 6-th sample.

3.3.1 Closure Tests

Closure tests were conducted to validate the method. The first test involved introducing various timing jitter values to the averaged pulse, ranging from -1.5 to 1.5 ns. Figure 3.11 depicts the reconstructed timing jitter against the simulated values. It is important to note the strong correlation observed in the central region of the scatter plot. This correlation illustrates the method's capability to accurately reconstruct the simulated timing jitter for small values of δt.

However, in the two regions to the left and right, where the simulated timing jitter is most consistent, some bias is evident. The bias can be mathematically explained, as the timing jitter weights are extracted through a linear approximation valid only for small timing jitter values, as shown in Equation (3.8). For a clearer representation of this bias, Figure 3.12 illustrates the bias ($\delta t - E[\delta t]$) as a function of the timing jitter δt. It becomes evident that the bias increases as the timing jitter becomes greater, which aligns with expectations.

Finally, a comprehensive examination was conducted using a dataset comprising 1000 independent events. Figure 3.13 presents the plot illustrating the relationship between the reconstructed timing jitter $E[\delta t]$ and the actual timing jitter δt for each pulse.

Additionally, within the vertical bands on the plot, the Root Mean Square (RMS) for each

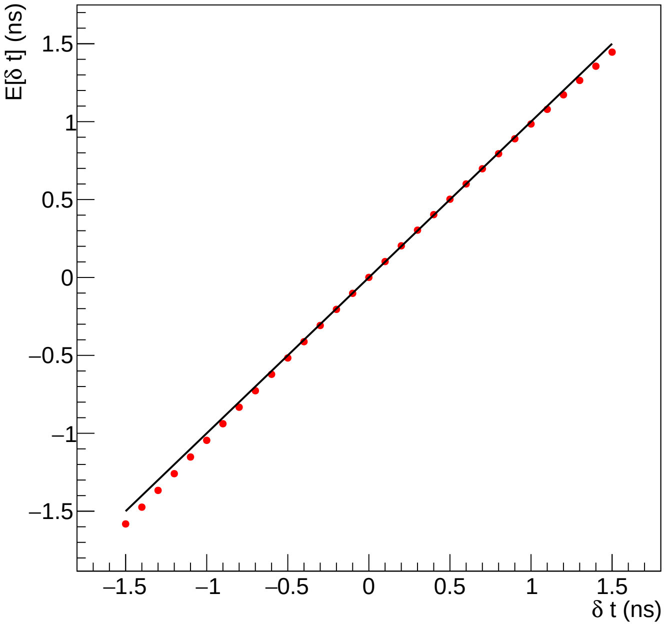


Figure 3.11: Reconstructed timing jitter versus the actual simulated timing jitter. Timing jitter values ranging from $\delta t = -1.5$ ns to $\delta t = 1.5$ ns were introduced to the averaged pulse. The red line represents the ideal case, $E[\delta t] = \delta t$.

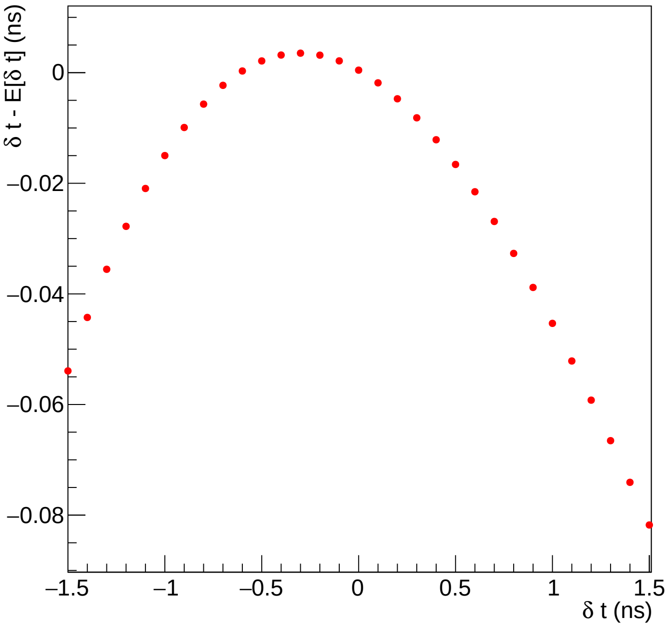


Figure 3.12: Time bias $\delta t - E[\delta t]$ as a function of the timing jitter δt.

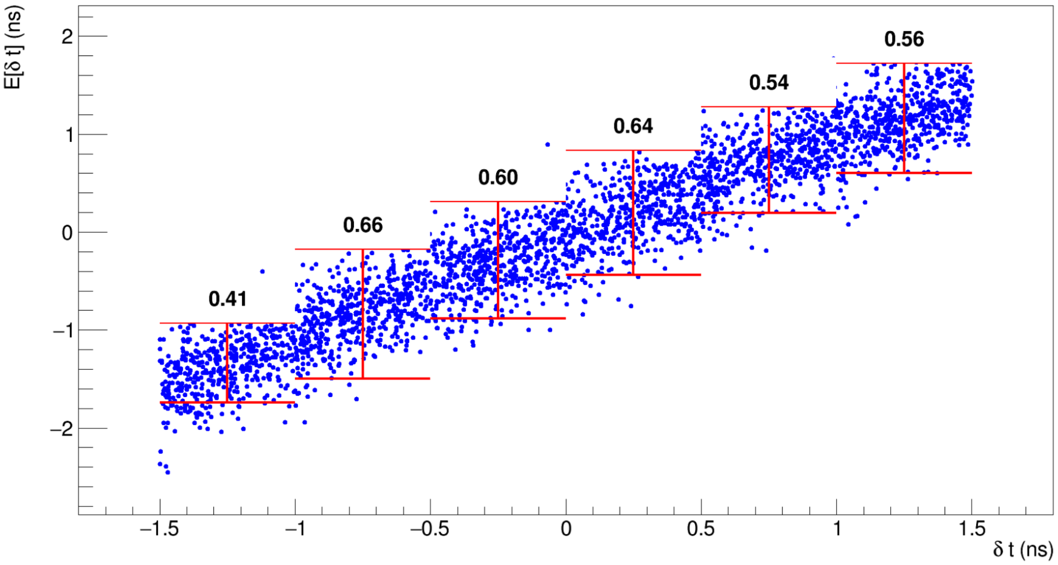


Figure 3.13: Relationship between reconstructed timing jitter ($E[\delta t]$) and actual timing jitter (δt) for pulses with maxADC $>$ 100 ADC counts in an independent 1000-event dataset. The red vertical bars represent the Root Mean Square (RMS) values for each $\delta t = 0.5\,\text{ns}$ interval, with the corresponding RMS values displayed above the bars.

timing jitter interval k of 0.5 ns was determined using the following formula:

$$\mathrm{RMS_k} = \sqrt{\frac{1}{M_k}\sum_{i=1}^{M_k}(E[\delta t]_i - \overline{E[\delta t]}_k)^2}. \tag{3.19}$$

Here is an elucidation of the notation used:

M_k denotes the total number of data points within interval k.

$E[\delta t]$ is the reconstructed timing jitter for each individual data point.

$\overline{E[\delta t]}_k$ represents the mean timing jitter within interval k.

As depicted in Figure 3.13, the RMS values consistently remain well below 68% across all δt intervals, implying a high degree of accuracy in the reconstructed timing jitter, indicating the effectiveness of the analysis and methodology.

In conclusion, the weights method has been successfully implemented for precise timing jitter reconstruction within ECAL pulses for Phase II. The method, as demonstrated by closure tests, proves its effectiveness.

Electroweak Production of $\mathrm{W}^+\mathrm{W}^-$ Pair Plus Two Jets in the Leptonic Channel

This chapter discusses a significant discovery at the LHC: the first observation of the EW production of $\mathrm{W}^+\mathrm{W}^-$ bosons in association with two jets in the fully leptonic channel. This study uses data collected by the CMS experiment from $\sqrt{s} = 13$ TeV proton-proton collisions in the years 2016 to 2018 (full Run 2 dataset). The signal is observed with a statistical significance of 5.6 standard deviations. The measured fiducial cross-section is 10.2 ± 2.0 fb, in agreement with the SM prediction. For a comprehensive exploration of the methods and results, the interested reader is referred to the publication [62] associated with this study.

Within the following sections of this thesis, a thorough approach is taken to dissect the details of this research. Section 4.1 elucidates the signal's topological features and the principal backgrounds encountered in this analysis. Section 4.2 details both the data and MC simulations employed. The reconstruction of the final state objects is reported in Section 4.3. Furthermore, Section 4.4 delineates the criteria employed for event selection, grouping the events into various categories crucial for the ultimate fit. To effectively address the substantial contamination from $\mathrm{t\bar{t}}$ production and the irreducible QCD-induced $\mathrm{W}^+\mathrm{W}^-$ background, a deep neural network, as described in Section 4.5, was trained. Control regions, highlighted in Section 4.6, were instrumental in monitoring the agreement between data and simulations for the various backgrounds. Section 4.7 delineates the methodologies applied to estimate background contamination in the signal region. Section 4.8 discusses the systematic uncertainties that impact the measurement. Section 4.9 explains the statistical procedures employed to extract the signal and presents the final results. Lastly, a summary of the results obtained and hints at future developments in this project are provided in Section 4.10.

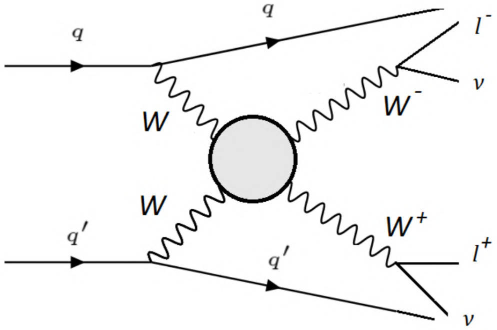


Figure 4.1: Schematic representation of the VBS W^+W^- process in the fully leptonic channel. The gray circle represents all possible interactions described in Section 1.2.

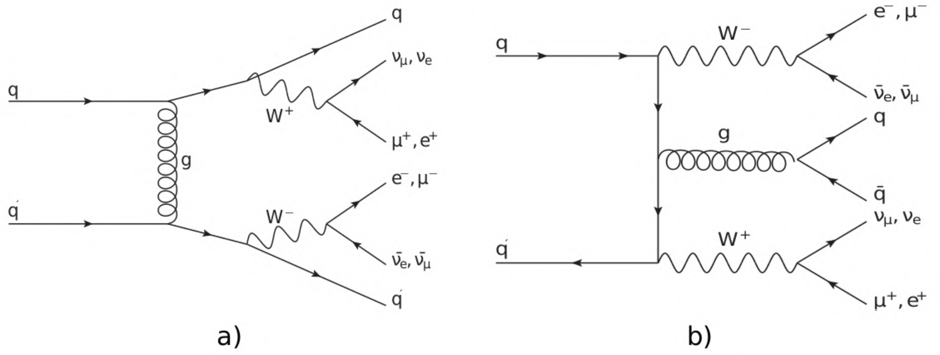
4.1 Signal Topology and Main Backgrounds

In this section, the signal topology and the primary background processes associated with VBS events are explored. VBS processes are characterised by the presence of two vector bosons in association with two high-energy and well-separated jets. Figure 4.1 illustrates a Feynman diagram representing the VBS W^+W^- process, with the gray circle encompassing all possible interactions described in Section 1.2. This analysis focuses on events where both W bosons decay leptonically, producing either an electron or a muon and the corresponding neutrino. Consequently, the signal final state consists of two jets, two leptons of opposite charge, and missing transverse energy. Events are categorised based on the flavour of the final state leptons into three distinct channels: ee, $\mu\mu$, and $e\mu$.

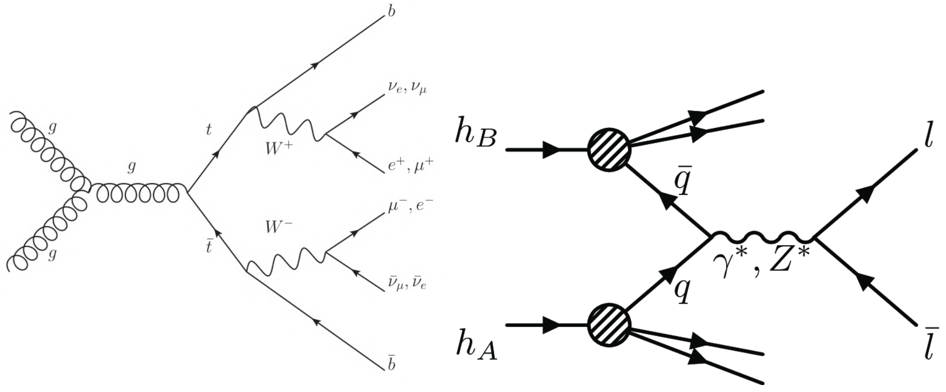
Illustrative Feynman diagrams depicting the main background processes, characterised by a final state similar to the signal, are presented in Figure 4.2. The primary contributor to background events originates from the decays of top quarks in $\mathrm{t\bar{t}}$ processes. A smaller contribution arises from single top (tW) production, in which the top quark decays into a b quark and a W boson, which can subsequently decay leptonically, mimicking the VBS final state. However, processes characterised by jets originating from b quarks (b-jets) can be identified using sophisticated “b-tagging” algorithms. Hence, a veto against b-jets is applied to mitigate the $\mathrm{t\bar{t}}$-tW background contamination in the signal region. As explained in Section 1.2, QCD-induced W^+W^- production shares the same final state as the signal. It can only be reduced by imposing criteria on the dijet invariant mass and the pseudorapidity separation of VBS jets.

Another source of contamination comes from DY processes ($pp \to Z/\gamma^* \to \ell\ell$), primarily due to detector effects. These processes result in a fake high missing transverse energy being reconstructed along with the two leptons and jets originating from Initial State Radiation (ISR). This background is significant in the ee and $\mu\mu$ channels but substantially reduced in the $e\mu$ channel. In the $e\mu$ final state, the DY contribution primarily arises from the leptonic decays of tau leptons in the $Z/\gamma^* \to \tau\tau \to \ell\ell$ processes.

Finally, another background contribution originates from W + jets production, where



(a) QCD WW processes at $\alpha_{EW}^4\alpha_S^2$.



(b) $t\bar{t}$ production.

(c) Drell-Yan process.

Figure 4.2: Exemplary Feynman diagrams for the main background processes resembling the signal final state.

one charged lepton arises from the W boson decay, while the second charged lepton (nonprompt lepton) results from the misidentification of a hadron. This background is referred to as the nonprompt background.

Higgs boson production and multiboson processes, including diboson (ZZ, WZ/γ^*, Zγ) and triboson (WWW, ZZZ, WWZ, WZZ) productions, are also considered in the analysis. However, they constitute minor contributions to the total background due to stringent requirements on VBS jets and leptons, as explained in Section 4.4.

Section 4.7 elaborates on the methods employed in this analysis to estimate the contributions from various sources of background in the signal region. Data-driven techniques are used to improve the modelling of main background processes such as $\mathrm{t\bar{t}}$-tW, DY, QCD-induced $\mathrm{W^+W^-}$, and nonprompt backgrounds, while the estimation of minor backgrounds relies entirely on MC simulations.

4.2 Data Set, Triggers, and Simulated Samples

The data used in this study corresponds to the recorded events from the 2016, 2017, and 2018 data-taking periods at the CMS experiment. These data sets are associated with integrated luminosities of $36.3\,\mathrm{fb}^{-1}$, $41.5\,\mathrm{fb}^{-1}$, and $59.7\,\mathrm{fb}^{-1}$, respectively [63, 64, 65].

4.2.1 Triggers

The selection of relevant events for this analysis is achieved through the application of a set of HLT criteria. Emphasis is placed on events in which two leptons, either electrons or muons, are featured in the final state. The HLT triggers are categorised into single lepton triggers and double lepton triggers, depending on whether they require one or two leptons, respectively.

For the single lepton triggers, specific p_T thresholds for electrons and muons are imposed, along with restrictions on their η. These triggers ensure that the events under consideration contain at least one high-energy lepton. The thresholds for these triggers vary from year to year, reflecting the evolving data-taking conditions.

The double lepton triggers, on the other hand, demand the presence of two leptons with lower p_T thresholds. Again, these leptons must satisfy identification and isolation criteria. The details of these triggers, including the p_T thresholds and acceptance criteria for each year, are summarised in Table 4.1.

4.2.2 Simulated Samples

MC simulations are used to model both signal and background processes. These simulations play a crucial role in understanding the expected outcomes and in evaluating the performance of these analysis techniques. Simulated samples are produced separately for each of the three data-taking years, accounting for changes in detector conditions, varying pileup conditions, and updates to reconstruction software. The event generation is

Single Lepton Triggers		
One Muon	2016	$p_T > 23\,\mathrm{GeV}$ in $\lvert\eta\rvert < 2.4$
	2017	$p_T > 27\,\mathrm{GeV}$ in $\lvert\eta\rvert < 2.4$
	2018	$p_T > 24\,\mathrm{GeV}$ in $\lvert\eta\rvert < 2.4$
One Electron	2016	$p_T > 25\,\mathrm{GeV}$ in $\lvert\eta\rvert < 2.1$ $p_T > 27\,\mathrm{GeV}$ in $\lvert\eta\rvert \in (2.1, 2.5)$
	2017	$p_T > 35\,\mathrm{GeV}$ in $\lvert\eta\rvert < 2.5$
	2018	$p_T > 32\,\mathrm{GeV}$ in $\lvert\eta\rvert < 2.5$

Double Lepton Triggers		
Two Muons	2016 – 2018	$p_{T,\ell_1} > 17\,\mathrm{GeV}$ & $p_{T,\ell_2} > 8\,\mathrm{GeV}$
Two electrons	2016 – 2018	$p_{T,\ell_1} > 17\,\mathrm{GeV}$ & $p_{T,\ell_2} > 8\,\mathrm{GeV}$
One muon and one electron	2016 begin	$p_{T,\ell_1} > 23\,\mathrm{GeV}$ & $p_{T,\ell_2} > 8\,\mathrm{GeV}$
	2016 end – 2018	$p_{T,\ell_1} > 23\,\mathrm{GeV}$ & $p_{T,\ell_2} > 12\,\mathrm{GeV}$

Table 4.1: High-Level Trigger (HLT) requirements for the single and double lepton triggers used in the analysis. Here, p_{T,ℓ_1} refers to the transverse momentum of the leading lepton, and p_{T,ℓ_2} refers to the transverse momentum of the trailing lepton.

accomplished using three primary MC generators: `Powheg v2` [66], `MadGraph5_aMC@NLO` v2.4.2 [67], and `mcfm` v7.0 [68, 69, 70]. The generator choice depends on the specific hard scattering process being simulated, ensuring high accuracy.
After the simulation of the hard scattering process, the subsequent parton shower and hadronisation processes need to be accounted for. For this purpose, `pythia` 8.226 (2016) and 8.230 (2017 and 2018) are used, being event generators capable of simulating the development of quarks and gluons into observable particles. These generators are configured to match the underlying event (UE) conditions specific to each year. In 2016, this configuration is denoted as `CUETP8M1` [71], while in 2017 and 2018 it is referred to as `CP5` [72]. The parton distribution functions (PDFs) for event generation differ between the 2016 sample (using `NNPDF` 3.0 [73, 74]) and the 2017-2018 samples (using `NNPDF` 3.1 [75]), reflecting improvements in the understanding of proton structure.

Simulated events are processed through the `Geant4` [76] package to mimic the interaction of particles with the CMS detector. `Geant4` is a software toolkit used to simulate the passage of particles through matter, in this case through the various detector components of CMS.
To ensure that simulated events align with the actual conditions of collision data, fine-tuning is necessary. This is achieved by applying a set of weights determined from data and by considering the efficiencies associated with lepton and jet kinematic selections. To account for the efficiencies of lepton reconstruction, identification, and isolation criteria,

a technique known as the “tag-and-probe” method [77] is employed. This approach uses events where a Z boson decays into two leptons (either electrons or muons). This approach is employed by utilizing events in which a Z boson decays into two leptons (either electrons or muons). The fundamental idea behind this method is to have a well-understood “tag” lepton, which is identified with high efficiency and purity, typically originating from the Z boson decay. The other lepton in the event, referred to as the “probe”, is then studied to evaluate the performance of the reconstruction, identification, and isolation criteria. By comparing the properties of the well-known Z boson with those of the “probe” lepton, the lepton efficiencies can be estimated. These estimated efficiencies are then applied as corrections to simulated events.

Furthermore, since the simulated events are not filtered by trigger criteria, all events are kept. To mimic the effect of triggers in the simulations, “trigger weights” derived from data are used. These weights depend on the p_T and η of the lepton candidates.

Additionally, “pileup reweighting factors” are applied to make simulated events match the pileup distribution observed in the real data. These factors are determined for each year.

In events involving b-tagging, it is crucial to understand how efficiently jets originating from b quarks can be identified. To achieve this, analyses are conducted using data samples enriched with events containing b-jets. The tagging efficiency in these enriched samples is measured, and weights are then applied to MC events based on various factors, including jet kinematics, flavour, and the b-tagging discriminator value.

To fine-tune the agreement between simulations and real data, specific weights are applied to particular distributions. For example, adjustments are made to the p_T distributions of the $\mathrm{t\bar{t}}$ process and the p_T distribution of the Z boson in DY samples.

4.2.3 Signal Characteristics

The signal of this study is simulated at LO using `MadGraph5_aMC@NLO`. The parton shower and hadronisation processes are handled by `Pythia 8` [78], employing the dipole recoil scheme to capture the kinematic characteristics of the VBS process accurately. Contributions from top quarks are explicitly subtracted from diagrams to isolate the VBS signal.

Background sources originating from on-shell Higgs boson production mechanisms are simulated using `Powheg v2` at NLO precision in QCD, with further reweighting to achieve next-to-NLO (NNLO) accuracy in the NNLOPS scheme [79] for the gluon-gluon fusion (ggF) production mode. The QCD-induced $\mathrm{W^+W^-}$ background is generated at NNLO accuracy in QCD using `Powheg v2`, with the second jet simulated at LO accuracy [80]. The interference term between the purely EW signal and the QCD-induced background is minimal and is not considered in the analysis due to its negligible contribution.

A comprehensive list of the processes simulated, along with the associated MC generators, is provided in Table 4.2. Careful studies were conducted to evaluate the impact of different parton shower algorithms on this analysis results. A comparison of two different parton shower algorithms was performed within the `MadGraph5_aMC@NLO` framework: the standard `pythia` 8 shower with a “global” recoil scheme and an alternative scheme known as the “local” recoil scheme, often referred to as the “dipole recoil” scheme [81]. The key

	Process	Generator
Signal	VBS $\mathrm{W^+W^-}$ $jj \to 2l2\nu$ no top quarks	`MadGraph5_aMC@NLO` (LO)
Backgrounds	$\mathrm{t\bar{t}}$, tW	`Powheg v2`
	DY $\to \ell\ell$ + jets	`MadGraph5_aMC@NLO` (LO)
	DY $\to \tau\tau \to \ell\ell$ + jets	`MadGraph5_aMC@NLO`
	Gluon fusion $\mathrm{W^+W^-}$	`mcfm v7.0`
	QCD-induced $\mathrm{W^+W^-}$	`Powheg v2` (NNLO)
	$Z + 2$ jets EW	`MadGraph5_aMC@NLO`
	$W\gamma$	`MadGraph5_aMC@NLO` (LO)
	$Z\gamma$	`MadGraph5_aMC@NLO`
	WZ/γ^*	`Powheg v2`
	$WZ \to 2\ell 2q$	`MadGraph5_aMC@NLO`
	$ZZ \to 2\ell 2q, 2l2\nu, 4l$	`Powheg v2`
	$ZZ \to 2l2q$ (2017/2018)	`MadGraph5_aMC@NLO`
	ZZZ, WZZ, WWW, WWW	`MadGraph5_aMC@NLO`
	H production: gluon fusion, VBF, . . .	`Powheg v2`

Table 4.2: List of signal and background processes simulated with MC generators. If not specified, the generators have NLO accuracy in QCD. The Pythia 8 generator is used for parton shower and hadronization simulation.

distinction lies in how these schemes handle the recoil generated during parton splittings[1]. In the global recoil scheme, any particle involved in the scattering process may absorb the recoil generated during parton splittings. In contrast, the dipole recoil scheme offers a more natural description of the VBS process, where colour flow occurs only between an initial-state quark and a final-state quark [82]. The dipole recoil scheme aligns more intuitively with this physics. In fact, the dipole recoil scheme considers the colour flow to be predominantly between an initial-state quark and a final-state quark. This approach is more in line with the physics of VBS, where a limited number of quarks play a dominant role in the interaction. In contrast, the global recoil scheme allows any particle involved in the scattering process to absorb the recoil generated during parton splittings, which may not be as well-suited for the specific colour flow patterns and limited quark involvement in VBS. So, the dipole recoil scheme is considered more natural and appropriate for VBS because it aligns with the physics of the process, where colour flow primarily involves the high-energy quarks responsible for VBS interactions.

Moreover, the global recoil scheme gives rise to unphysical results for the third jet, as evidenced by the fact that, when the third jet observables are described at NLO, the huge enhancement disappears, and predictions align with those obtained using the dipole recoil. This observation emphasises the inadequacy of the global recoil scheme for accurately capturing the physics of the third jet in VBS interactions.

To assess the consequences of these choices, the impact on various observables was analysed. Figure 4.3 presents a visual comparison of some key variables of interest in signal samples generated with these two different parton shower settings. Among the variables reported in this picture, the so-called Zeppenfeld variable [83] ($Z_{\ell\ell}$) is defined as:

$$Z_{\ell\ell} = \frac{1}{2}|Z_{\ell_1} + Z_{\ell_2}|, \tag{4.1}$$

where

$$Z_{\ell_i} = \eta_{\ell_i} - \frac{1}{2}(\eta_{j_1} + \eta_{j_2}), \tag{4.2}$$

measuring the centrality of the dilepton system with respect to the tagging jets. The observables most affected are those related to the third jet. The dipole recoil scheme, by its nature, mitigates the production of the third jet. As a result, only minor differences are observed in inclusive variables such as m_{jj} and $\Delta\eta_{jj}$. This analysis primarily focuses on observables unrelated to the third jet. Therefore, the influence of the parton shower setting on this analysis remains relatively modest.

The comparison between the global and dipole recoil samples was conducted using selections similar to the signal region, as described in Table 4.3 of Section 4.4. All the final state flavour categories are merged applying criteria of $m_{\ell\ell} > 50$ GeV and $p_{T,\text{miss}} > 20$ GeV.

The 2017 and 2018 signal samples are generated through `MadGraph5_aMC@NLO` (v.2.6.5), while the 2016 one uses `MadGraph5_aMC@NLO` (v.2.4.2). The cross-sections differ slightly,

---

[1] Parton showers often involve the formation of colour dipoles between a scattered parton and a beam remnant. When these remnants are allowed to act as endpoints for dipoles and absorb the recoil during parton branchings, some fraction of the remnant energy needs to be considered. This energy redistribution is accounted for by adjusting the momentum fraction assigned to the incoming scattered parton.

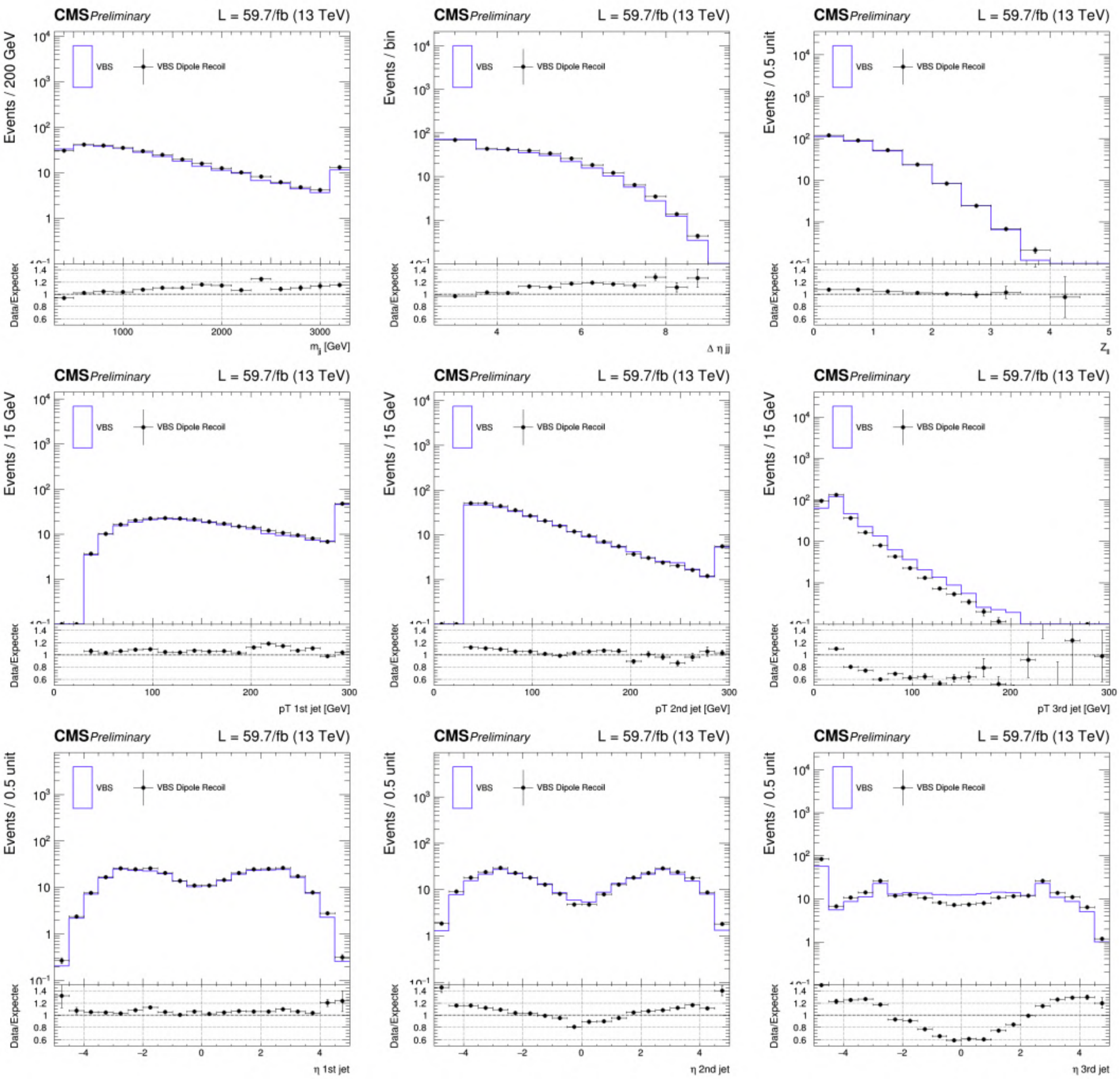


Figure 4.3: Comparison of parton shower algorithms applied to 2018 `MadGraph5_aMC@NLO` signal samples. The figure displays various observables: the dijet mass of the two tagging jets, the pseudorapidity gap of the two tagging jets, and $Z_{\ell\ell}$ (defined in Equation (4.1)) in the first row; the transverse momentum (p_T) of the first, second, and third jet in the second row; and the pseudorapidity (η) of the first, second, and third jet in the third row. The second panel in each plot presents the ratio between dipole recoil and global recoil signal samples. Underflow events and overflow events are folded in the first and last bin, respectively. The peak in the first bin of the η of the 3rd jet is due to events without a 3rd jet.

with values of 0.09283±0.00016 pb for samples produced with `MadGraph5_aMC@NLO` (v.2.4.2) and 0.09217 ± 0.00020 pb for those produced with `MadGraph5_aMC@NLO` (v.2.6.5).

Given the dipole recoil scheme superior alignment with the physics of this analysis, it was selected for generating the signal samples employed in this study.

4.3 Physical Objects Identification

The fundamental objective of the CMS PF [48], as discussed in Section 2.3, is to harness information from various CMS subdetectors to reconstruct kinematic observables and identify the particles involved in a collision event.

4.3.1 Electrons

Electron reconstruction is a critical process that involves connecting tracks in the silicon tracker to energy deposits in the ECAL. For consideration in the analysis, electrons must fall within the tracker acceptance region ($|\eta| < 2.5$). However, identifying prompt electrons [51] poses a challenge due to potential sources of background. These background sources include photon conversions, semileptonic decays of heavy quarks (such as b and c quarks), or jets with a significant electromagnetic component that can be mistakenly reconstructed as electrons.

To effectively distinguish prompt electrons from these background sources, a range of observables is employed. Calorimetric characteristics, such as the dimensions of the ECAL shower and the ratio of energy deposits between the ECAL and the HCAL, play an important role in discriminating between electromagnetic and hadronic showers. Hadronic showers tend to be more extensive and deeper than their electromagnetic counterparts. Considering information from reconstructed tracks allows for further discrimination between electrons and charged hadrons. Additionally, variables that compare the momentum measured in the tracker with the energy measured by ECAL, along with observables assessing the compatibility between the track and the ECAL supercluster position, contribute to electron identification.

Of particular importance is the concept of relative isolation, denoted as $\mathrm{ISO}_{\mathrm{rel}}$. This variable quantifies the isolation of the electron candidate and is defined as the sum of p_T of all PF candidates within a cone of radius $R = 0.3$ centred on the electron candidate position. The sum includes contributions from charged hadrons originating from the primary vertex, neutral particles (including hadrons and photons), and an estimate of neutral hadrons from pileup vertices, which is subtracted to account for pileup contributions. The relative isolation is defined as follows:

$$\mathrm{ISO}_{\mathrm{rel}} = \frac{\sum p_T^{\mathrm{charged}} + \max(0, \sum p_T^{\mathrm{neutral}} + \sum p_T^{\gamma} - \frac{1}{2}\sum p_T^{\mathrm{PU}})}{p_T}. \tag{4.3}$$

Here, $\sum p_T^{\mathrm{charged}}$, $\sum p_T^{\mathrm{neutral}}$, and $\sum p_T^{\gamma}$ represent the p_T sums of charged hadrons from the primary vertex and neutral particles (comprising hadrons and photons) within the isolation cone. Instead, $\sum p_T^{\mathrm{PU}}$ denotes an estimate of the contribution from hadrons originating from pileup vertices. The factor $\frac{1}{2}$ approximately to the ratio of neutral particle to charged hadron production in inelastic proton-proton collisions, as estimated from simulation [48]. To optimise electron identification, a single discriminator variable employing a multivariate technique, such as a Boosted Decision Tree (BDT), is trained. This BDT combines information from various observables, including $\mathrm{ISO}_{\mathrm{rel}}$, to generate a single discriminator output [51]. By selecting different threshold values for this discriminator output (working points), one can achieve different levels of signal efficiency and background rejection. In the analysis, a working point is used to ensure a 90% signal efficiency. Moreover, it is essential to apply a stringent cut on the $\mathrm{ISO}_{\mathrm{rel}}$ of the electron candidate, requiring it to be less than 6%. This measure helps preventing confusion between electrons and other particles, such as jets, or real electrons originating from semileptonic decays of heavy quarks within a jet.

4.3.2 Muons

Muon identification relies on reconstructing muons using information from both the inner tracker and the outer muon system, as elaborated in Section 2.3. In the analysis, muon candidates must meet specific criteria, which ensure the high-quality muons reconstruction while minimising contamination from other sources. First of all, the muon must have $p_T > 10\,\mathrm{GeV}$ and fall within the region of $|\eta| < 2.4$.
Moreover, stringent selection criteria, as described in Ref. [22], are applied. These criteria encompass requirements on the number of hits in the muon tracks, such as a minimum of six hits in the tracker layers, one hit in the pixel detector, and two hits in the muon system. This criterion ensures that the muon candidate trajectory is well-reconstructed and consistent across different detector components. Additionally, to assess the goodness of fit for the muon track, the ratio of the track fit $\frac{\chi^2}{\mathrm{dof}}$ was examined. This ratio must be less than 10, indicating that the track fit provides a satisfactory representation of the muon path through the detectors. To reduce contamination from muons originating from secondary vertices, constraints on the muon impact parameter with respect to the primary vertex are imposed. Specifically, the impact parameter on the transverse plane ($|d_{xy}|$) must be less than 0.2 cm, while the impact parameter along the longitudinal direction ($|d_z|$) must be less than 0.5 cm. These requirements ensure that the muon trajectory in both the transverse and longitudinal direction is closely associated with the primary vertex, reducing the likelihood of it coming from a secondary vertex.

