ETP-KA/2025-05

Karlsruher Institut für Technologie

Searching for same-sign WW vector boson scattering in the semi-leptonic decay channel at $\sqrt{s} = 13$ TeV with the CMS experiment by designing and employing an innovative jet charge tagger

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
(Dr. rer. nat.)

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

DISSERTATION

von

M.Sc. Komal Tauqeer

aus Islamabad

Mündliche Prüfung: 25.04.2025
Referent: Prof. Dr. Thomas Müller
Korreferent: Prof. Dr. Abideh Jafari

Eidesstattliche Versicherung gemäß § 13 Absatz 2 Ziffer 3 der Promotionsordnung des Karlsruher Instituts für Technologie (KIT) für die KIT-Fakultät für Physik:

1. Bei der eingereichten Dissertation zu dem Thema

   Searching for same-sign WW vector boson scattering in the semi-leptonic decay channel at $\sqrt{s} = 13$ TeV with the CMS experiment by designing and employing an innovative jet charge tagger

   handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Karlsruhe, den 25. April 2025

....
Komal Tauqeer

Printed and published with the support of the German Academic Exchange Service

Contents

1. Introduction

Since the discovery of the Higgs boson in 2012 [1, 2], physicists have been striving to understand its properties with high precision. While this discovery was a major triumph for the standard model (SM), many fundamental questions remain unanswered, such as the nature of dark matter, the matter-antimatter asymmetry, and neutrino oscillations. Numerous efforts have been made to uncover signs of new particles and interactions that could address these mysteries. One approach is to search for deviations in known SM processes, particularly in interactions involving the Higgs boson.

A key process studied in this context is vector boson scattering (VBS), which plays a crucial role in electroweak interactions. Before the discovery of the Higgs boson, VBS posed a significant challenge to the SM, because at energies larger than the vector boson mass and in the absence of the Higgs-mediated process, the scattering amplitude grows indefinitely with centre-of-mass energy, violating unitarity. To restore unitarity, physicists hypothesized the existence of a new particle and interaction that would cancel out the divergent terms. The Higgs boson discovered at the Large Hadron Collider (LHC) exhibited properties consistent with this hypothesized particle, resolving the immediate unitarity problem. However, if additional Higgs-like particles exist or if the Higgs boson interacts differently than predicted by the SM, the VBS cross section could deviate significantly from expectations. This makes VBS an important process to study for potential signs of new physics.

This thesis presents two novel studies: the first is the development of a state-of-the-art artificial intelligence (AI)-powered jet charge tagger, and the second is a search for same-sign WW VBS in the semi-leptonic decay channel. The latter employs the former to analyse proton-proton collision data at 13 TeV collected by the Compact Muon Solenoid (CMS) experiment at CERN. The analysis is based on the full Run 2 dataset, corresponding to an integrated luminosity of 138 fb^{-1} recorded during 2016–2018.

Traditionally, same-sign WW VBS is studied in the fully leptonic final state, as it provides a cleaner experimental signature and a reliable charge identification of leptons exists for isolating same-sign events. In contrast, in the hadronic or semi-leptonic channels, same-sign WW events are indistinguishable from opposite-sign WW and WZ processes because the charge of the vector boson decaying hadronically is not straightforward to infer — unlike in the leptonic channel, where the charge information is directly accessible. This thesis overcomes this limitation by introducing a jet charge tagger, a novel technique that reconstructs the electric charge of jets originating from hadronic vector boson decays using jet constituents. By leveraging the charge information of the reconstructed jet and the lepton in the final state, the tagger enables the isolation of same-sign WW scattering events, which would otherwise be indistinguishable. The semi-leptonic final state, compared to the fully leptonic one, offers larger branching fractions and allows for a more reliable reconstruction of the W boson due to the presence of only one neutrino. In contrast, the fully leptonic channel involves two neutrinos, making the reconstruction of both W bosons significantly more challenging. Beyond this analysis, the jet charge tagger provides a generalizable tool for the future CMS analyses requiring jet charge identification.

A statistical analysis is performed to measure the significance of the same-sign WW VBS process, and 95% confidence intervals are obtained for the cross section times branching fraction. The best fit value of the same-sign WW VBS signal strength is found to be 1.63 $^{+0.40}_{-0.32}$(syst)$^{+0.59}_{-0.57}$(stat), with observed(expected) significance of 2.8(1.8) standard deviations with respect to the background-only hypothesis. The measured cross section times the branching fraction is found to be 204^{+90}_{-82} fb, with the theory value of 125 fb. An event display of a collision event recorded by the CMS detector in 2017, reconstructed as a candidate same-sign WW scattering event in real data, is shown in Figure 1.1.

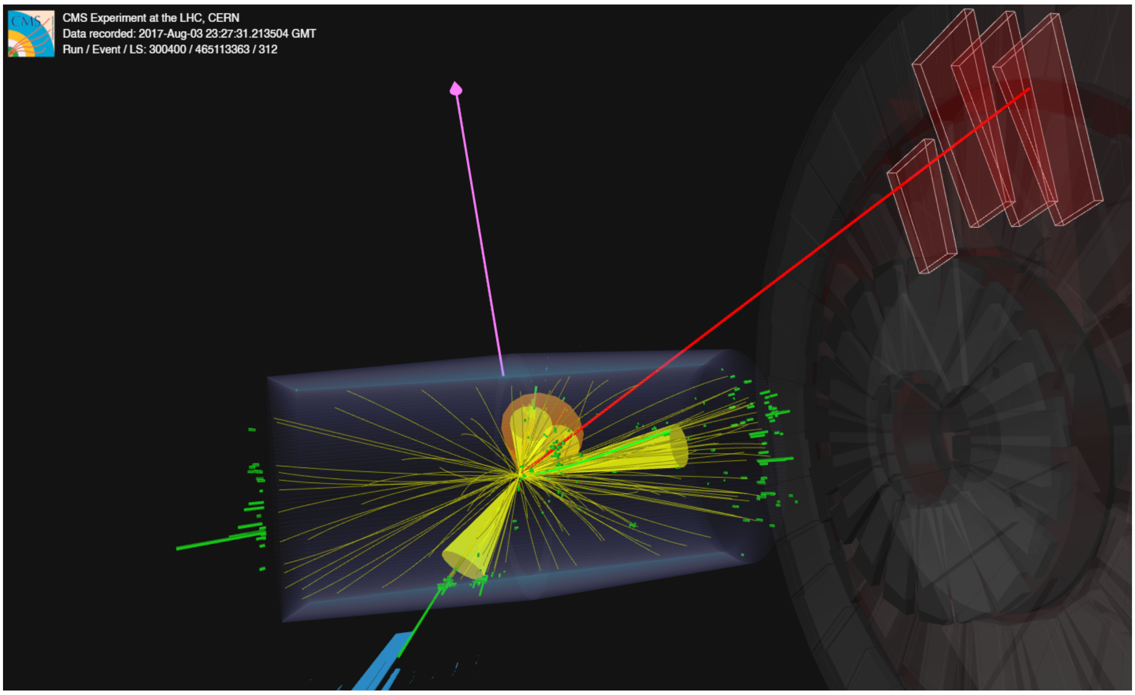

CMS Experiment at the LHC, CERN
Data recorded: 2017-Aug-03 23:27:31.213504 GMT
Run / Event / LS: 300400 / 465113363 / 312

Figure 1.1: An event display of a collision event recorded by the CMS detector during the 2017 data-taking period. The event features two highly energetic jets, shown as yellow cones, positioned in the forward and backward regions of the detector with a large pseudorapidity separation — characteristic of vector boson scattering. Two collinear jets are reconstructed as a single large-radius jet, represented by an orange cone, which is expected to originate from the hadronic decay of a boosted W boson. A muon, depicted as a red line, extends outward to the muon detectors of the CMS detector. The presence of a neutrino is inferred from missing transverse momentum, represented by a pink arrow.

2. Theoretical foundations and the physics of massive gauge bosons

This chapter describes the theoretical foundations of the analysis presented in this thesis. The foundation of modern particle physics is laid by the standard model (SM). It describes the elementary particles and their interactions in nature. The SM, developed in the 1960s and 1970s, is considered as the most successful theory of particle physics. It has been tested for decades by experimentalists. The SM has correctly predicted the existence of several fundamental particles and their properties, which were discovered years later. The Higgs boson, predicted in 1964, was the last missing piece of the SM, which was discovered by the ATLAS and CMS experiments at the Large Hadron Collider at CERN in 2012 [1, 2]. The SM has an exceptional prediction power, but it does not describe the full picture of our universe. It cannot explain some phenomena that have been observed in experiments. The central theme of today's research in particle physics is to do precision measurements of the SM processes and look for areas that can potentially reveal new physics. The search presented in this thesis looks for new physics by directly measuring a low cross section electroweak process, same-sign WW ($W^{\pm}W^{\pm}$) vector boson scattering. The cross section of this process can be significantly enhanced if new particle(s) or interaction(s) exist in nature.

The content of this chapter is organized in three sections, with several subsections in them. Section 2.1 presents a bigger picture of the SM. Fermions and bosons, the two types of elementary particles, are described in subsections 2.1.1, and 2.1.2, respectively. The three types of fundamental forces considered in the SM are described under the mathematical framework of the SM in subsection 2.1.3. Section 2.1 comes to an end with subsection 2.1.4, which explains how different particles get their masses through the Higgs mechanism and spontaneous symmetry breaking. Section 2.2 presents the theoretical shortcomings of the SM, explaining some phenomena that have been observed in experiments, for which the SM has no explanation. Some beyond the standard model (BSM) theories, which can potentially explain these phenomena, and are relevant in the context of the study presented in this thesis, are also mentioned. Section 2.3 is dedicated to the physics of massive gauge bosons, in particular, the scattering of the same-sign W bosons. The experimental signature of this VBS process at the LHC in the semi-leptonic decay channel is described in subsection 2.3.1. This subsection is followed by a theoretical subsection 2.3.2, which explains how the Higgs boson preserves unitarity in VBS processes, with a particular example of same-sign WW VBS. Out of many VBS processes, the same-sign VBS is special, and is considered as the "golden channel" in VBS studies. The reasons for this are explained in subsection 2.3.3. In the end, experimental challenges to study same-sign WW VBS in the semi-leptonic decay channel are stated in subsection 2.3.4 that also describes the necessity to develop an algorithm to identify jet charge, which plays a critical role in this analysis.

2.1. The standard model

The standard model of particle physics is a theory, which explains fundamental interaction forces between the elementary particles of our universe. It uses the mathematical framework of quantum field theory. It is the most important achievement of high energy physics and one of the most successful theories in physics.