In the analysis, even stricter conditions on the impact parameters are implemented for certain p_T ranges to enhance the purity of the muon selection:

$|d_{xy}| < 0.01$ cm for $p_T < 20\,\mathrm{GeV}$

$|d_{xy}| < 0.02$ cm for $p_T \geq 20\,\mathrm{GeV}$

$|d_z| < 0.1$ cm

Similar to electrons, the concept of relative isolation is applied to muons. It involves summing the transverse momenta of PF candidates within a cone of radius $R = 0.4$ centred on the muon direction. It is essential to set a condition on the ISO_{rel} to be less than 15% to exclude real muons within jets and to mitigate the risk of mistaking jets as muons.

Additionally, an MVA discriminator is employed [84]. A threshold cut at an output value greater than 0.8 is applied in conjunction with all previously described selections.

4.3.3 Jets

Jets are crucial objects in particle physics, representing collimated sprays of particles originating from the hadronisation process. In the analysis, jet reconstruction begins with the grouping of PF candidates using the anti-k_T [53, 54] algorithm with a distance parameter of $R = 0.4$. This choice of algorithm ensures robustness against issues related to collinear and soft divergences. Charged particles from non-primary vertices are removed before the clustering through the Charged Hadron Subtraction (CHS) method.

Jets considered for analysis must satisfy certain criteria, including falling within the pseudorapidity range of $|\eta| < 4.7$ and having $p_T > 30\,\text{GeV}$. To account for possible discrepancies between data and MC events, jet energy corrections (JEC) are applied to calibrate the jet energies. Furthermore, jet energy resolution (JER) scale factors are employed to enhance jet transverse momentum resolution. These JER scale factors effectively smear the jet p_T only in 2017 and 2018 MC events, since in 2016 they led to a worsening of data-MC agreement.

Jet reconstruction also encounters background sources that arise either from electronic noise in the calorimeters (noise jets), or from particles originating from non-primary vertices (pileup jets). To distinguish between physically meaningful jets and these spurious contributions, identification criteria based on jet-related observables are used. These observables include the fractions of hadronic (both charged and neutral) and electromagnetic energies within the jet, as well as the multiplicity of PF candidates contained within the jet. To maintain high-quality jet reconstruction, all jets with $p_T > 30\,\text{GeV}$ must pass the tight working point of the jet identification criteria, as defined in [85]. Additionally, a Boosted Decision Tree (BDT) trained on jet shape and tracking observables [86] is employed. This BDT plays a crucial role in discriminating against pileup jets.

For jets with $p_T < 50\,\text{GeV}$, a specific requirement to satisfy the loose working point of the pileup jet identification [85] is introduced. Furthermore, selected jets must not overlap with previously identified leptons within a distance ΔR of 0.4.

4.3.4 B-tagged Jets

The identification of b jets is of paramount importance in the analysis, particularly for the rejection of $\text{t}\bar{\text{t}}$-tW background events. Various b tagging algorithms exist, each exploiting different aspects of a jet characteristics (track information, secondary vertex information, and soft lepton information) to produce an output discriminator, which indicates the likelihood of a jet originating from a b quark.

These algorithms typically recommend three working points: loose, medium, and tight. These working points offer varying b tagging efficiencies and misidentification probabilities for light-flavor jets, with the tight working point providing the highest purity. In particular, the three misidentification probabilities are around 10%, 1%, and 0.1%, respectively. In the analysis, jets are defined as b jets if they surpass the loose working point of the DeepJet algorithm [87, 88], an algorithm based on a neural network. The specific discriminator values corresponding to the loose working point vary for different years: 0.0614, 0.0521, and 0.0494 for 2016, 2017, and 2018, respectively.

4.3.5 Missing Transverse Momentum

The concept of missing transverse momentum (p_T^{miss}) [56] plays a crucial role in many physics analyses, including ours. It quantifies the imbalance in transverse momentum in an event, indicative of particles that escape detection or measurement. To calculate p_T^{miss}, the negative vector sum of transverse momenta of all PF candidates reconstructed in the event is considered.

To improve the accuracy of p_T^{miss} and reduce contributions from pileup interactions, the Pileup Per Particle Identification (PUPPI) algorithm [89] is employed. PUPPI assigns a weight to each PF candidate based on the likelihood that it originates from the primary interaction vertex. The weight assigned to each PF candidate by the PUPPI algorithm is subsequently used to adjust the contribution of each candidate in the calculation of p_T^{miss}. Specifically, candidates with higher weights are given more importance in the p_T^{miss} calculation because they are more likely to be associated with the primary interaction of interest and less likely to be due to pileup, thereby enhancing the accuracy of p_T^{miss} by reducing the influence of pileup interactions on the measurement. This approach results in better resolution and enhances the agreement between data and simulated events. Furthermore, to address discrepancies between data and MC events, the JEC are propagated to the PUPPI-based p_T^{miss}, ensuring that the same corrections applied to jet energy measurements are also considered in the p_T^{miss} calculation.

4.4 Selections

In this section, the fundamental selection criteria, referred to as preselections, applied to identify events with a VBS-like signature are detailed. Such events are characterised by the presence of two oppositely charged leptons (either electrons or muons), at least two jets, and a certain amount of missing transverse energy (p_T^{miss}). These first selection criteria for the analysis are detailed as follows:

- Set thresholds for the transverse momenta (p_T) of the leading and trailing leptons to 25 GeV and 13 GeV, respectively. Reject any event featuring an additional loosely identified lepton with $p_T > 10$ GeV to minimise contamination from multiboson interactions.

Observable	Selection
Leptons pairs	e^+e^-, $\mu^+\mu^-$, $e^+\mu^-$, $e^-\mu^+$
p_T of leading lepton	$p_{T,\ell_1} > 25\,\text{GeV}$
p_T of trailing lepton	$p_{T,\ell_2} > 13\,\text{GeV}$
p_T of additional leptons	$p_{\ell_3} < 10\,\text{GeV}$
p_T of VBS jets	$p_T^{VBS} > 30\,\text{GeV}$
Missing transverse momentum	$p_T^{\text{miss}} > 20\,\text{GeV}$
Invariant mass of leptons pair	$m_{\ell\ell} > 50\,\text{GeV}$
p_T of leptons pair	$p_{\ell\ell} > 30\,\text{GeV}$
Dijet invariant mass	$m_{jj} > 300\,\text{GeV}$
Pseudorapidity gap	$\Delta\eta_{jj} > 2.5$

Table 4.3: Preselections applied to all events to define a loose VBS-like phase space.

- Select events that contain at least two jets with $p_T > 30$ GeV. Identify the two jets with the highest p_T as the "VBS jets" since these are the jets most likely originating from the VBS process (see Figure 4.1).
- Set a threshold on p_T^{miss} to 20 GeV to account for the presence of neutrinos.

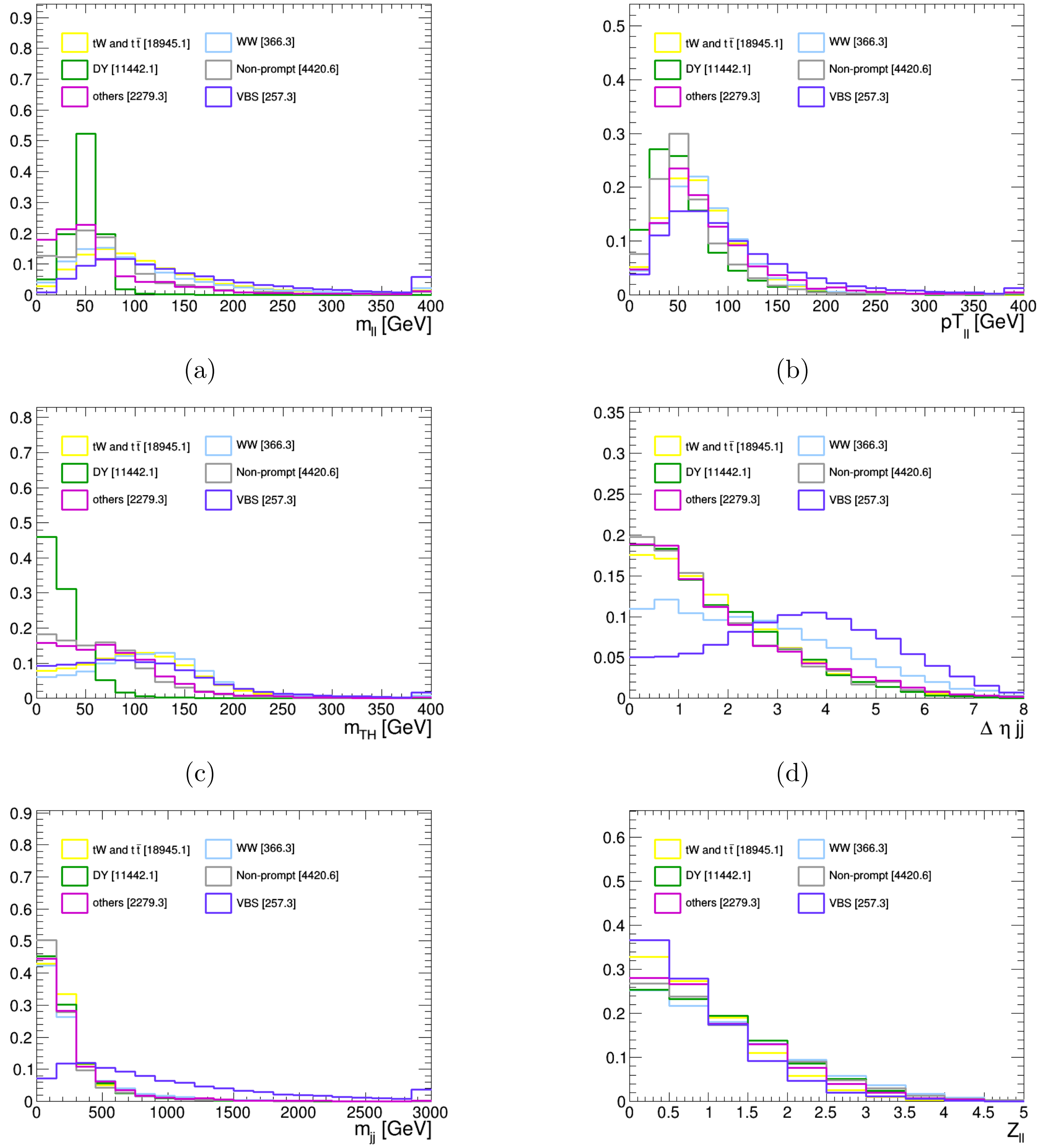
The shape comparison of several kinematic variables for the main backgrounds and the signal is reported in Figure 4.4 for different flavour final state and in Figure 4.5 for same flavour final state, with all the previously mentioned cuts applied.

To further reduce background contamination, additional selection criteria are imposed:

- Require the dilepton invariant mass ($m_{\ell\ell}$) to be greater than 50 GeV to minimise contamination from Higgs boson decays to two W bosons (H→WW).
- Impose a minimum requirement on the dilepton transverse momentum ($p_{T,\ell\ell}$) of at least 30 GeV to reduce the presence of leptons originating from DY production.
- Apply a requirement that the invariant mass of the VBS jets (m_{jj}) be greater than 300 GeV.
- Demand that the pseudorapidity gap between the VBS jets ($\Delta\eta_{jj}$) must be at least 2.5.

These selections on the last two observables effectively reduce contamination from various background sources, in particular from the QCD-induced background. Further details and the rationale behind these criteria can be found in Section 1.2.1.
A summary of all the preselection criteria is presented in Table 4.3.

In addition to these preselections, further requirements to optimise the signal-to-background ratio in the signal region are applied, as explained in the following paragraphs.

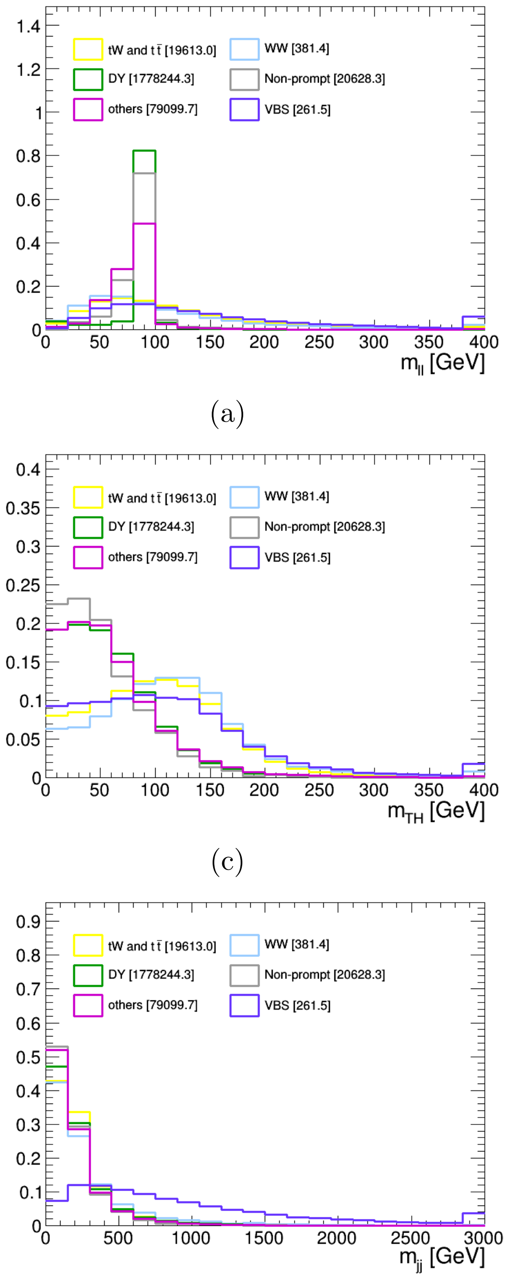


(a) (b)

(c) (d)

(e) (f)

Figure 4.4: Shape comparison of the major backgrounds and signal processes for different flavour final state ($e\mu/\mu e$) in 2018. The comparison is performed in the signal region.



(a) (b)

(c) (d)

(e) (f)

Figure 4.5: Shape comparison of the major backgrounds and signal processes for same flavour final state ($ee/\mu\mu$) in 2018. The comparison is performed in the signal region.

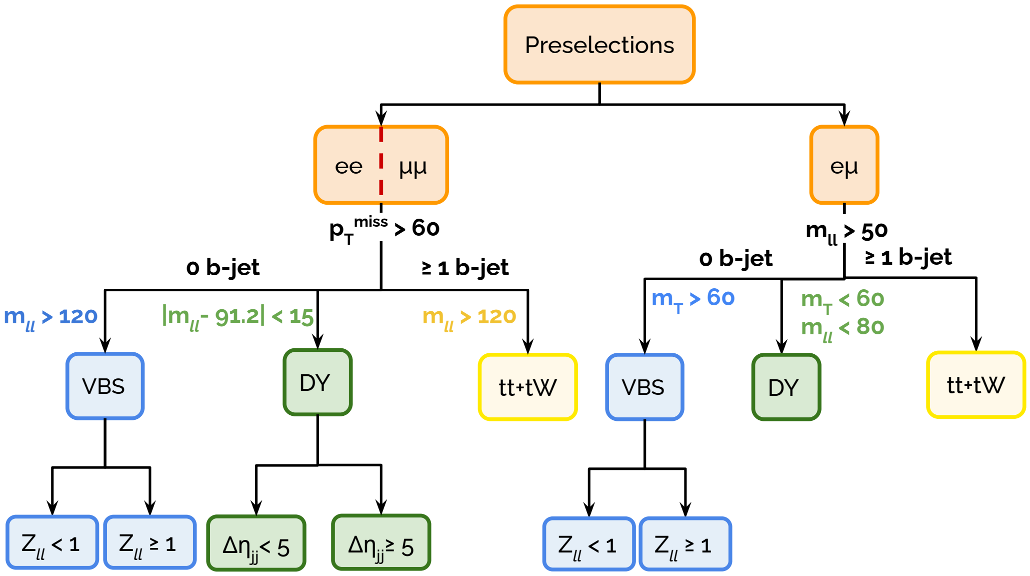


Figure 4.6: Summary of phase space categorisation, starting from the preselections listed in Table 4.3. In this chart, "VBS" depicts the signal region, while "DY" and "top" the two respective control regions.

A visual representation of these selections is depicted in Figure 4.6. Within preselected events, two control regions are defined, aiding in studying the agreement between data and simulation, and providing constraints for normalising top quark processes and DY production. Control regions are enriched in either of these two backgrounds through carefully chosen selections that ensure orthogonality with the signal region. Each control region is further categorised based on lepton flavour: two electrons (ee), two muons ($\mu\mu$), or one electron and one muon ($e\mu$).

The signal region is primarily defined by excluding the presence of b-jets (b veto) to minimise contamination from $\mathrm{t\bar{t}}$-tW events. In the $e\mu$ category, an additional requirement is imposed on the transverse mass (m_T), formed by the combination of $p_{T,\ell\ell}$ and p_T^{miss}, to be above 60 GeV. In the ee and $\mu\mu$ categories, no specific selection is applied to m_T, but more stringent cuts on p_T^{miss} and $m_{\ell\ell}$ are imposed. Specifically, the p_T^{miss} threshold is raised to 60 GeV, and $m_{\ell\ell}$ is required to be greater than 120 GeV to discard events associated with Z boson production.

To further enhance sensitivity and optimise significance, the signal region is divided into two subregions. This categorisation is based on the dilepton Zeppenfeld variable ($Z_{\ell\ell}$), defined in Equation (4.1). Categories with $Z_{\ell\ell} < 1$ are enriched in signal events and experience reduced background contamination, as illustrated in Figure 4.7.

Control regions dedicated to $\mathrm{t\bar{t}}$-tW events are defined in a manner similar to the signal regions, but with the b veto inverted, requiring the presence of at least one b-jet with

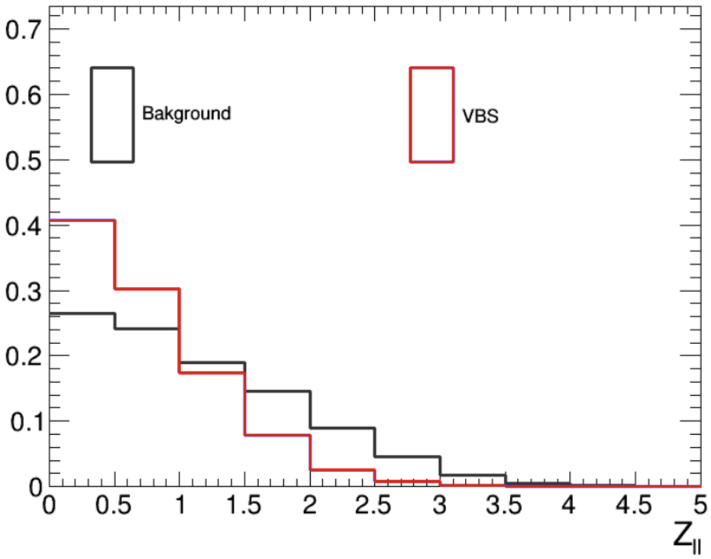


Figure 4.7: Normalised distributions of the VBS signal (red line) and the background contributions (black line) in the $Z_{\ell\ell}$ variable, as defined in Equation (4.1). The plots refer to the 2018 data sample in the $e\mu$ final state.

$p_T > 20\,\text{GeV}$ in the final state. In these control regions, the fraction of $\mathrm{t\bar{t}}$-tW processes is approximately 95%.

For the control regions targeting DY events, the b veto is applied, similar to the signal region, to maintain orthogonality with the $\mathrm{t\bar{t}}$-tW control region. Different selections are employed for $e\mu$ and $ee/\mu\mu$ final states to ensure orthogonality with their respective signal regions. In the $e\mu$ DY category, the m_T cut is reversed compared to the $e\mu$ signal region and introduce a window of $50\,\text{GeV} < m_{\ell\ell} < 80\,\text{GeV}$.

For the ee and $\mu\mu$ final states, $m_{\ell\ell}$ is restricred to a region near the Z boson mass peak ($|m_{\ell\ell} - m_Z| < 15\,\text{GeV}$). Additionally, in the ee and $\mu\mu$ final states, the DY control region is further divided into two $\Delta\eta_{jj}$ bins, each enriched in a different component of the DY background, as explained in Section 4.7. The fraction of DY events in the $e\mu$ DY control regions is approximately 64%, while it reaches approximately 91% in the ee and $\mu\mu$ categories.

Table 4.4 provides an overview of the various analysis categories and their respective selection criteria. As thoroughly explained in Sections 4.8 and 4.9, the observed significance is extracted including all the regions in the likelihood profiling such that the fit jointly determines the background contribution importance in the signal region in the form of nuisances. In particular, the normalisation of the top and DY control regions are derived by the fit, while their shape is taken from simulation.

VBS	eμ/μe	$Z_{\ell\ell} < 1$	$m_T > 60\,\text{GeV}$ $m_{\ell\ell} > 50\,\text{GeV}$ no b-jet with $p_T > 20\,\text{GeV}$
		$Z_{\ell\ell} \geq 1$	
	ee	$Z_{\ell\ell} < 1$	$m_{\ell\ell} > 120\,\text{GeV}$ $p_T^{\text{miss}} > 60\,\text{GeV}$ no b-jet with $p_T > 20\,\text{GeV}$
		$Z_{\ell\ell} \geq 1$	
	$\mu\mu$	$Z_{\ell\ell} < 1$	
		$Z_{\ell\ell} \geq 1$	
top	eμ/μe		$m_{\ell\ell} > 50\,\text{GeV}$ at least one b-jet with $p_T > 20\,\text{GeV}$
	ee		$m_{\ell\ell} > 120\,\text{GeV}$ $p_T^{\text{miss}} > 60\,\text{GeV}$ at least one b-jet with $p_T > 20\,\text{GeV}$
	$\mu\mu$		
DY	eμ/μe		$m_T < 60\,\text{GeV}$ $50\,\text{GeV} < m_{\ell\ell} < 80\,\text{GeV}$ no b-jet with $p_T > 20\,\text{GeV}$
	ee	$\Delta\eta_{jj} < 5$	$\lvert m_{\ell\ell} - m_Z \rvert < 15\,\text{GeV}$ $p_T^{\text{miss}} > 60\,\text{GeV}$ no b-jet with $p_T > 20\,\text{GeV}$
		$\Delta\eta_{jj} \geq 5$	
	$\mu\mu$	$\Delta\eta_{jj} < 5$	
		$\Delta\eta_{jj} \geq 5$	

Table 4.4: Summary of phase space categorisation, following the preselections in Table 4.3. The signal region ("VBS") is divided into six sub-categories based on flavour composition in the final state and dilepton centrality with respect to the two jets.

4.5 DNN for the $e\mu$ Category

In the $e\mu$ signal region, a feed-forward Deep Neural Network (DNN) serves as a discriminator to distinguish signal events from background. The primary objective is binary classification, categorising events as either signal (1) or background (0). Consequently, the DNN produces a continuous output ranging from 0 to 1. Its training focuses on identifying the VBS signal in the presence of two major backgrounds: the $\mathrm{t\bar{t}}$ process and QCD-induced $\mathrm{W^+W^-}$ production. Within the $e\mu$ signal region (as outlined in Section 4.4), two distinct models were developed and trained optimised for different phases of $Z_{\ell\ell}$ – one for $Z_{\ell\ell} < 1$ and another for $Z_{\ell\ell} \geq 1$. Both models share an identical architectural framework and are provided with nine carefully selected discriminating variables. To increase the signal discrimination power, both simple kinematics variables of the events and more complex variables with promising discrimination results constructed from the event physics objects were used.

In the following sections, a brief overview of the DNN concept is provided, and the specifics of the two models used in the analysis are delved into.

4.5.1 Deep Neural Networks

The fundamental unit of a neural network is the perceptron, a computational node that receives N inputs and produces a single output through a weighted linear combination, governed by a chosen activation function. The output of a perceptron is mathematically expressed as:

$$y_j(\vec{x}) = \phi\left(\sum_{i=1}^{N} w_{ij}x_i + b_j\right). \tag{4.4}$$

In this equation, x_i represents the input values, w_{ij} are the weights associated with each input, b_j denotes the bias, and ϕ signifies the selected activation function. The network parameters encompass the weights and biases of all individual perceptrons.

DNNs can adopt versatile architectures customised for specific tasks. For binary classification tasks, the widely employed structure is the feed-forward neural network. This architecture organises neurons into layers, including an input layer, one or more hidden layers, and an output layer. Information flows sequentially from the input layer through the hidden layers before reaching the final output layer. In binary classification scenarios, the output layer comprises a single node, and its result serves as the classification test statistic.

The architecture of a DNN is primarily determined by defining several tunable hyperparameters that must be set before the training process commences. Hyperparameters play a pivotal role in configuring DNNs. It is essential to distinguish these hyperparameters from the parameters introduced in Equation (4.4) (i.e., the weights and biases of perceptrons). The latter are optimised during the training process, with initial values assigned randomly. The three key hyperparameters mentioned thus far include the activation functions for each layer, the number of layers, and the number of neurons per layer.

Additional hyperparameters, some of which have been optimised in the context of this project, will be elaborated upon below. Unlike parameters, which evolve during training, hyperparameters remain fixed throughout the process, profoundly influencing the network performance and behaviour.

Activation Functions

Activation functions are fundamental components of neural networks that introduce non-linearity, enabling these models to capture intricate data relationships. They act as mathematical "gates" attached to each neuron, determining whether the neuron should be activated based on the relevance of its input to the model prediction. The choice of activation functions plays a pivotal role in neural network design.

Two widely used activation functions are the Rectified Linear Unit (ReLU) and the Sigmoid function, as depicted in Figure 4.8. ReLU is defined as $f(x) = \max(\mathrm{x}, 0)$, and is favoured for input and hidden layers due to its simplicity and effectiveness. It transforms inputs by outputting the input value for positive inputs and zero for negative inputs, providing linear behaviour for positive values. This characteristic helps mitigate the vanishing gradient problem, which arises when gradients become very small during backpropagation in DNNs, hindering learning.

To understand this, consider that when the input to a ReLU neuron is positive, it allows the gradient to flow backward unchanged during backpropagation. In contrast, traditional activation functions like the sigmoid can squash their input into a small range, resulting in very small gradients (close to zero) in the backpropagation process for positive or negative inputs, respectively. These small gradients can cause the vanishing gradient problem, making it difficult for the network to update the weights effectively in deep layers.

For the output layer, the Sigmoid activation function is commonly used. It maps inputs to values between 0 and 1, making it suitable for binary classification tasks. The Sigmoid function produces probabilities, which can be interpreted as the likelihood of an event belonging to a particular class. This makes it well-suited for producing the final classification output in neural networks.

4.5.2 Training and Validation

In supervised learning, DNNs are trained using labeled datasets containing both signal and background events. The aim is to adjust the neural network parameters, such as the weights and biases of each perceptron, by minimising a loss function. This loss function quantifies the disparity between predicted scores and actual labels. For instance, predicting a score close to 0 when the true label is 1 results in a high loss value. Consequently, achieving optimal classification performance involves minimising this loss function. In essence, during training, the network predictions for input events are compared to their true labels to evaluate prediction errors and fine-tune the DNN parameters.

A commonly used loss function for binary classification tasks is the binary cross-entropy

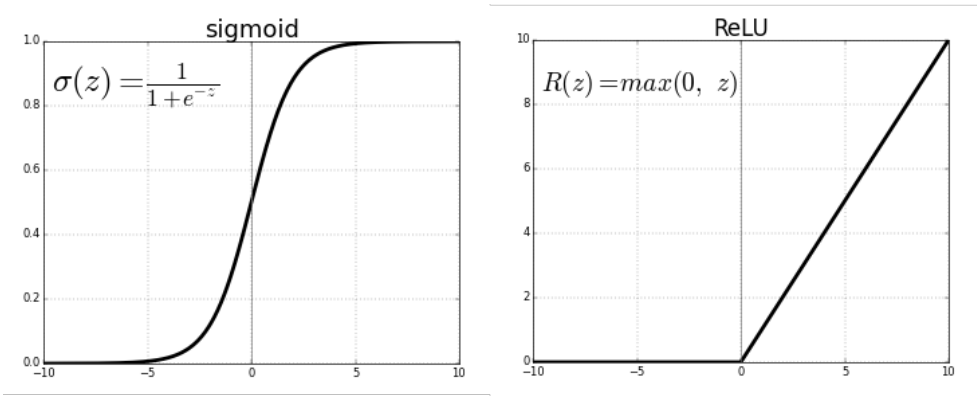


Figure 4.8: Activation functions in Neural Networks: Rectified Linear Unit (ReLU) for input and hidden layers (right) and Sigmoid for the output layer (left).

loss function, expressed as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \cdot \log(\hat{y_i}) + (1 - y_i) \cdot \log(1 - \hat{y_i})\Big]. \tag{4.5}$$

In this equation, N represents the total number of events, y_i denotes actual event labels (0 for background, 1 for signal), and $\hat{y_i}$ represents continuous predicted values.

During training, parameters are updated using the backpropagation, a procedure that propagates corrections from the output node back to previous layers, and in which parameters are adjusted proportionally to their impact on the final results. One pivotal hyperparameter is the learning rate (η), which controls the magnitude of parameter changes in each iteration. The update formula for each parameter θ_{ij} is given by:

$$\theta_{ij} \to \theta_{ij} - \eta \cdot \frac{\partial L}{\partial \theta_{ij}}. \tag{4.6}$$

Here, L signifies the loss function, and η is the learning rate. The gradient $\frac{\partial L}{\partial \theta_{ij}}$ measures each weight influence on the final outcomes.

Training occurs over several iterations, or epochs. This gradient update is performed at the end of each epoch, and the total number of epochs that a network must perform before finishing the training process is a customisable hyperparameter. However, the network does not use the entire training dataset in each gradient update. Instead, the dataset is divided into batches, and the "batch size" is another tunable hyperparameter. When all the samples have been used, the dataset is shuffled and another epoch of the training is started with new batches.

Training continues until the loss converges to a small value. Subsequently, the classifier performance is assessed using an independent dataset, comprising both background and signal samples.

Another crucial hyperparameter requiring careful selection is the choice of metrics. Metrics offer numerical insights into the performance of the DNN model, playing a pivotal role in both its evaluation and optimisation. Further details regarding the metrics are elaborated in Section 4.5.7.

Every hyperparameter described (activation function, number of layers, node count per layer, loss function, learning rate, batch size, metrics) requires careful selection to determine the optimal network configuration that yields the best performance.

Overtraining

In the context of neural networks, the term “deep” is used to describe networks with numerous nodes per layer and more than two hidden layers. Increasing the number of parameters in a neural network enhances its capacity to tackle complex classification tasks by strengthening its approximation capabilities. However, training DNNs becomes a time-consuming and challenging process, necessitating careful consideration of the issue of overtraining. Overtraining can be thought of as the network learning the details of the training sample. This phenomenon can be detected by assessing the performance gap between the training dataset and the one determined as an independent validation dataset. Typically, the training loss continues to decrease with each epoch, while the validation loss initially reaches a minimum and then begins to rise. To attain optimal performance, a prudent training technique called “early stopping” [90] is employed, which involves halting the training when the validation reaches its minimum and plateaus.

Another effective approach to mitigate the risk of overtraining is known as “dropout”. The fundamental idea behind dropout is to randomly deactivate nodes along with their connections within the neural network during each training step. This technique significantly reduces overtraining. The dropout fraction is another hyperparameter that needs to be tuned.

Batch Normalisation and L2 Regularisation

Batch normalisation [91] plays a crucial role in DNNs by normalising the input of a layer before applying the activation function. This normalisation ensures consistent scaling and distribution of inputs during training, addressing two major challenges: vanishing and exploding gradients. When gradients become extremely small during backpropagation, learning becomes slow or halts altogether. Batch normalisation combats this issue by maintaining activations within a reasonable range, preventing them from diminishing too drastically. As a result, gradients are less likely to vanish, allowing the network to learn effectively. Conversely, exploding gradients can destabilise training when gradients become exceptionally large. Batch normalisation acts as a safeguard, keeping activations stable and preventing sudden spikes. This, in turn, limits the growth of gradients, contributing to a stable training process. In addition to its gradient-controlling capabilities, batch normalisation enhances generalisation by restraining overtraining. Furthermore, it accelerates training, often reducing the number of required epochs for convergence.

L2 regularisation, also known as weight decay, is a regularisation technique applied to the network weights. It works by adding a penalty term to the loss function, discouraging the model from learning overly complex patterns in the training data. This penalty term is represented as:

$$\text{L2 Regularisation Term} = \lambda \sum_i w_i^2. \tag{4.7}$$

Here, λ is the hyperparameter that controls the strength of the regularisation, and w_i represents individual weights in the network. As a result, L2 regularisation encourages the network to keep its weight values small, effectively limiting the network capacity to fit noise in the training data. This regularisation technique is particularly useful in preventing overtraining. Applying L2 regularisation helps strike a balance between fitting the training data well and maintaining a model that can generalise effectively to new, unseen data.

Optimiser Algorithms

Optimiser algorithms play a critical role in training DNNs by iteratively adjusting their attributes, such as weights and learning rates, to minimise loss functions. This study explores two different optimisers: Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (Adam) [92]. Below, a concise explanation of these algorithms is provided, along with a discussion of the customisation of the learning rate values employed during the analysis with these two optimisers.

Stochastic Gradient Descent (SGD) is a widely-used optimisation method that aims to minimise the loss function L by updating the model parameters θ in the opposite direction of the gradient of the loss function with respect to the parameters. In essence, the SGD algorithm follows the slope of the loss function surface until it converges to the minimum. In particular, SGD performs a parameter update for each training sample $\vec{x}_i$ and label y_i as follows:

$$\theta \to \theta - \eta \cdot \nabla_\theta L(\theta|\vec{x}_i, y_i). \tag{4.8}$$

SGD is favored for its efficiency, as it updates the model parameters rapidly.

During the optimisation of hyperparameters, it was observed that achieving optimal performance with the SGD optimizer often requires the use of a learning rate scheduler. This scheduler adjusts the learning rate during the training process, enhancing convergence. The "step-decay" learning rate scheduler, for example, reduces the learning rate according to a predefined schedule. If the initial learning rate is 0.1 and it is halved every two epochs, the function for this learning rate scheduler can be defined as:

$$\eta = 0.1 \cdot 0.5^{\left\lfloor \frac{1+\text{epoch}}{2} \right\rceil}. \tag{4.9}$$

On the other hand, Adam is an optimisation algorithm that adapts the learning rate for each parameter individually. It performs smaller updates (i.e., low learning rates) for parameters associated with frequently occurring variables and larger updates (i.e., high learning rates) for parameters linked to infrequent variables.

While SGD updates all parameters θ simultaneously with a uniform learning rate η, Adam employs a unique learning rate for each parameter θ_i and computes adaptive learning rates based on the historical gradient information.

4.5.3 Preparation of Samples

In the training phase of the DNN, simulated events for both signal and background processes, specifically $\mathrm{t\bar{t}}$ and QCD-induced $\mathrm{W^+W^-}$ production, are used. Details of the simulated samples used are provided in Table 4.5.
To ensure the reliability of the model, the training samples differ from those employed in the analysis (refer to Section 4.2.2). Data from the years 2016, 2017, and 2018, are incorporated weighting each sample according to its integrated luminosity. Merging these samples expands the training dataset. A single neural network is then constructed for all three years for both $Z_{\ell\ell} < 1$ and $Z_{\ell\ell} \geq 1$ models.

The training dataset is divided into two subsets: a training data set (80%) and a validation data set (20%). The training data set is used to train the DNN, while the validation data set monitors the training process.

During training, each event is assigned a weight based on its cross-section, integrated luminosity, and corrections, as explained in Section 4.2.2. The loss computed for each sample is multiplied by the weight associated with it. In this way, the back propagation will behave differently depending on the weight of the events, giving more importance to the events with a higher weight.
However, after applying these weights, background events ($\mathrm{t\bar{t}}$ and QCD-induced $\mathrm{W^+W^-}$) become more influential due to their higher cross-sections.

To mitigate this bias, the goal is to balance the number of weighted events in the signal and background datasets. Specifically, the aim is to have the total number of weighted events in the signal dataset match the combined total of the background datasets. To achieve this balance, the weights assigned to the signal samples during training are increased, using the following as the new weight:

$$\text{signal weight} = \frac{\text{weight}}{\text{mean(weights)}}. \tag{4.10}$$

While to balance the background, a weight computed as follows is employed:

$$\text{background weight} = \frac{\text{weight}}{\sum \text{weights}} \cdot \mathrm{n_S}. \tag{4.11}$$

In these equations, weight represents the individual weights of each event, accounting for the cross-section, luminosity, and corrections. The summation $\sum$ weights is performed over the entire set of unweighted signal or background events. Meanwhile, n_S denotes the number of unweighted simulated signal events that meet the selection criteria (refer to Section 4.4 for details).

4.5.4 Optimisation of Phase Space Selection

In the quest to define the most suitable phase space for training the DNN, various sets of selections were investigated. The optimisation process began with a stricter phase space than the one detailed in Section 4.4, requiring the following criteria:

Process	Generators
VBS $W^+W^-jj \rightarrow 2l2\nu$	`MadGraph5_aMC@NLO`
QCD-induced $W^+W^- \rightarrow 2l2\nu$	`powheg v2`
$t\bar{t}$+ jets (2016)	`MadGraph5_aMC@NLO` (LO)
$t\bar{t}$ (2017)	`powheg v2`
$t\bar{t}$ (2018)	`MadGraph5_aMC@NLO`, `powheg v2`

Table 4.5: Simulated samples used in the DNN training. The table lists the simulated processes and the corresponding generators. All simulated samples are interfaced with Pythia 8 for parton shower and hadronisation.

$$m_{jj} > 500 \text{ GeV}$$

$$\Delta\eta_{jj} > 3.5$$

The selections on m_{jj} and $\Delta\eta_{jj}$ were then gradually relaxed to expand the number of events included in the phase space used for DNN training. Enlarging the phase space allowed for more complexity in the neural network architecture, including additional nodes, layers, and input variables. Consequently, the neural network gained greater information processing capabilities, resulting in improved signal-to-background discrimination.

Specifically, the m_{jj} threshold was decreased in intervals of 100 GeV and relaxed $\Delta\eta_{jj}$ by 0.5 for every 100 GeV decrease in m_{jj}. Ultimately, the optimised phase space selections for the DNN analysis were set as follows:

$$m_{jj} > 300 \text{ GeV}$$

$$\Delta\eta_{jj} > 2.5$$

As a result, the training phase space matched the signal region described in Section 4.4. Further lowering the thresholds of these two variables did not yield any significant improvement in the network performance.

4.5.5 Optimisation of Architecture and Input Variables

For the analysis, two separate DNN models were trained, one for the $Z_{\ell\ell} < 1$ phase space and another for the $Z_{\ell\ell} \geq 1$ phase space. The results were extracted through a differential likelihood fit, as detailed in Section 4.9.

Creating the DNNs required careful consideration to determine the optimal input variables and the best architecture. Increasing the network dimensions, such as the number

of layers and nodes, or adding more input variables can enhance the model discriminatory power. However, an overly complex structure may lead to overtraining.
To find the ideal configuration, initiation began with a minimal DNN consisting of 2 hidden layers with 20 neurons each and only 2 input variables. Subsequently, experimentation involved varying architectures (adjusting the number of nodes and layers) while maintaining a constant number of input variables, aiming to identify the most effective configuration.

After optimising the network structure, the introduction of additional variables was undertaken to enhance its performance. A comprehensive set of observables (refer to Table 4.6), including kinematic and angular variables of leptons and jets, were considered as potential inputs for the DNN.
To guide the selection of variables, differences in correlation matrices between signal and background for these observables were assessed, as illustrated in Figure 4.9. The DNN relies on these differences to distinguish signal events from the background, making variables with pronounced differences, like Z_{ℓ_1} or Z_{ℓ_2}, preferable. New variables were iteratively added until no further improvement in separation power was achieved.

The performance of each model was assessed by comparing the ROC curve obtained with the DNN evaluated on the analysis samples and that of m_{jj}, keeping into account the different efficiencies of the selections.

The optimisation process initially focused on the challenging $Z_{\ell\ell} \geq 1$ region due to limited signal events. Surprisingly, the same architecture that performed well in the $Z_{\ell\ell} \geq 1$ phase space also proved to be the best choice for the $Z_{\ell\ell} < 1$ phase space, utilising the same input variables.

4.5.6 Final Architecture and Input Variables

Both DNN models, corresponding to the $Z_{\ell\ell} < 1$ and $Z_{\ell\ell} \geq 1$ phase spaces, were implemented using Keras [93]. An identical structure is employed for both models, featuring 5 hidden layers with 128, 128, 64, 64, and 64 neurons, respectively. A schematic representation of the network can be found in Figure 4.10. Given that the primary task of these networks is binary classification, the most appropriate choice for the loss function is binary cross-entropy, defined in Equation (4.5). The Rectified Linear Unit (ReLU) activation function was used for the hidden layers and the sigmoid function for the output layer, as shown in Section 4.5.1. To enhance network performance and prevent overtraining, two key techniques were employed: batch normalisation and L2 regularisation, explained in detailed in Section 4.5.2.

Early stopping, as defined in Section 4.5.2, was employed to monitor and prevent overtraining by evaluating the loss on the validation dataset. Training ceased when the difference between loss values calculated on the validation data for two consecutive epochs was less than 0.0002 after 30 epochs.

In this project, the Adam optimiser was used, as described in Section 4.5.2, with an initial learning rate set to $\eta = 10^{-3}$.

Nine variables were carefully selected as inputs for both networks, following the procedure outlined in Section 4.5.5. These variables encompass a subset of the observables

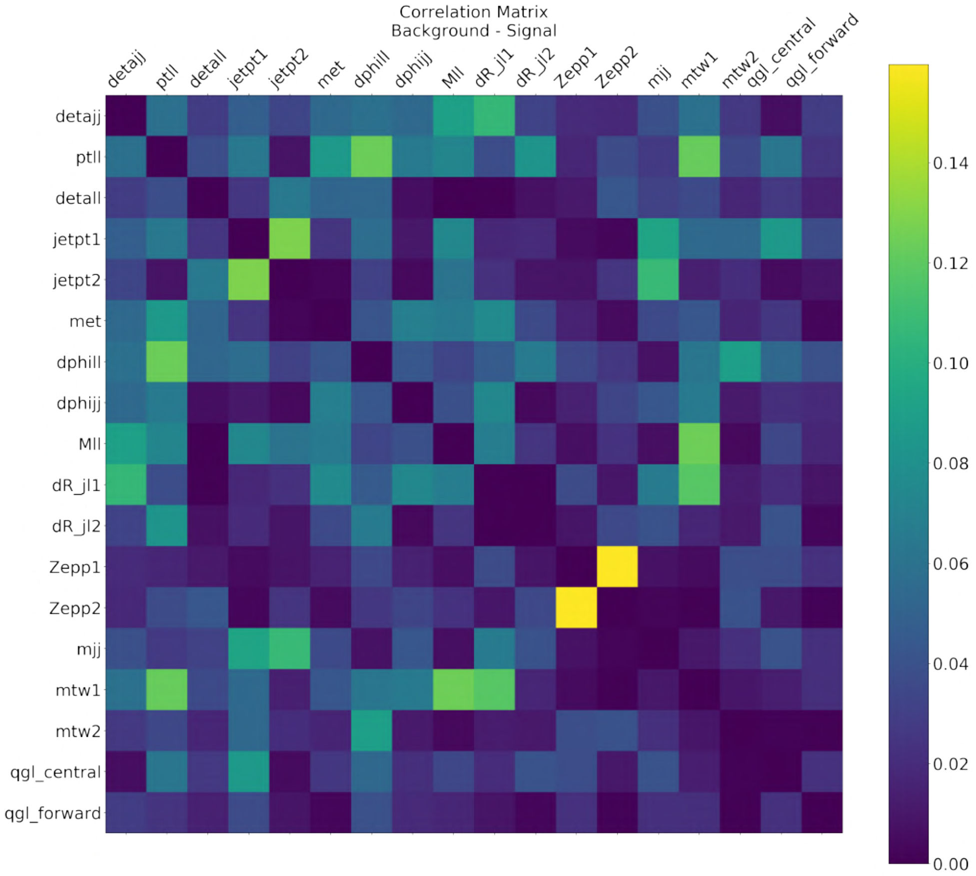


Figure 4.9: Difference between the correlation matrices of input variables for backgrounds and signal events in the $Z_{\ell\ell} \geq 1$ phase space. This analysis aided in the selection of suitable training variables for the DNN. Variables with noticeable differences, such as Z_{ℓ_1} or Z_{ℓ_2}, were favoured as they enable the DNN to learn and distinguish signal from background effectively.

Variable	Description
$\Delta\eta_{jj}$	Pseudorapidity interval between the two jets.
$p_{T,\ell\ell}$	Transverse momentum of the dilepton system.
$\Delta\eta_{\ell\ell}$	Pseudorapidity interval between the two leptons.
$p_{T,j1}$	Transverse momenta of the leading jet.
$p_{T,j2}$	Transverse momenta of the trailing jet.
$p_{\mathrm{T}}^{\mathrm{miss}}$ (MET)	Missing transverse momentum.
$\Delta\phi_{\ell\ell}$	Difference in azimuth angle ϕ between the two leptons.
$\Delta\phi_{jj}$	Difference in azimuth angle ϕ between the two jets.
$m_{\ell\ell}$	Invariant mass of the dilepton system.
$\Delta R_{j,\ell_1}$	Distance of the leading lepton from the nearest jet in the $\phi - \eta$ plane.
$\Delta R_{j,\ell_2}$	Distance of the trailing lepton from the nearest jet in the $\phi - \eta$ plane.
Z_{ℓ_1}	Centrality of the leading lepton as defined in the Equation (4.2).
Z_{ℓ_2}	Centrality of the trailing lepton as defined in the Equation (4.2).
m_{jj}	Invariant mass of the dijet system.
$m_{\mathrm{T}}^{\mathrm{W}_1}$	Transverse mass of the $(p_{\mathrm{T},\ell_1},\, p_{\mathrm{T}}^{\mathrm{miss}})$ system.
$m_{\mathrm{T}}^{\mathrm{W}_2}$	ransverse mass of the $(p_{\mathrm{T},\ell_2},\, p_{\mathrm{T}}^{\mathrm{miss}})$ system.
qgl central	Quark gluon likelihood discriminator of the first most central jet.
qgl forward	Quark gluon likelihood discriminator of the forward jet.

Table 4.6: Set of variables among the most discriminating ones, tested to be input of DNN. Their correlation matrix for the $Z_{\ell\ell} \geq 1$ is reported in Figure 4.9.

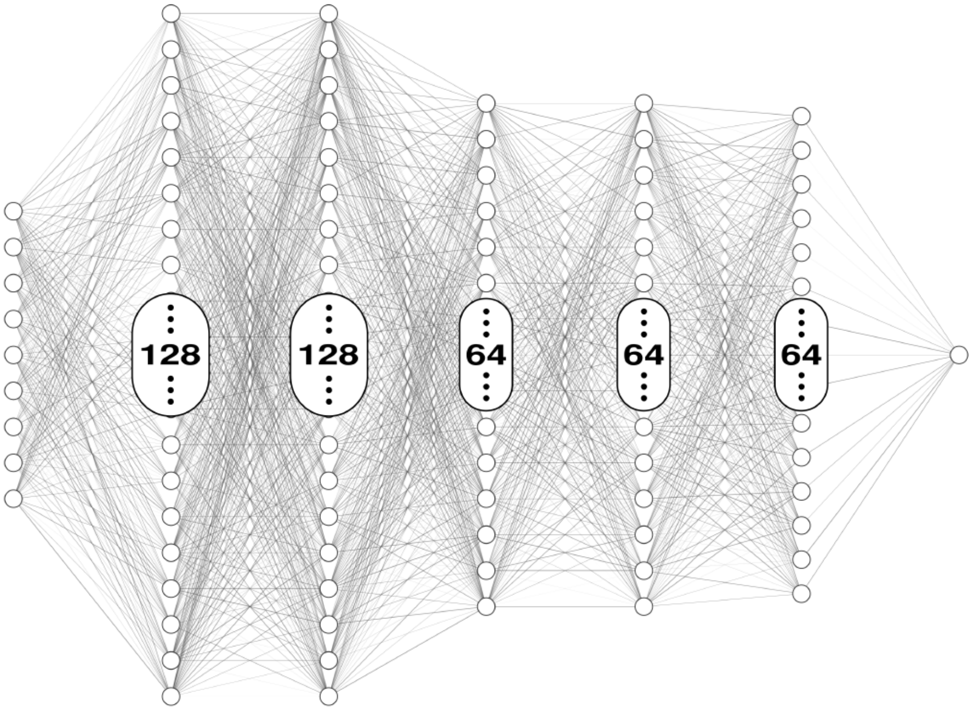


Figure 4.10: A schematic overview of the DNN architecture, featuring one input layer, five hidden layers, and an output layer. The input layer has nine neurons, one for each of the input variables. The five dense layers comprise varying numbers of neurons. The output layer has one neuron, that will give one number, namely the DNN score $\in [0, 1]$.

Variable	Description
m_{jj}	Invariant mass of the dijet system.
$\Delta\eta_{jj}$	Pseudorapidity interval between the two jets.
$p_{T,j1}$	Transverse momenta of the leading jet.
$p_{T,j2}$	Transverse momenta of the trailing jet.
$\Delta\phi_{\ell\ell}$	Difference in azimuth angle ϕ between the two leptons.
Z_{ℓ_1}	Centrality of the leading lepton as defined in the Equation (4.2).
Z_{ℓ_2}	Centrality of the trailing lepton as defined in the Equation (4.2).
$m_{\mathrm{T}}^{\mathrm{W}_1}$	Transverse mass of the (p_{T,ℓ_1}, $p_{\mathrm{T}}^{\mathrm{miss}}$) system.

Table 4.7: Set of variables used as inputs to the DNN for both $Z_{\ell\ell} < 1$ and $Z_{\ell\ell} \geq 1$ models.

detailed in Table 4.6 and are listed with their descriptions in Table 4.7. This subset of variables is carefully selected to encompass all the kinematic properties of the final state objects. Figure 4.11 illustrates the shapes of the final input variables in the $Z_{\ell\ell} \geq 1$ phase space for the EW $\mathrm{W}^+\mathrm{W}^-$ signal (violet), $\mathrm{t\bar{t}}$–tW (yellow), and QCD-induced $\mathrm{W}^+\mathrm{W}^-$ backgrounds (light blue). These observed differences in variable shapes between signal and background provide insights into their discriminatory power. Indeed, the DNN leverages these shape distinctions in input observables to make predictions regarding the event being signal or background.

4.5.7 Training Monitoring and Evaluation

Training monitoring involves evaluating metrics on both the training and validation datasets. This process helps assess the network training progress, detect signs of overtraining, and evaluate model performance.

The first metric chosen to compare how the network behaves on the training and validation dataset is the loss function used to train the network itself; the chosen loss function is the binary cross entropy, defined in Equation (4.5). Figure 4.12 displays the loss functions for the two DNN models: $Z_{\ell\ell} < 1$ on the left and $Z_{\ell\ell} \geq 1$ on the right. Overtraining is often indicated when the loss function on the training data steadily decreases while the loss on the validation data starts to increase, as explained in Section 4.5.2. In this case, as can be seen in Figure 4.12, both training (blue line) and validation (orange line) loss functions consistently decrease during training, suggesting no signs of overtraining.

After training, the models are assessed on testing samples used in the analysis. Figure 4.13 displays the DNN output distributions for training and testing samples. The overlap between the blue-azure (background) and red-orange (signal) curves indicates that the models generalise well and maintain predictive power on new events.

As discussed in Section 4.5.5, experimentation was conducted with different numbers of variables and network structures. Model performance in terms of discriminating power was evaluated using Receiver Operating Characteristics (ROC) [94] curves, comparing them

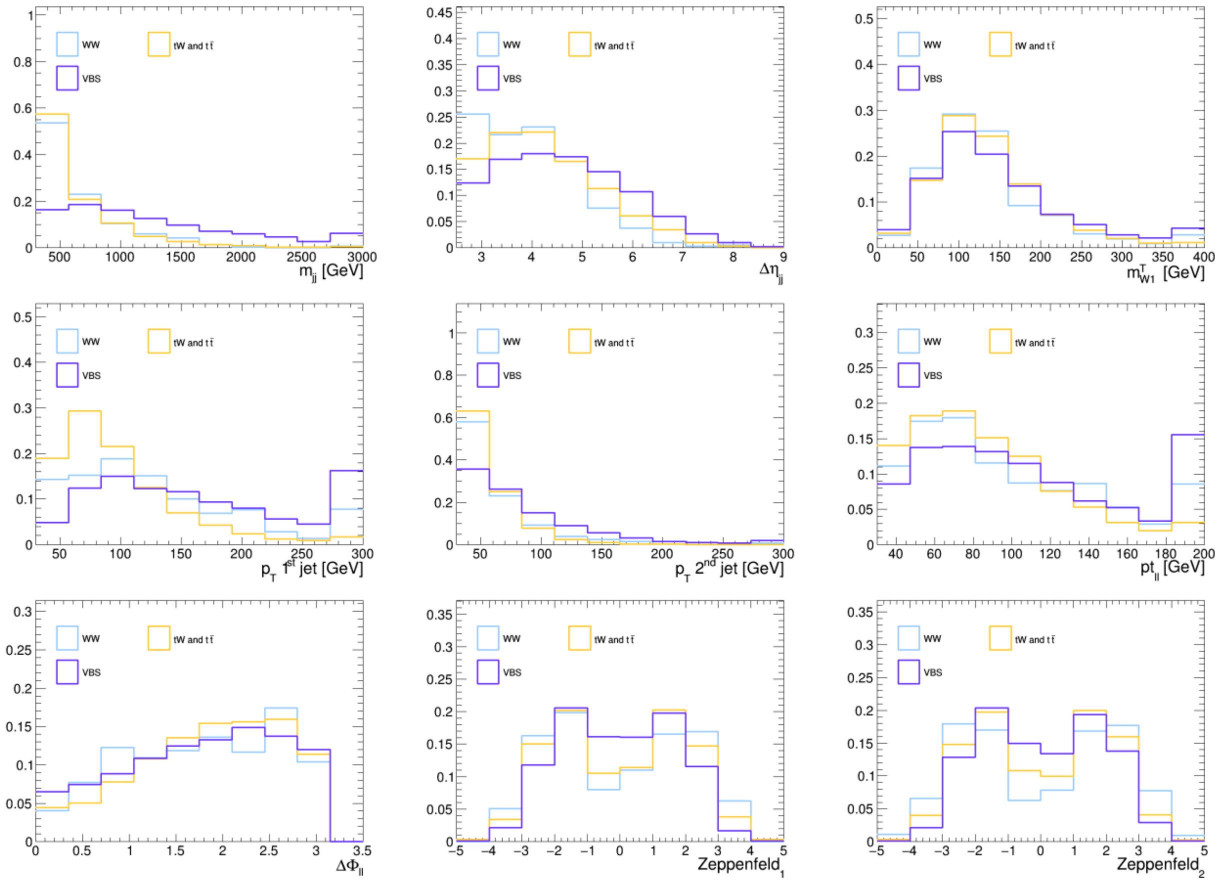


Figure 4.11: Input variable shapes in the $Z_{\ell\ell} \geq 1$ signal region for the signal (violet), $t\bar{t}$–tW (yellow), and QCD-induced W^+W^- (light blue) backgrounds. The histograms are normalised to one.

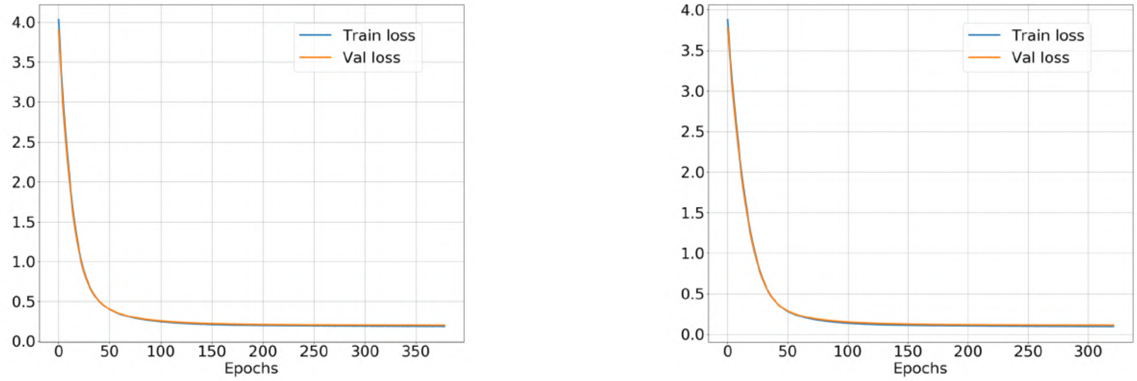


Figure 4.12: Loss functions for the $Z_{\ell\ell} < 1$ (left) and $Z_{\ell\ell} \geq 1$ (right) DNN models. Both loss functions, training (blue line) and validation (orange line), consistently decrease during training, suggesting no signs of overtraining.

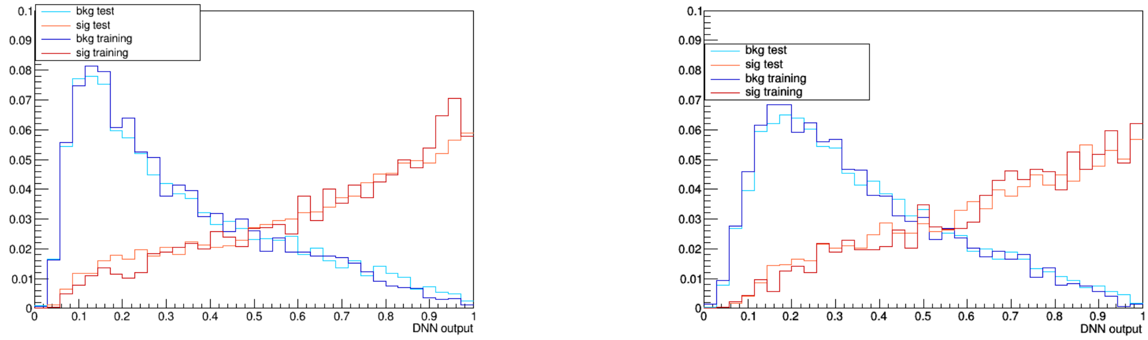


Figure 4.13: DNN output distributions for training and testing samples for $Z_{\ell\ell} < 1$ (left) and $Z_{\ell\ell} \geq 1$ (right).

to the m_{jj} variable as a benchmark. The ROC curves (Figure 4.14), obtained using the test samples, show that the DNN consistently outperforms m_{jj}, even in the low-efficiency region. For a fixed level of background contamination, the DNN consistently achieves higher signal purity than m_{jj}, implying an improvement in analysis performance when using the DNN output as a discriminator.

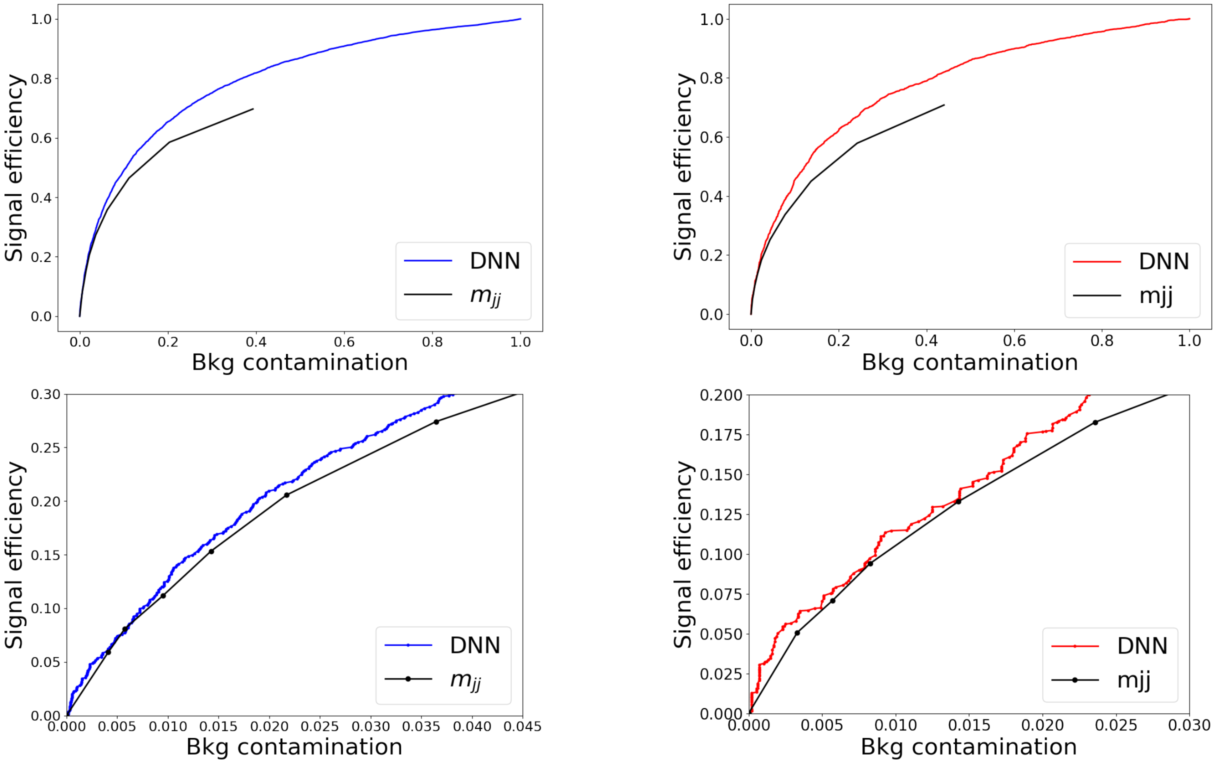


Figure 4.14: ROC curves for m_{jj} and DNN for the full Run 2 dataset, with a zoomed-in view of the low-efficiency region. On the left: $Z_{\ell\ell} < 1$. On the right: $Z_{\ell\ell} \geq 1$.

4.6 Control Regions and Data-MC Comparisons

This chapter focuses on the establishment of control regions and the subsequent data-MC comparisons. As discussed in Section 4.4, the primary goal is to assess the agreement between simulations and observed data. To achieve this, control regions are strategically defined to be orthogonal to the signal regions. These control regions are designed to be enriched with events originating from two primary background sources: $\mathrm{t\bar{t}}$-tW and DY processes.

4.6.1 Top Control Region

Figure 4.15 presents the comparisons of observation and prediction in the top enriched phase space for the different flavour category $e\mu$, while figures 4.16 and 4.17 refer to the same flavour categories ee and $\mu\mu$.

The data/MC agreement is consistently strong across all variables of interest, spanning three years of data-taking. Specifically, key discriminating variables are analyzed, including the di-jet invariant mass (m_{jj}), the pseudorapidity gap ($\Delta\eta_{jj}$), and the Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}). These variables are of particular importance as they also serve as inputs for the DNN analysis, as discussed in detail in Section 4.5.

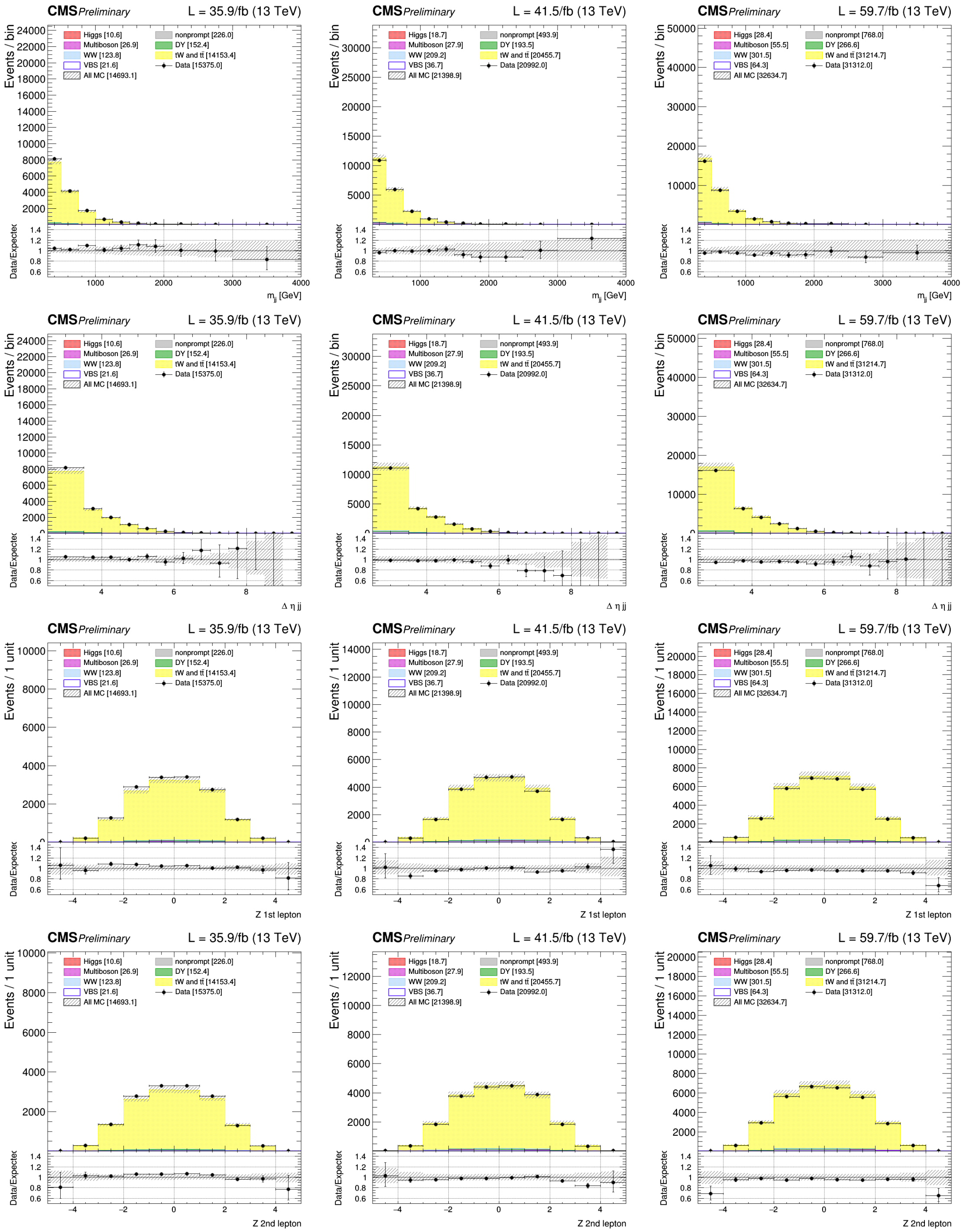


Figure 4.15: Comparison of observation and prediction in the top $e\mu$ control region from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

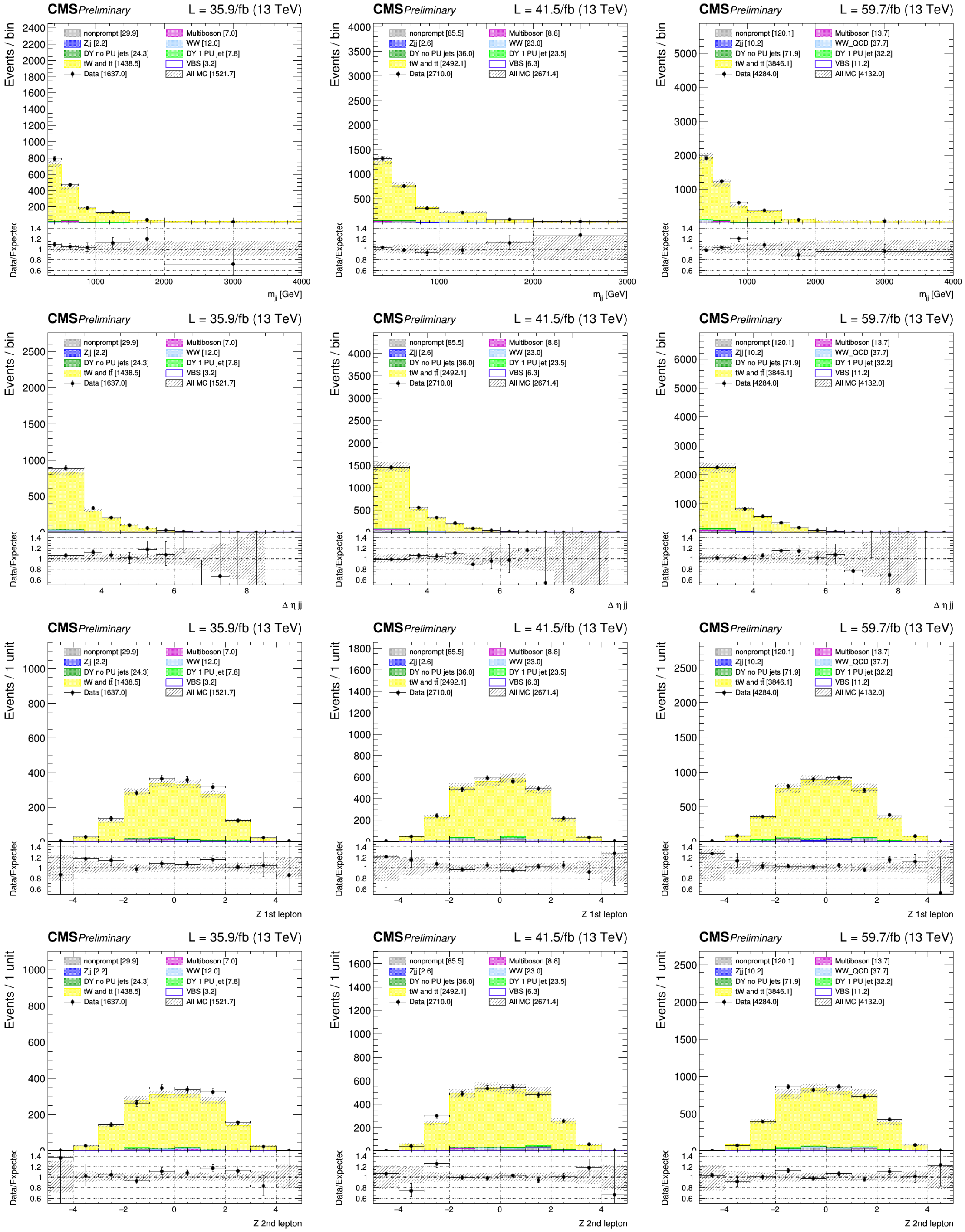


Figure 4.16: Comparison of observation and prediction in the top ee control region from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

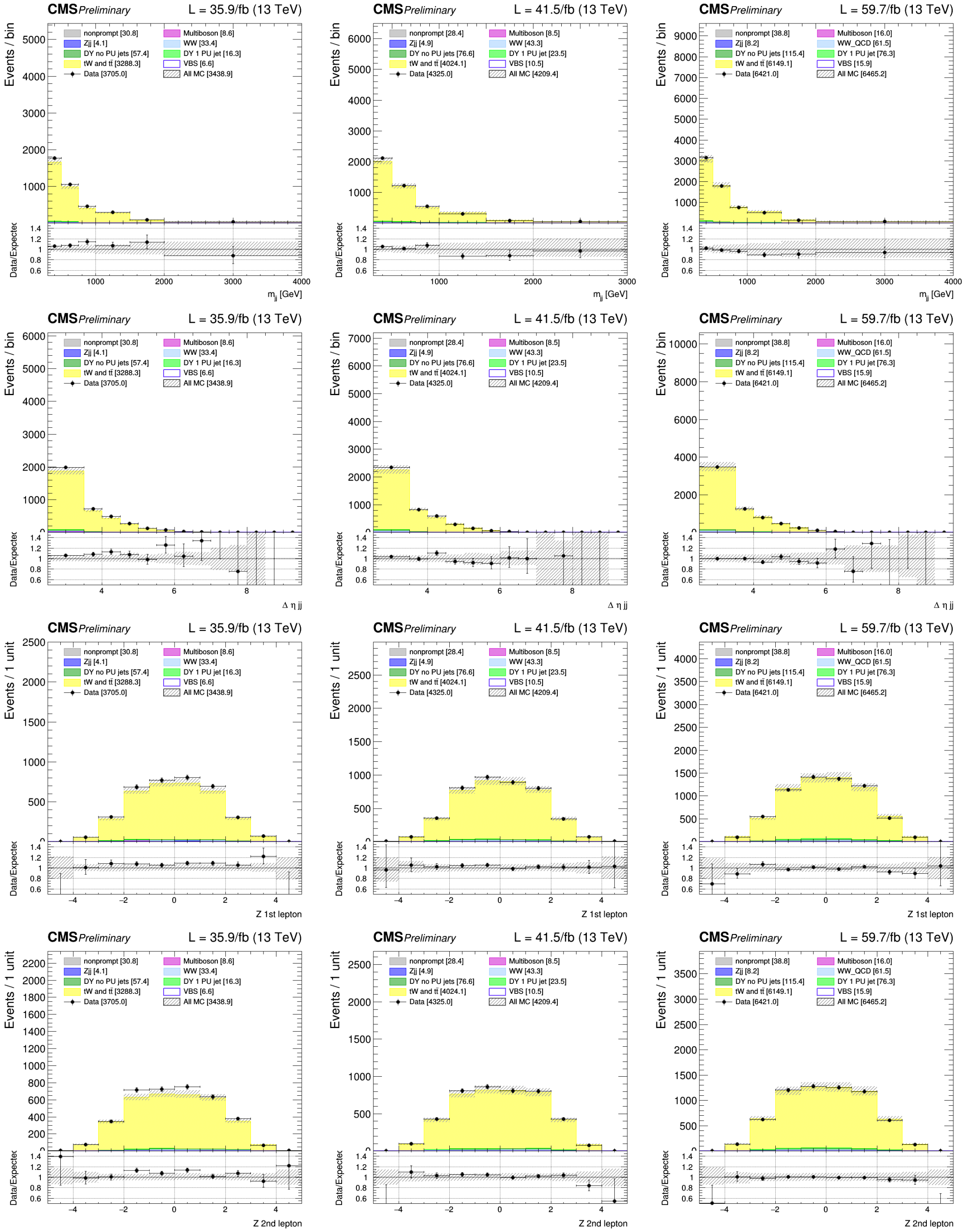


Figure 4.17: Comparison of observation and prediction in the top $\mu\mu$ control region from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

4.6.2 Drell-Yan Control Region

A first study of the DY control region began by selecting events based on the preselection criteria for p_T^{miss}, as detailed in Table 4.3, where a loose requirement of $p_T^{\text{miss}} > 20\,\text{GeV}$ was imposed. Figures 4.18-4.26 were generated using this preliminary selection. The figures under consideration, namely Figures 4.18-4.26, will now be described in details.