The SM includes three of the four fundamental interaction forces: the strong force, the electromagnetic force, and the weak force. Gravity is not described in the SM, because on the typical mass scales considered in particle physics, gravity appears to be much weaker than the other forces. The fundamental particles in the SM are grouped based on their spin: fermions and bosons. Fermions, described in the first part of this section, are the particles with half-integer spin values, while bosons have integer spin values, described in the second part of this section. Fermions follow Fermi-Dirac statistics [3, 4], while bosons follow Bose-Einstein statistics [5, 6]. As the SM uses the mathematical framework of relativistic quantum field theory, the fields are considered as fundamental quantities and individual particles corresponds to excitations in these fields. The whole framework is encoded in a compact description called "Lagrangian". The interaction of quantum fields is expressed in the Lagrangian density of the theory, which derives the dynamics of the fields. In the third part of this section, the three fundamental forces considered in the SM are described using the quantum field theory formulation, followed by a dedicated subsection on the Higgs mechanism and the spontaneous electroweak symmetry breaking.

2.1.1. Fermions

Fermions, spin half-integer particles, obey the Pauli exclusion principle [7], which forbids the presence of more than one identical particle in a single quantum state. This condition ensures that matter made up from fermions, like an atom made of electrons, does not collapse into an extremely dense state. Instead, electrons in a system occupy successive orbitals around the nucleus of an atom to build its structure.

Fermions in the SM are classified into two groups: quarks and leptons, based on their charge and how they interact. A fermion can have several types of charges, which determines with how many forces it can interact. In the SM, there are six quarks carrying colour and electromagnetic charge, namely up, down, charm, strange, top, and bottom quarks. They interact with other matter particles via both strong and electromagnetic interaction forces. Leptons, on the other hand, do not interact via the strong force, as they do not carry colour charge. There are two types of leptons: electrically charged and neutral. The neutral leptons, called neutrinos, only interact with other particles via the weak interaction force, while the electrically charged leptons participate in electromagnetic interactions as well. The fermions of the SM along with their properties are listed in the Table 2.1.

Within each fermion group, the pair of fermions sharing similar quantum numbers are grouped in the same generation. Each group consists of three generations. The first generation of fermions are stable particles and do not decay. Nearly all matter that we see around us is made up of the first generation fermions, called baryonic matter. Fermions of the first generation are the lightest of all. Each subsequent generation contains heavier fermions, which makes them unstable and decay into first generation fermions. On the other hand, neutrinos of all generations, do not decay and barely interact with the surrounding matter.

All quarks and electrically charged leptons have a weak isospin quantum number, and therefore they interact with the weak force as well. As the name suggests, the intensity of the weak interaction is much lower than the other two forces. All fermions have an identical antifermion particle, with the same mass and spin, but opposite electric charge. Particles have another property called chirality, determined by whether the particle transforms in a right- or left-handed representation of the Poincaré group [8]. Particle physicists have only observed left-chiral fermions and right-chiral antifermions participating in the charged weak interactions. Right-chiral fermions have weak isospin of zero and do not engage in the charged weak interactions. For massless particles, like neutrinos in the SM, chirality is the same as helicity (handedness), which is the sign of the projection of the spin vector of a particle onto its momentum vector. The SM does not include right-handed neutrinos as they are massless, colorless, have a weak isospin and electric charge of zero. On the other hand, the left-handed neutrinos exist with a non-zero weak isospin number. Although the SM considers neutrinos as massless particles, experiments have confirmed that they have a non-zero mass because of the neutrino oscillation phenomenon that has been observed [9, 10]. The neutrinos can change their flavour and move from one generation to the other, this can only be explained if neutrinos are considered as massive. The study of such a phenomenon is an active area of research, and the best constraints on the upper limits of the neutrino masses are shown in the Table 2.1.

Table 2.1: Fermions: quarks and leptons, of the standard model, grouped in the three generations based on their physical properties. The values of their electric charge (in the units of electron charge), weak isospin (T_3), colour charge, and mass are shown. Masses are taken from Reference [11].

particle	gen.	electric charge	T_3	color	mass
up (u)	1	2/3	1/2	r,g,b	$2.2^{+0.5}_{-0.3}$ MeV
down (d)	1	-1/3	-1/2	r,g,b	$4.7^{+0.5}_{-0.2}$ MeV
electron neutrino (ν_e)	1	0	1/2	none	< 2 eV
electron (e)	1	-1	-1/2	none	0.511 MeV
charm (c)	2	2/3	1/2	r,g,b	1.27 ± 0.02 GeV
strange (s)	2	-1/3	-1/2	r,g,b	93^{+11}_{-5} MeV
muon neutrino (ν_μ)	2	0	1/2	none	< 0.19 MeV
muon (μ)	2	-1	-1/2	none	105.7 MeV
top (t)	3	2/3	1/2	r,g,b	172.76 ± 0.30 GeV
bottom (b)	3	-1/3	-1/2	r,g,b	$4.18^{+0.03}_{-0.02}$ GeV
tau neutrino (ν_τ)	3	0	1/2	none	< 18.2 MeV
tau (τ)	3	-1	-1/2	none	1776.86 ± 0.12 MeV

Table 2.2: Fundamental bosons of the SM listed along with their properties. The electric charge is given in the units of electron charge. The mass values are taken from Reference [11].

particle	couples to	electric charge	spin	mass (GeV)
gluon (g)	colour charge	0	1	0
photon (γ)	electric charge	0	1	0
W bosons ($W^{\pm}$)	weak isospin	±1	1	80.379 ± 0.012
Z boson (Z)	weak isospin	0	1	91.188 ± 0.002
Higgs boson (H)	mass	0	0	125.10 ± 0.14

2.1.2. Bosons

Bosons are elementary particles with integer spin values. In the SM, there are five elementary bosons namely gluon, photon, $W^{\pm}$ bosons, Z boson, and the Higgs boson. These bosons have a special role in particle physics. They act as a mediator of the fundamental forces. The strong force is mediated by the force carrier gluon, the electromagnetic force is mediated by the photon, and the weak force is mediated by the $W^{\pm}$ and Z bosons. All aforementioned bosons have a spin value of 1 and are called vector bosons or gauge bosons. The Higgs boson is special and has a unique role in the SM. It is a scalar particle, meaning that its spin value is 0. The Higgs boson is responsible for giving masses to the particles through the Higgs mechanism and the spontaneous electroweak symmetry breaking, as discussed in the Section 2.1.4. The properties of the scalar boson and the four vector bosons of the SM are listed in the Table 2.2.

The $W^{\pm}$ and Z vector bosons are the main focus of the analysis presented in this thesis. They are amongst the heaviest elementary particles. Their high mass values limit the range of the weak interaction force. When a particle emits or absorbs a W^{+} or a W^{-} boson, its electric charge and spin gets altered by one unit. At the same time, the particle changes its flavour by emitting or absorbing $W^{\pm}$ bosons, e.g. an up-type quark can change to a down-type quark by emitting a W^{-} boson. The phenomena of changing flavour of a particle by emitting or absorbing a charged vector boson is called flavour changing charged currents. This explains the β-decay process, in which a neutron (udd) converts to a proton (udu) by emitting a W^{-} boson and vice versa by emitting a W^{+} boson. The Z boson, on the other hand, does not change the flavour of the particle, as it is electrically neutral. The exchange of a Z boson between the two particles transfers the spin, momentum, and energy, however, quantum numbers like electric charge, flavour etc. remain unaffected.

2.1.3. Mathematical framework of the standard model

The theory of particle physics — the standard model, relies on the mathematical formalism of relativistic quantum field theory (QFT). In QFT, the fundamental quantities are fields. The SM accommodates 17 fundamental quantum fields. This includes: 12 fields for fermions — six quarks and six leptons, 4 fields for vector bosons — the gluon (g), the photon (γ), the W boson, and the Z boson, and 1 field for the scalar Higgs boson. All particles of the SM are imagined as excitations of their corresponding fields, which fill up all space uniformly. The properties of the particles follow from the imposed symmetry conditions on the fields. In the Lagrangian formalism of QFT, the Lagrangian density, defined as the function of spacetime dependent fields Φ(x) and their first derivative, as in Eq. 2.1, holds the complete description of the fields and their dynamics.

$$\mathcal{L} = \mathcal{L}(\Phi(x), \partial_\mu(\Phi(x))), \tag{2.1}$$

where, x is the spacetime four-vector, and $\partial_\mu = \frac{\partial}{\partial x^\mu}$ with $\mu = 0, 1, 2, 3$, is the derivative operator. The Lagrangian of a system is given by:

$$L = \int dx^4 \mathcal{L}. \tag{2.2}$$

Based on the principle of least action, one can get the equations of motion of the fields by using the Euler-Lagrange equation, which is defined as:

$$\frac{\partial \mathcal{L}}{\partial \Phi} - \partial_\mu \frac{\partial \mathcal{L}}{\partial(\partial_\mu \Phi)} = 0. \tag{2.3}$$

As an example, consider the Lagrangian density of a non-interacting scalar field $\phi(x)$ given by:

$$\mathcal{L} = \frac{1}{2}(\partial_\mu \phi)(\partial^\mu \phi) - \frac{1}{2} m^2 \phi^2. \tag{2.4}$$

The Lagrangian density above contains the kinetic term and the mass term for the scalar field. Application of the Euler-Lagrange equation on this Lagrangian, results in the equation of motion, generally known as Klein-Gordon equation:

$$(\partial_\mu \partial^\mu + m^2)\phi = 0. \tag{2.5}$$

The Lagrangian of many successful field theories, which explains the dynamics of elementary particles, are invariant under some symmetry transformation groups. If the Lagrangian of the theory is invariant under local transformations of a symmetry group, the theory is referred as gauge theory. If the symmetry group is commutative, the theory is called abelian gauge theory. If the symmetry group is non-commutative, the theory is referred as non-abelian gauge theory. Quantum electrodynamics, a relativistic field theory, which explains how light and matter interact, is an abelian gauge theory with symmetry group $U(1)$. The SM is a non-abelian gauge theory with the symmetry group $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$. For each symmetry group or gauge group, there exists a set of group generators, which defines the transformation of that group under which the Lagrangian of the theory is invariant. For the SM symmetry group, the transformation is defined as:

$$\psi(x) \to \psi'(x) = U(x) \cdot \psi(x), \tag{2.6}$$

where, $\psi(x)$ represents the fermionic fields or Dirac fields and,

$$U(x) = e^{\mathrm{i}\beta(x)} \cdot e^{\frac{\mathrm{i}}{2}\sum_{j=1}^{3}\alpha^j(x)\sigma^j} \cdot e^{\frac{\mathrm{i}}{2}\sum_{k=1}^{8}\epsilon^k(x)\lambda^k}. \tag{2.7}$$

Here, $\beta(x)$, $\alpha^j(x)$, and $\epsilon^k(x)$ are arbitrary real-valued functions; σ^j are Pauli matrices, generators of the symmetry group $SU(2)_L$; and λ^k are Gell-Mann matrices, generators of the symmetry group $SU(3)_C$. In the following subsections, the gauge groups and the Lagrangian of the standard model will be discussed, which incorporates electromagnetism (quantum electrodynamics), strong interactions (quantum chromodynamics), and electroweak interactions.