Figures 4.18 and 4.19 to 4.22 depict general comparisons between observed and predicted data in the DY enriched phase space, covering both different and same flavour categories, respectively.

While the agreement between the data and simulations for the 2017 and 2018 datasets is reasonable, there are noticeable discrepancies in the 2016 configuration, particularly in the $\Delta\eta_{jj}$ distribution within the same flavour channels. To further investigate the origin of this disagreement, a decision was made to subdivide the DY same flavour control region based on the $Z_{\ell\ell}$ variable, similar to the approach used in the signal categories. This division will help determined whether the observed differences between data and MC simulations can be attributed to specific Zeppenfeld selections.

For the 2016 dataset and considering the m_{jj} and $\Delta\eta_{jj}$ variables, the Figures 4.23 and 4.24 were generated.

Since the source of disagreement is evident in both categories, particularly at low values of $\Delta\eta_{jj}$, the decision has been made to employ a Next-to-Leading Order (NLO) DY sample to investigate whether this discrepancy could be attributed to a lack of QCD precision. Figures 4.25 and 4.26 illustrate this new comparison, presenting data in both the m_{jj} and $\Delta\eta_{jj}$ variables. While it is observable an overall improved agreement with the data, the $\Delta\eta_{jj}$ distribution still exhibits some disparities.

Such NLO DY samples cannot be employed in the signal region either. This limitation arises due to their extremely low MC statistics, rendering them unreliable for estimating this background within the constraints of a tight selection, as outlined in the analysis.

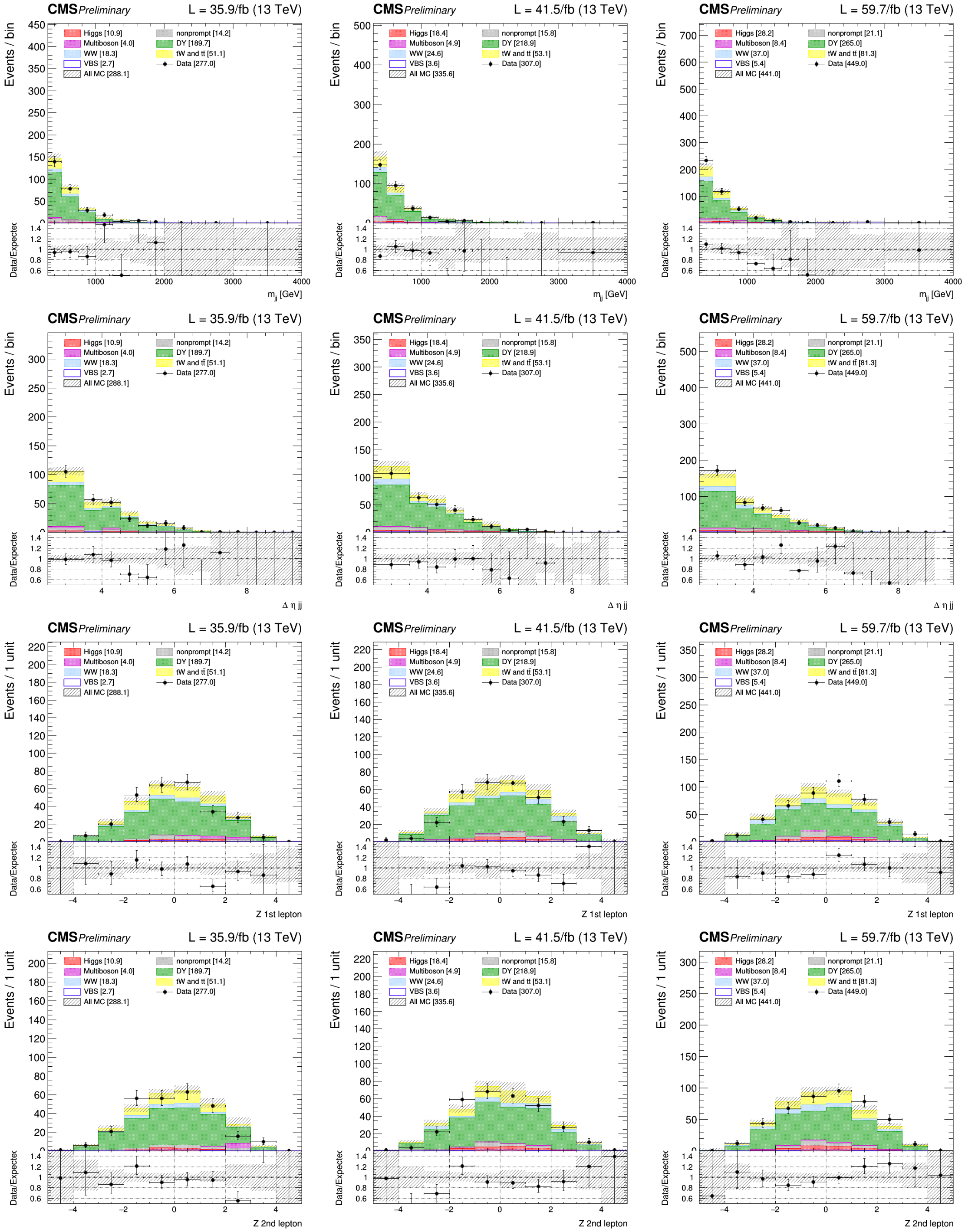


Figure 4.18: Comparison of observation and prediction in the DY $e\mu$ control region from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

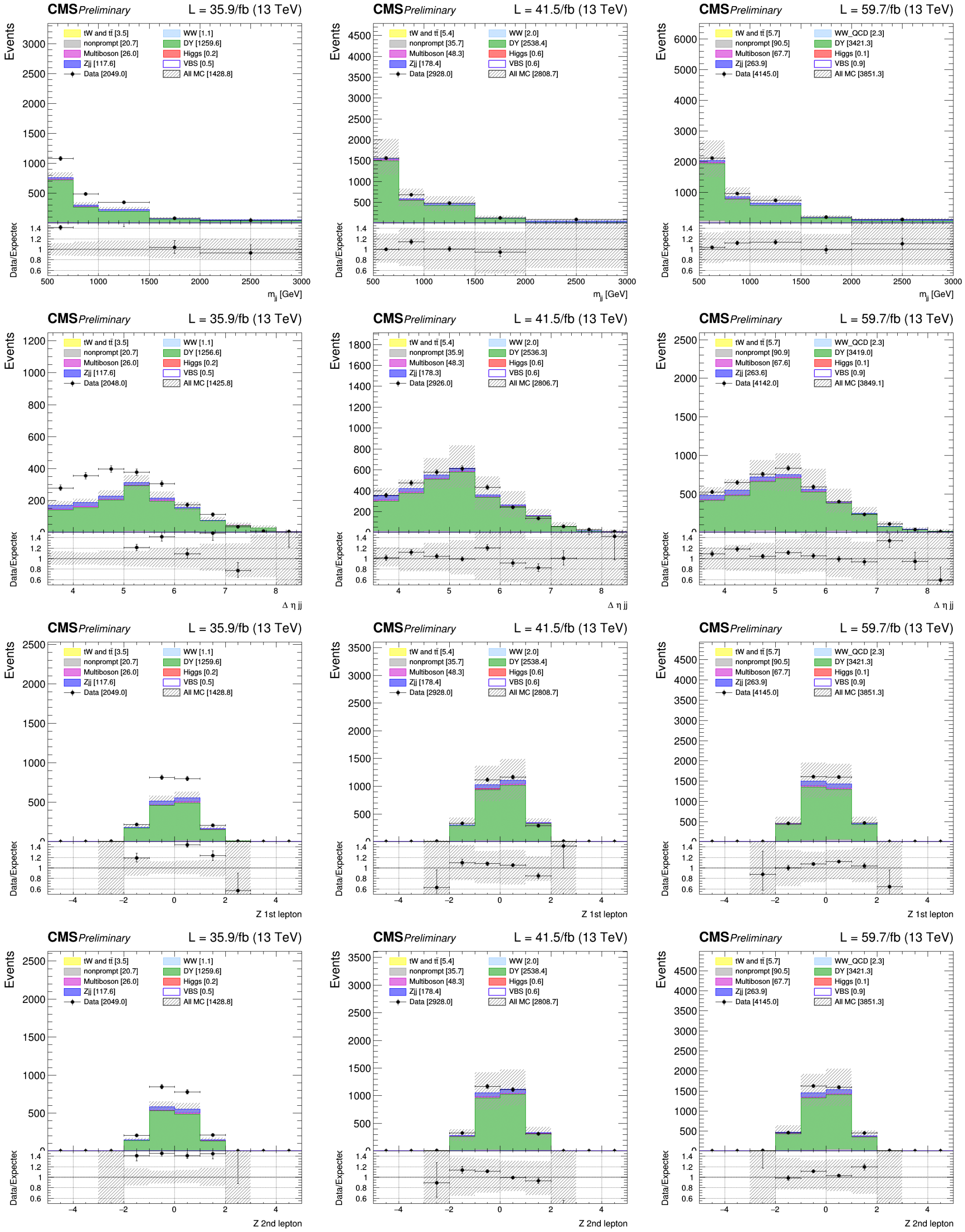


Figure 4.19: Comparison of observation and prediction in the DY ee ($Z_{\ell\ell} < 1$) control region from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

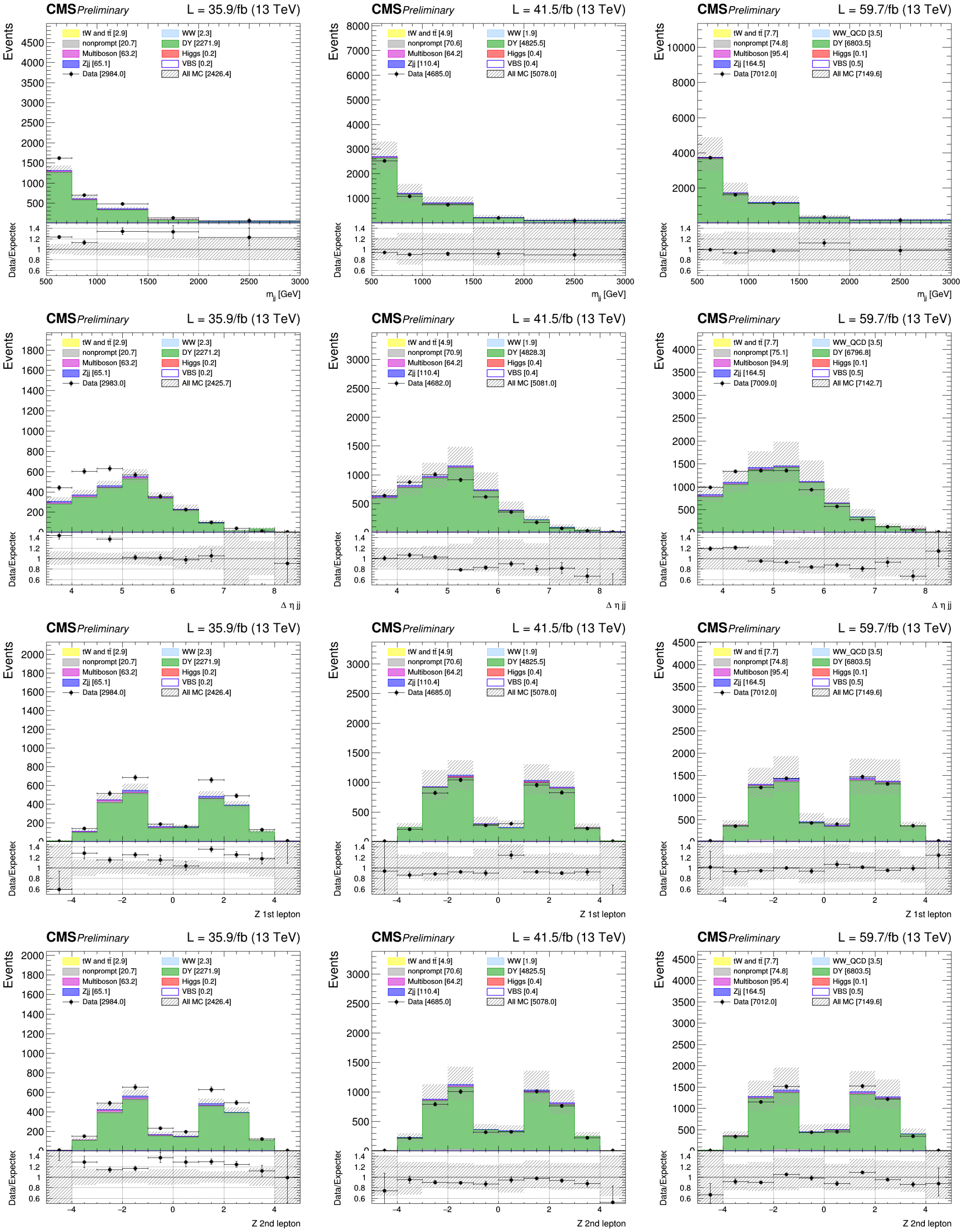


Figure 4.20: Comparison of observation and prediction in the DY *ee* ($Z_{\ell\ell} \geq 1$) control region from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

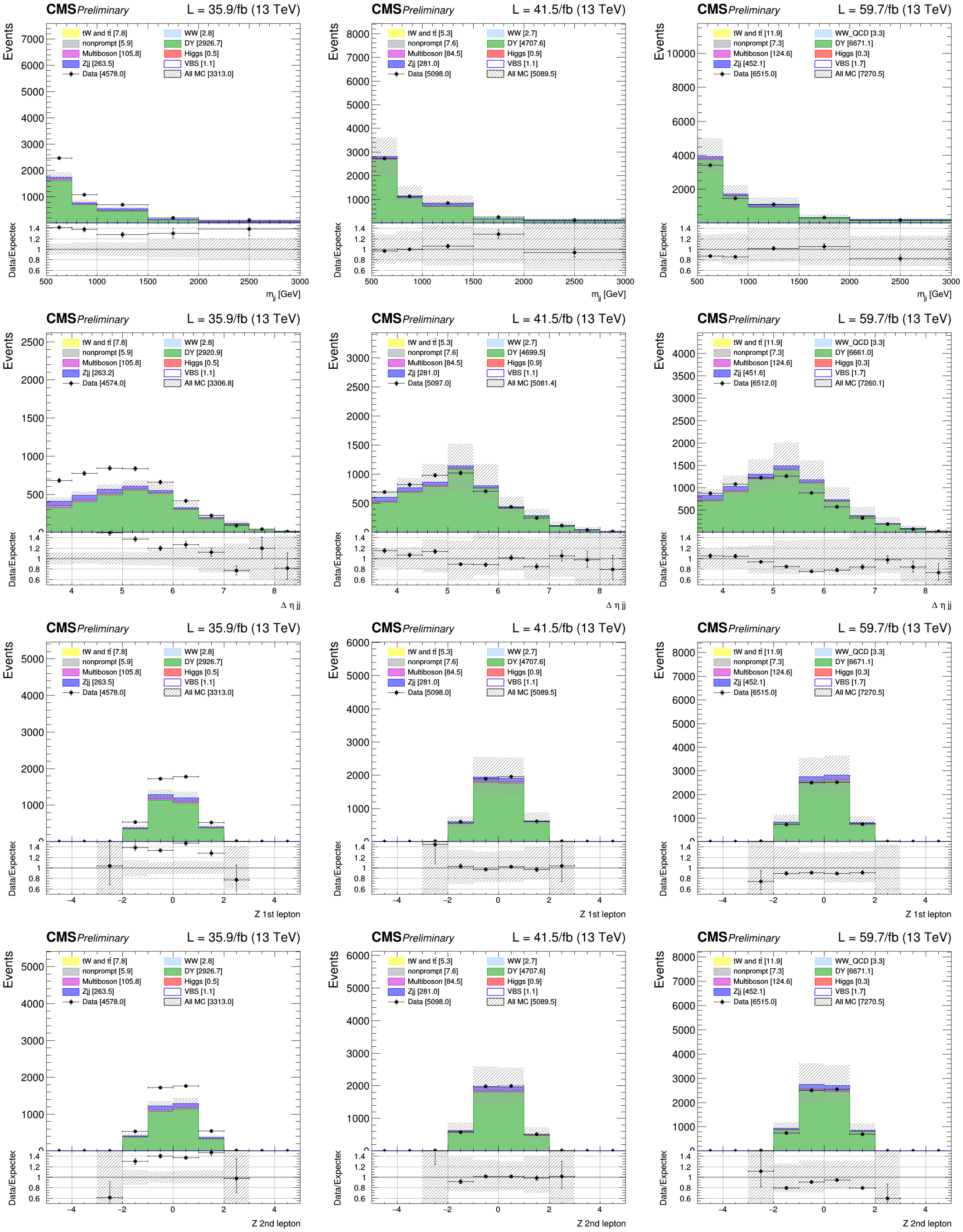


Figure 4.21: Comparison of observation and prediction in the DY $\mu\mu$ ($Z_{\ell\ell} < 1$) control region from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

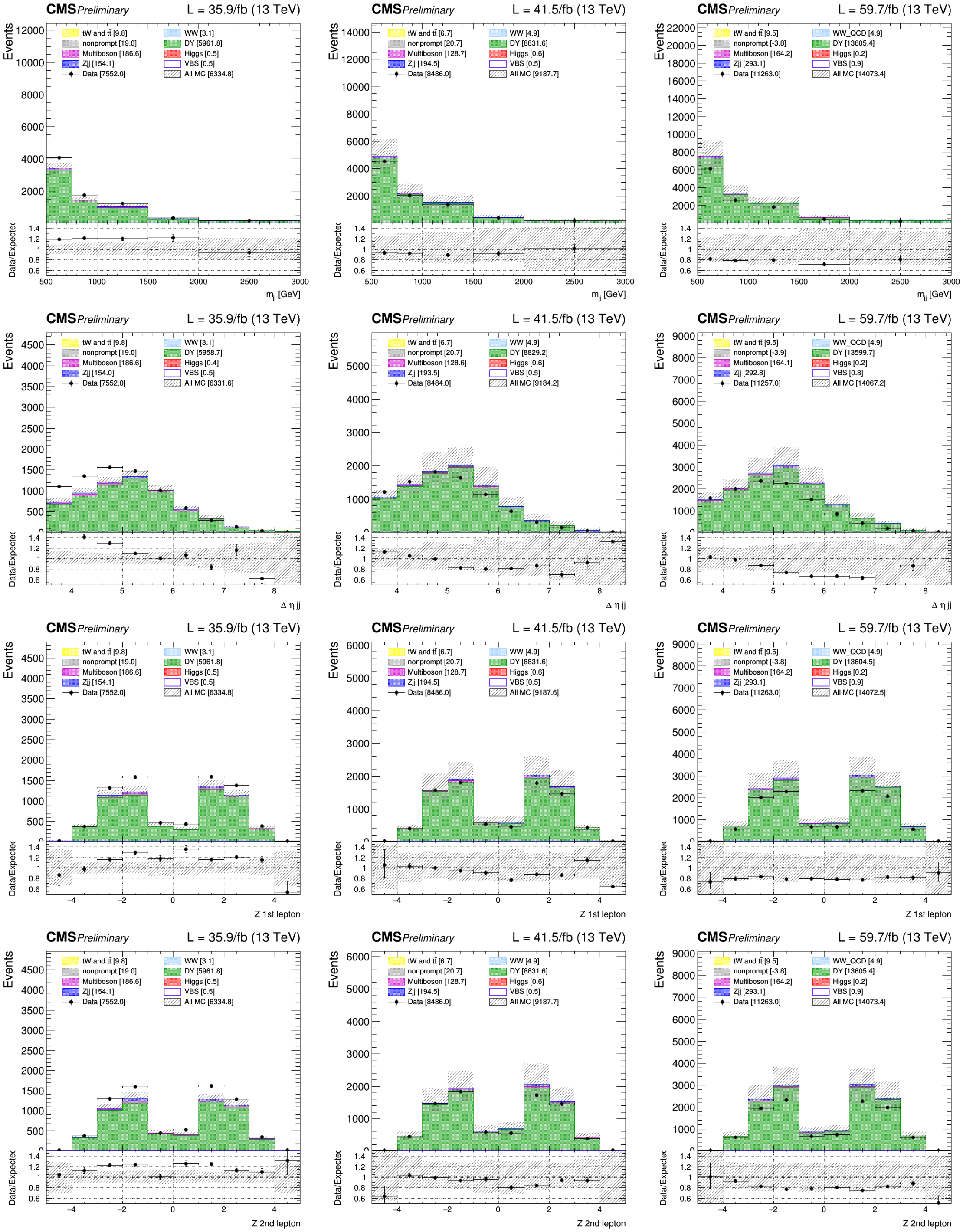


Figure 4.22: Comparison of observation and prediction in the DY $\mu\mu$ ($Z_{\ell\ell} \geq 1$) control region from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

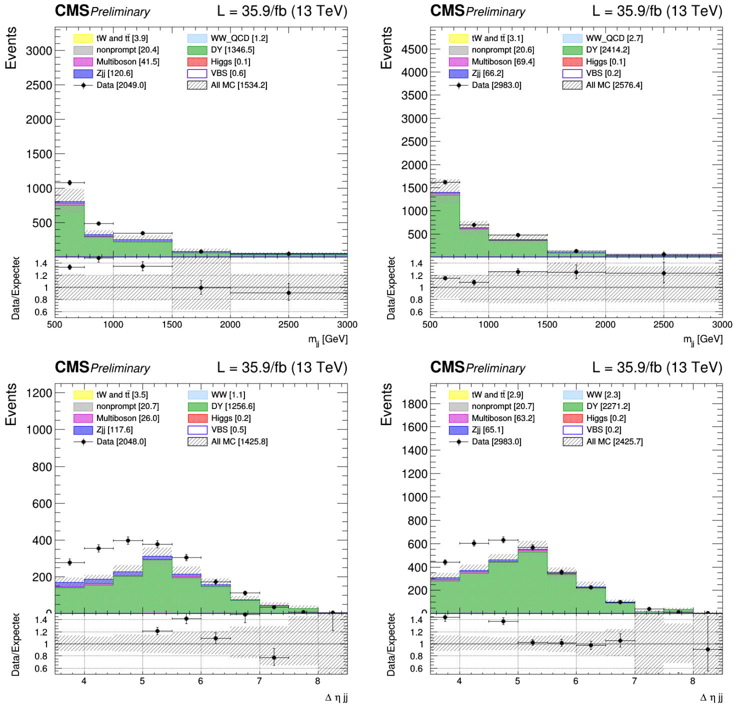


Figure 4.23: Comparison of observation and prediction in the DY ee control region for the 2016 samples, split in $Z_{\ell\ell} < 1$ (left) and $Z_{\ell\ell} \geq 1$ (right).

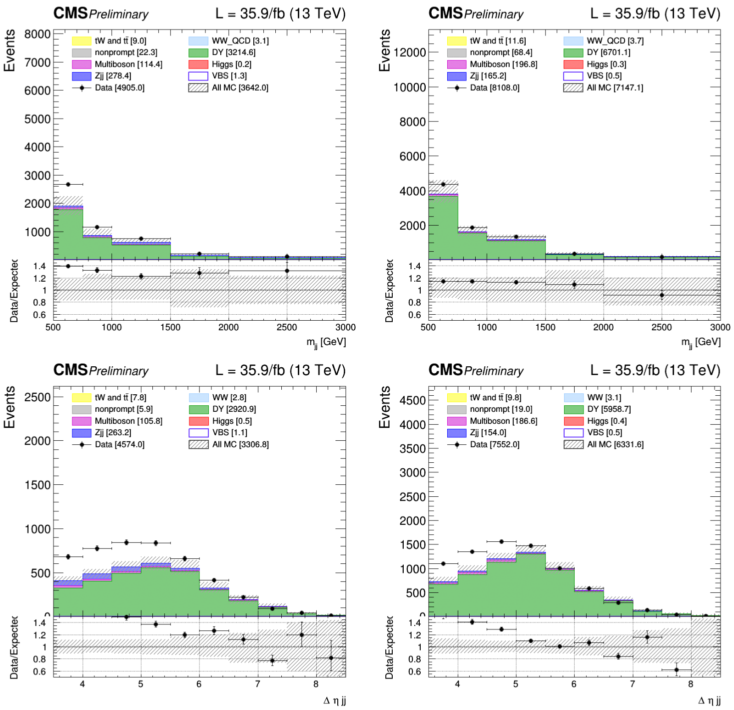


Figure 4.24: Comparison of observation and prediction in the DY $\mu\mu$ control region for the 2016 samples, split in $Z_{\ell\ell} < 1$ (left) and $Z_{\ell\ell} \geq 1$ (right).

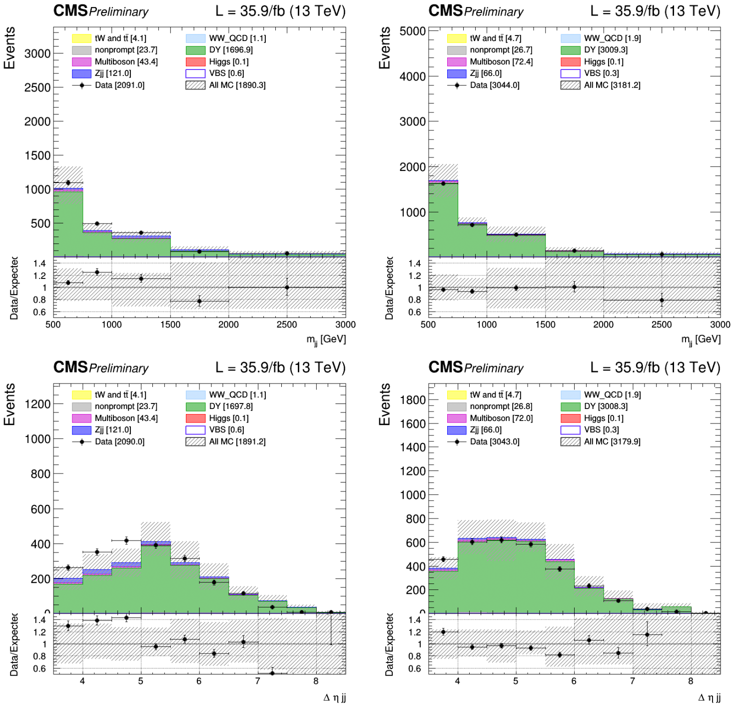


Figure 4.25: Comparison of observation and prediction (NLO sample) in the DY *ee* control region for the 2016 samples, split in $Z_{\ell\ell} < 1$ (left) and $Z_{\ell\ell} \geq 1$ (right).

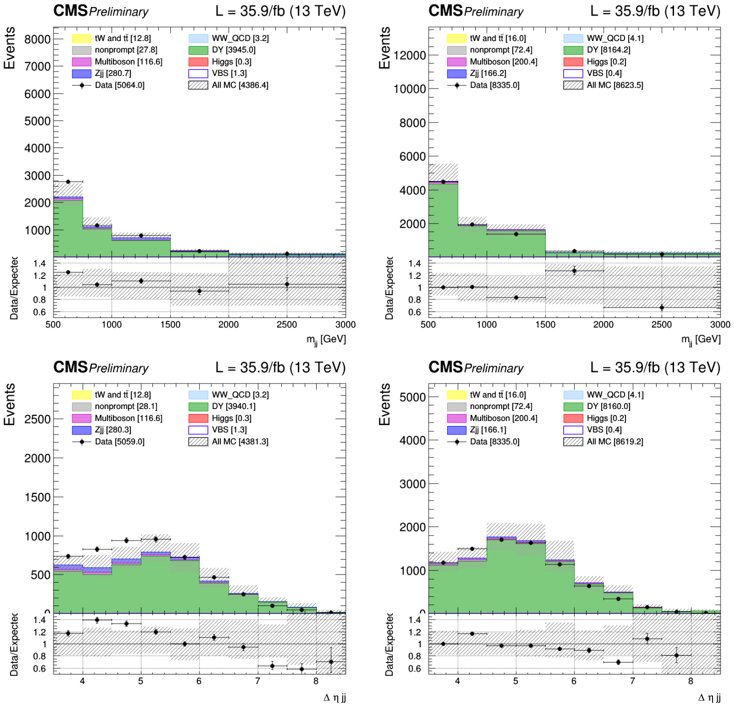


Figure 4.26: Comparison of observation and prediction (NLO sample) in the DY $\mu\mu$ control region for the 2016 samples, split in $Z_{\ell\ell} < 1$ (left) and $Z_{\ell\ell} \geq 1$ (right).

In response to the discrepancies observed in the DY control region, a first consideration was made about the selections defining the DY control region, specially the ones constraining p_T^{miss}. As previously mentioned, an initial selection in the DY control region employed a loose criterion for p_T^{miss} ($p_T^{\text{miss}} > 20\,\text{GeV}$), while a tighter p_T^{miss} requirement ($p_T^{\text{miss}} > 60\,\text{GeV}$) was enforced in the same flavour signal region to effectively reduce the DY contribution. By employing this approach, the range between $20\,\text{GeV}$ and $60\,\text{GeV}$ assumed a pivotal role in characterising the DY background in the analysis. This reliance can be somewhat inconvenient because the p_T^{miss} shape for the DY process cannot be entirely trusted. In reality, there is no intrinsic p_T^{miss} from DY diagrams, and any observed p_T^{miss} is solely a result of detector effects. In response to these considerations, the selection criteria for the p_T^{miss} in the DY control region have been synchronised with that of the same flavour signal region, with a requirement of $p_T^{\text{miss}} > 60\,\text{GeV}$.

To address these discrepancies further, a second consideration was given to the distinction between DY events with at least one pile-up jet ("DY PU") and the hard DY process, which encompasses events where no pile-up jets are present, with jets originating exclusively from the primary interaction vertex ("DY hard"). It is anticipated that DY PU will dominate in regions where $\Delta\eta_{jj} > 5$. In fact, in these regions the absence of tracker information makes it challenging to discern the origin of particles. Conversely, it is expected that DY hard will prevail in the region where $\Delta\eta_{jj} < 5$.
In response to these considerations, the following steps were also taken to tackle the discrepancies in the DY control region, as can be seen in Figure 4.27:

- Two distinct DY sub-samples have been defined:
    - The "DY PU", represented by the light green histogram, encompasses events in which at least one of the two leading jets originates from a secondary vertex, indicative of pile-up activity.
    - The "DY hard", depicted by the dark green histogram, includes events with jets exclusively stemming from the primary interaction vertex.
- Instead of dividing the DY control region according to the $Z_{\ell\ell}$ variable, it has been split based on the $\Delta\eta_{jj}$ variable. This approach enhances sensitivity to pile-up effects.

The inclusive DY same-flavour control region for the 2016 data set is presented, maintaining the $p_T^{\text{miss}} > 20\,\text{GeV}$ cut in place. The two DY sub-samples exhibit markedly different shapes in both m_{jj} and $\Delta\eta_{jj}$, as clearly illustrated in Figure 4.27. Specifically, DY PU exhibits a peak around $\Delta\eta_{jj} \sim 5$, while DY hard $\Delta\eta_{jj}$ distribution rapidly decreases.

Through the fitting process of the two contributions (DY PU and DY hard) to data within the $\Delta\eta_{jj}$ bins of the control region, a substantial reduction in data/MC discrepancies is expected to be achieved. This is due to the rescaling of DY contributions by the two distinct factors that are determined during the fit procedure. To demonstrate the effectiveness of this approach, the MC simulations were actually fitted to the data within the DY control region. This fitting process substantially reduces data/MC discrepancies, confirming the validity of the strategy. An illustration of this effect is presented in Figure 4.28, where the two DY contributions have been rescaled by factors obtained through log-likelihood minimisation.

Finally, Figures 4.29-4.32 depict the inclusive ee and $\mu\mu$ DY same flavour categorisation for all three years data-taking sets, with the new strategy implemented.

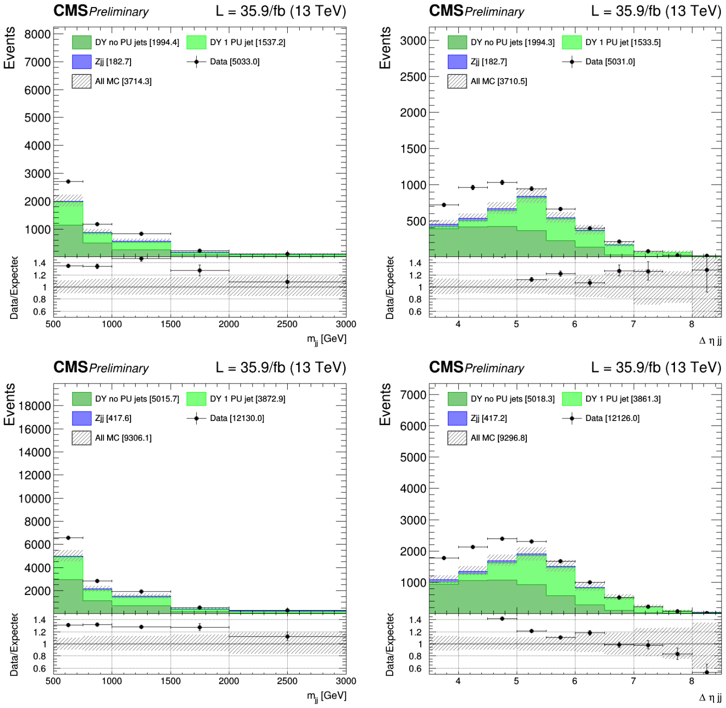


Figure 4.27: Data/MC comparison in the DY same flavour inclusive control region for the 2016 samples, where the $p_T^{\text{miss}} > 20\,\text{GeV}$ cut is still applied.
From top to bottom: ee final state, $\mu\mu$ final state.
DY PU events (light green) exhibit a peak $\Delta\eta_{jj} \sim 5$, while the DY hard (dark green) $\Delta\eta_{jj}$ distribution rapidly decreases.