2.1.3.1. Quantum electrodynamics

Quantum electrodynamics (QED) is a relativistic quantum field theory, which explains the interaction between the electrically charged fermions and photons. QED is an abelian gauge theory defined in Minkowski space, whose Lagrangian is invariant under the symmetry group $U(1)$ transformations. QED is the first successful quantum field theory, whose results fully agree with both the special relativity and the quantum mechanics.

We will construct the mathematical formalism of QED, starting from the Lagrangian density of a free fermion with mass m. It is given by the Dirac Lagrangian:

$$\mathcal{L}_D = \bar{\psi}(x)(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi(x), \tag{2.8}$$

here, γ^μ are the Dirac matrices and the fermion field is represented by four-component

Dirac-spinors $\psi(x)$. Using the Euler-Lagrange equation, we can get the equation of motion of a free fermion field, which is the famous Dirac equation:

$$(\mathrm{i}\gamma^{\mu}\partial_{\mu} - m)\psi(x) = 0. \tag{2.9}$$

The SM is a gauge theory, as mentioned earlier, and its Lagrangian should be gauge invariant under the local gauge transformations. For $U(1)$ symmetry group, the gauge transformations are defined as:

$$\psi(x) \to e^{-\mathrm{i}q\alpha(x)}\psi(x), \tag{2.10}$$

and

$$\bar{\psi}(x) \to e^{+\mathrm{i}q\alpha(x)}\bar{\psi}(x), \tag{2.11}$$

where, q is the electric charge and $\alpha(x)$ represents an arbitrary phase. The Lagrangian density of Eq. 2.8 is not invariant under this local transformation. To make it invariant, the derivative ∂_μ is replaced by the covariant derivative:

$$\partial_{\mu} \to D_{\mu} = \partial_{\mu} + \mathrm{i}qA_{\mu}(x). \tag{2.12}$$

Doing so, a new gauge field A_μ has been added to ensure the local gauge invariance. This forces us to introduce a new term in the QED Lagrangian:

$$\mathcal{L}_I = -q\bar{\psi}(x)\gamma^{\mu}A_{\mu}\psi(x). \tag{2.13}$$

This term represents the interaction between the fermionic fields and the gauge field of QED. Specifically, it is the interaction of a fermion field $\psi(x)$, an antifermion field $\bar{\psi}(x)$, and the gauge field A_μ, which in the case of QED represents a photon field. The electric charge q represents the strength with which a fermion and an antifermion couples to a photon.

The newly introduced gauge field A_μ is a vector field and has to transform as:

$$A_{\mu}(x) \to A_{\mu}(x) + \partial_{\mu}\alpha(x). \tag{2.14}$$

The dynamics of such a vector field is described by the following Lagrangian density:

$$\mathcal{L}_A = -\frac{1}{16\pi}F^{\mu\nu}F_{\mu\nu} + \frac{1}{8\pi}m^2 A^\nu A_\nu, \tag{2.15}$$

here, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, is the electromagnetic field tensor, also known as the curvature of the gauge field. Application of Euler-Lagrange equation on the Lagrangian density $\mathcal{L}_A$ gives the Proca equation of the vector field A_μ:

$$(\partial_\mu \partial^\mu + m^2)A^\nu = 0. \tag{2.16}$$

Requiring the gauge invariance condition for the Lagrangian density $\mathcal{L}_A$, implies that the gauge boson associated with the vector field A_μ has to be massless, and indeed the photon is a massless particle. Hence, the final QED Lagrangian is given by:

$$\begin{aligned}\mathcal{L}_{\text{QED}} &= \mathcal{L}_D + \mathcal{L}_I + \mathcal{L}_A \\ &= \bar{\psi}(x)(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi(x) - q\bar{\psi}(x)\gamma^\mu A_\mu \psi(x) - \frac{1}{16\pi}F^{\mu\nu}F_{\mu\nu}\end{aligned} \tag{2.17}$$

The first term describes the dynamics of the fermions, the third term describes the kinematics of the photon, and the second term describes the interaction between the two.

2.1.3.2. Quantum chromodynamics

Quantum chromodynamics (QCD) is a relativistic quantum field theory, which describes the interactions between the colour charged fermions (quarks) and gluons — mediator of the strong interaction force. The strong interaction force described by QCD, is a non-abelian gauge theory with symmetry group $SU(3)_C$, unlike QED which is an abelian gauge theory. The mathematical formalism of QCD will follow the same steps as in QED, described in Section 2.1.3.1, but with an additional requirement on the fields to respect the non-abelian nature of the gauge group $SU(3)_C$. Quarks have both electric charge and colour charge. Just like in QED, quark fields will be represented by the Dirac spinors $\psi(x)$ but for each of the three colour charges (r, g, b). The notation for a quark field and an antiquark field will be:

$$\Psi(x) = \begin{pmatrix} \psi_{r(x)} \\ \psi_{g(x)} \\ \psi_{b(x)} \end{pmatrix}, \text{ and } \bar{\Psi}(x) = \begin{pmatrix} \bar{\psi}_{r(x)} & \bar{\psi}_{g(x)} & \bar{\psi}_{b(x)} \end{pmatrix}, \tag{2.18}$$

respectively. The Dirac Lagrangian density for quarks will take the form:

$$\mathcal{L}_D = \sum_{f=1}^{6} \bar{\Psi}_f(x)(\mathrm{i}\gamma^\mu \partial_\mu - m)\Psi_f(x), \tag{2.19}$$

where the summation is over all flavours of the quarks. The generators of the gauge group $SU(3)_C$ are the eight Gell-Mann matrices λ_a, which defines the local transformation for the fields as:

$$\Psi_f(x) \to e^{-\mathrm{i}\frac{g_s}{2}\sum_{a=1}^{8}\lambda_a\alpha_a(x)}\Psi_f(x), \tag{2.20}$$

and

$$\bar{\Psi}_f(x) \to e^{+\mathrm{i}\frac{g_s}{2}\sum_{a=1}^{8}\lambda_a\alpha_a(x)}\bar{\Psi}_f(x). \tag{2.21}$$

To make the Lagrangian density in Eq. 2.19 invariant under the local gauge transformation of $SU(3)_C$, the derivative ∂_μ is replaced by the covariant derivative:

$$\partial_\mu \to D_\mu = \partial_\mu + \mathrm{i}\frac{g_s}{2}\sum_{a=1}^{8}\lambda_a G_{\mu,a}(x). \tag{2.22}$$

Requiring gauge invariance has added eight new gauge fields $G_{\mu,a}(x)$, called gluon fields. Replacing ∂_μ by covariant derivative D_μ in Eq. 2.19 gives rise to an interaction term:

$$\mathcal{L}_I = -\frac{g_s}{2}\sum_{f=1}^{6}\sum_{a=1}^{8}\bar{\Psi}_f\gamma^\mu\lambda_a G_{\mu,a}(x)\Psi_f(x), \tag{2.23}$$

which turns out to be not invariant under local gauge transformations because the Gell-Mann matrices do not commute, $[\lambda_a, \lambda_b] = i2f_{abc}\lambda_c$. To make the interaction

Lagrangian invariant, the gluon fields $G_{\mu,a}$ have to follow the following transformation rule:

$$G_{\mu,a}(x) \to G_{\mu,a}(x) - \frac{1}{g_s}\partial_\mu \alpha_a(x) + f_{abc}\alpha^b(x)G^c_\mu, \tag{2.24}$$

here, f_{abc} are the structure constants of the gauge group $SU(3)_C$. The term above shows the interaction between six flavours of quarks and antiquarks, and eight types of gluons. The appearance of the term $f_{abc}\alpha^b(x)G^c_\mu$ is due to the non-abelian nature of the gauge group. From this term follows the other remarkable properties of the strong interaction, such as quark confinement [12] and asymptotic freedom [13].

The dynamics of the gluon fields can be described by defining the gluon field tensor, which respects the non-abelian nature of the gauge group, as:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_s f_{abc} G^b_\mu G^c_\nu. \tag{2.25}$$

Using the gluon field tensor, we can define the kinetic term of gluons as:

$$\mathcal{L}_G = -\frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a. \tag{2.26}$$

A mass term for gluon fields in the Lagrangian is not allowed as it will violate the gauge invariance, hence the gluon fields are considered as massless in the theory.

The final Lagrangian of QCD is given by combining Equations 2.19, 2.23, and 2.26:

$$\mathcal{L}_{\text{QCD}} = \sum_{f=1}^{6} \bar{\Psi}_f(x)(\mathrm{i}\gamma^\mu\partial_\mu - m)\Psi_f(x) - \frac{g_s}{2}\sum_{f=1}^{6}\sum_{a=1}^{8} \bar{\Psi}_f \gamma^\mu \lambda_a G_{\mu,a}(x)\Psi_f(x) - \frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a. \tag{2.27}$$

Using Einstein summation notation, we can rewrite the above Lagrangian as:

$$\mathcal{L}_{\text{QCD}} = \bar{\Psi}(x)(\mathrm{i}\gamma^\mu\partial_\mu - m)\Psi(x) - \frac{g_s}{2}\bar{\Psi}\gamma^\mu\lambda_a G^a_\mu(x)\Psi(x) - \frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a. \tag{2.28}$$

The first term in the above Lagrangian describes the quark-quark interaction, the third term describes the gluon-gluon interaction, and the second term describes

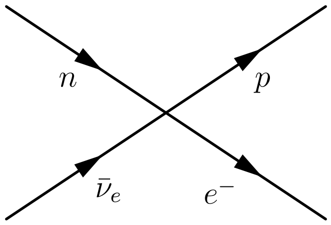


Figure 2.1: Fermi's four-point interaction for β-decay process.

the interaction between the quarks and gluons. g_s is the coupling constant of the theory and describes the coupling strength. Using g_s, the strong coupling constant of the theory is defined as $\alpha_s = g_s^2/4\pi$. The QCD Lagrangian looks very similar to the QED Lagrangian 2.17, but the gluon field tensors imply completely new features: the gluon self-interactions via tri-linear couplings proportional to g_s and four-linear coupling proportional to g_s^2. This reflects the fact the gluons carry colour charge themselves and can interact with each other. The gauge invariance condition determines the structure of gluon self-interactions and forbids any high-order gluon couplings.

2.1.3.3. Weak interaction and electroweak theory

The development of the theory that describes the weak interaction force in our nature was purely experimental driven. All fermions can feel weak force. Neutrinos are special particles, in a sense that they only experience weak force. The weak interaction is responsible for radioactive decays of nuclei. For instance, consider the β-decay process:

$$n \to p + e^- + \bar{\nu}_e, \tag{2.29}$$

which is purely a weak interaction process. The first attempt to develop a theory of weak interaction was made by Fermi in 1932, where he imagined, in analogy to electromagnetic interaction, β-decay as a four-point interaction at a single point in space-time as shown in Figure 2.1.