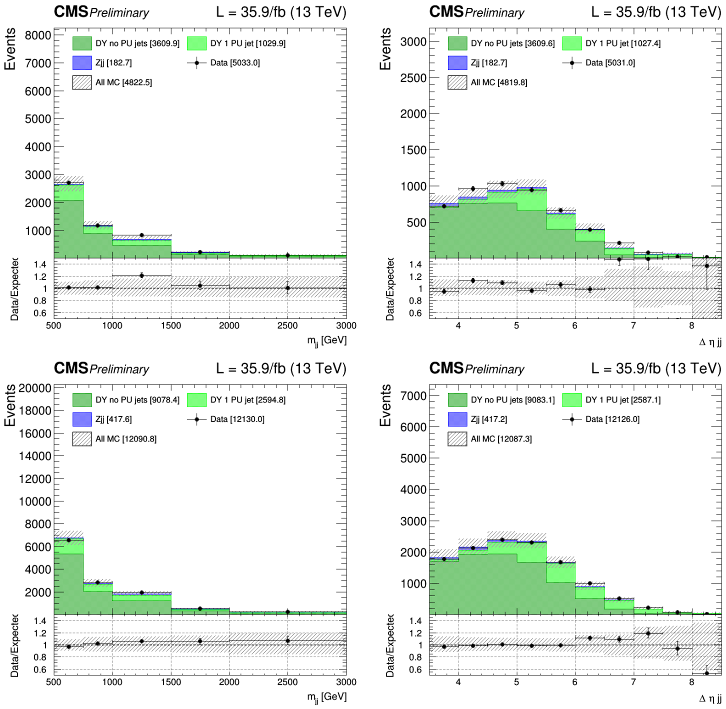


Figure 4.28: Same plots as above but DY contributions have been rescaled by the factor determined during the fit procedure.

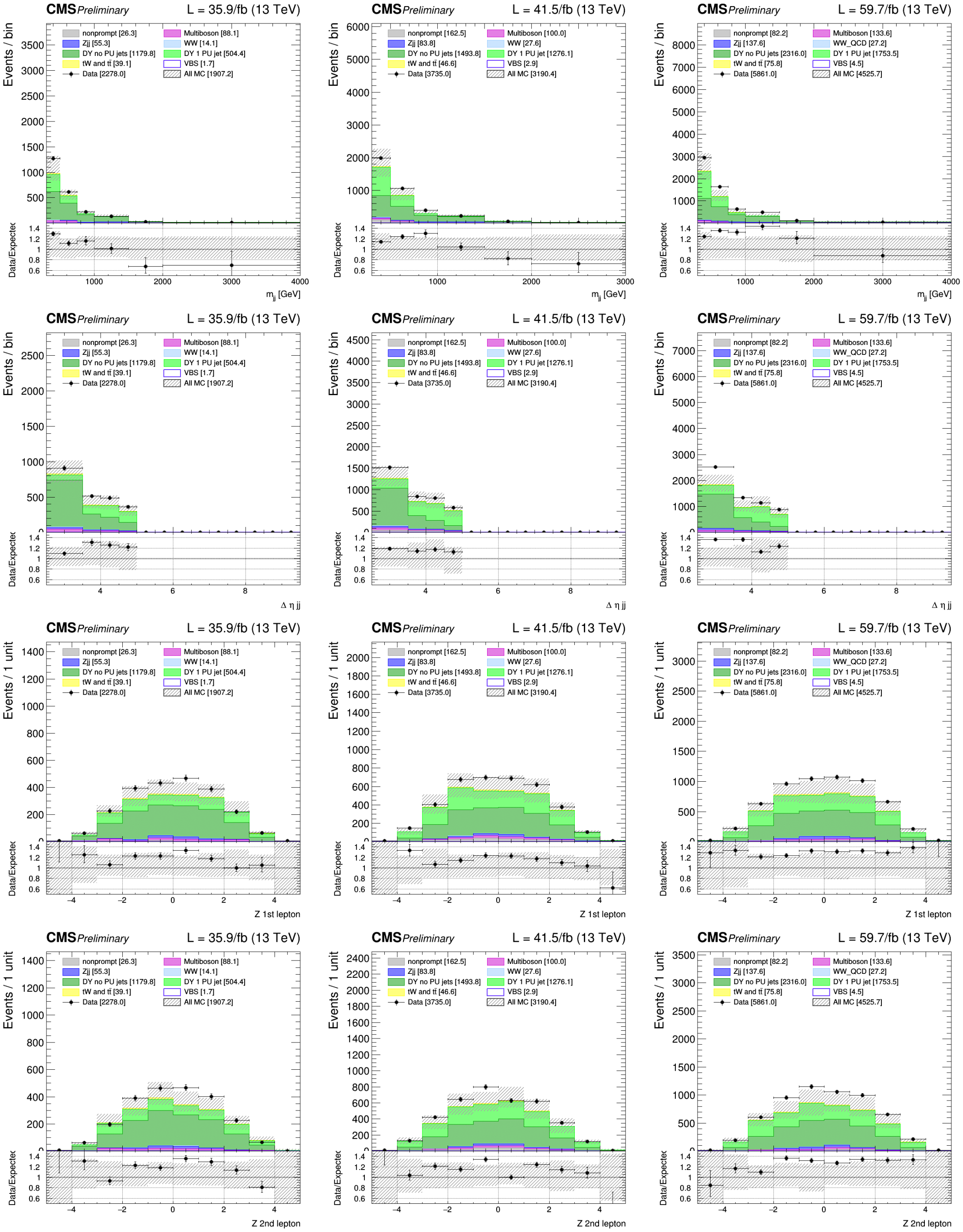


Figure 4.29: Comparison of observation and prediction in the DY ee control region ($\Delta\eta_{jj}$ < 5) from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

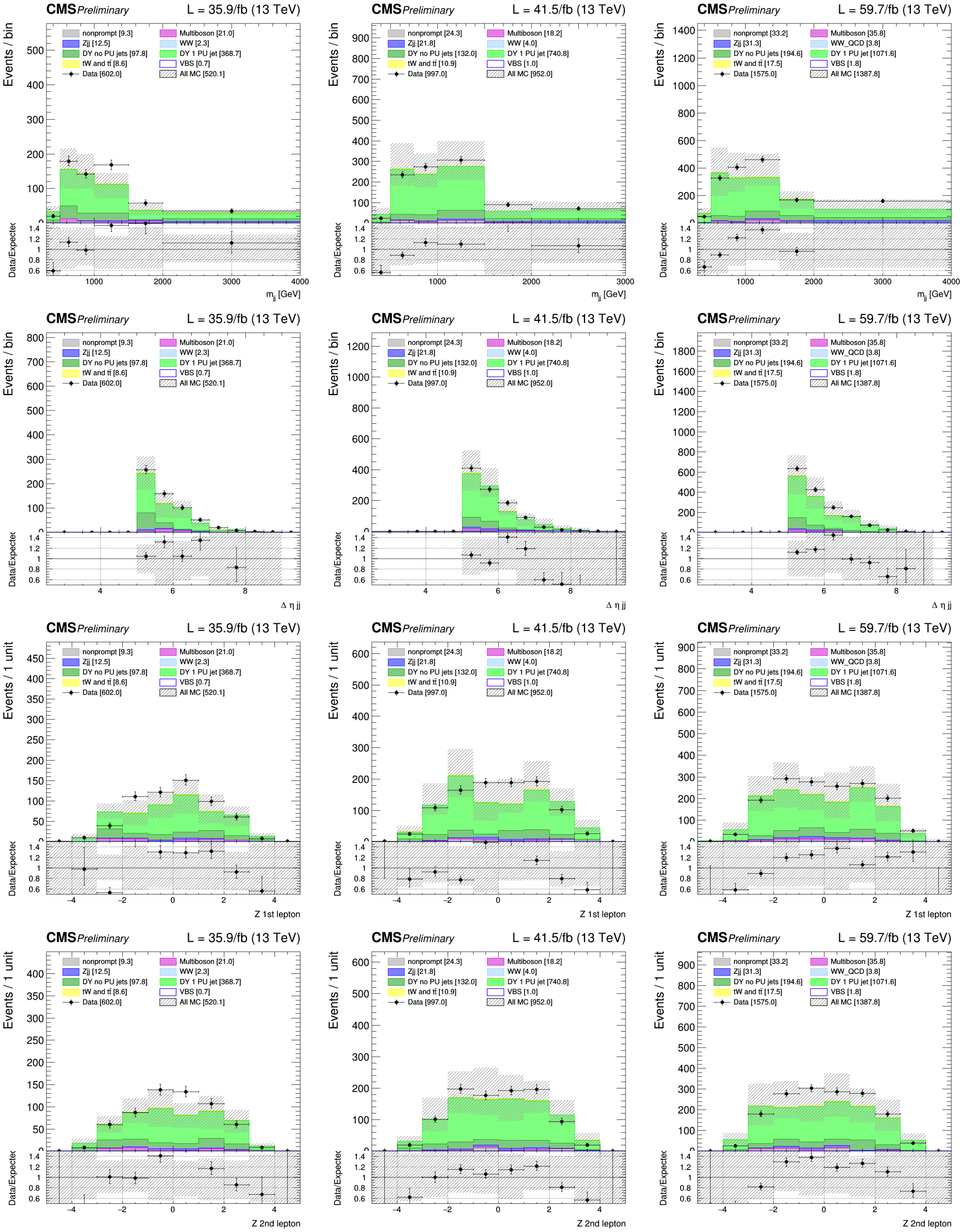


Figure 4.30: Comparison of observation and prediction in the DY ee control region ($\Delta\eta_{jj}$ ≥ 5) from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

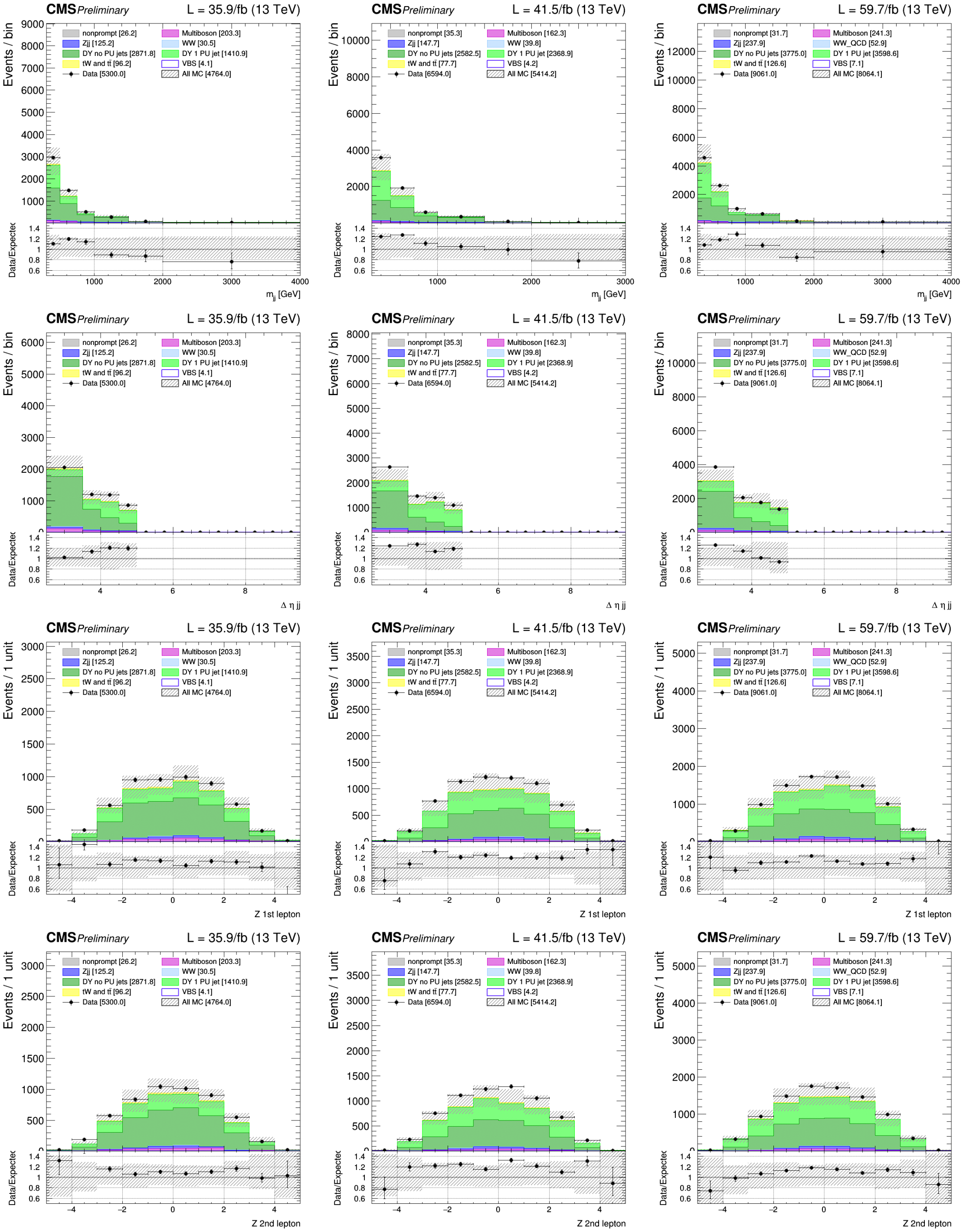


Figure 4.31: Comparison of observation and prediction in the DY $\mu\mu$ control region ($\Delta\eta_{jj} < 5$) from the 2016 (left), 2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

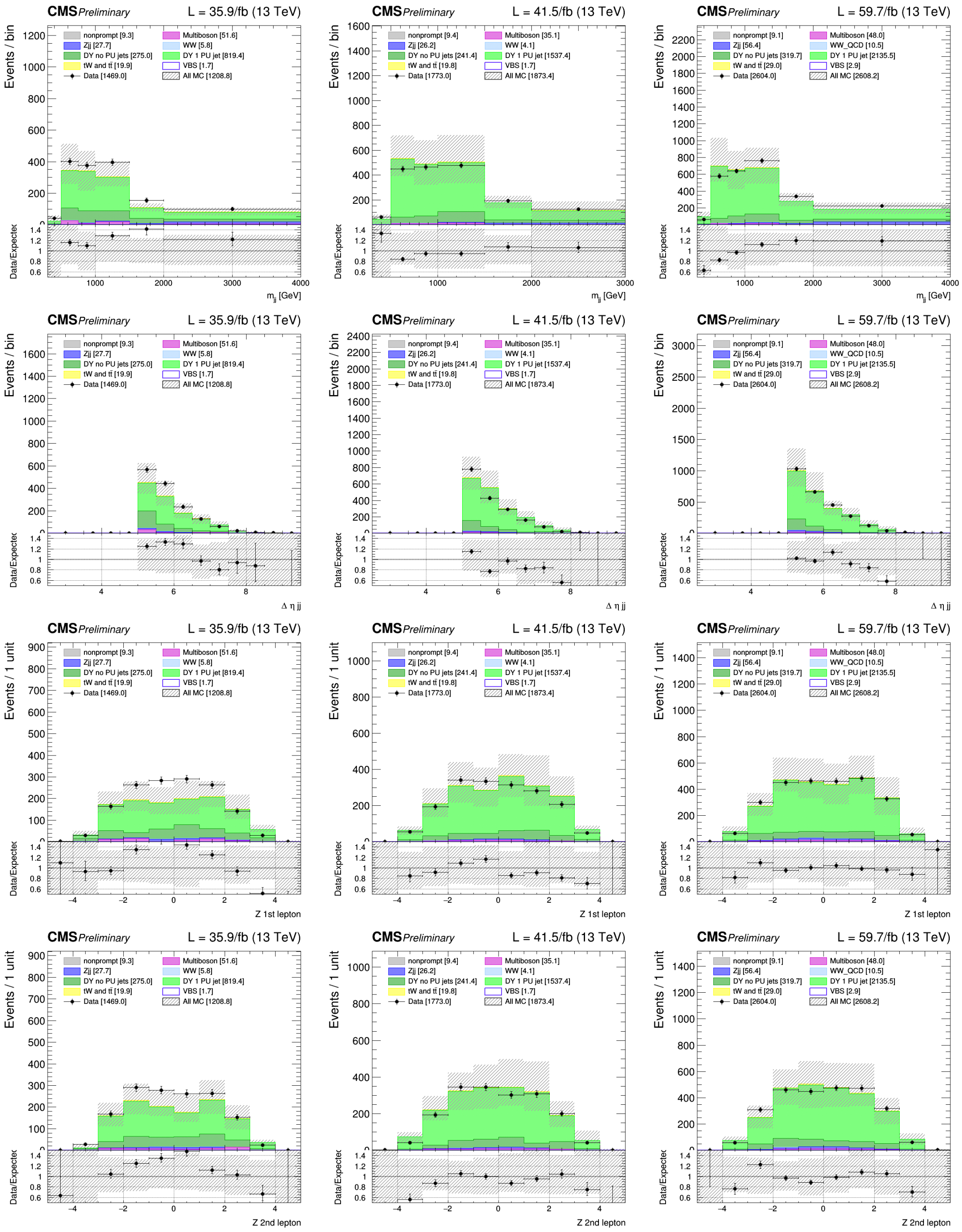


Figure 4.32: Comparison of observation and prediction in the DY $\mu\mu$ control region ($\Delta\eta_{jj} \geq 5$) from the 2016 (left), 2017 (2017 (centre) and 2018 (right) samples. From top to bottom: di-jet mass, pseudorapidity gap of the two jets, and Zeppenfeld variables (Z_{ℓ_1} and Z_{ℓ_2}).

4.7 Backgrounds Estimation

Top Background

The primary background process in the study arises from $\mathrm{t\bar{t}}$ events. To estimate its contribution within the signal region, a data-driven technique is employed. In the analysis, a control region specifically enhanced with $\mathrm{t\bar{t}}$-tW events, as detailed in Section 4.4, is incorporated as part of the fit procedure.

DY Background

In the same flavour categories, DY events represent a significant background source. The DY sample is divided into two sub-contributions, one involving jets originating from a pileup vertex and the other with the two leading jets radiated from the initial state ("DY PU" and "DY hard"). The DY control region is further categorised into $\Delta\eta_{jj}$ bins for each channel, defining independent scaling parameters to measure the normalisation of these contributions, determined through log-likelihood minimisation. The control region with $\Delta\eta_{jj} < 5$ is enriched with "hard DY" events, while the region with $\Delta\eta_{jj} \geq 5$ primarily contains events with jets from pileup interactions, as explained in Section 4.6.

In the $e\mu$ category, the dominant component of the DY background is represented by DY $\tau\tau$ events, where τ leptons undergo leptonic decays. This relatively minor contribution is estimated through a so-called "embedded sample", which is generated using data-driven techniques. Instead of using traditional MC simulations to model the behaviour of τ leptons, the data deriving from $Z \to \mu\mu$ in which the two muons are replaced with simulated tau leptons possessing the same four-momentum characteristics as the original muons they replace. This process ensures that the kinematic properties, such as the energy and momentum, of the tau leptons introduced into the events accurately mimic the behaviour of real tau leptons. The rationale behind using an embedded sample lies in its ability to provide a more faithful representation of the experimental data compared to traditional MC simulations, especially for observables that are highly sensitive to the detector's response and calibration, like p_T^{miss}. For a comprehensive understanding of this technique, refer to [95]. Additionally, for the remaining DY events that access the different flavour phase space but fail to fire the $e\mu$ trigger (explained in Table 4.1), they are simulated using standard MC techniques.

Finally, the normalisation of the DY $\tau\tau$ background is constrained in the fit procedure by utilising the DY $e\mu$ control region.

QCD-induced $\mathrm{W^+W^-}$ Background

The normalisation of the non-resonant QCD-induced $\mathrm{W^+W^-}$ background is not fixed, as is the case for the $\mathrm{t\bar{t}}$-tW and DY backgrounds. Instead, it is determined dynamically during the fit procedure.

Unlike the $\mathrm{t\bar{t}}$-tW and DY backgrounds, it is challenging to establish a sufficiently pure phase space that exclusively isolates this particular background due to overlap with other

processes such as $\mathrm{t\bar{t}}$-tW, W+jets, and DY (primarily in ee and $\mu\mu$ channels).
Consequently, the normalisation of the $\mathrm{W^+W^-}$ background is constrained in the fit without the availability of a dedicated control region that corresponds directly to this specific background process.

Nonprompt Leptons Background

Another source of background estimated directly from data is the nonprompt leptons background, originating mainly from W+jets events. The estimate is performed with a control sample enriched in W+jets events, and a separated control region enriched in QCD di-jet events.
The W+jets events are selected in data requiring one charged lepton with tight isolation and identification criteria (consistent with those applied in the main analysis), and a second charged lepton is chosen with loose requirements, that fails the tight ones.
Simultaneously, a control region is established, enriched in QCD di-jet data. These data are selected with dedicated triggers, requiring the presence of either one electron and one jet or a single muon. These triggers employ loose selection criteria to enhance the presence of nonprompt leptons in the sample. Additionally, charged leptons selected in this context are required to exhibit substantial isolation from the highest-p_T jet.
The probability that nonprompt events from the di-jet control region pass the tight requirements of the analysis is evaluated in function of the η and p_T of the nonprompt leptons. To avoid biases in this estimate, potentially due to the contribution from real leptons, events with $p_T^{\text{miss}} < 20\,\text{GeV}$ are excluded to suppress contributions from W bosons leptonic decays. Furthermore, simulated DY events are subtracted from the data to eliminate any influence from the EW production of Z+jets.
Finally, the calculated probability is applied to weight data in the W+jets control sample, to predict the number of W+jets events that are expected to enter the signal region.

Other Minor Backgrounds

Additional minor background sources in the analysis include the Higgs boson production and the multiboson processes, both of which are entirely derived from simulation. The production of Higgs bosons is effectively minimised by implementing a stringent mass cut, while the multiboson processes are significantly diminished by the exclusion of events containing a third additional lepton.

4.8 Systematic Uncertainties

The precision of the measurements is influenced by various sources of uncertainty. Specifically, the integrated luminosity exhibits uncertainties of 1.2%, 2.3%, and 2.5% for the years 2016, 2017, and 2018, respectively [63, 64, 65], resulting in a cumulative 2.1% effect on the signal cross-section. Additionally, uncertainties associated with trigger efficiency primarily affect the normalisation of the templates, amounting to less than 1% in magnitude.

Conversely, uncertainties linked to lepton reconstruction and identification efficiencies induce shape variations, with an impact of approximately 1.5% for electrons and 2% for muons [96, 97]. Prefiring corrections, which involve additional adjustments made specifically for the 2016 and 2017 datasets, collectively contribute a 3.3% uncertainty to the signal yield. Prefiring occurs when particles in a collision event are erroneously detected as triggers in the calorimeter before the actual collision takes place. These corrections aim to account for such prefire events and mitigate their impact on the measurements. The lepton momentum scale, jet momentum scale and resolution uncertainties [98] introduce modifications to both the shape and normalisation of the templates. These changes are quantified by adjusting the p_T of the leptons and jets by one standard deviation from their nominal values. The impact on the signal cross-section is minimal, with deviations of less than 1% attributed to lepton-related uncertainties and approximately 2.6% attributed to jet-related uncertainties. Furthermore, uncertainties associated with p_T^{miss} [99], excluding those linked to lepton and jet momentum scales, are evaluated by varying the momentum of each PF candidate not associated with a lepton or a jet. These variations result in a 2.4% impact on the signal cross-section. The corrections related to b-tagging [87] include multiple sources of uncertainty, collectively contributing to a 3.5% impact on the signal yield. Lastly, the uncertainty associated with the pileup reweighting procedure is assessed to be approximately 1.8%.

Theoretical uncertainties stemming from the choice of QCD renormalisation and factorisation scales are assessed in the analysis. This assessment involves independently varying these scales upwards and downwards by a factor of two with respect to their nominal values. As elucidated in Section 4.7, the normalisation of the main background processes is constrained using data. Similarly, the cross-section of the signal is directly measured within the fit process. When examining the QCD scales, the focus is solely on the effects on the shapes of the distributions. The VBS W^+W^- signal is a purely EW process at LO, and so any variations in the normalisation scale, which affects the coupling of α_S, do not manifest as discernible effects in the signal. Consequently, for the signal, the analysis primarily considers shape-related effects resulting from changes in the factorisation scale. The PDF uncertainties are computed as recommended by the NNPDF prescription [75]. PDF uncertainties may introduce shape variations in both the signal and background processes, since their normalisation are constrained in the fit. A thorough examination of the signal shape of the DNN output and m_{jj}, has revealed no significant deviations attributable to PDF uncertainties, and hence they are not considered. For $t\bar{t}$ and DY backgrounds, since normalisation effects have no impact in the fit, PDF uncertainties can only affect the ratio of the expected yields between the signal and the control region. Such uncertainties are included in the control regions and estimated to be 1% and 2% for $t\bar{t}$ and DY backgrounds, respectively.

Table 4.8 presents an overview of the most significant systematic uncertainties affecting the measurement of the signal cross-section. Uncertainties with values below 1% are not detailed.

The most influential ones are theoretical: the QCD factorisation scale for the VBS signal, and the QCD scale (including renormalisation and factorisation) of the $t\bar{t}$ background.

Uncertainty source	Value
QCD-induced W^+W^- normalisation	5.3%
$t\bar{t}$ QCD scale	5.1%
QCD factorisation scale for VBS signal	5.0%
$t\bar{t}$ normalisation	4.9%
b tagging	3.5%
Trigger corrections	3.3%
DY normalisation	2.9%
Jet energy scale + resolution	2.6%
p_T^{miss} energy scale	2.4%
QCD-induced W^+W^- QCD scale	2.1%
Integrated luminosity	2.1%
Muon efficiency	2.0%
Pileup	1.8%
Electron efficiency	1.5%
Underlying event	1.3%
Parton shower	1.0%
Other	< 1%
Total systematic uncertainty	13.1%
Total statistical uncertainty	14.9%
Total uncertainty	**19.8%**

Table 4.8: Summary of Systematic Uncertainties Impacting the signal cross-section measurement (%). Uncertainties exceeding 1% are displayed. The table includes statistical uncertainty, the sum of systematic contributions, and the overall uncertainty.

Among the experimental uncertainties, the ones associated with b-tagging are particularly noteworthy because they have a significant influence. The b-veto directly influences the count of events entering the signal region.
Despite the presence of systematic uncertainties, statistical uncertainty remains the predominant contributor to the overall uncertainty. Specifically, the root sum of squares of all systematic sources results in a relative uncertainty of 13.1% on the cross-section, while the statistical component is estimated to be 14.9%. This estimate of statistical uncertainty is derived by setting all systematic uncertainty components to their respective best-fit values. Ultimately, the total relative uncertainty on the signal cross-section amounts to 19.8%. As with other VBS analyses, this study is expected to benefit from the accumulation of a more extensive dataset in the future.

4.9 Signal Extraction

In this section, the process of signal extraction is explored, delving into the statistical procedures employed to evaluate hypotheses, with a specific focus on the significance of the results. Additionally, the results of the analysis, particularly the observed significance of the signal, are discussed.

4.9.1 Statistical Procedure

In this section, the statistical procedure used to evaluate hypotheses is delved into, with a specific focus on the extraction of signal strength (μ), as defined in Equation (4.12), and the significance of the results. This procedure is fundamental in distinguishing between the null hypothesis (H_0), which assumes no signal, and alternative hypotheses (H_μ) that allow for the presence of a signal.

A modified frequentist methodology is employed, based on a maximum likelihood fit of differential distributions, to extract the parameter μ associated with VBS. This parameter μ is then used to adjust the expected cross-section and determine the observed cross-section value. Let's take a closer look at what this means.

The signal strength parameter μ is defined as the ratio of the observed signal cross-section (σ_{obs}) to the signal cross-section predicted by a specific hypothesis (σ_{exp}):

$$\mu = \frac{\sigma_{\text{obs}}}{\sigma_{\text{exp}}}. \tag{4.12}$$

Each hypothesis H_μ introduces a signal with signal strength μ. The null hypothesis H_0 corresponds to $\mu = 0$, signifying a scenario where no signal events are present. The expected event yields per a certain hypothesis can be defined as:

$$N = \mu \cdot s + b, \tag{4.13}$$

where s is the number of expected signal events and b of background event.

To quantitatively assess a hypothesis, all the information from the data $\vec{x}$ is summarised into a single meaningful variable called the test statistic $t(\vec{x})$. By comparing the observed test statistic with values computed under different hypotheses, their level of agreement can be quantified. The test statistic used in this analysis is the profile likelihood ratio:

$$t(\vec{x}|\mu) = \frac{\mathcal{L}(\vec{x}|\mu, \hat{\hat{\theta}})}{\mathcal{L}(\vec{x}|\hat{\mu}, \hat{\theta})}, \tag{4.14}$$

where:

$\hat{\hat{\theta}}$ represents the best fit of the nuisance parameters for a fixed μ,

$\hat{\mu}$ and $\hat{\theta}$ are the best fit values when both are allowed to vary in the likelihood, maximising it.

p_0	Significance σ
$15.8 \cdot 10^{-2}$	1
$2.3 \cdot 10^{-2}$	2
$1.3 \cdot 10^{-3}$	3
$3.2 \cdot 10^{-5}$	4
$2.9 \cdot 10^{-7}$	5

Table 4.9: Significance levels corresponding to p-values for hypothesis testing.

This test statistic ranges from 0 to 1, with values closer to 1 favouring the tested μ and values closer to 0 indicating no significant deviations from backgrounds.
To assess the null hypothesis (H_0), the discovery test statistic [100] is computed:

$$\lambda_0 = -2\ln(t(\vec{x}|0)). \tag{4.15}$$

A high value of λ_0 suggests poor agreement between the null hypothesis and the data.
When analysing data collected by the LHC, a single value of λ_0, denoted as λ_0^{obs}, is obtained. If the experiment were to be repeated multiple times, a probability density function (pdf) for this random variable could be generated. In this case, such a pdf can be built through MC simulations, assuming the null hypothesis H_0, denoted as $f(\lambda_0|H_0)$.
The p-value [101] is then introduced as:

$$p_0 = \int_{\lambda_0^{\text{obs}}}^{+\infty} f(\lambda_0|H_0). \tag{4.16}$$

Under the assumption of H_0, the p-value p_0 represents the probability of obtaining a result as incompatible with the model or worse than the observed one (i.e., with λ_0 equal or larger).

In particle physics, it is customary to express results in terms of significance, which relates to the area under the tail of a Gaussian distribution. The threshold for discovery corresponds to 5 standard deviations. The p-value is converted into an equivalent significance (σ) using a standard Gaussian distribution ($G(0,1)$):

$$p_0 = \int_{\sigma}^{+\infty} G(0,1). \tag{4.17}$$

Different p-values correspond to various levels of significance, as shown in Table 4.9. To extract the expected significance σ, various input variables could be used. In this project, a DNN discriminator is utilized in the $e\mu$ phase space, along with the distribution of m_{jj} for the same flavour phase spaces. A likelihood for signal and background is constructed based on the distribution of event yields with respect to these variables. This approach is also known as shape analysis.
The histograms of a distribution is divided into N bins, each with an expected number of events E_i:

$$E_i(\mu, \theta) = \mu \cdot s_i(\theta) + b_i(\theta). \tag{4.18}$$

Here, $s_i(\theta)$ represents the expected signal yield for the i-th bin, and $b_i(\theta)$ is the expected background yield. The likelihood for the entire histogram $\vec{n} = (n_1, n_2, \ldots, n_N)$ is defined as:

$$\mathcal{L}(\vec{n}|\mu, \theta) = p(\theta) \cdot \prod_{i=1}^{N} \frac{(E_i(\mu, \theta))^{n_i}}{n_i!} e^{-E_i(\mu,\theta)}, \tag{4.19}$$

where $p(\theta)$ represents the pdf for the nuisance parameters, and the second term accounts for the product of Poisson probabilities describing the observation of E_i events in each bin. With this likelihood, the profile likelihood ratio is computed as defined in Equation (4.14), along with the corresponding p-value (Equation (4.16)) and significance (Equation (4.17)) to assess the hypothesis.

4.9.2 Results

In the $e\mu$ signal region, the signal phase space is partitioned into bins based on the DNN output, defined in Section 4.5. However, in the ee and $\mu\mu$ categories, distinct signal phase space regions are defined based on two key variables: m_{jj} and $\Delta\eta_{jj}$. These regions are defined as follows:

1. $300 < m_{jj}$ [GeV] < 500 and $2.5 < \Delta\eta_{jj} < 3.5$;
2. $m_{jj} > 500\,\text{GeV}$ and $2.5 < \Delta\eta_{jj} < 3.5$;
3. $300 < m_{jj}$ [GeV] < 500 and $\Delta\eta_{jj} > 3.5$;
4. $m_{jj} > 500\,\text{GeV}$ and $\Delta\eta_{jj} > 3.5$.

The number of events in each bin of the templates included in the likelihood function is modeled as a Poisson random variable, with a mean value that is the sum of the contributions from all processes.
The last region, characterised by the highest signal-to-background ratio, is further subdivided into bins based on m_{jj}. Irrespective of the final state under consideration, the control regions are designed such that the number of events is fit to the data. This fit process serves to constrain the normalisation of the $\mathrm{t\bar{t}}$-tW and DY backgrounds, ensuring an accurate representation of the data in the analysis.

Now, here is where the concept of "fiducial volumes" comes into play: using the measurement of μ_{EW} obtained from the fit, the production cross-section of the EW $\mathrm{W^+W^-}$ process within two distinct fiducial volumes can be estimated. These fiducial volumes are designed to correspond reasonably closely to what is experimentally accessible by the detector and reconstruction/identification algorithms. The exact definition of this phase space is somewhat arbitrary, but in any case it is designed to define the measurement in a maximally model-independent way, so that theory and experiment can be compared without needing to rely on extrapolation beyond the experimental acceptance. Fiducial measurements have the advantage of being more model-independent than extrapolating ones. In simpler terms, μ_{EW} is used to quantify how strong the signal of interest (EW W^+W^-) is within these carefully selected volumes of the data.

The Figures Figure 4.33 through Figure 4.36 display the best-fit values (left column) and the impacts (right column) of the initial 30 influential nuisance parameters within the fit process. These figures cover fits to the datasets from 2016, 2017, 2018, and the full Run 2 dataset.

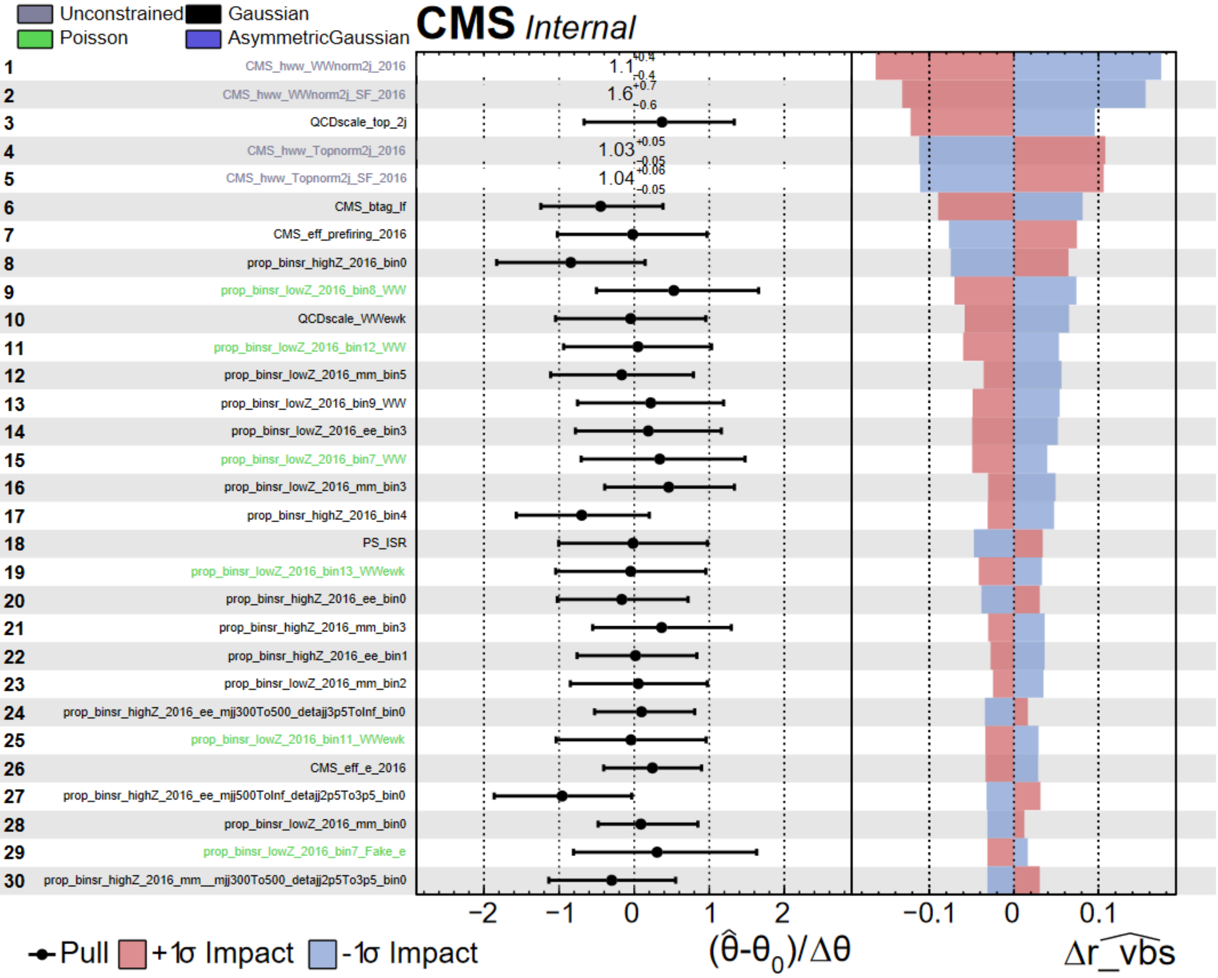


Figure 4.33: Best-fit values and impacts of the first 30 influential nuisance parameters in the combined fit with 2016 dataset. Best-fit values are depicted in the figure for nuisance parameters (black point) and for normalisation parameters (grey) for the $t\bar{t}$-tW, DY and QCD-induced W^+W^- backgrounds, which are directly constrained from the fit to data.