The cross section of such a four-point interaction process was found to be linearly rising with the energy. This means that at high energies, Fermi's model break down. To overcome this issue, there was a need to add a propagator of the weak force at the interaction vertex, as in the Figure 2.2. We know that the weak interactions are mediated by W and Z bosons, which are massive. At high energies, the mass of the W boson stops the cross section of such a process from going to infinity. Hence, the presence of massive W bosons is required to explain β-decay process.

In 1956, T.D. Lee and C.S. Yang, while trying to solve a mysterious puzzle of two

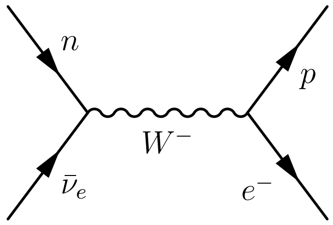


Figure 2.2: β-decay mediated by W boson.

strange mesons decays [14], the τ and the θ, proposed that the parity can be violated in the weak interactions, which was conserved in the strong and the electromagnetic interactions. Shortly after, C.S. Wu devised a beautiful experiment to test the parity (non-)conservation in the β-decays using polarized cobalt-60 nuclei, known as the famous Wu experiment [15]. Her experiment concluded that the electrons emitted in the β-decay process prefer a very specific direction of decay, specifically opposite to the direction of the nuclear spin, which is a clear sign of parity violation. This preference doesn't change by changing the initial polarization of the nuclei, hence the parity was observed to be maximally violated in the weak interactions. It was a surprising new feature of weak interaction, as parity conservation was long considered as a fundamental symmetry of nature. Followed by this discovery, some physicists proposed to check whether the combined charge and parity, known as the CP, is conserved in weak interactions. It turns out, it is conserved, but in rare cases, roughly 0.3% of the weak interactions, the CP symmetry is also violated. This concludes that in nature, there is a true violation of mirror symmetry and there is a difference between our world and the mirror world. This could be the reason why we have a universe which is matter-dominated.

To develop a gauge theory of weak interactions based on experimental evidences, we will follow the same steps from QED. Starting with the Dirac Lagrangian for free fermion field:

$$\mathcal{L}_D = \bar{\psi}(x)(\mathrm{i}\gamma^\mu\partial_\mu - m)\psi(x), \tag{2.30}$$

where, $\psi(x)$ are the four component Dirac spinors. Since we are developing a chiral theory for weak interactions, the chiral components of the Dirac spinors for fermions and antifermions can be obtained by using the chiral projection operators as:

$$\psi^{R/L}(x) = \frac{1}{2}(1 \pm \gamma^5)\psi(x),\ \bar{\psi}^{R/L}(x) = \frac{1}{2}(1 \mp \gamma^5)\bar{\psi}(x), \tag{2.31}$$

where,

$$\gamma^5 = \frac{\mathrm{i}}{4!}\epsilon_{\mu\nu\alpha\beta}\gamma^\mu\gamma^\nu\gamma^\alpha\gamma^\beta, \text{ and } \{\gamma^\mu, \gamma^5\} = 0, \tag{2.32}$$

with ϵ being the Levi-Civita symbol. Using the projection operators $\frac{1}{2}(1 \pm \gamma^5)$, the Dirac spinor can be factorized into two components, left chiral and right chiral, and can be written as:

$$\psi(x) = \begin{pmatrix} \psi_L(x) \\ \psi_R(x) \end{pmatrix}, \text{ and } \bar{\psi}(x) = \begin{pmatrix} \bar{\psi}_L(x) & \bar{\psi}_R(x) \end{pmatrix}, \tag{2.33}$$

for fermions and antifermions, respectively. From experiments, it was known that only left-handed fermions and right-handed antifermions interact with the charged weak interactions. Moreover, the W bosons interact with left-handed charged leptons and neutrinos of the same lepton flavours, strictly. This is not true for quarks, where the W bosons can interact with quarks of different generations. The strength of these cross generation interactions with quarks are given by the Cabibo-Kobayashi-Maskawa matrix [11]. The Z boson is different from W boson as it can interact with right chiral particles. It behaves similar to photons and can couple to any two fermions of opposite charge, irrespective of their chirality. The only difference is that Z boson can also couple to neutrinos, while photons can not. In fact, the first evidence of the existence of Z boson was observed in a neutrino-electron scattering process using a bubble chamber experiment at CERN [16], which was only possible if a neutral weak gauge boson exists.

The symmetry group for weak interaction is $SU(2)_L$ generated by the Pauli matrices, with the third component of weak isospin (T^3) conserved. The left-handed fermions, which interacts with weak force, are represented as isospin doublets ($|T^3| = \frac{1}{2}$). The right-handed fermions, which do not interact with the weak force, are represented as isospin singlets ($T^3 = 0$). The fermions of the SM under $SU(2)_L$ symmetry group representation looks like:

$$\text{Quarks}: \begin{pmatrix} \mathrm{u} \\ \mathrm{d} \end{pmatrix}_{\mathrm{L}} \otimes (\mathrm{u})_{\mathrm{R}} \otimes (\mathrm{d})_{\mathrm{R}}, \begin{pmatrix} \mathrm{c} \\ \mathrm{s} \end{pmatrix}_{\mathrm{L}} \otimes (\mathrm{c})_{\mathrm{R}} \otimes (\mathrm{s})_{\mathrm{R}}, \begin{pmatrix} \mathrm{t} \\ \mathrm{b} \end{pmatrix}_{\mathrm{L}} \otimes (\mathrm{t})_{\mathrm{R}} \otimes (\mathrm{b})_{\mathrm{R}} \tag{2.34}$$

$$\text{Leptons}: \begin{pmatrix} \mathrm{e} \\ \nu_{\mathrm{e}} \end{pmatrix}_{\mathrm{L}} \otimes (\mathrm{e})_{\mathrm{R}}, \begin{pmatrix} \mu \\ \nu_\mu \end{pmatrix}_{\mathrm{L}} \otimes (\mu)_{\mathrm{R}}, \begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix}_{\mathrm{L}} \otimes (\tau)_{\mathrm{R}} \tag{2.35}$$

It is worth noting that in Eq. 2.35, the neutrinos don't have a right chiral particle representation. This is because the neutrinos can only be produced in charged weak interactions, which only couples to left chiral particles. Hence, even if the right chiral neutrinos exist in nature, they do not interact with the fundamental forces considered in the SM. Therefore, right chiral neutrinos are not included in the SM.

In terms of left and right chiral components, the Lagrangian density for the first generation leptons can be written as:

$$\mathcal{L}_{D,e} = \bar{\psi}_e^{\mathrm{L}} \mathrm{i} \partial\!\!\!/ \psi_e^{\mathrm{L}} + \bar{\psi}_e^{\mathrm{R}} \mathrm{i} \partial\!\!\!/ \psi_e^{\mathrm{R}} - m \bar{\psi}_e^{\mathrm{L}} \psi_e^{\mathrm{R}} - m \bar{\psi}_e^{\mathrm{R}} \psi_e^{\mathrm{L}} + \bar{\psi}_{\nu_e}^{\mathrm{L}} \mathrm{i} \partial\!\!\!/ \psi_{\nu_e}^{\mathrm{L}} \tag{2.36}$$

In the above expansion of Dirac Lagrangian, the coordinate expansion $\psi(x)$ is omitted, and a new notation has been introduced $\partial\!\!\!/ = \gamma^\mu \partial_\mu$. The left-chiral fermion field has been treated as an isospin doublet and the right-chiral field as an isospin singlet. The neutrinos are considered as massless. The Dirac Lagrangian density in Eq. 2.36 is not invariant under the local gauge symmetry of the $SU(2)_L$ group, because of the fermion mass terms. If the fermions were massless, one could introduce massless gauge fields in the form of covariant derivative, to make the Lagrangian density gauge invariant. But, we know that the fermions are not massless nor the gauge bosons of weak interactions. So, the technique from QED and QCD will not work in this case and the mass terms of fermions and bosons will disturb the local gauge invariance. To overcome this problem, first we define the massless theory of weak interaction and give masses to the gauge bosons and fermions through spontaneous symmetry breaking and the Higgs mechanism as explained in detail in the Section 2.1.4.

For the case of massless fermions, the Dirac Lagrangian becomes:

$$\mathcal{L}_D = \bar{\Psi}_{\mathrm{L}} \mathrm{i} \partial\!\!\!/ \Psi_{\mathrm{L}} + \bar{\psi}_{\mathrm{R}} \mathrm{i} \partial\!\!\!/ \psi_{\mathrm{R}} \tag{2.37}$$

where Ψ and ψ represents the doublets and singlets representations of the fermions, according to the $SU(2)_L$ gauge group. It is worth noting that the Lagrangian density, for the case of massless fermions, in Eq. 2.37 is invariant under the global $SU(2_L)$ transformations and global $U(1)$ transformations defined as:

$$\psi_{\mathrm{R}} \to \psi_{\mathrm{R}} \ , \ \ \Psi_{\mathrm{L}} \to e^{\frac{\mathrm{i}}{2}\alpha_j \sigma^j} \Psi_{\mathrm{L}} \tag{2.38}$$

and

$$\psi_{\mathrm{R}} \to e^{\mathrm{i}\beta Y}\psi_{\mathrm{R}}\ ,\ \ \Psi_{\mathrm{L}} \to e^{\mathrm{i}\beta' Y}\Psi_{\mathrm{L}} \tag{2.39}$$

for $SU(2)_L$ and $U(1)$, respectively. In Eq. 2.38 and 2.39, α_j and β are the arbitrary real-valued functions of spacetime, σ^j and Y are the Pauli's matrices — the generators of $SU(2)_L$ gauge group, and weak hypercharge corresponding to the $U(1)$ gauge group, respectively. The weak hypercharge is defined as:

$$Y = 2(Q - T_3) \tag{2.40}$$

Here, hypercharge Y connects the electric charge Q and the weak isospin T_3 of a particle. This invariance of Lagrangian density under $SU(2)_L \times U(1)$ suggests that the theory describes both electromagnetic and weak interactions together, hereby referred to as electroweak theory.

To make the Lagrangian density of Eq. 2.37 invariant under the local transformations of $SU(2)_L \times U(1)$, we now define the covariant derivative and introduce four massless gauge fields as:

$$D_\mu = \partial_\mu - \frac{\mathrm{i}}{2} g \alpha_j W_\mu^j - \frac{\mathrm{i}}{2} g' Y B_\mu \tag{2.41}$$

We replace the derivative in the Lagrangian density of Eq. 2.37 by the covariant derivative defined above and add the kinematic terms of the newly added gauge fields, to construct a Lagrangian density for electroweak interactions as following:

$$\mathcal{L}_{\mathrm{EW}} = \bar{\Psi}_{\mathrm{L}} \mathrm{i} \not{D} \Psi_{\mathrm{L}} + \bar{\psi}_{\mathrm{R}} \mathrm{i} \not{D} \psi_{\mathrm{R}} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W_{\mu\nu}^j W_j^{\mu\nu} \tag{2.42}$$

The mass terms of the gauge terms are not included, as they will disturb the gauge invariance. The masses to the gauge bosons will be given using a different mechanism — the Higgs mechanism, described in the next section. The field tensors $B_{\mu\nu}$ and $W_{\mu\nu}^j$ are defined as:

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu, \tag{2.43}$$

and

$$W_{\mu\nu}^j = \partial_\mu W_\nu^j - \partial_\nu W_\mu^j + g \epsilon^{jkl} W_\mu^k W_\nu^l. \tag{2.44}$$

It is worth noticing that $W^j_{\mu\nu}$ has a quadratic term, which will introduce terms for cubic and quadratic self-couplings of the gauge fields in the electroweak Lagrangian.

2.1.4. The Higgs mechanism and spontaneous electroweak symmetry breaking

The requirement that the electroweak Lagrangian should be gauge invariant, as described in the Section 2.1.3.3, has imposed the condition that the gauge fields should be massless. However, it is well known from the experiments that the electroweak gauge bosons are not massless but carries a significant mass as listed in the Table 2.2. This conflict between the experimental results and the theory predictions was resolved by a method, called the Higgs mechanism, introduced by Peter Higgs, Robert Brout, Francois Englert in 1964 [17, 18]. The method introduces a complex scalar field defined as:

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \begin{pmatrix} \phi_1 + \mathrm{i}\phi_2 \\ \phi_3 + \mathrm{i}\phi_4 \end{pmatrix} \tag{2.45}$$

which acts like a doublet under $SU(2)_L$ gauge group. The Lagrangian density corresponding to this complex scalar field can be written as:

$$\mathcal{L}_\phi = \mathcal{L}_{\text{Higgs}} = (D_\mu\phi)^\dagger(D_\mu\phi) - V(\phi^\dagger\phi) \tag{2.46}$$

where, D_μ is the covariant derivative defined in Eq. 2.41 and the potential $V(\phi^\dagger\phi)$ is defined as:

$$V(\phi^\dagger\phi) = \mu^2(\phi^\dagger\phi) + \frac{\lambda}{2}(\phi^\dagger\phi)^2 \tag{2.47}$$

here, μ and λ are arbitrary real parameters. For $\mu^2 < 0$ and $\lambda > 0$, the potential has infinite ground states, each with non-zero field. The two-dimensional analogue of this potential is illustrated in Figure 2.3, generally referred to as "Mexican hat potential". The potential is symmetric and has a ring minimum at:

$$\phi_0 = \sqrt{\frac{-\mu^2}{\lambda}} e^{\mathrm{i}\theta}, \;\; \theta \in [0, 2\pi] \tag{2.48}$$

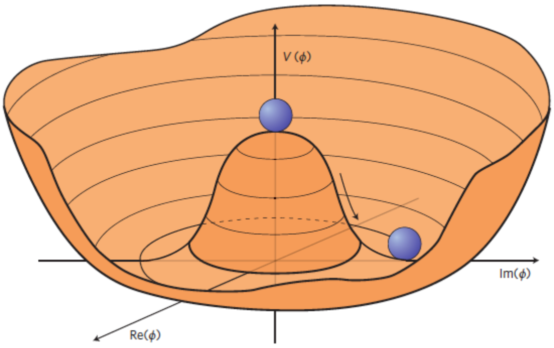


Figure 2.3: An illustration of the Higgs field potential for the case $\mu^2 < 0$ and $\lambda > 0$ as defined in the Eq. 2.47. The potential has infinite ground states, each with non-zero scalar field ϕ. Choosing a specific ground state spontaneously breaks the rotational symmetry. Picture taken from the Reference [19].

Out of infinite choices, one must make a specific choice of a ground state for the vacuum state. This choice spontaneously breaks the symmetry of the electroweak gauge group $SU(2)_L \times U(1)$. The choice of a ground state is completely arbitrary and doesn't change the physics, as all ground states are connected to each other through a global phase transition. The ground state is often chosen to be:

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v \end{pmatrix}, \text{ with } v \equiv \sqrt{\frac{-\mu^2}{\lambda}} \tag{2.49}$$

The magnitude of the minima of the Higgs potential (v) is referred to as the vacuum expectation value of the Higgs field, and has a value of 246 GeV [11]. The scalar field ϕ can be expanded in a perturbation series around the ground state as:

$$\phi = \begin{pmatrix} \eta_1 + \mathrm{i}\eta_2 \\ v + h + \mathrm{i}\eta_3 \end{pmatrix} \tag{2.50}$$

The field h is called as the Higgs field and gauge boson corresponding to this field is called as the Higgs boson. The fields represented by η corresponds to the massless scalar fields and bosons corresponding to these fields are called Goldstone bosons, which are non-physical [20]. These Goldstone bosons can be eliminated with the help of $U(1)$ transformations, but as a consequence the degrees of freedom of Goldstone bosons are passed to the gauge fields B_μ and W_μ^j, which gives masses to the electroweak gauge bosons. The gauge fields B_μ and W_μ^j are not actually observed in experiments as gauge bosons fields. They are given by the linear combination of B_μ and W_μ^j as:

$$W^{\pm}_{\mu} = \frac{1}{\sqrt{2}}(W^1_{\mu} \mp \mathrm{i}W^2_{\mu}), \tag{2.51}$$

and

$$Z_{\mu} = \frac{1}{\sqrt{g^2 + g'^2}}(gW^3_{\mu} - g'B_{\mu}) \tag{2.52}$$

Since, the theory represents electroweak interactions, the massless photon fields are given by:

$$A_{\mu} = \frac{1}{\sqrt{g^2 + g'^2}}(g'W^3_{\mu} + gB_{\mu}) \tag{2.53}$$

Noting the structure of Z_{μ} and A_{μ}, in Eq. 2.52 and 2.53, we can introduce the weak mixing angle, also known as Weinberg angle, represented by θ_W ($\sin^2\theta_W = 0.232$ [11]). It measures the angle by which the spontaneous symmetry breaking rotates the plane of B_{μ} and W^3_{μ} to produce the massive Z boson and the photon. It is easy to identify the following relations:

$$\cos\theta_W = \frac{g}{\sqrt{g^2 + g'^2}}, \quad \text{and} \quad \sin\theta_W = \frac{g'}{\sqrt{g^2 + g'^2}} \tag{2.54}$$

Furthermore, the mass relations for the Higgs boson, the W boson, and the Z boson can be written as:

$$\begin{aligned} m_{\mathrm{H}} &= \sqrt{\lambda v^2}, \\ m_{\mathrm{W}} &= \frac{gv}{2}, \\ m_{\mathrm{Z}} &= \frac{m_{\mathrm{W}}}{\cos\theta_W}. \end{aligned}$$

After getting the masses of the gauge bosons, we can also construct the mass terms for the fermions through the Higgs field. The Lagrangian density corresponding to this is given as:

$$\mathcal{L}_{\mathrm{Yukawa}} = -g_f(\bar{\Psi}_{\mathrm{L}}\phi\psi_{\mathrm{R}} + \bar{\psi}_{\mathrm{R}}\phi^{\dagger}\Psi_{\mathrm{L}}) \tag{2.55}$$

where, g_f is the coupling strength of a fermion to the Higgs field, known as Yukawa coupling. The mass of fermions is directly proportional to the Yukawa coupling strength.

Hence, by adding a complex scalar field to the SM theory, it is possible to trigger the spontaneous electroweak symmetry breaking, which gives masses to the gauge bosons of weak interactions ($W^{\pm}$ and Z) and fermions, while preserving the gauge symmetries of the theory. This mechanism also predicts the existence of a new boson — the Higgs boson, which comes from the remainder of the scalar field. The quest of finding this boson culminated in 2012, almost 50 years after its prediction, by the CMS and the ATLAS experiment at the LHC at CERN [1, 2].

2.2. Beyond the standard model

After the discovery of the Higgs boson, the SM is complete in terms of its particle content and shows consistent results in experiments. However, there are many experimental observations that remain unexplained by the SM. Few of them are:

Matter-antimatter asymmetry: which corresponds to the question of why our universe is matter dominated, rather than having matter and anti-matter in equal amount.

The existence of dark matter and dark energy: as observed by cosmological experiments that our universe is made of only 5% of baryonic matter and energy, the rest is non-luminous matter and energy referred to as dark matter and dark energy.

Neutrino oscillations: The phenomenon in which a neutrino created with a specific lepton flavour can spontaneously change flavour. This phenomenon has been experimentally observed [21], and can only be explained if neutrinos have a non-zero mass, unlike the SM case, where neutrinos are considered as massless particles.

The hierarchy problem: The value of the Higgs boson mass as measured by the experiments is 125.10 ± 0.14 GeV [11]. Theoretically, the mass-squared parameter of the Higgs boson should get large radiative corrections, which would make the bare mass huge, unless there is a precise cancellation between the quantum corrections and the bare mass. This kind of fine-tuning is considered as unnatural by many physicists, and the problem is referred as the hierarchy problem [22].

The strong CP problem: Why is the CP-symmetry conserved in the strong interactions? There is no experimental evidence of CP violation in strong interactions, and there does not exist any reason why the CP needs to be conserved in QCD. This problem is also considered as a fine-tuning problem known as the strong CP problem.

To explain these phenomena, new theories are needed, as the SM clearly has no explanation. Many theorists hypothesize extensions of the SM, either by predicting new particle(s), or by expecting the parameters of the SM to deviate from the SM predictions. Popular theories beyond the standard model (BSM) include: supersymmetry, extra dimensions, grand unified theories, string theory etc. None of these have been experimentally observed yet.

One of the BSM theories is the two Higgs doublet model (2HDM) [23], which proposes two Higgs doublets instead of one that was introduced in Section 2.1.4. The 2HDM model can potentially solve the hierarchy problem, and explain matter-antimatter asymmetry. The model proposes existence of five physical Higgs bosons including: two neutral CP-even scalars h and H, one neutral CP-odd scalar A, and two charged scalars H^+ and H^-. One of the CP-even scalars is the SM-like Higgs boson, which has been discovered. Presence of additional Higgs bosons in nature can alter the SM physics that we know today. In the context of this thesis, additional Higgs bosons can significantly alter the cross section of same-sign WW VBS from its expected SM value, providing a hint for new physics.

Another BSM model, relevant in the context of same-sign WW VBS, is the Georgi-Machacek model [24], which hypothesizes the existence of additional scalar fields in the form of scalar triplets. The neutral component of scalar multiplets, like H_5^0 (typically considered as heavy) or H_3^0 (considered as light), can mediate the same-sign WW VBS process. If these additional scalar particles exist, experiments can see deviations in the cross section values of same-sign WW VBS or possibly new signatures depending on the type of interactions.

2.3. Vector boson scattering

The discovery of the Higgs boson has opened new aspects in the field of particle physics. Our goal is to understand the electroweak sector, measure the couplings with which the Higgs boson interacts with other particles of the SM, look for divergences and new experimental signatures that might hint towards new physics beyond the standard model. Among many processes, which gain attention from theory and experiment sides, vector boson scattering is a prominent one. It probes two key aspects of the electroweak sector together: self-gauge interactions, and the interactions of vector bosons with the Higgs boson.