In Table 4.10, the post-fit yields for various processes in both the VBS $e\mu$ and $ee/\mu\mu$ signal regions can be found. The categories involving the $e\mu$ final state leptons significantly contribute to the overall significance. This is because they exhibit a superior signal-to-background ratio when compared to the same flavour channels. In contrast, the purity of

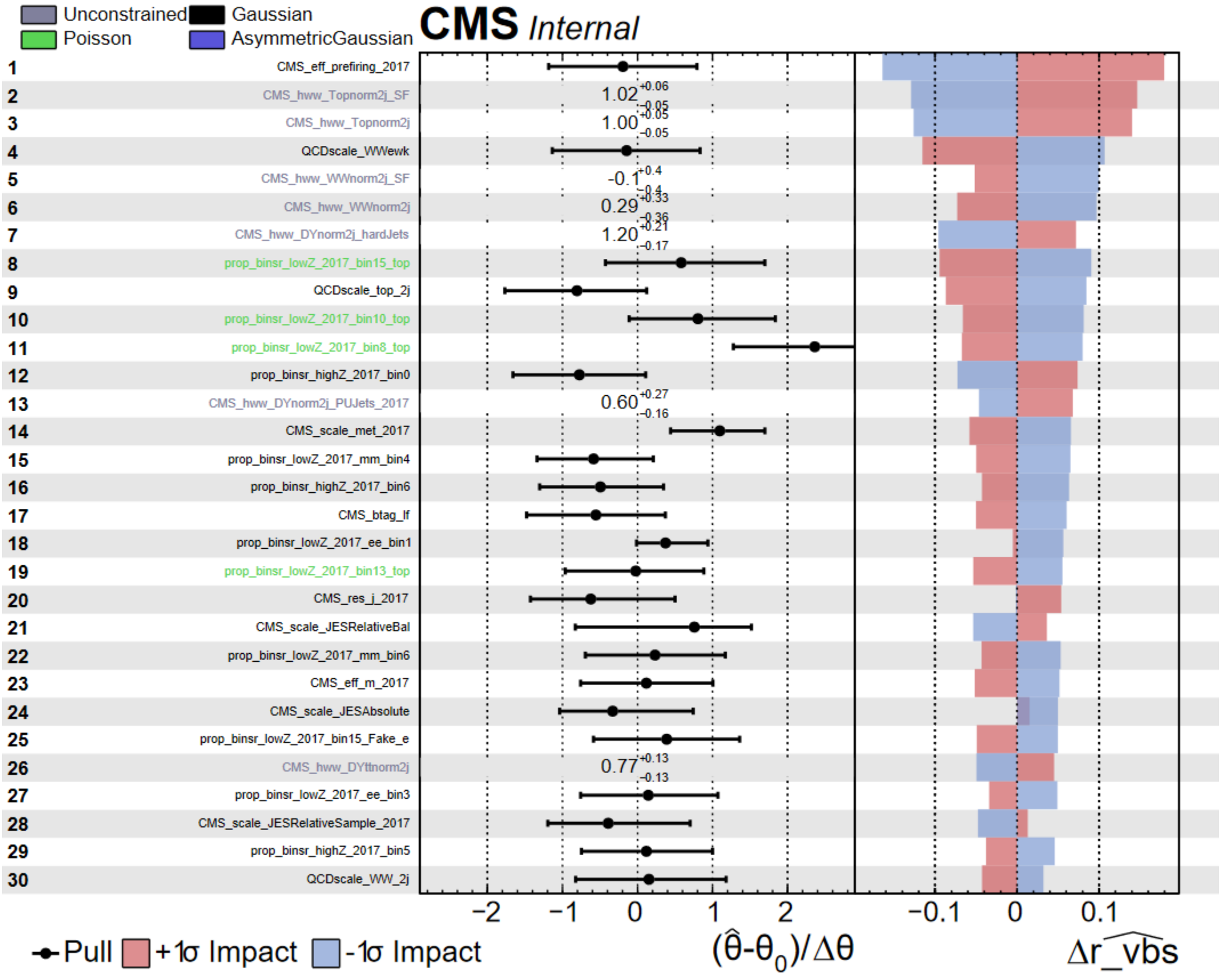


Figure 4.34: Best-fit values and impacts of the first 30 influential nuisance parameters in the combined fit with 2017 dataset. Best-fit values are depicted in the figure for nuisance parameters (black point) and for normalisation parameters (grey) for the $\mathrm{t\bar{t}}$-tW, DY and QCD-induced $\mathrm{W^+W^-}$ backgrounds, which are directly constrained from the fit to data.

the same flavour channels is somewhat compromised due to the substantial presence of DY production. Nonetheless, these same flavour channels play a crucial role in enhancing the overall performance of the analysis. Their inclusion helps us reach the critical 5 standard deviations threshold required for an observation. Figures 4.37 and 4.38 display the post-fit distributions observed in the full Run 2 dataset. These figures present data binning based on the DNN output for the $e\mu$ category, as well as binning according to m_{jj} and $\Delta\eta_{jj}$ for the ee and $\mu\mu$ categories. Furthermore, Figure 4.39 provides insight into the post-fit event counts within the control regions.

The investigation of the difference between data and MC in the last bin of the DNN output for $Z_{\ell\ell} < 1$ in Figure 4.37 involved a thorough examination of the input variables

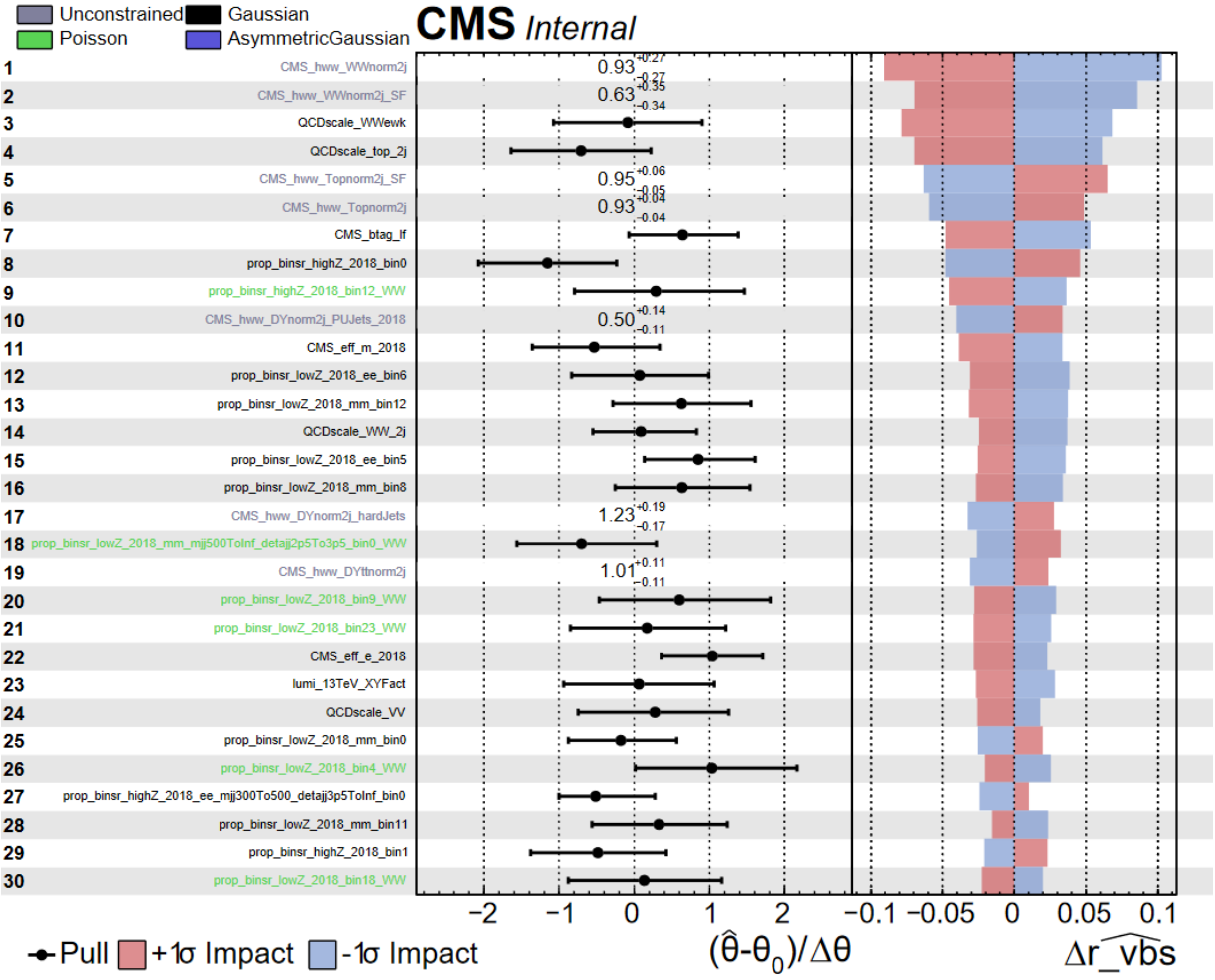


Figure 4.35: Best-fit values and impacts of the first 30 influential nuisance parameters in the combined fit with 2018 dataset. Best-fit values are depicted in the figure for nuisance parameters (black point) and for normalisation parameters (grey) for the $t\bar{t}$-tW, DY and QCD-induced W^+W^- backgrounds, which are directly constrained from the fit to data.

of the DNN. This examination was carried out to ensure that the input variables aligned reasonably well with data when focusing on events with the DNN output greater than 0.88. The observed discrepancy was not confined to specific bins within these distributions. Given the strong agreement in modelling the top background, the discrepancy is considered as consistent with a statistical under-fluctuation of the data.

Figure 4.40 presents the background-subtracted plot specifically for the $e\mu$ categories. This plot serves as a valuable tool for visualising the presence of the signal over the background. In essence, it provides a direct comparison between the observed data minus the total post-fit background yield (black dots), and the post-fit signal (red line). In the high-score bins of the DNN output, the signal becomes distinctly visible. This visibility

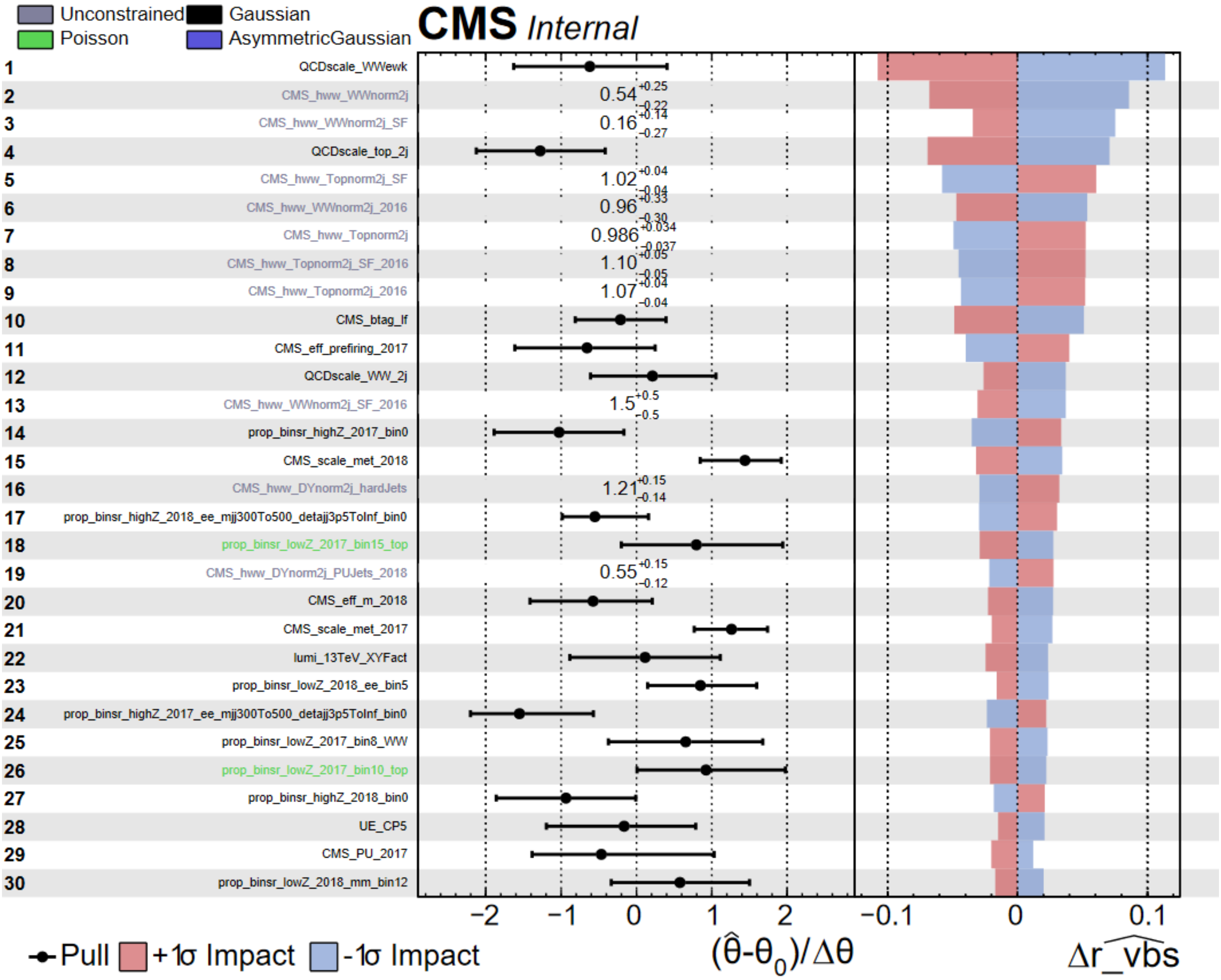


Figure 4.36: Best-fit values and impacts of the first 30 influential nuisance parameters in the combined fit with the full Run 2 dataset. Best-fit values are depicted in the figure for nuisance parameters (black point) and for normalisation parameters (grey) for the $t\bar{t}$-tW, DY and QCD-induced W^+W^- backgrounds, which are directly constrained from the fit to data.

is particularly pronounced, as anticipated, in the $Z_{\ell\ell} < 1$ category, where the signal stands out more prominently against the background, underlining the effectiveness of the analysis in this specific category. The EW W^+W^- signal is observed with a significance of 5.6 standard deviations. This measurement notably exceeds the anticipated significance of 5.2 standard deviations based on the Asimov data set [100], which is a constructed dataset used as a reference, designed to represent an idealised scenario used as a benchmark to assess the expected significance of the signal under optimal conditions.

The measurement of the cross-section is presented in two distinct fiducial phase spaces, each with varying levels of inclusivity and constraints. Firstly, the inclusive fiducial phase

Process	VBS $e\mu$	VBS $ee/\mu\mu$
$\mathrm{W^+W^-}jj$	238.9 ± 21.9	136.3 ± 6.9
$\mathrm{Z}jj$	—	60.8 ± 4.4
$\mathrm{t\bar{t}}$-tW	3081.9 ± 99.7	1410.3 ± 19.6
DY_PUJets	—	560.1 ± 46.9
DY_hardJets	—	667.7 ± 32.2
DY	171.2 ± 7.3	—
QCD-WW	736.3 ± 98.8	232.6 ± 24.0
Higgs	46.6 ± 1.8	—
Nonprompt	216.8 ± 24.6	62.0 ± 6.4
Multiboson	143.3 ± 9.8	121.1 ± 7.5

Table 4.10: Post-fit yields of the signal process (first row) and background processes in the signal regions. The yields are based on the merging of $Z_{\ell<1}$ and $Z_{\ell\geq1}$.

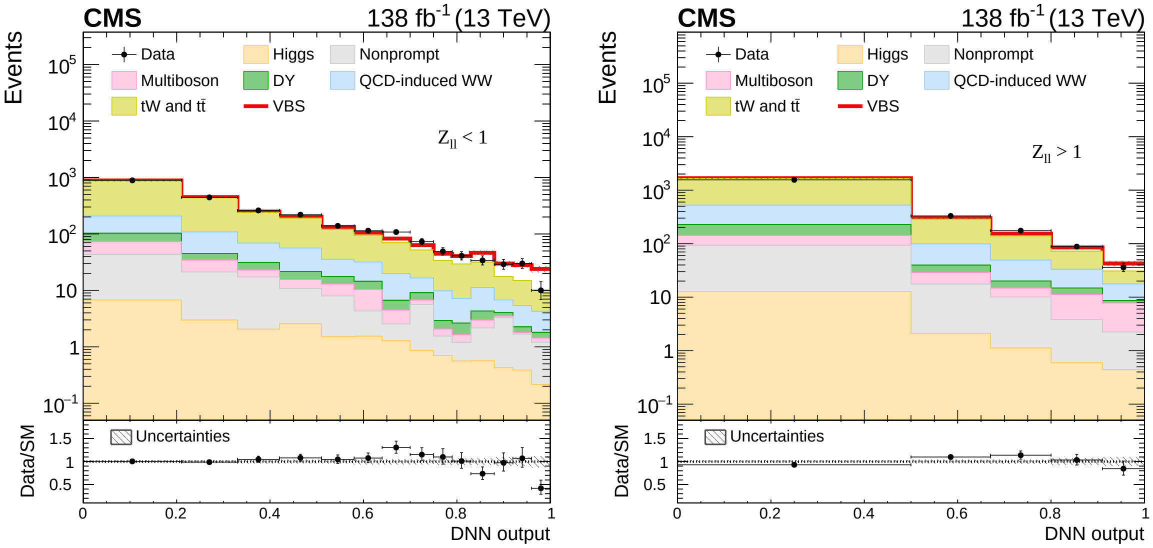


Figure 4.37: Post-fit DNN output distribution in $e\mu$ signal regions for $Z_{\ell\ell} < 1$ (left) and $Z_{\ell\ell} \geq 1$ (right) categories. Background and signal (red line) contributions are represented as stacked histograms, with systematic uncertainties depicted as dashed gray bands. Data points are displayed with asymmetric Poisson vertical bars to maintain proper statistical coverage across the spectrum.

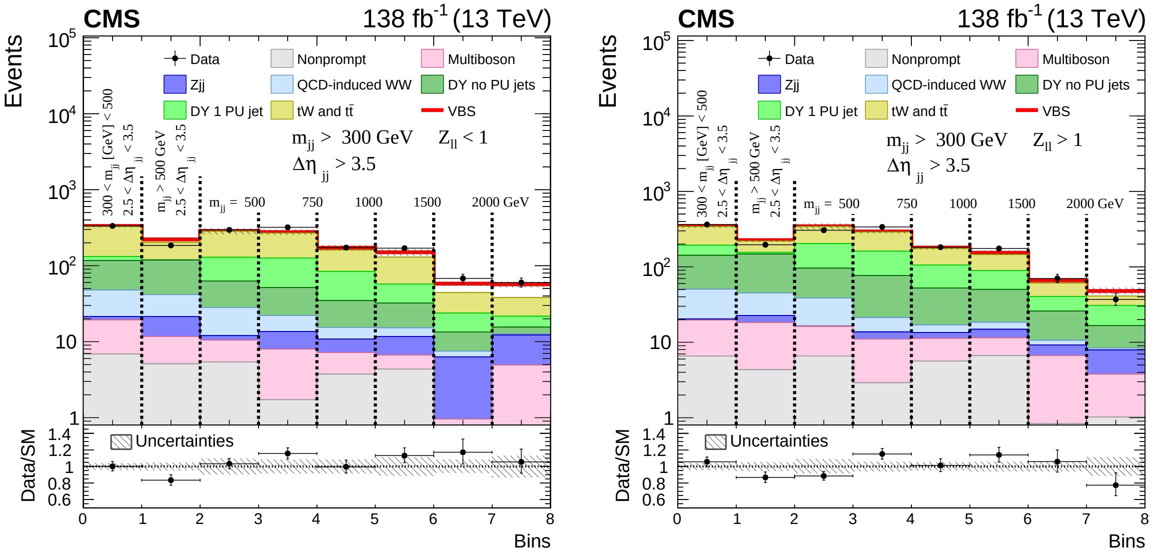


Figure 4.38: Post-fit m_{jj} distribution in $ee/\mu\mu$ signal regions for $Z_{\ell\ell} < 1$ (left) and $Z_{\ell\ell} \geq 1$ (right) categories.

The first two bins contain the number of events in the selected region (as reported in the plots themselves). The third bin contains the number of events in the $300 < m_{jj}\,[\mathrm{GeV}] < 500$ and $\Delta\eta_{jj} > 3.5$ regions and, for display purposes, is included in the m_{jj} distribution, shown in the last five bins.

Background and signal (red line) contributions are represented as stacked histograms, with systematic uncertainties depicted as dashed gray bands. Data points are displayed with asymmetric Poisson vertical bars to maintain proper statistical coverage across the spectrum.

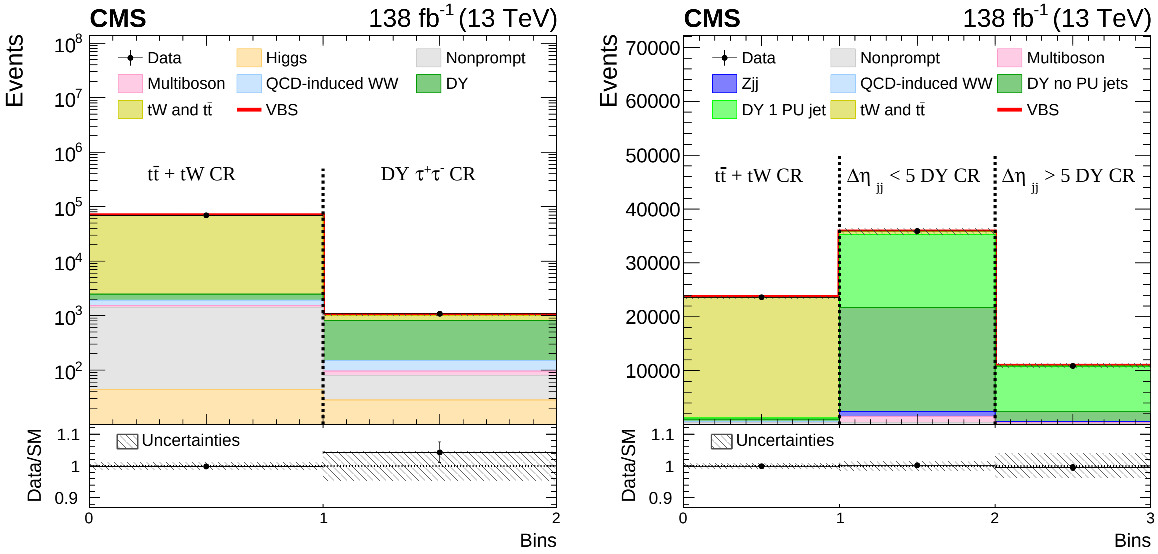


Figure 4.39: Post-fit number of events in $e\mu$ (left) and $ee/\mu\mu$ (right) control regions. In the left plot, the first bin contains the number of events in the t$\bar{\text{t}}$-tW $e\mu$ control region, and the second bin those in the DY$\tau\tau$ control region. In the right plot, the first bin contains the number of events in the t$\bar{\text{t}}$-tW $ee/\mu\mu$, the second bin those in the $\Delta\eta_{jj} < 5$ DY control region, and the third bin those in the $\Delta\eta_{jj} > 5$ DY control region. Background and signal (red line) contributions are represented as stacked histograms, with systematic uncertainties depicted as hatched gray bands. Data points are displayed with asymmetric Poisson vertical bars to maintain proper statistical coverage across the spectrum.

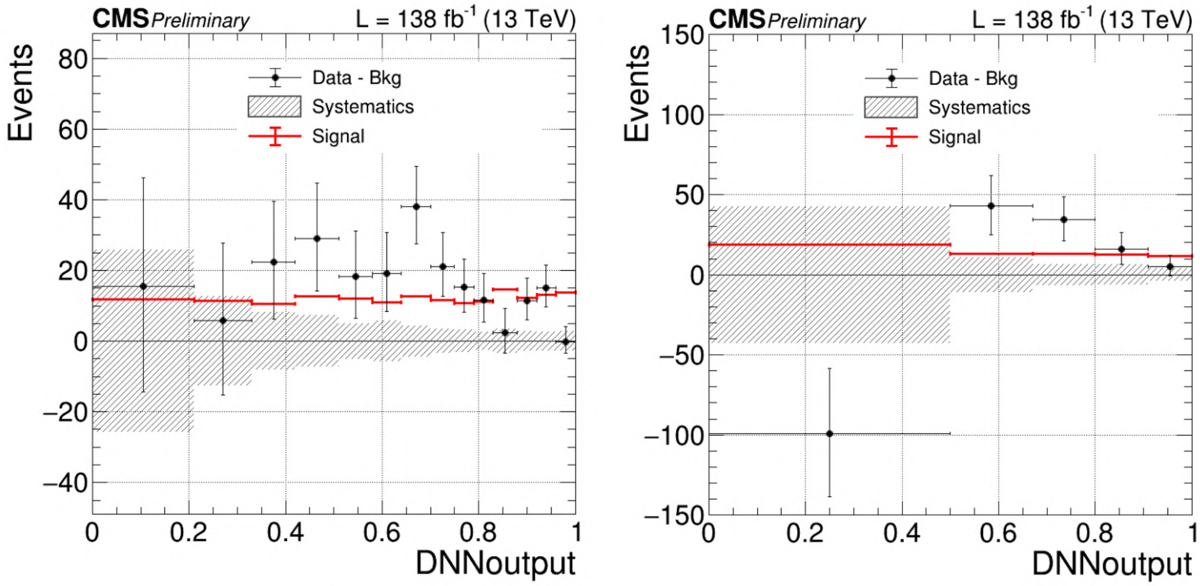


Figure 4.40: The background-subtracted DNN output distribution for the $e\mu$ $Z_{\ell\ell} < 1$ (on the left) and the $Zll \geq 1$ (on the right) categories. The black dots represent the yield of data minus total post-fit background, while the post-fit signal is drawn as the red line. The dashed band represents the total post-fit uncertainty.

space is characterised by relatively relaxed criteria applied to parton-level variables. Specifically, the two outgoing partons are required to have $p_T > 10\,\text{GeV}$ and $m_{qq'} > 100$ GeV. Within this phase space, the measured cross-section is found to be $99 \pm 20\,\text{fb}$, while the LO prediction stands at $89 \pm 5\,\text{fb}$. The assessment of the theoretical error on the expected cross-section involves variations in the QCD factorisation scale for the EW signal.

A tighter, exclusive fiducial volume is defined through selections applied at the generator level, which closely mimic the preselections outlined in Table 4.3. These criteria are summarised in Table 4.11. Within this exclusive fiducial volume, "dressed leptons" and "clean jets" are used. Dressed leptons incorporate the p_T of all photons found within a $\Delta R < 0.1$ radius around the lepton, effectively accounting for photon emissions associated with the lepton. Additionally, jets that contain at least one dressed lepton within a cone of $\Delta R = 0.4$ are excluded from consideration.

The measured cross-section in this more stringent fiducial phase space is determined to be $10.2 \pm 2.0\,\text{fb}$, while the LO theoretical prediction stands at $9.1 \pm 0.6\,\text{fb}$. The measured cross-sections align with the predictions of the SM, affirming the agreement between experimental observations and theoretical expectations.

Objects	Requirements
Leptons	e^+e^-, $\mu^+\mu^-$, $e^+\mu^-$, $e^-\mu^+$ $p_{T,\ell} = p_{T,\text{bare }\ell} + \sum_i p_{T,\gamma_i}$ if $\Delta R(\ell,\gamma_i) < 0.1$ $p_{T,\ell_1} > 25\,\text{GeV}$, $p_{T,\ell_2} > 13\,\text{GeV}$, $p_{T,\ell_3} < 10\,\text{GeV}$ $\lvert\eta\rvert < 2.5$, $p_{T,\ell\ell} > 30\,\text{GeV}$, $m_{\ell\ell} > 50\,\text{GeV}$
Jets	$p_{T,j} > 30\,\text{GeV}$, $\lvert\eta\rvert < 4.7$, $\Delta R(j,\ell) > 0.4$, At least 2 jets, no b jets $m_{jj} > 300\,\text{GeV}$, $\Delta\eta_{jj} > 2.5$
p_T^{miss}	$p_T^{\text{miss}} > 20\,\text{GeV}$

Table 4.11: Definition of the exclusive fiducial volume. The selections are applied on generator-level observables.

4.10 Summary and Outlook

This chapter reports the first observation of EW W^+W^- boson production in association with two jets at the LHC. The measurement uses data corresponding to an integrated luminosity of $138\,\text{fb}^{-1}$ collected by the CMS detector during 2016-2018 LHC proton-proton collisions at 13 TeV center-of-mass energy. The analysis focused on events characterised by the presence of two leptons (electrons or muons), two jets, and a significative amount of missing transverse energy. These events were further categorised based on the flavour of the final-state leptons, leading to the formation of distinct $e\mu$, ee, and $\mu\mu$ categories.

The challenges posed by the $t\bar{t}$ background and the irreducible QCD-induced W^+W^- production necessitated the deployment of advanced machine learning techniques. One notable innovation in the approach was the utilisation of a DNN, trained to discern the signal within the $e\mu$ categories, effectively separating it from the $t\bar{t}$ and QCD-induced W^+W^- backgrounds. This approach substantially enhanced the sensitivity of the analysis. Additionally, despite significant contamination from DY production, the ee and $\mu\mu$ categories are included to derive the final result.

The VBS W^+W^- signal was observed with a statistical significance of 5.6 standard deviations, exceeding the expected 5.2 standard deviations.

Furthermore, cross-sectional measurements were conducted within two distinct fiducial volumes: one inclusive and another closely resembling the signal region definition. In the inclusive phase space, the fiducial cross-section was determined to be $99 \pm 20\,\text{fb}$, in agreement with the LO theoretical prediction of $89 \pm 5\,\mathrm{fb}$. In more stringent phase space, the fiducial cross-section for the EW $\mathrm{W}^{+}\mathrm{W}^{-}$ signal was measured at $10.2 \pm 2.0\,\mathrm{fb}$, with the LO theoretical prediction $9.1 \pm 0.6\,\mathrm{fb}$. The cross-sectional measurements in the two fiducial volumes reveal agreement with SM predictions. However, their precision is still strongly limited by the statistical uncertainty.

Looking ahead, this research has promising prospects. The LHC started collecting new data in 2022, with the Run 3, and this will significantly increase the integrated luminosity available for analysis, thereby reducing statistical uncertainties. At the end of Run 3, the integrated luminosity accumulated by the CMS detector is expected to be 300 fb^{-1}. This analysis is expected to benefit from the larger data set collected as the contribution from the statistical uncertainty will be reduced. Furthermore, in 2026, the HL-LHC will start, featuring proton-proton collisions at a centre-of-mass energy of 14 TeV and an instantaneous luminosity three times greater than that of Run 2. This phase promises to provide an order-of-magnitude increase in integrated luminosity (3000 fb^{-1}), offering exciting opportunities for further exploration.

The understanding of VBS processes has grown significantly, transitioning from the first discovery with the 2016 Run 2 data sample to the achievement of the first precise measurements with the 2016-2018 data sets. The HL-LHC represents an ideal platform for expanding VBS analyses, including investigations into polarised cross-sections and EFT searches. These efforts hold the potential to provide valuable insights into the SM consistency, with a focus on longitudinally polarised scattering and the EW symmetry-breaking mechanism. Additionally, research into dimension-6 and dimension-8 EFT operators has already started, as will be seen in Chapter 5.

VBS analyses have emerged as a robust means of constraining both SM and BSM physics. While the era of groundbreaking discoveries may have passed, a new and even more exciting phase, focused on precise measurements, awaits at the LHC.

Sensitivity Study of W^+W^- VBS to dimension-6 EFT operators

This chapter delves into the analysis of the W^+W^- VBS process within the framework of dimension-6 EFT. The primary objective is to investigate indirect searches to uncover new physics phenomena that extend BSM. A detailed description of the W^+W^- VBS topology can be found in Section 1.2.

Firstly, the event generation process for both the SM and various EFT components is addressed (Section 5.1). Subsequently, some kinematic effects of the SMEFT operators under study are shown, along with their impact on the VBS cross-section (Section 5.2). Then, the analysis strategy is elaborated, including the construction of the likelihood function for the results extraction (Section 5.3). Finally, the resulting one-dimensional and two-dimensional fits are presented, along with the effect of the systematics on the individual constraints (Section 5.4).

5.1 Events Generation

Events are simulated using `Madgraph5_aMCNLO` (v. 2.6.5) [67], interfaced to the SMEFTsim package (v. 3) [32, 33]. The simulations were conducted for proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$. The NNLO parton distribution functions, provided by the NNPDF collaboration [75], were employed. The strong coupling constant was set to $\alpha_S = 0.118$, and the calculations were carried out within the framework of the four-flavour scheme (LHAPDF identification code 325500). The choice of normalisation and factorisation scales was determined by the MC generator, specifically as the transverse mass of the $2 \to 2$ scattering following k_T clustering.

The EFT components were extracted using two distinct techniques:

a) Direct Event Simulation: In this approach, dedicated event simulations were performed for each contribution. This included the computation of the SM signal (N_{SM}), the pure interference ($N^{\text{int}}_{\alpha,\text{vert.}}$), the quadratic term ($N^{\text{quad}}$) for individual operators, and the mixed quadratic contributions ($N^{\text{mix}}_{\alpha,\beta}$). The interaction-order syntax in the

`Madgraph5_aMCNLO` and SMEFTsim software packages facilitated the direct generation of these components.

b) Reweighting Technique: Alternatively, a reweighting method, integrated within `Madgraph5_aMCNLO`, was employed. This method involved generating events once, for a specific point in parameter space, and subsequently adjusting their weights to align with different sets of Wilson coefficients. This approach relies on the calculation of matrix elements at fixed phase-space points, ensuring an exact dependence of the weights on the Wilson coefficients. While N_α^{int} and N_α^{quad} could be determined directly, isolating the quantities $N_{\alpha,\beta}^{\text{mix}}$ often required algebraic combinations of results obtained at two or three different parameter points. The reweighting technique offers computational advantages and reduces statistical uncertainties. However, it could yield less reliable results in sparsely populated regions of phase space. To address this, both techniques were employed to estimate and cross-validate the signal dependence on the Wilson coefficients.

Linear corrections arising from SMEFT insertions at vertices and within the propagators of the W, Z, and Higgs bosons were simulated independently. Propagator corrections exhibited identical forms for all contributing operators.

In this study, consideration is limited to a specific subset of five bosonic operators of the Warsaw basis (described in detail in Section 1.3):

$$\begin{aligned} Q_W &= \varepsilon^{ijk} W_\mu^{i\nu} W_\nu^{j\rho} W_\rho^{k\mu}, \\ Q_{HW} &= (H^\dagger H) W_{\mu\nu}^i W^{i\mu\nu}, \\ Q_{HWB} &= (H^\dagger \sigma^i H) W_{\mu\nu}^i B^{\mu\nu}, \\ Q_{HD} &= (H^\dagger D_\mu H)(H^\dagger D^\mu H), \\ Q_{H\Box} &= (H^\dagger H)\Box(H^\dagger H), \end{aligned} \tag{5.1}$$

where:

- H represents the Higgs doublet,
- $W_{\mu\nu}$ and $B_{\mu\nu}$ denote the field strengths associated with the $SU(2)$ and $U(1)$ gauge symmetries, respectively,
- σ_i represents the Pauli matrices associated with $SU(2)$.