This section establishes the theoretical and experimental groundwork necessary for a search for vector boson scattering, with a specific focus on same-sign WW scattering in the semi-leptonic decays.

2.3.1. Experimental signature

A typical signature of VBS at the Large Hadron Collider involves two incoming quarks that radiate two vector bosons. These vector bosons then scatter and decay into final-state particles, along with the two scattered quarks, which are detected as jets. Depending on the type of vector bosons decay, there are three possible VBS signatures: 2 jets and 4 leptons (*fully leptonic*), 4 jets and 2 leptons (*semi-leptonic*), or 6 jets (*fully hadronic*). At tree level, the same-sign WW ($W^{\pm}W^{\pm}$) VBS process can be mediated in three ways: the four-W contact interaction, t-channel with photon or Z boson exchange, and t-channel with the SM Higgs boson exchange. The representative Feynman diagrams for these processes in the semi-leptonic decay channel are shown in Figure 2.4.

The most important and distinctive feature of any VBS process is the presence of two highly energetic forward jets coming from the scattered quarks. These jets are often called as VBS jets or tagging jets, and are identified by their large invariant mass (m_{jj}) and large rapidity difference (Δy_{jj}). These kinematic features form the basic VBS selections and are used to suppress the contributions from irreducible backgrounds coming from interference and/or QCD induced vector boson production, such as nonresonant diboson production shown in Figure 2.5. The difference between different types of contributions (electroweak, interference, QCD background) is illustrated in Figure 2.6, with the help of two-dimensional differential distributions, as a function of invariant mass and rapidity of the VBS jets. The basic VBS selections, i.e. requiring large invariant mass and large rapidity difference, is instrumental in isolating EW VBS signal from the bulk of the background. At the LHC, VBS has a very distinctive signature in the detectors. A schematic showing a typical VBS signature in a detector for the semi-leptonic decay mode consisting of forward and backward VBS jets, along with the other final state particles, is shown in Figure 2.7.

2.3.2. Higgs boson preserving unitarity in VBS

In the Section 2.1.4, the introduction of the Higgs boson in the SM was primarily motivated by the need to restore the gauge invariance in the chiral Lagrangian of the electroweak sector and to provide masses to other particles. Another key aspect of introducing the Higgs boson was to restore the unitarity in the SM VBS processes. The unitarity condition is defined as the sum of probability of all possible final states originating from a particular initial state must equal to 1. Due to the fact that the vector bosons possess mass, they must have non-zero longitudinal polarization component along with the transverse polarization components. The general expression for the three polarization components of a boson with mass M, energy E, three-momentum p_z directed along the z-axis, can be written as [26]:

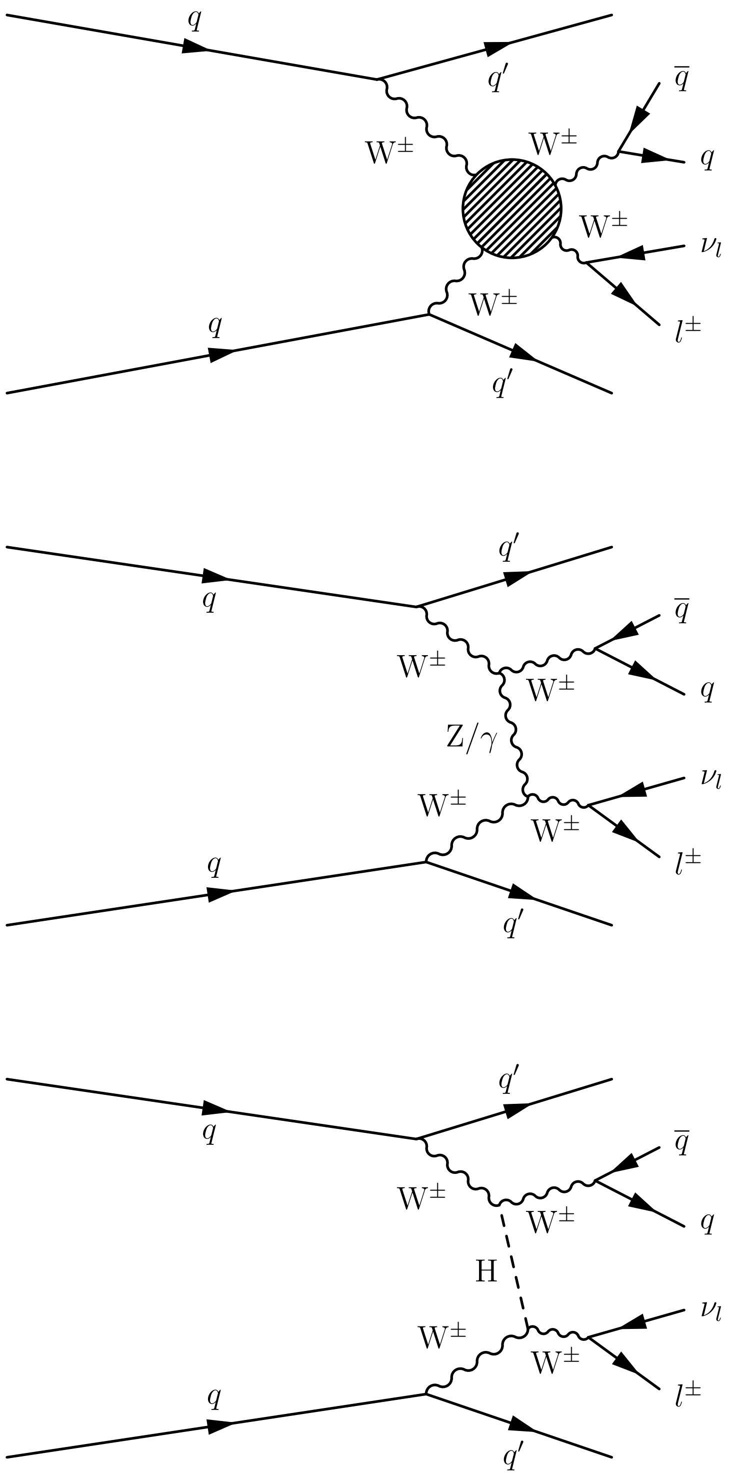


Figure 2.4: Representative Feynman diagrams for same-sign WW VBS in the semi-leptonic decay channel.

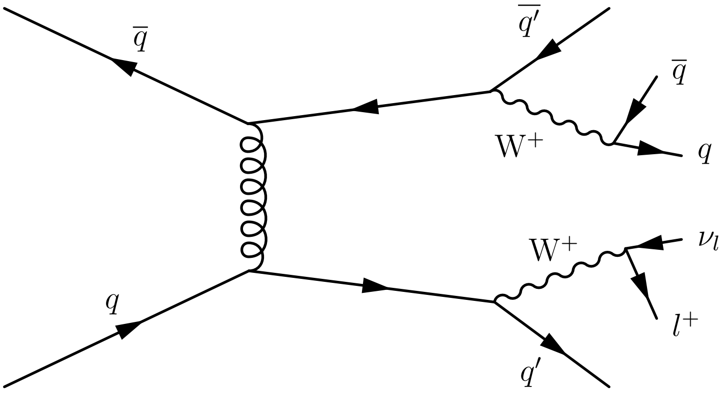


Figure 2.5: Nonresonant diboson production, contributing as an irreducible background to same-sign WW VBS final state at the $\mathcal{O}(\alpha_s^2\alpha^4)$.

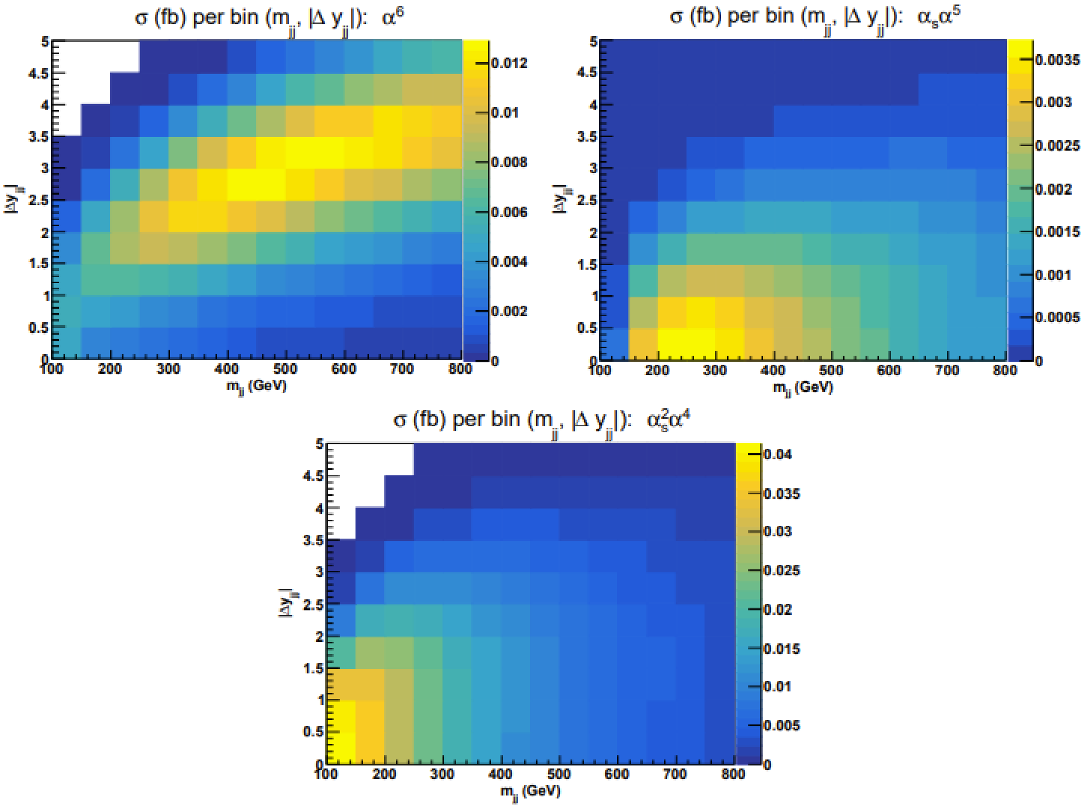


Figure 2.6: Double-differential cross sections as a function of the invariant mass (m_{jj}) and rapidity difference (Δy_{jj}) for the three leading order contributions, EW (top left), interference (top right), and irreducible QCD background (bottom). The EW contribution is characterized by large m_{jj} and Δy_{jj}. The contributions from interference and QCD background are very low in this kinematic phase-space. This figure is taken from Reference [25].

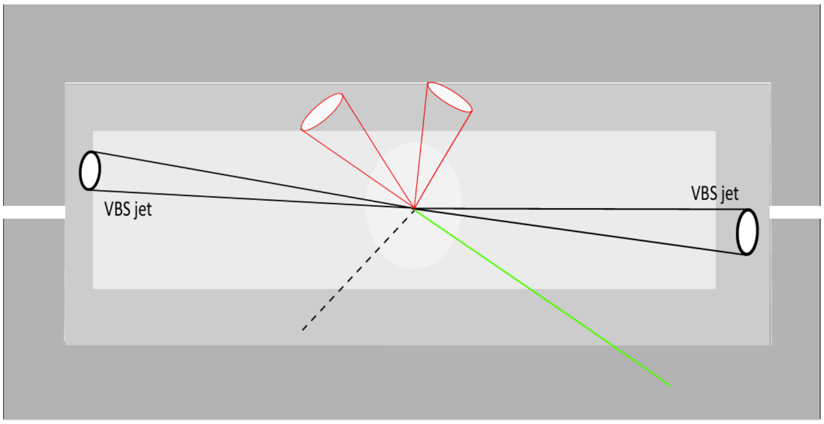


Figure 2.7: A schematic drawing illustrating a typical VBS process at the LHC, characterized by the presence of two highly energetic forward and backward jets labelled as VBS jets. The schematic is depicting a semi-leptonic final state with the presence of a lepton (green line), two additional jets from the hadronic decay of one of the vector bosons (red cones), and a dashed line representing the presence of a neutrino.

$$\epsilon^{\mu}_{+} = \frac{1}{\sqrt{2}}(0, 1, \mathrm{i}, 0), \tag{2.56}$$

$$\epsilon^{\mu}_{-} = \frac{1}{\sqrt{2}}(0, 1, -\mathrm{i}, 0), \tag{2.57}$$

$$\epsilon^{\mu}_{\mathrm{L}} = \frac{1}{\mathrm{M}}(p_z, 0, 0, \mathrm{E}). \tag{2.58}$$

It is important to note here that the longitudinal component depends on the energy. At energies much larger than the boson mass M, it grows indefinitely with energy. To elucidate the problem, let us consider a particular VBS process. Consider an example of longitudinally polarized same-sign W bosons scattering.

$$\mathrm{W}^+_{\mathrm{L}}\mathrm{W}^+_{\mathrm{L}} \to \mathrm{W}^+_{\mathrm{L}}\mathrm{W}^+_{\mathrm{L}} \tag{2.59}$$

At tree-level and in the absence of the Higgs boson, three subprocesses contribute to this process: the four-W contact interaction, and a t-channel process mediated by a Z boson or a photon. The Feynman diagrams for these processes are shown in the Figure 2.8.

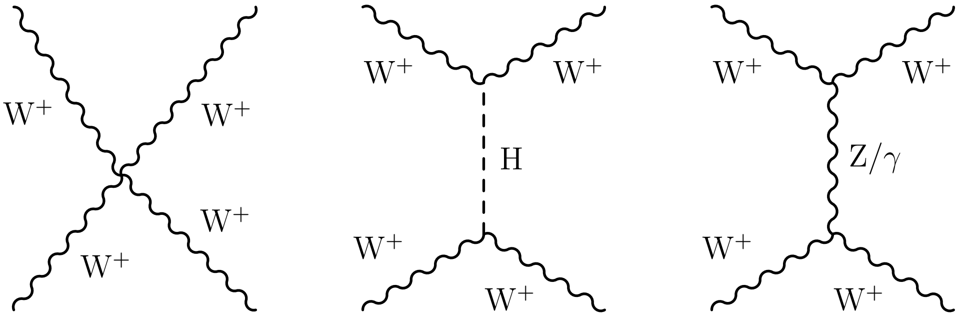


Figure 2.8: Feynman diagrams for $W^+W^+ \rightarrow W^+W^+$ standard model process.

The cross section amplitude of the contact interaction part is given as:

$$\mathcal{M} \sim \epsilon_L \epsilon_L \epsilon_L \epsilon_L \sim s^2 \tag{2.60}$$

and diverges like a fourth power of energy. Here, s is the centre of mass energy squared. The cross section amplitude of the t-channel diagram mediated by the Z boson or a photon is found to be [26]:

$$\mathcal{M}_{\text{gauge}} = -g^2 \frac{s}{4{\text{M}_\text{W}}^2} + \mathcal{O}(s^0) \tag{2.61}$$

The cross section amplitudes in Eq. 2.60 and 2.61 imply unitarity violation and non-renormalizability because they increase with energy. In particle physics, such problems are usually addressed by postulating new particles and interactions that would create counter-terms to cancel divergences. It was proposed that the inclusion of a scalar particle H that can be exchanged between the two W bosons can introduce additional terms like:

$$\mathcal{M}_\text{H} = g^2_\text{HWW} \frac{s}{{\text{M}_\text{W}}^4} + \mathcal{O}(s^0) \tag{2.62}$$

By looking at the leading terms of $\mathcal{M}_{\text{gauge}}$ and $\mathcal{M}_\text{H}$, one can easily notice that the leading order divergences can cancel only if the couplings follow the relation: $g_\text{HWW} = g\text{M}_\text{W}$. Recall from Section 2.1.4 that the coupling constant g itself depends on M_W, this means that the scalar particle H must couple to the W boson with a coupling strength $g_\text{HWW} \sim {\text{M}_\text{W}}^2$. In the standard model, there already exists a scalar particle that fulfils these conditions, the Higgs boson. Inclusion of a subprocess for WW scattering mediated by the Higgs boson cancels exactly the divergent terms and restores unitarity. Figure 2.9 shows the total cross section of $W^+W^+ \rightarrow W^+W^+$ as a function of centre-of-mass energy for different initial (and final) polarization states, without the presence of the Higgs boson and with the presence of Higgs bosons for different mass values.

This in turn completes the standard model from unitarity perspective. Theorists knew that a Higgs boson is necessary before the energy scale of the unitarity viola-

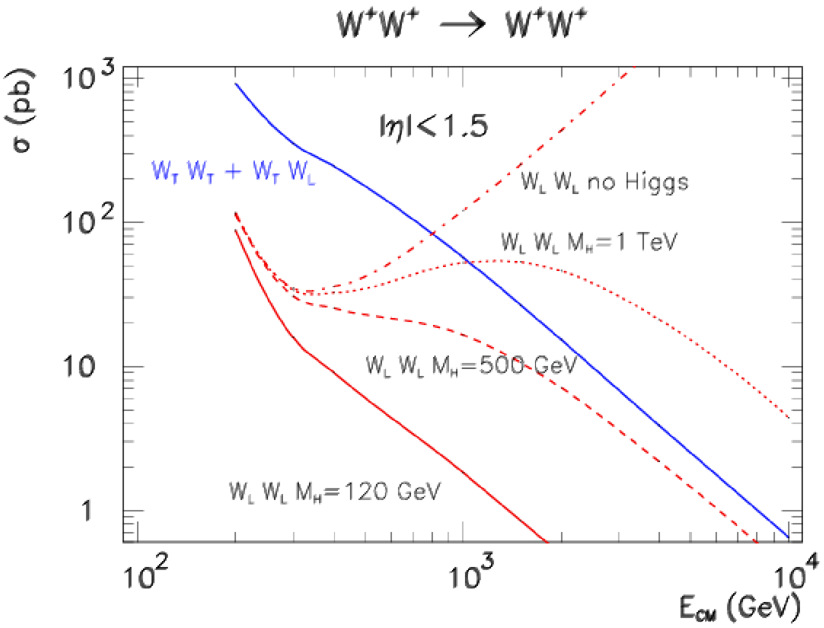


Figure 2.9: The total W^+W^+ VBS cross section as a function of the centre-of-mass energy for different initial (and final) state polarizations and for different Higgs boson masses, including the limiting Higgsless case scenario, that violate the unitarity. This graph is taken from Reference [26].

tion ($\sim$ 1.2 TeV) to restore unitarity. Considering all these facts, upper bounds on the Higgs boson mass were derived way before its actual discovery [27].

The unitarity in WW → WW scattering is preserved by the Higgs boson, with its coupling to the W boson precisely given by $g_{\mathrm{HWW}} = g\mathrm{M_W}$. Any changes to the coupling values would disturb the cancellation of divergences, exhibiting unitarity violation. Also, in models with extended Higgs sector, such as 2HDM and Georgi-Machacek, stated in Section 2.2, it is not necessary that a single scalar boson ensure unitarity. It is only necessary that the sum rules of scalar bosons h_i couplings to the W bosons are fulfilled. For models with doublet Higgs field, these sum rules are given as:

$$\sum_i g^2_{h_iWW} = g^2_{HWW} \tag{2.63}$$

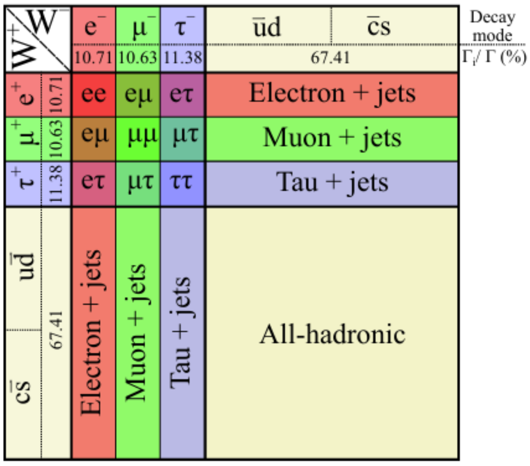
2.3.3. Importance of same-sign WW VBS

The same-sign WW ($W^\pm W^\pm$) VBS process is considered as the "golden channel" in VBS studies. The reason for this is that it has the best cross section ratio of the electroweak component to the QCD component, of the order of 4-6 in typical fiducial regions, while for other processes like opposite-sign ($W^\pm W^\mp$), ZZ, and $W^\pm Z$

VBS, it is typically $\ll 1$. It is due to the charge conservation, which prevents the gluon initiated processes in the QCD background. In addition, basic VBS selections, as explained in the subsection 2.3.1, further enrich the phase-space with the EW component by significantly rejecting the QCD contributions. For this reason, the $W^{\pm}W^{\pm}$ VBS process is the most sensitive channel to search for BSM physics.

2.3.4. Experimental challenge to study semi-leptonic final state of $W^{\pm}W^{\pm}$ VBS and the need to develop a jet charge tagger

The branching fraction of the W boson decaying into a charged lepton and its corresponding neutrino is roughly $10.86 \pm 0.09\%$ [11] for each lepton flavour, and the rest is for hadronic decays. The values of branching fraction for W^+ and W^- bosons are shown in Figure 2.10.



W^+ \ W^-		e^-	μ^-	τ^-	$\bar{u}d$ / $\bar{c}s$	Decay mode
		10.71	10.63	11.38	67.41	Γ_i/Γ (%)
e^+	10.71	ee	eμ	eτ	Electron + jets	
μ^+	10.63	eμ	μμ	μτ	Muon + jets	
τ^+	11.38	eτ	μτ	ττ	Tau + jets	
$u\bar{d}$ / $c\bar{s}$	67.41	Electron + jets	Muon + jets	Tau + jets	All-hadronic	

Figure 2.10: Branching fractions for W^+ and W^- bosons are shown. The figure is taken from Reference [28].