This collection includes the operators that manifest through adjustments to the EW input parameters (Q_{HD}, Q_{HWB}), along with a selection of operators that make substantial contributions to all VBS processes, particularly once experimental selection criteria are applied. These modifications are primarily induced by alterations in trilinear gauge couplings between vector bosons (Q_W) and HVV couplings (Q_{HD}, Q_{HW}, Q_{HWB}). This set of operators is a pragmatic choice for the study as it enables the examination of all the previously mentioned categories of SMEFT effects and the exploration of how VBS processes can constrain EFT parameters, all while avoiding the complexities of an exceedingly high-dimensional parameter space.

For a more complete EFT interpretation, the inclusion of pure fermionic operators, other bosonic operators like $Q_{HB} = H^\dagger H B_{\mu\nu} B^{\mu\nu}$, and contact interactions between two quarks and two leptons, would be necessary. However, for this preliminary investigation of VBS process sensitivity to EFT effects, the consideration of these five operators is deemed sufficient, with a more comprehensive analysis reserved for future research.

The selection of these five operators is based on the expectation that they might influence significantly the dynamics of the VBS under study. This choice was informed by an analysis of the shapes of some kinematic variables, as discussed in the following section.

5.2 Kinematic Effects of SMEFT Operators

In this section, the shapes of the most important kinematic observables for the pure SM and the BSM samples are presented. The EFT components for these observables are obtained using the five Wilson coefficients detailed in Equation (5.1). The objective here is to demonstrate the influence of these Wilson coefficients on the kinematic variables for the signal under consideration.

As illustrative examples, the influence of two Wilson coefficients, namely c_W and c_{HW}, set at $c_\alpha/\Lambda^2 = 1\,\mathrm{TeV}^{-2}$, on several key kinematic variables and the DNN output is depicted in Figures 5.1 to 5.3 for the low-$Z_{\ell\ell}$ signal region. Let's provide a detailed commentary on the distributions presented in the figure.

The operator Q_W is expected to have an impact on the high-energy regions, leading to alterations in both the cross-section and the shapes of kinematic variables, such as the invariant mass of the dilepton system ($m_{\ell\ell}$) or the dijet system (m_{jj}). The other operators involved in EFT modifications are more likely to primarily influence the overall cross-section without significant changes to the distributions.

Concerning the DNN, it does not categorise the EFT contributions as background processes. In fact, the DNN was trained to identify the SM VBS process as the signal. The EFT contributions have the potential to preserve the kinematics observed in the SM process. However, in the case of operators such as Q_W, they can “augment” or “boost” certain kinematic attributes. For instance, this operator may exhibit a tendency to generate highly energetic jets, thus enhancing those specific kinematic characteristics, such as m_{jj}.

The estimated cross-section for the BSM predictions keeping into account the five SMEFT operators under study, will now be considered. For each operator, the corresponding Wilson coefficient is set at $c_\alpha/\Lambda^2 = 1\,\mathrm{TeV}^{-2}$, and its cross-section will be referred to as $\sigma(\mathrm{SM} + c_\alpha)$. In contrast, the estimated cross-section for the SM prediction will be denoted as $\sigma(\mathrm{SM})$. The ratios between these two cross-section values, computed in the combined signal region (which includes both the low-$Z_{\ell\ell}$ and high-$Z_{\ell\ell}$ categories) will now

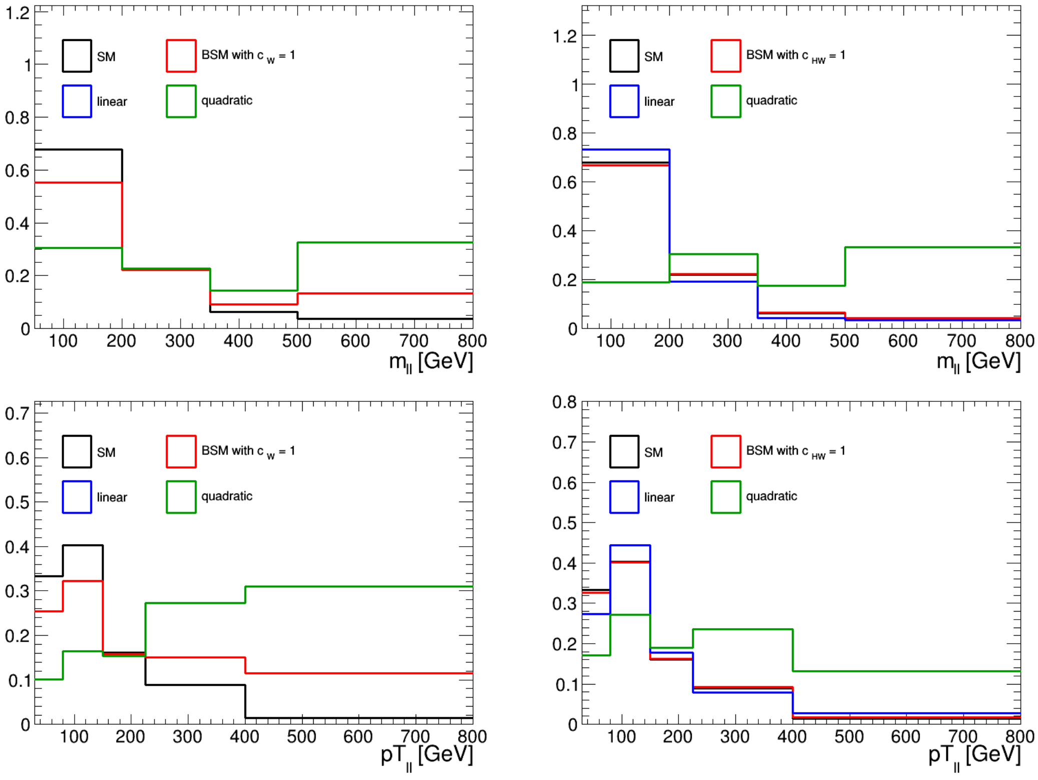


Figure 5.1: Comparison of normalised $m_{\ell\ell}$ (top) and $p_{T,\ell\ell}$ (bottom) distributions in the low-$Z_{\ell\ell}$ signal regions. The black line shows the SM prediction, the red line represents the BSM prediction with Wilson coefficients c_W (left) and c_{HW} (right), both set at $c_\alpha/\Lambda^2 = 1\,\text{TeV}^{-2}$. The linear and quadratic components are shown with blue and green lines, respectively. The last bin includes overflow events.

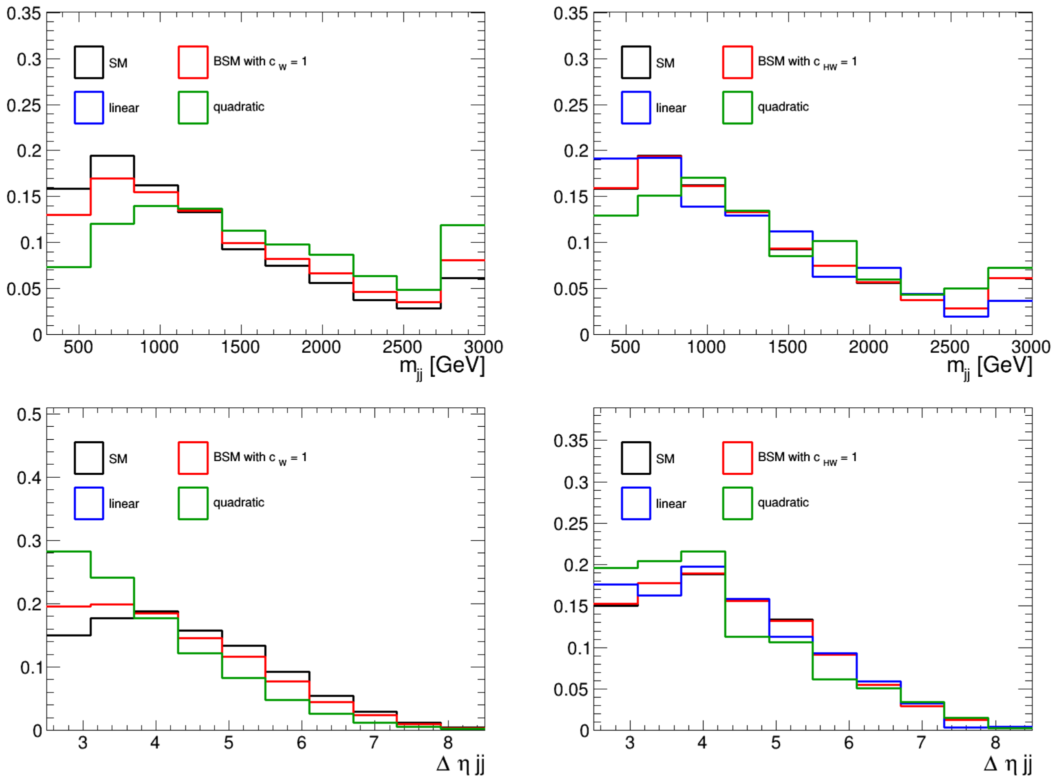


Figure 5.2: Comparison of normalised m_{jj} (top) and $\Delta\eta_{jj}$ (bottom) distributions in the low-$Z_{\ell\ell}$ signal regions. The black line shows the SM prediction, the red line represents the BSM prediction with Wilson coefficients c_W (left) and c_{HW} (right), both set at $c_\alpha/\Lambda^2 = 1\,\text{TeV}^{-2}$. The linear and quadratic components are shown with blue and green lines, respectively. The last bin includes overflow events.

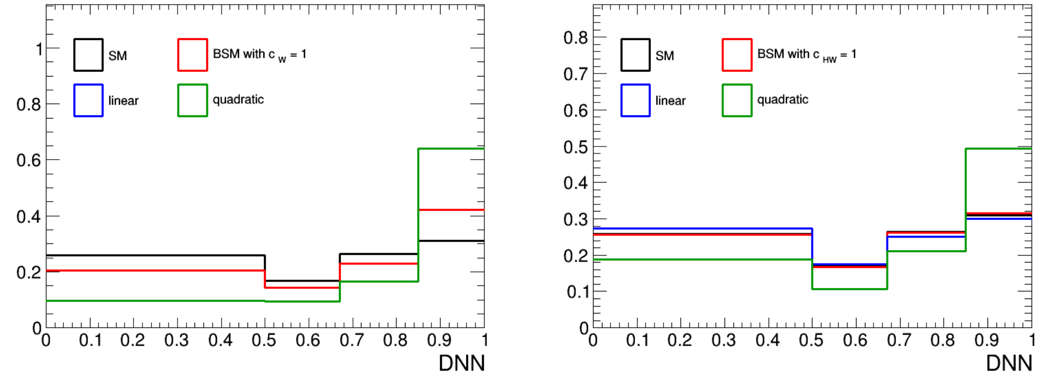


Figure 5.3: Comparison of normalised DNN output distributions in the low-$Z_{\ell\ell}$ signal regions. The black line shows the SM prediction, the red line represents the BSM prediction with Wilson coefficients c_W (left) and c_{HW} (right), both set at $c_\alpha/\Lambda^2 = 1\,\text{TeV}^{-2}$. The linear and quadratic components are shown with blue and green lines, respectively.

be examined. This ratio, denoted as $r(c_\alpha)$, is defined as:

$$r(c_\alpha) = \frac{\sigma(\mathrm{SM} + c_\alpha)}{\sigma(\mathrm{SM})}. \tag{5.2}$$

The computed ratios for each of the five Wilson coefficients under study are as follows:

$$\begin{aligned} r(c_W) &= 1.57, \\ r(c_{HW}) &= 1.06, \\ r(c_{HWB}) &= 0.93, \\ r(c_{HD}) &= 0.96, \\ r(c_{H\Box}) &= 0.98. \end{aligned} \tag{5.3}$$

Values below 1 are due to the linear interference between the SM amplitude and the EFT terms.

5.3 Analysis Strategy

In this section, the analysis strategy is outlined. The first step involves isolating a phase space region where the VBS signal becomes more prominent due to the suppression of background contributions. To accomplish this, the kinematic cuts already described in Section 4.4, used for the SM $\mathrm{W}^+\mathrm{W}^-$ VBS analysis, are employed.
This analysis is focused exclusively on the $e\mu$ phase space and on the 2018 data-taking year. The 2016-2017 data-taking years, as well as the ee and $\mu\mu$ phase spaces, will be addressed in upcoming studies.

The core of this analysis relies on two bidimensional phase spaces, namely the DNN output plotted against two key observables: the invariant mass ($m_{\ell\ell}$) and the transverse momentum ($p_{T,\ell\ell}$) of the dilepton system. The DNN output, explained in detail in Section 4.5, significantly aids in reducing two of the main background sources: top quark pair production ($\mathrm{t\bar{t}}$) and QCD-induced $\mathrm{W}^+\mathrm{W}^-$ production. On the other hand, $m_{\ell\ell}$ and $p_{T,\ell\ell}$ offer the strongest limit on the Wilson coefficients studied in this work. By combining the discriminating power of the DNN output with the precision offered by $m_{\ell\ell}$ and $p_{T,\ell\ell}$, a comprehensive strategy is created for the analysis, enabling the effective identification of the signal, suppression of background, and extraction of meaningful information about the Wilson coefficient.

In Figures 5.4 and 5.5, a comparison is presented, illustrating the distribution of the SM, the backgrounds and the BSM predictions in the signal region for $m_{\ell\ell}$, $p_{T,\ell\ell}$, and two 2D variables, DNN vs.$m_{\ell\ell}$ and DNN vs.$p_{T,\ell\ell}$. The BSM predictions for this example are generated setting $c_W = 1\,\mathrm{TeV}^{-2}$. The binning for $p_{T,\ell\ell}$ and $m_{\ell\ell}$ have been loosely optimised.

Figures 5.6 and 5.7 present the comparisons of observation and prediction in the $\mathrm{t\bar{t}}$ and DY control regions, for $m_{\ell\ell}$, $p_{T,\ell\ell}$, and two 2D variables, DNN vs.$m_{\ell\ell}$ and DNN vs.$p_{T,\ell\ell}$. The data/MC agreement is consistently strong across the two 2D variables of interest.

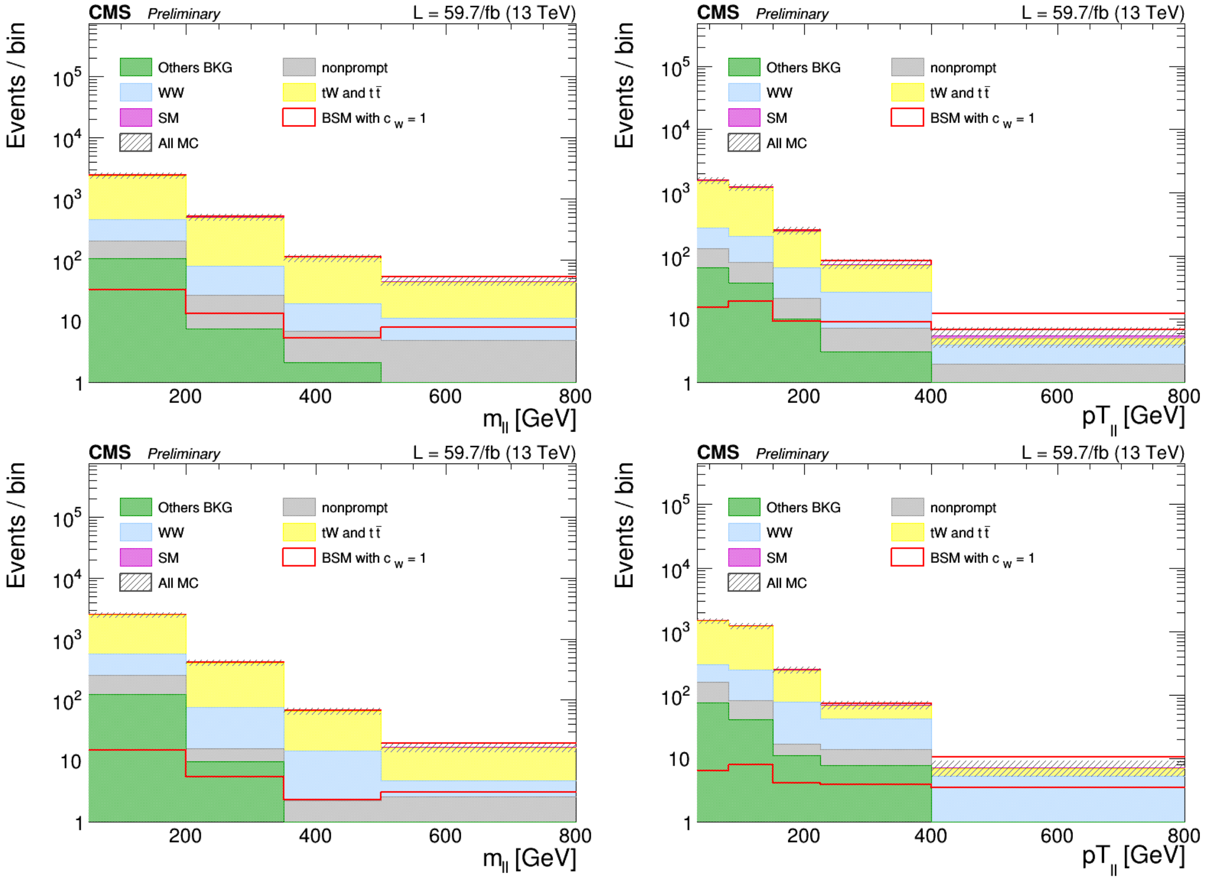


Figure 5.4: Number of events predicted in the low-$Z_{\ell\ell}$ (top) and high-$Z_{\ell\ell}$ (bottom) signal regions from 2018 samples for $m_{\ell\ell}$ (left) and $p_{T,\ell\ell}$ (right) are shown. The collective BSM effect with $c_W = 1\,\mathrm{TeV}^{-2}$ is indicated with a red line, both stacked and superimposed. The last bin includes overflow events.

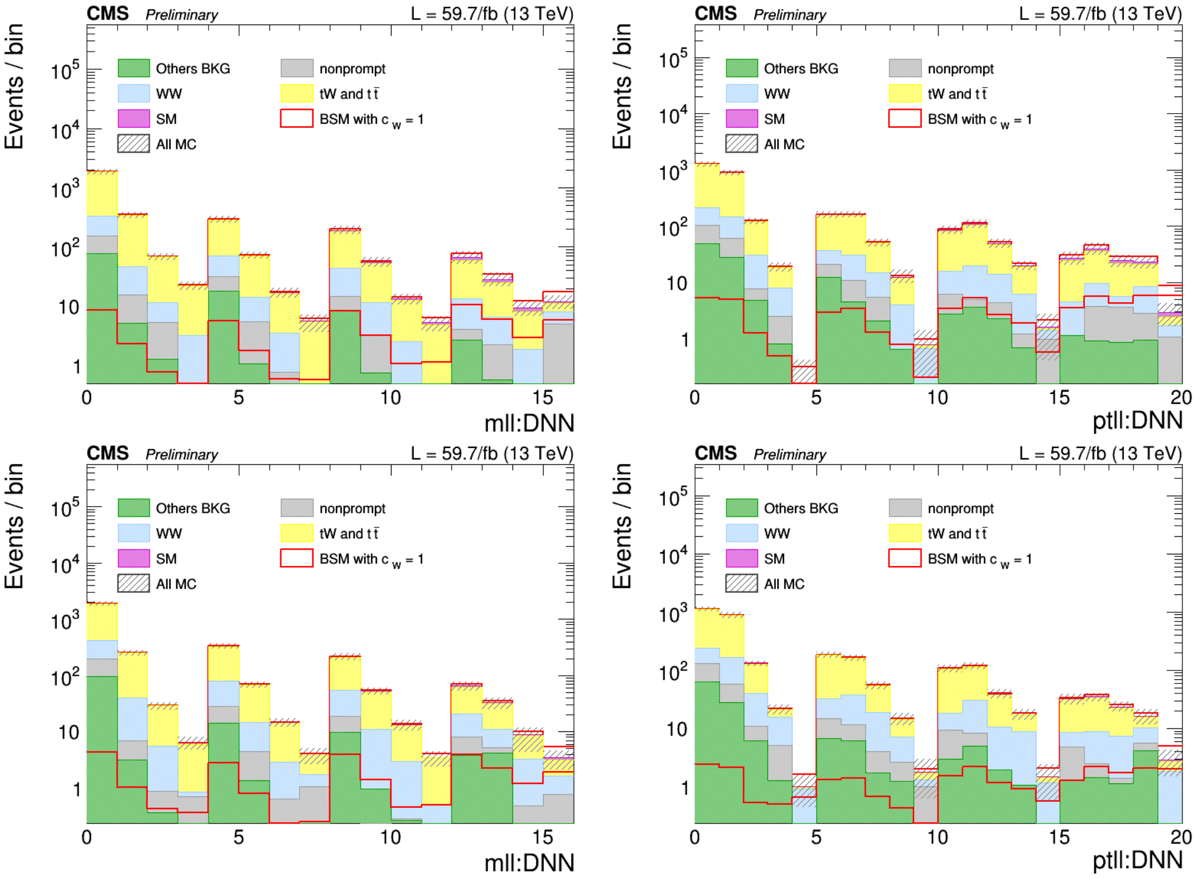


Figure 5.5: Number of events predicted in the low-$Z_{\ell\ell}$ (top) and high-$Z_{\ell\ell}$ (bottom) signal regions from 2018 samples. The 2D variables DNN vs.$m_{\ell\ell}$ (left) and DNN vs.$p_{T,\ell\ell}$ (right) are shown. The collective BSM effect with $c_W = 1\,\text{TeV}^{-2}$ is indicated with a red line, both stacked and superimposed. The x-axis bin counter represents the 4 bins in DNN output, the 4 bins in $m_{\ell\ell}$ (for a total of 16 bins), and the 5 bins in $p_{T,\ell\ell}$ (for a total of 20 bins).

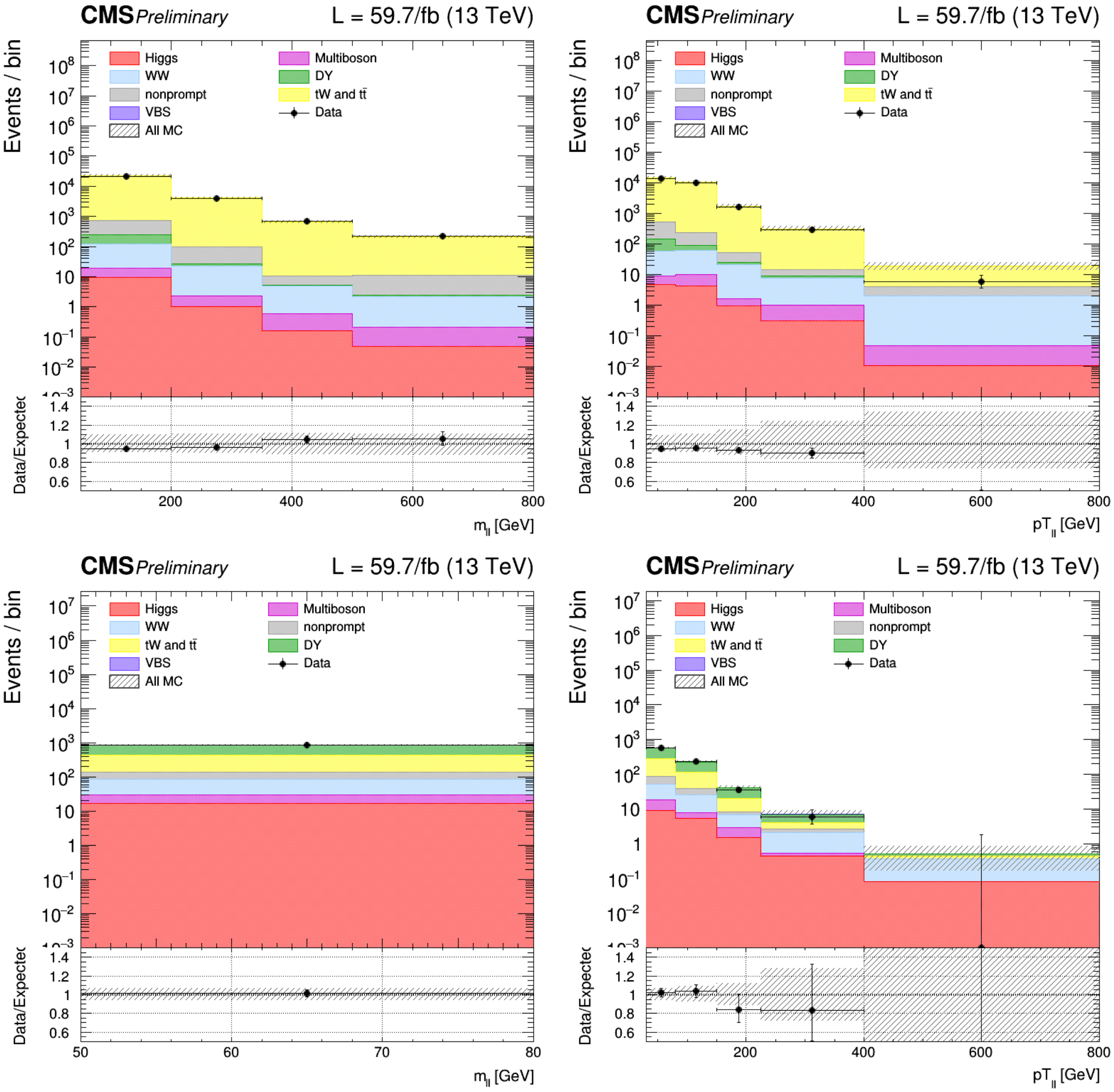


Figure 5.6: Comparison of observation and prediction in the $t\bar{t}$ (top) and DY (bottom) control regions from 2018 samples. The variables $m_{\ell\ell}$ (left) and $p_{T,\ell\ell}$ (right) are shown. The last bin includes overflow events.

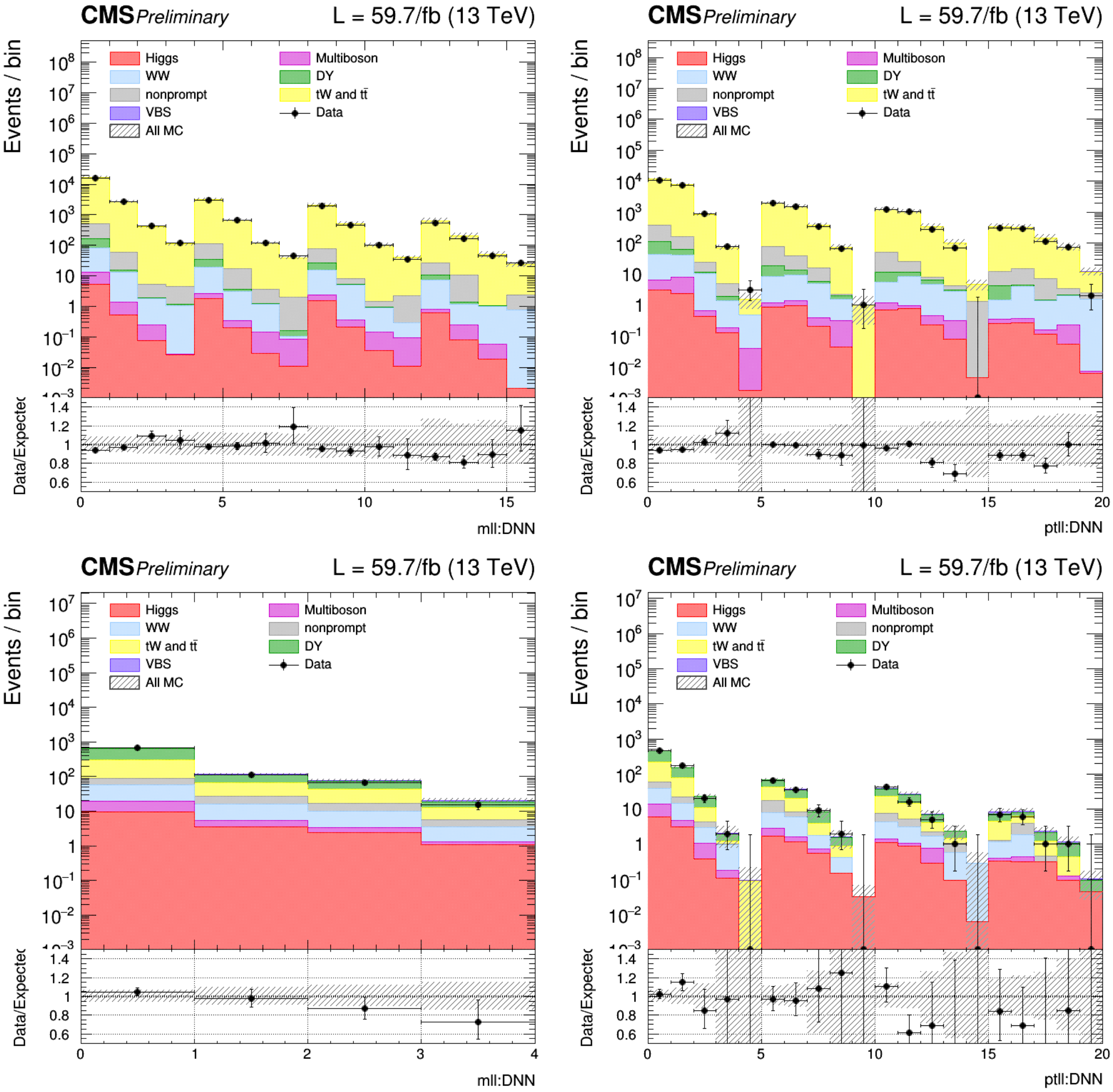


Figure 5.7: Comparison of observation and prediction in the $t\bar{t}$ (top) and DY (bottom) control regions from 2018 samples. The 2D variables DNN vs.$m_{\ell\ell}$ (left) and DNN vs.$p_{T,\ell\ell}$ (right) are shown. The last bin includes overflow events.

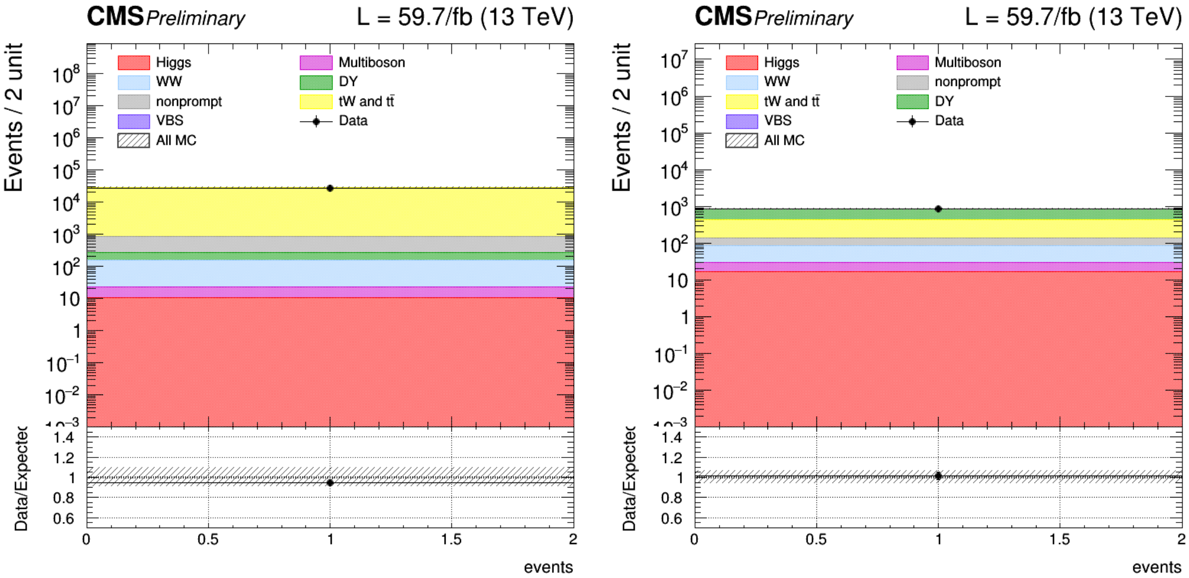


Figure 5.8: Comparison of observation and prediction for the number of events in the $t\bar{t}$ (left) and DY (right) control regions from 2018 samples.

For the fit, the number of events in the control regions was used to estimate the number of events of the background processes in the signal region. Figure 5.8 presents the comparison of observation and prediction in the $t\bar{t}$ and DY control regions for such observable. The data/MC agreement is consistent.

5.3.1 Likelihood Construction

The likelihood function, denoted as $\mathcal{L}$, is used to assess how well the data aligns with the EFT model. It is constructed based on Equation (1.21), which determines the expected number of events, $N_k(\mathbf{c})$, surviving the data analysis selections. These events can be considered either collectively or within individual bins (k-th distribution bin). The likelihood function is defined as:

$$\mathcal{L}(\mathbf{c}|\theta) = p(\theta) \cdot \prod_k \frac{N_k(\mathbf{c},\theta)^{n_k}}{n_k!} e^{-N_k(\mathbf{c},\theta)}. \tag{5.4}$$

In this equation, $\mathbf{c}$ represents the vector of free Wilson coefficients used in the fitting process, while N_k corresponds to the expected number of events, defined as in Equation (1.21). Additionally, n_k is defined as $n_k \equiv N_k(\mathbf{0})$, representing the expected number of events in the SM. The systematic uncertainties previously discussed in Section 4.8 are also considered in this analysis through the parameter θ.

The expected sensitivity to the Wilson coefficients is evaluated using the likelihood profile, already discussed in 4.9. The 68% and 95% confidence level (c.l.) intervals for coefficient estimates are determined by the conditions $-2\Delta \log \mathcal{L} < 1$ and $-2\Delta \log \mathcal{L} < 3.84$,

Operator	68% for DNN vs.$m_{\ell\ell}$	68% for DNN vs.$p_{T,\ell\ell}$
Q_W	$[-0.88, 0.69]$	$[-0.63, 0.54]$
Q_{HW}	$[-1.22, 1.11]$	$[-1.20, 1.09]$
Q_{HWB}	$[-1.51, 1.42]$	$[-1.41, 1.39]$
Q_{HD}	$[-1.42, 1.35]$	$[-1.40, 1.38]$
$Q_{H\Box}$	$[-1.37, 1.41]$	$[-1.44, 1.31]$

Table 5.1: Individual expected constraints on Wilson coefficients at the 68% confidence level.

respectively [5]. Here, $\Delta \log \mathcal{L}$ measures how much the logarithm of the likelihood function changes from its maximum value as it moves away from the best-fit value of the Wilson coefficients.

5.4 Results

In this section, the results of likelihood scans of the five SMEFT operators under study, reported in Equation (5.1), are presented.

5.4.1 One-Dimensional Constraints

Figures 5.9 and 5.10 present the profiled $-2\Delta \log \mathcal{L}$ as a function of individual Wilson coefficients for the two 2D variables under investigation, namely DNN vs. $m_{\ell\ell}$ and DNN vs. $p_{T,\ell\ell}$, respectively. The influence of the nuisance parameters on the fit is illustrated through a comparison between the results with and without the consideration of systematic effects.

For all operators, except for Q_W, there is a substantial disparity between the limits when considering the inclusion of systematics versus when not including them. However, Q_W appears to be relatively insensitive to the effects of systematics. This insensitivity can be attributed to its more pronounced impact on the high-energy tails of the distributions, where low statistics are observed. In these cases, statistical fluctuations are likely to dominate the results for the Q_W operator.

The resulting limits at 68% confidence-level with the systematics included are reported in Table 5.1. The most-stringent limits are posed by Q_W, as expected.

5.4.2 Two-Dimensional Constraints

In this section, the constraints obtained by allowing two operators to vary simultaneously while fixing the remaining ones to zero are discussed. The analytical approach closely resembles the strategy employed in the individual studies. The optimal observable used for each operators pair is DNN vs. $p_{T,\ell\ell}$. Figure 5.11 present a subset of the likelihood

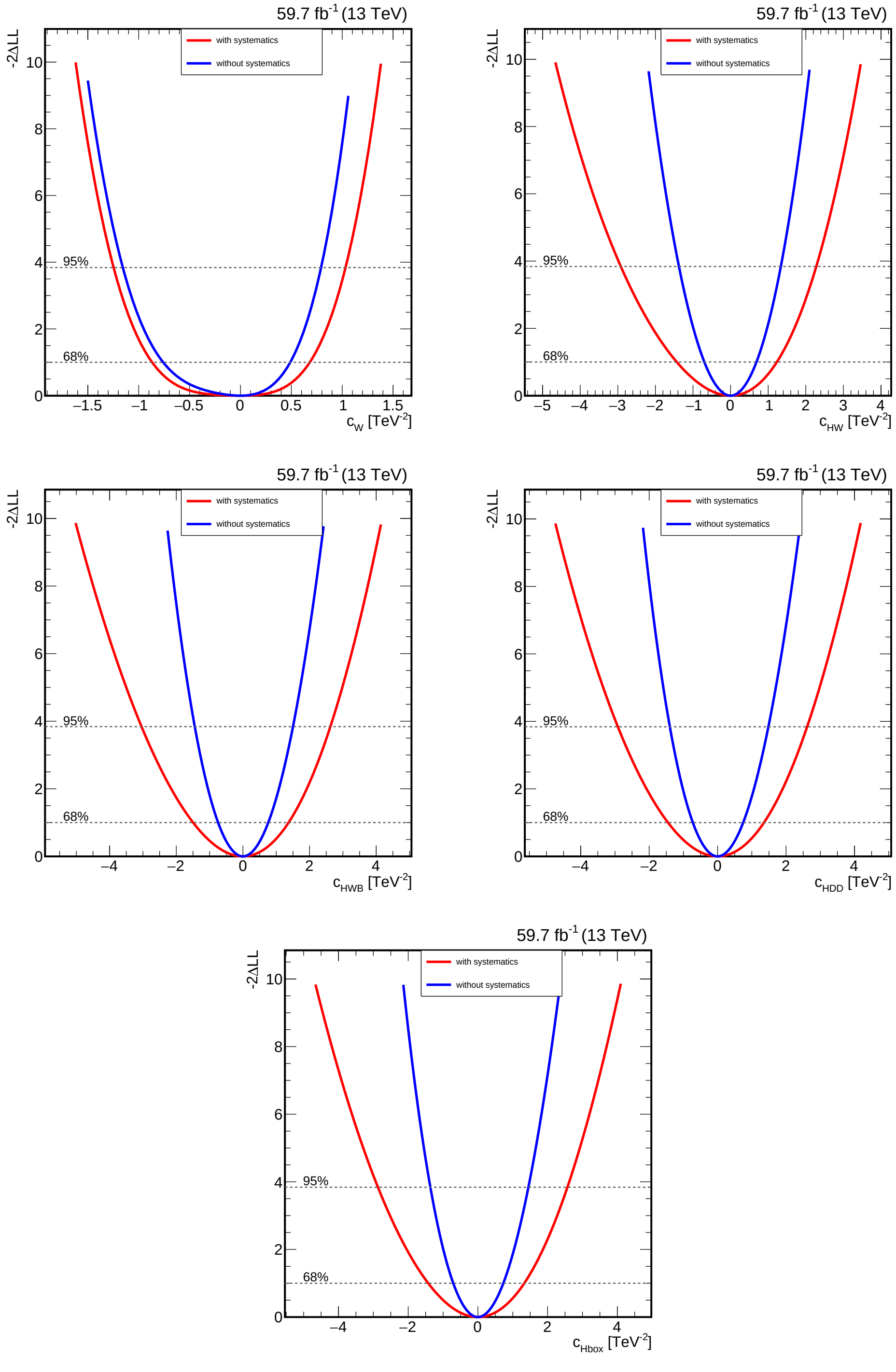


Figure 5.9: Profiles of $-2\Delta\log\mathcal{L}$ are presented for DNN vs. $m_{\ell\ell}$ as a function of the Wilson coefficients. Horizontal dashed lines indicate the 68% and 95% confidence levels, corresponding to $-2\Delta\log\mathcal{L} = 1$ and $-2\Delta\log\mathcal{L} = 3.84$. Only the displayed Wilson coefficient is varied in each case, with the others set to 0. The influence of the nuisance parameters on the fit is illustrated through a comparison between the results with (red line) and without (blue line) the consideration of systematic effects.

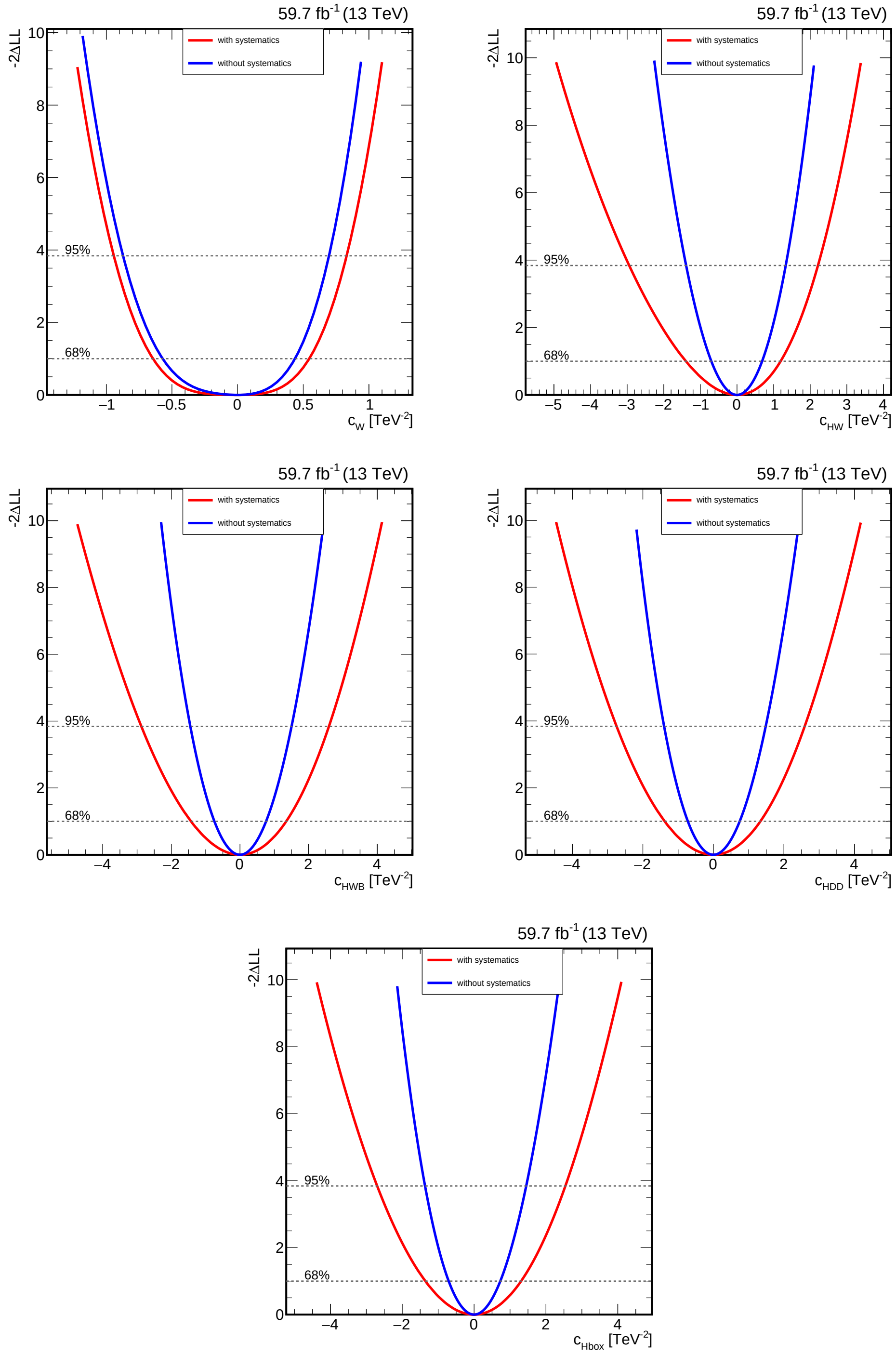


Figure 5.10: Profiles of $-2\Delta \log \mathcal{L}$ are presented for DNN vs. $p_{T,\ell\ell}$ as a function of the Wilson coefficients. Horizontal dashed lines indicate the 68% and 95% confidence levels, corresponding to $-2\Delta \log \mathcal{L} = 1$ and $-2\Delta \log \mathcal{L} = 3.84$. Only the displayed Wilson coefficient is varied in each case, with the others set to 0. The influence of the nuisance parameters on the fit is illustrated through a comparison between the results with (red line) and without (blue line) systematic effects.

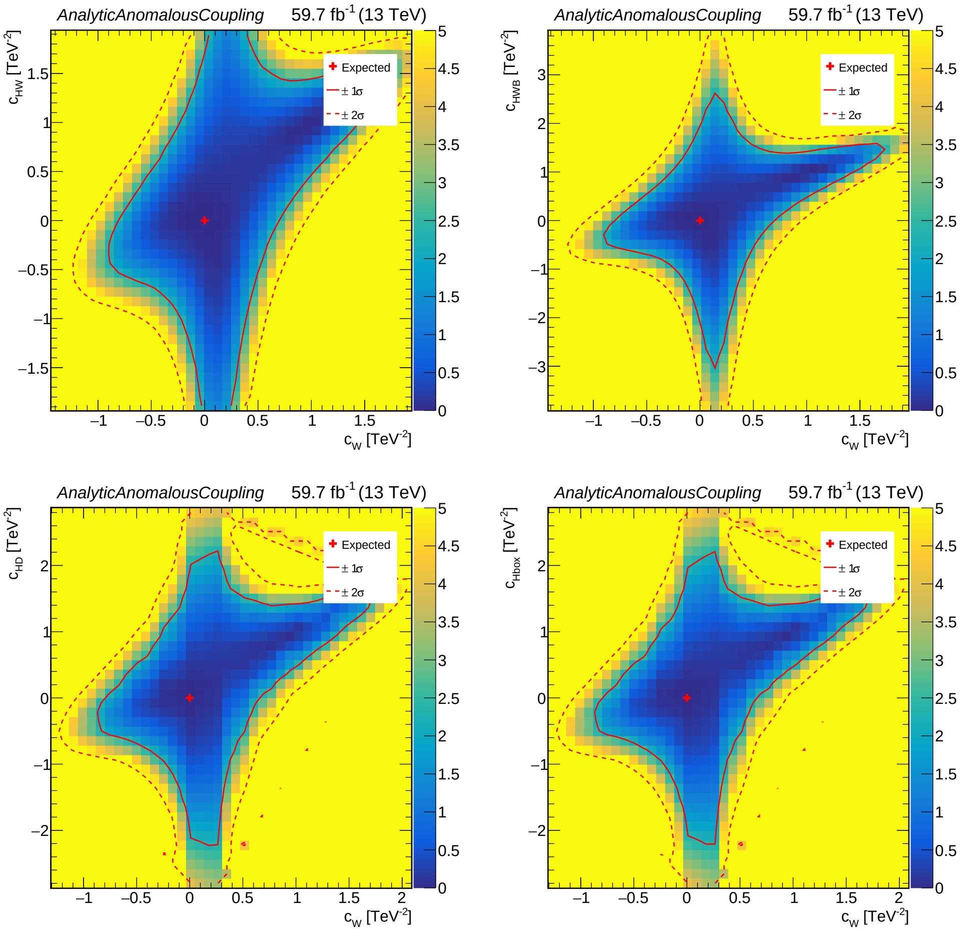


Figure 5.11: Bidimensional likelihood contours are depicted at 68% and 95% confidence level. Two Wilson coefficients are simultaneously varied, while the others are held at 0.

scans obtained, with the c_W Wilson coefficient varying in the x-axis, and the other four Wilson coefficients varying on the y-axis. The systematics are considered in the likelihood fit, ensuring a comprehensive analysis of the constraints.

The star-like shape of these distributions is likely a result of the flattened shape of the one-dimensional limit for c_W, as depicted in Figure 5.10. Furthermore, the mixed terms, which were not explicitly addressed in this study, may play a role in shaping this distribution by introducing correlations between the two operators. The impact of these mixed terms will be the subject of future investigations.

5.4.3 Floating EFT Parameters

In this section, fits to single Wilson coefficients are conducted by allowing all other EFT coefficients to vary freely. Particular attention is given to preventing convergence to local minima. Figures 5.12 depict the profiled $-2\Delta \log \mathcal{L}$ as a function of individual Wilson

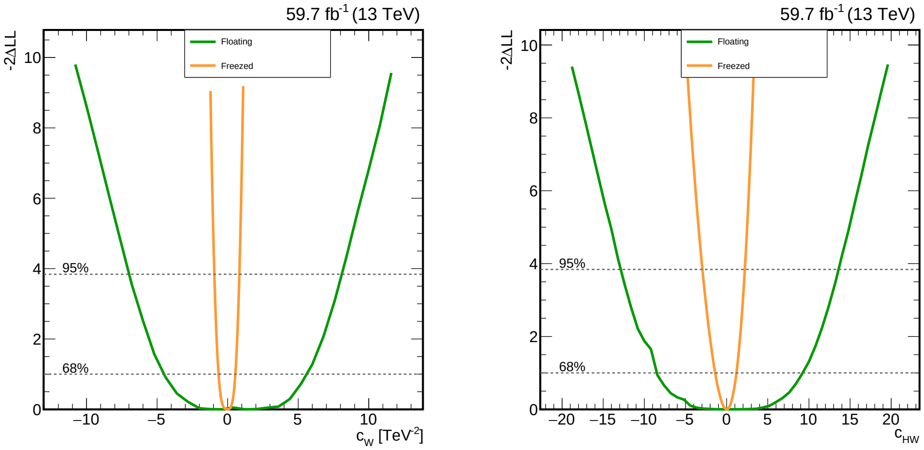


Figure 5.12: Profiles of $-2\Delta\log\mathcal{L}$ are displayed for DNN vs. $p_{T,\ell\ell}$ for c_W and c_{HW} as a function of individual Wilson coefficients. The other EFT parameters are allowed to vary freely. Horizontal dashed lines mark the 68% and 95% confidence levels, which correspond to $-2\Delta\log\mathcal{L} = 1$ and $-2\Delta\log\mathcal{L} = 3.84$. The impact of the floating EFT parameters on the fit is illustrated through a comparison between the results with fixed (orange line) and floating (green line) EFT parameters.

coefficients for DNN vs. $p_{T,\ell\ell}$. In particular, the effect on c_W and c_{HW} in shown. The impact of letting the other EFT parameters float in the fit is demonstrated through a comparison between the profiled $-2\Delta\log\mathcal{L}$ with the other coefficients fixed at 0, and with the other Wilson coefficients allowed to vary. The systematics are considered in the likelihood fit, ensuring a comprehensive analysis of the constraints. It is anticipated that there would be a decrease in performance compared to straightforward one-dimensional fits, where only the EFT parameter of interest is modified.

The augmentation of the dataset by including the 2016-2017 datasets, incorporating the same-flavour category, and potentially utilising the Run 3 dataset in the future is expected to significantly improve the constraints. Moreover, for future research, the investigation of the other three operators will be undertaken.

5.4.4 Future Perspectives

In conclusion, the findings presented in this chapter represent a preliminary result. To progress towards an internal review within the CMS Collaboration and, ultimately, a publication, several important steps are on the horizon. The work accomplished in this thesis has laid a foundational framework for measuring BSM effects using dimension-6 EFT in the context of W^+W^- VBS. While this framework is a significant step forward, it remains incomplete: the inclusion of datasets from 2016 and 2017, along with the incorporation of same-flavour categories ee and $e\mu$, is paramount. Additionally, expanding the analysis to include other bosonic operators within the Warsaw basis, sensitive to the

studied channel, will be a critical step.
Exploring the possibility of training a dedicated neural network to distinguish SM VBS from EFT contributions, as a means to identify new physics, is worth considering.
Finally, implementing Machine Learning techniques for the reconstruction of the invariant mass of the W^+W^- boson pair could be a crucial advancement in the ability to detect EFT contributions.

Conclusions

The first observation of the electroweak (EW) W^+W^- boson pair production in association with two jets in the fully leptonic final state was presented in this thesis. Data collected by the Compact Muon Solenoid (CMS) detector from proton-proton collisions at a centre-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$ at the Large Hadron Collider (LHC) between 2016 and 2018 were used. The process under study belongs to the rare Vector Boson Scattering (VBS) processes, with cross-sections in the fb range. The main sources of background in this analysis came from $t\bar{t}$ production, which could be reduced by vetoing jets containing b quarks. Additionally, QCD-induced W^+W^- production and the Drell-Yan (DY) production contributed to background. The signal has been observed with a statistical significance of 5.6 standard deviations. This achievement was greatly attributed to the use of a Deep Neural Network (DNN), which has been trained to distinguish between the signal and background from both $t\bar{t}$ and QCD-induced W^+W^- production in the different flavour channels. Due to the remarkable significance revealed by this analysis, the cross-section was successfully measured in this channel, obtaining a value of $99 \pm 20, \text{fb}$. This measurement represented the first determinations of the purely EW W^+W^- cross-section, and serves to corroborate the SM within the confines of the associated uncertainties.

In the second phase of this project, the primary objective was to evaluate the expected sensitivities to five Standard Model Effective Field Theory (SMEFT) operators within the context of W^+W^- VBS. This analysis hinged on the utilisation of two distinct 2D phase spaces, each constructed by plotting the DNN output against two critical observables: the invariant mass ($m_{\ell\ell}$) and transverse momentum ($p_{T,\ell\ell}$) of the dilepton system. This approach not only allowed the DNN output to effectively suppress the dominant background contributions but also, in tandem, enabled the two kinematic variables to impose stringent constraints on the Wilson coefficients associated with the SMEFT operators under investigation. This analysis studied one-dimensional and two-dimensional constraints of the SMEFT operators under scrutiny. In particular, a comparative examination of the limits was undertaken by considering the impact of systematic uncertainties and excluding them for the one-dimensional fits. Furthermore, the impact of fixing EFT parameters versus allowing them to float during the fitting process was explored.

The thesis additionally presented the research efforts directed towards enhancing the energy and timing reconstruction capabilities of the CMS Electromagnetic Calorimenter

(ECAL) in view of the high-luminosity phase of the LHC: when particles interact with crystals in the ECAL, a reconstruction algorithm is used to extract key information about the incident particles energy. The project presented in this thesis was focused on the refinement and development of the reconstruction algorithm employed for the estimation of signal amplitudes. This estimation is by calculating them through weighted combinations of time samples. This reconstruction algorithm is analytically efficient, robust and well-suited for scenarios where insensitivity to pedestal variations is a priority, making it a favourable choice for Phase II, ensuring consistency and adaptability in the reconstruction process. The modules developed for this project has been successfully integrated into the official CMS software.

Bibliography

[1] I. Bejar Alonso et al. *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. Tech. rep. CERN, 2020. DOI: `10.23731/CYRM-2020-0010`.

[2] Francis Halzen and Alan D. Martin. *Quarks and Leptons: An Introductory Course in Modern Particle Physics*. John Wiley & Sons Inc, 1984.

[3] Daniel Schroeder and Michael Peskin. *An Introduction To Quantum Field Theory*. Westview Press, 1995.

[4] R. L. Workman et al. "Review of Particle Physics". In: *PTEP* 2022 (2022), p. 083C01. DOI: `10.1093/ptep/ptac097`.

[5] P. A. Zyla et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). ISSN: 2050-3911. DOI: `10.1093/ptep/ptaa104`.

[6] CMS Collaboration. *Measurement of azimuthal correlations among jets and determination of the strong coupling in pp collisions at 13 TeV*. CMS Physics Analysis Summary CMS-PAS-SMP-22-005. 2022.

[7] Sheldon L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: `https://doi.org/10.1016/0029-5582(61)90469-2`.

[8] Steven Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (21 Nov. 1967), pp. 1264–1266. DOI: `10.1103/PhysRevLett.19.1264`.

[9] S. L. Glashow, J. Iliopoulos, and L. Maiani. "Weak Interactions with Lepton-Hadron Symmetry". In: *Phys. Rev. D* 2 (7 Oct. 1970), pp. 1285–1292. DOI: `10.1103/PhysRevD.2.1285`.

[10] Sirunyan et al. "Measurement of the weak mixing angle using the forward–backward asymmetry of Drell–Yan events in *pp* collisions at 8 TeV". In: *The European Physical Journal C* 78.9 (2018), p. 701. ISSN: 1434-6052. DOI: `10.1140/epjc/s10052-018-6148-7`.

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

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

[13] CMS collaboration. *CMS Standard Model summary plots.* URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsCombined`.

[14] "Observation of the Mass Difference between Neutral Charm-Meson Eigenstates". In: *Phys. Rev. Lett.* 127 (11 Sept. 2021), p. 111801. DOI: `10.1103/PhysRevLett.127.111801`.

[15] Céline Degrande et al. "Effective field theory: A modern approach to anomalous couplings". In: *Annals of Physics* 335 (2013), pp. 21–32. ISSN: 0003-4916. DOI: `https://doi.org/10.1016/j.aop.2013.04.016`.

[16] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (2012). DOI: `10.1016/j.physletb.2012.08.021`.

[17] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (2012). DOI: `10.1016/j.physletb.2012.08.020`.

[18] Michael Rauch. "Vector-Boson Fusion and Vector-Boson Scattering". In: (Oct. 2016). arXiv: `1610.08420 [hep-ph]`.

[19] ATLAS Collaboration. *Standard Model Summary Plots June 2021.* Tech. rep. CERN, July 2021. URL: `http://cds.cern.ch/record/2777014`.

[20] C. Quigg Benjamin W. Lee and H. B. Thacker. "Strength of Weak Interactions at Very High Energies and the Higgs Boson Mass". In: *Phys. Rev. Lett.* 38 (1977). DOI: `10.1103/PhysRevLett.38.883`.

[21] Wolfgang Kilian Ana Alboteanu and Jürgen Reuter. "Resonances and unitarity in weak boson scattering at the LHC". In: *Journal of High Energy Physics* 2008.11 (2008). DOI: `10.1088/1126-6708/2008/11/010`.

[22] CMS Collaboration. "Observation of electroweak production of same-sign W boson pairs in the two jet and two same-sign lepton final state in proton-proton collisions at $\sqrt{s}$ =13 TeV". In: *Phys. Rev. Lett.* 120.8 (2018). DOI: `10.1103/PhysRevLett.120.081801`.

[23] CMS Collaboration. "Measurements of production cross sections of WZ and same-sign WW boson pairs in association with two jets in proton-proton collisions at $\sqrt{s}$ =13 TeV". In: *Phys. Lett. B* 809 (2020). DOI: `10.1016/j.physletb.2020.135710`.

[24] CMS Collaboration. "Measurements of production cross sections of polarized same-sign W boson pairs in association with two jets in proton-proton collisions at $\sqrt{s}$ =13 TeV". In: *Phys. Lett. B* 812 (2020). DOI: `10.1016/j.physletb.2020.136018`.

[25] CMS Collaboration. “Evidence for electroweak production of four charged leptons and two jets in proton-proton collisions at $\sqrt{s}$ =13 TeV”. In: *Phys. Lett. B* 812 (2021). DOI: `10.1016/j.physletb.2020.135992`.

[26] CMS Collaboration. “Measurement of the electroweak production of Zγ and two jets in proton-proton collisions at $\sqrt{s}$ =13 TeV and constraints on anomalous quartic gauge couplings”. In: *Phys. Rev. D* 104 (2021). DOI: `10.1103/PhysRevD.104.072001`.

[27] CMS Collaboration. *Search for vector boson scattering at the LHC Run 2 with CMS data in the semi-leptonic lνqq final state*. Tech. rep. CERN, 2021. URL: `https://cds.cern.ch/record/2776799`.

[28] Ilaria Brivio and Michael Trott. “The standard model as an effective field theory”. In: *Physics Reports* 793 (2019). The standard model as an effective field theory, pp. 1–98. ISSN: 0370-1573. DOI: `https://doi.org/10.1016/j.physrep.2018.11.002`.

[29] R. Bellan et al. “A sensitivity study of VBS and diboson WW to dimension-6 EFT operators at the LHC”. In: *Journal of High Energy Physics* 2022.5 (May 2022). DOI: `10.1007/JHEP05(2022)039`.

[30] N. Castro et al. *LHC EFT WG Report: Experimental Measurements and Observables*. 2022. arXiv: `2211.08353 [hep-ph]`.

[31] B. Grzadkowski et al. “Dimension-six terms in the Standard Model Lagrangian”. In: *Journal of High Energy Physics* 2010.10 (2010). DOI: `10.1007/jhep10(2010)085`.

[32] Ilaria Brivio, Yun Jiang, and Michael Trott. “The SMEFTsim package, theory and tools”. In: *Journal of High Energy Physics* 2017 (Sept. 2017). DOI: `10.1007/JHEP12(2017)070`.

[33] I. Brivio. “SMEFTsim 3.0 — a practical guide”. In: *Journal of High Energy Physics* 04 (2021). arXiv: `2012.11343 [hep-ph]`.

[34] Raquel Gomez-Ambrosio. “Studies of Dimension-Six EFT effects in Vector Boson Scattering”. In: *Eur. Phys. J. C* 79 (May 2019). DOI: `10.1140/epjc/s10052-019-6893-2`.

[35] Barbara Jager, Alexander Karlberg, and Giulia Zanderighi. “Electroweak ZZjj production in the Standard Model and beyond in the POWHEG-BOX V2”. In: *Journal of High Energy Physics* 2014.3 (2014), p. 141. ISSN: 1029-8479. DOI: `10.1007/JHEP03(2014)141`.

[36] Jacob J. Ethier et al. “SMEFT analysis of vector boson scattering and diboson data from the LHC Run II”. In: *The European Physical Journal C* 81.6 (2021), p. 560. ISSN: 1434-6052. DOI: `10.1140/epjc/s10052-021-09347-7`.

[37] Thomas Sven Pettersson and P Lefèvre. *The Large Hadron Collider: conceptual design*. Tech. rep. 1995. URL: `https://cds.cern.ch/record/291782`.

[38] Esma Mobs. *The CERN accelerator complex. Complexe des accélérateurs du CERN.* General Photo. 2016. URL: `https://cds.cern.ch/record/2197559`.

[39] CMS collaboration. *Public CMS Luminosity Information.* URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`.

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

[41] CMS Collaboration. *CMS detector design.* `https://cms.cern/news/cms-detector-design`.

[42] A Dominguez et al. *CMS Technical Design Report for the Pixel Detector Upgrade.* Tech. rep. CERN, 2012. URL: `https://cds.cern.ch/record/1481838`.

[43] Q. Ingram et al. "Energy resolution of the barrel of the CMS Electromagnetic Calorimeter". In: *Journal of Instrumentation* 2.04 (2007). DOI: `10.1088/1748-0221/2/04/p04004`.

[44] J Mans et al. *CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter.* Tech. rep. CERN, 2012. URL: `https://cds.cern.ch/record/1481837`.

[45] CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software.* Tech. rep. CERN, 2006. URL: `http://cds.cern.ch/record/922757`.

[46] CMS Collaboration. "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of Instrumentation* 15 (2020). DOI: `10.1088/1748-0221/15/10/P10017`.

[47] CMS Collaboration. "The CMS trigger system". In: *Journal of Instrumentation* 12 (2017). DOI: `10.1088/1748-0221/12/01/P01020`.

[48] CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *Journal of Instrumentation* 12.10 (2017). DOI: `10.1088/1748-0221/12/10/p10003`.

[49] CMS Collaboration. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of Instrumentation* 13.06 (2018). DOI: `10.1088/1748-0221/13/06/P06015`. arXiv: `1804.04528 [hep-ex]`.

[50] Wolfgang Adam et al. *Track Reconstruction in the CMS tracker.* Tech. rep. CERN, 2006. URL: `https://cds.cern.ch/record/934067`.

[51] CMS Collaboration. "Performance of electron reconstruction and selection with the CMS Detector in proton-proton collisions at $\sqrt{s} = 8$ TeV". In: *Journal of Instrumentation* 10.06 (2015). DOI: `10.1088/1748-0221/10/06/P06005`. arXiv: `1502.02701 [hep-ex]`.

[52] K. Ito and K. Xiong. "Gaussian filters for nonlinear filtering problems". In: *IEEE Transactions on Automatic Control* 45.5 (2000), pp. 910–927. DOI: `10.1109/9.855552`.

[53] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-kt jet clustering algorithm". In: *JHEP* 04 (2008). DOI: `10.1088/1126-6708/2008/04/063`.

[54] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet user manual". In: *Eur. Phys. J. C* 72 (2012). DOI: `10.1140/epjc/s10052-012-1896-2`.

[55] CMS Collaboration. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *Journal of Instrumentation* 12.02 (2017). DOI: `10.1088/1748-0221/12/02/p02014`.

[56] CMS Collaboration. "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector". In: *Journal of Instrumentation* 14.07 (2019). DOI: `10.1088/1748-0221/14/07/P07004`.

[57] G. Apollinari et al. "High Luminosity Large Hadron Collider (HL-LHC)". In: *CERN Yellow Reports*. Vol. 5. CERN, 2015. URL: `https://e-publishing.cern.ch/index.php/CYR/article/view/191`.

[58] *The Phase-2 Upgrade of the CMS Barrel Calorimeters*. Tech. rep. This is the final version, approved by the LHCC. Geneva: CERN, 2017. URL: `https://cds.cern.ch/record/2283187`.

[59] Federico Ferri. "The CMS ECAL Phase-2 upgrade for high precision energy and timing measurements". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 958 (2020). Proceedings of the Vienna Conference on Instrumentation 2019. ISSN: 0168-9002. DOI: `https://doi.org/10.1016/j.nima.2019.04.113`.

[60] CMS Collaboration. "Reconstruction of signal amplitudes in the CMS electromagnetic calorimeter in the presence of overlapping proton-proton interactions". In: *Journal of Instrumentation* 15.10 (2020). DOI: `10.1088/1748-0221/15/10/P10002`.

[61] CMS Collaboration. "Reconstruction of the signal amplitude of the CMS electromagnetic calorimeter". In: *Eur. Phys. J. C* 46.s01 (2006). DOI: `10.1140/epjcd/s2006-02-002-x`.

[62] CMS Collaboration. "Observation of electroweak W^+W^- pair production in association with two jets in proton-proton collisions at $\sqrt{s}$ =13 TeV". In: *Physics Letters B* 841 (2023), p. 137495. ISSN: 0370-2693. DOI: `https://doi.org/10.1016/j.physletb.2022.137495`.

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

[64] CMS Collaboration. *CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV. CMS Physics Analysis Summary CMS-PAS-LUM-17-004.* Tech. rep. 2017. URL: `https://cds.cern.ch/record/2621960`.

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

[66] Simone Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *JHEP* 06 (2010). DOI: `10.1007/JHEP06(2010)043`.

[67] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014). DOI: `10.1007/JHEP07(2014)079`. URL: `https://arxiv.org/abs/1405.0301`.

[68] John M. Campbell and R. Keith Ellis. "Update on vector boson pair production at hadron colliders". In: *Phys. Rev. D* 60 (1999). DOI: `10.1103/PhysRevD.60.113006`.

[69] John M. Campbell, R. Keith Ellis, and Ciaran Williams. "Vector boson pair production at the LHC". In: *Journal of High Energy Physics* 2011.7 (2011). DOI: `10.1007/jhep07(2011)018`.

[70] John M. Campbell, R. Keith Ellis, and Walter T. Giele. "A multi-threaded version of MCFM". In: *Eur. Phys. J. C* 75 (2015). DOI: `10.1140/epjc/s10052-015-3461-2`.

[71] Vardan Khachatryan et al. "Event generator tunes obtained from underlying event and multiparton scattering measurements". In: *Eur. Phys. J. C* 76 (2016). DOI: `10.1140/epjc/s10052-016-3988-x`.

[72] CMS Collaboration. "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements". In: *Eur. Phys. J. C* 80 (2020). DOI: `10.1140/epjc/s10052-019-7499-4`.

[73] NNPDF Collaboration. "Unbiased global determination of parton distributions and their uncertainties at NNLO and at LO". In: *Nucl. Phys. B* 855 (2012). DOI: `10.1016/j.nuclphysb.2011.09.024`.

[74] NNPDF Collaboration. "Parton distributions with QED corrections". In: *Nucl. Phys. B* 877 (2013). DOI: `10.1016/j.nuclphysb.2013.10.010`.

[75] NNPDF Collaboration. "Parton distributions from high-precision collider data". In: *Eur. Phys. J. C* 77 (2017). DOI: `10.1140/epjc/s10052-017-5199-5`.

[76] S. Agostinelli et al. "GEANT4: a simulation toolkit". In: *Nucl. Instrum. Meth. A* 506 (2003). DOI: `10.1016/S0168-9002(03)01368-8`.

[77] CMS Collaboration. "Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Lett. B* 791 (2019). DOI: `10.1016/j.physletb.2018.12.073`.

[78] Torbjörn Sjöstrand et al. "An Introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191 (2015). DOI: `10.1016/j.cpc.2015.01.024`.

[79] Keith Hamilton, Paolo Nason, and Giulia Zanderighi. "Finite quark-mass effects in the NNLOPS POWHEG+MiNLO Higgs generator". In: *JHEP* 05 (2015). DOI: `10.1007/JHEP05(2015)140`.

[80] Emanuele Re, Marius Wiesemann, and Giulia Zanderighi. "NNLOPS accurate predictions for W^+W^- production". In: *Journal of High Energy Physics* 2018.12 (2018). DOI: `10.1007/jhep12(2018)121`.

[81] Baptiste Cabouat and Torbjörn Sjöstrand. "Some dipole shower studies". In: *The European Physical Journal C* 78.3 (Mar. 2018). ISSN: 1434-6052. DOI: `10.1140/epjc/s10052-018-5645-z`.

[82] Barbara Jäger et al. "Parton-shower effects in Higgs production via vector-boson fusion". In: *The European Physical Journal C* 80.8 (2020). DOI: `10.1140/epjc/s10052-020-8326-7`.

[83] D. Rainwater, R. Szalapski, and D. Zeppenfeld. "Probing color-singlet exchange in Z + 2-jet events at the CERN LHC". In: *Physical Review D* 54.11 (1996). DOI: `10.1103/physrevd.54.6680`.

[84] CMS Collaboration. "Measurement of the Higgs boson production rate in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV". In: *Eur. Phys. J. C* 81 (2021). DOI: `10.1140/epjc/s10052-021-09014-x`.

[85] CMS Collaboration. *Jet algorithms performance in 13 TeV data.* Tech. rep. 2017.

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

[87] E. Bols et al. "Jet flavour classification using DeepJet". In: *Journal of Instrumentation* 15.12 (2020). ISSN: 1748-0221. DOI: `10.1088/1748-0221/15/12/p12012`.

[88] CMS Collaboration. *Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13 TeV with Phase 1 CMS detector.* CMS-DP-2018-058. CMS Detector Performance Note, 2018. URL: `http://cds.cern.ch/record/2646773`.

[89] Daniele Bertolini et al. "Pileup per particle identification". In: *JHEP* 10 (2014). DOI: `10.1007/JHEP10(2014)059`.

[90] Lutz Prechelt. "Early Stopping — But When?" In: *Neural Networks: Tricks of the Trade: Second Edition.* Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Springer Berlin Heidelberg, 2012. DOI: `10.1007/978-3-642-35289-8_5`.

[91] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *CoRR* abs/1502.03167 (2015). URL: `https://arxiv.org/abs/1502.03167`.

[92] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization.* 2017. URL: `https://arxiv.org/abs/1412.6980`.

[93] François Chollet. *keras*. `https://github.com/fchollet/keras`. 2015.

[94] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006), ROC Analysis in Pattern Recognition. DOI: `https://doi.org/10.1016/j.patrec.2005.10.010`.

[95] CMS Collaboration. "An embedding technique to determine $\tau\tau$ backgrounds in proton-proton collision data". In: *Journal of Instrumentation* 14.06 (2019). DOI: `10.1088/1748-0221/14/06/p06032`.

[96] Vardan Khachatryan et al. "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV". In: *JINST* 10 (2015). DOI: `10.1088/1748-0221/10/06/P06005`.

[97] A. M. Sirunyan et al. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JINST* 13 (2018). DOI: `10.1088/1748-0221/13/06/P06015`.

[98] Vardan Khachatryan et al. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12 (2017). DOI: `10.1088/1748-0221/12/02/P02014`.

[99] Albert M Sirunyan et al. "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector". In: *JINST* 14 (2019). DOI: `10.1088/1748-0221/14/07/P07004`.

[100] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (2011). DOI: `10.1140/epjc/s10052-011-1554-0`.

[101] Luv Demortier. "P Values and Nuisance Parameters". In: *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*. 2008. DOI: `10.5170/CERN-2008-001.23`.