The largest branching fraction is for final states consisting of hadrons only. However, this final state is heavily affected by significant background from QCD interactions. On the other hand, the cleanest final state is the one involving only leptons, but this final state has the smallest branching fraction. The final state in which one vector boson decays into hadrons and the other into leptons offers a good balance between a reasonable branching fraction and manageable background contamination. This makes it an attractive option for analysis. Consequently, this study focuses on the final state where one vector boson decays hadronically, while the other decays leptonically.

The VBS is a pure electroweak process and is extremely rare at the LHC, having cross section $\mathcal{O}(10^{-14})$ of that of proton-proton collisions. Only in recent years, it became possible to study this process with the large enough dataset collected by the LHC experiments during Run 2. Both the ATLAS and CMS experiments have reported the observation of $W^{\pm}W^{\pm}$ VBS process in association with two jets in the fully-leptonic final state [29, 30]. The physics community from both theory and experiment sides has shown renewed interest in measuring this process in all possible final states. A very recent study from the CMS experiment has reported the measurement of $W^{\pm}W^{\pm}$ VBS in association with one hadronic tau final state [31].

Studying $W^{\pm}W^{\pm}$ VBS in other final states, such as semi-leptonic or hadronic, could improve the overall sensitivity of the electroweak signal, when combined with the leptonic final states. However, the main hurdle in the semi-leptonic final state of $W^{\pm}W^{\pm}$ VBS is to identify the charge of the W boson that decays hadronically. Requiring a total electric charge of ± 2 in the final state is easily met in the leptonic final states by requiring two same-sign leptons. But in the semi-leptonic final state, when one of the vector boson decays into a pair of hadrons and detected as a jet(s), the information of the charge of the originating boson is lost. This makes $W^{\pm}W^{\pm}$ VBS process indistinguishable from $W^{\pm}W^{\mp}$, and $W^{\pm}Z$ VBS. This is one of the reasons why the semi-leptonic final states are usually studied by considering WW/WZ VBS as a combined signal process. The evidence of WW/WZ VBS in semi-leptonic decay channel using full Run 2 data is reported by the CMS experiment in Reference [32].

To experimentally distinguish between $W^{\pm}W^{\pm}$, $W^{\pm}W^{\mp}$, and $W^{\pm}Z$ VBS processes in the semi-leptonic decay channel, advanced reconstruction techniques and novel methods that can identify jet charge are essential. The development of such technology is not merely desirable but crucial for this task. The analysis presented in this thesis is the first of its kind to develop and utilize a jet charge tagger to study $W^{\pm}W^{\pm}$ VBS in the semi-leptonic final state. Details of the jet charge tagger are provided in Chapter 7. This innovative technology is not limited to this analysis; it has broader applications in other analyses where measuring jet charge is vital for identifying a signal process.

3. The LHC and the CMS experiment

The standard model provides the theoretical foundation to explain the elementary particles and their interactions in nature. To test the predictions of the standard model, we need to conduct experiments. In high energy physics, we typically look at the smallest possible length scales to probe the tiniest particles. This requires extremely high energy environments, which can only be achieved in particle accelerators. The world's largest particle accelerator is the Large Hadron Collider located at CERN (the European Organization for Nuclear Research), Geneva, Switzerland.

In the first half of this chapter, the experimental setup necessary for accelerating and colliding the particles at the LHC is described, along with a brief introduction of all experiments at the LHC. In the second half, a specific LHC experiment — the Compact Muon Solenoid, is explained whose data have been used for the analysis described in this thesis.

3.1. The Large Hadron Collider

The Large Hadron collider is the world's largest and powerful particle accelerator. It is conceived as a groundbreaking tool for exploring the fundamental constituents of matter. The LHC has become a cornerstone of modern particle physics. It represents a monumental achievement in both engineering and scientific research, pushing the boundaries of human understanding of the universe.

The LHC is a circular accelerator with a circumference of 26.7 kilometres, buried approximately 100 meters underground. It is designed to accelerate proton beams in opposite directions around the accelerator ring. Once the beams reach a specific energy, they are collided at specific points along the ring, where different experiments are installed to collect the collision information and detect newly produced particles.

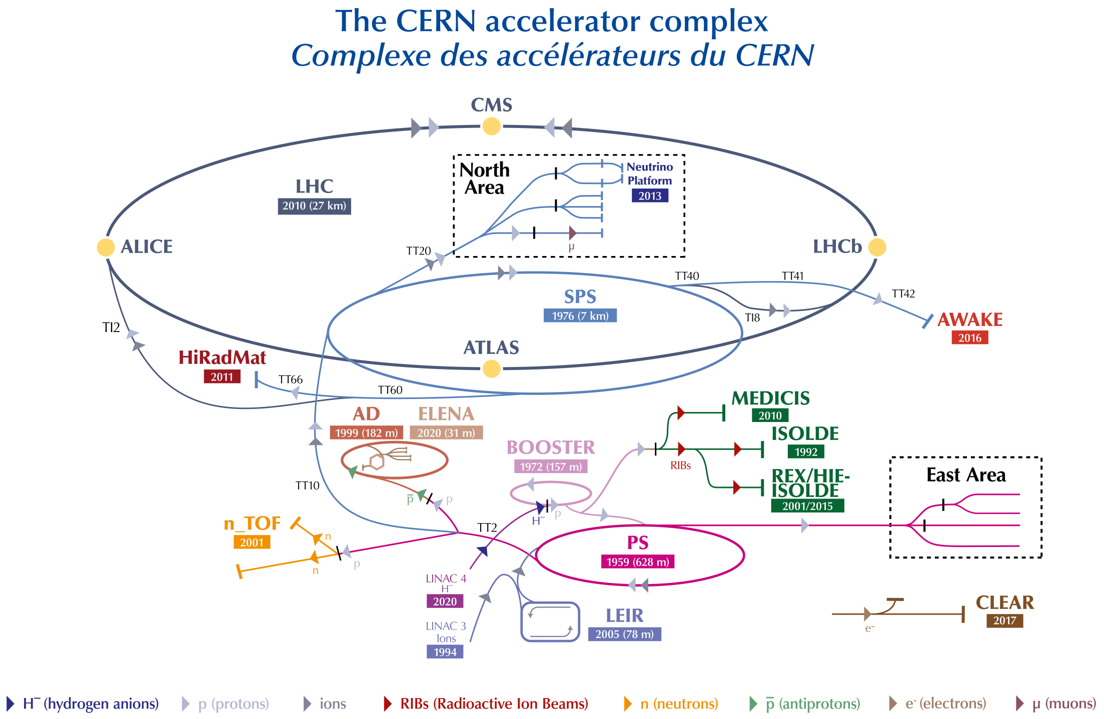
The LHC started its first operation in 2010, at 7 TeV centre-of-mass energy, referred to as Run 1. It continued for three years, and the centre-of-mass energy was raised to 8 TeV in 2012. The Higgs boson was discovered using the Run 1 data [1, 2].

In 2013 and 2014, there was a long shutdown phase to upgrade and improve the experimental setup of the LHC. The LHC started its second operation in 2015, referred to as Run 2 at 13 TeV centre-of-mass energy. Run 2 continued until the end of 2018, followed by another long shutdown of three years for phase 2 upgrades. The LHC is currently recording data in the Run 3 phase, started in 2022, with the new record centre-of-mass energy of 13.6 TeV. Run 3 operations are expected to continue until the end of 2025. In the next few years, the LHC will go through major upgrades for the High Luminosity Large Hadron Collider (HL-LHC) era, which will enable the machine to record many more collisions at much higher energies.

3.1.1. Accelerator complex

The CERN's accelerator complex is a succession of machines that accelerate the particles to increasingly high energies. Each machine boosts the energy of the particles before sending them to the next one in a sequence. A schematic of the LHC accelerator chain, along with its main experiments, is shown in Figure 3.1. The complex chain starts from a linear accelerator (Linac4 [33]), successor of Linac2 [34], which acts as a source of proton beams at CERN. The process begins by accelerating the negative hydrogen ions to 160 MeV energy in Linac4 using radio-frequency (RF) cavities [35], which prepares them to enter the Proton Synchrotron Booster (PSB) [36]. Before the injection of the ions to PSB, they are stripped of their two electrons, leaving behind protons. The PSB is a circular booster consisting of four rings of circumference 157 m. Each booster ring circulates one fourth of the proton packet, hence increasing the beam intensity. The PSB accelerates the protons up to 2 GeV energy to inject them to Proton Synchrotron (PS) [37]. The PS is a circular accelerator with circumference of 628 m and is operational since 1959. It has 100 dipole magnets to keep the proton beams in their circular path while increasing their energy to 26 GeV. After the PS, protons are sent to Super Proton Synchrotron (SPS) [38], a 7 km circular ring, where they reach to an energy of 450 GeV. The SPS accelerator is historically important because the W and Z bosons were discovered by the UA1 and the UA2 collaborations [39–42] using the proton and antiproton beams accelerated by the SPS.

After the SPS, the proton beams are finally sent to the LHC ring, which has a circumference of 26.7 km. It consists of two beam pipes. One beam pipe circulates the beam in clockwise direction, while the other in anticlockwise direction. It takes almost 4 minutes and 20 seconds to fill up each beam pipe, and 20 minutes for the proton beams to reach to an energy of 6.5 TeV. 16 radio-frequency cavities are utilized along the ring to accelerate the protons. In order to keep the proton beams in the circular orbit, strong magnetic fields are required. Such strong magnetic fields are generated from superconducting magnets made from niobium-titanium (Nb-Ti) wires, which are cooled by superfluid liquid helium at 1.9 K. There are 1232 dipole magnets, each 15 meters long, along the length of the LHC ring, generating a magnetic field of 8.34 T. A cross section of the LHC beam pipe, showing beam cavities,



[redacted]

Figure 3.1: Schematic showing the accelerator complex at CERN and experimental set up at various points. The largest ring at the centre is the LHC, which is supported by various small accelerators. The proton beams are injected in the LHC ring after sequentially travelling through various small accelerators. This figure is taken from Reference [43].

the dipole magnets, and the cooling and support system is shown in Figure 3.2.

The proton beams are made up of 2808 proton bunches, which themselves contain roughly 10^{11} protons each. The protons in the bunches experience electrostatic repulsion. To counteract this, beams have to be focused to keep the proton bunches intact. Focusing the beams allow the height and width of the proton bunches to be as close as possible to the initial bunch size. This is achieved by using quadrupole magnets. Two quadrupole magnets work together, one focuses in the horizontal plane and defocuses in the vertical plane, and vice versa. There are 858 quadrupole magnets at the LHC, which are laid in FODO pattern, where the "F" in FODO stands for focusing, "D" for defocusing, and "O" is for drift spaces or bending magnets. The LHC uses 23 such FODO structures along its arcs to keep the beam focused. In addition, there are eight sets of inner triplet magnets, with the job to focus the particle beams at the four collision points, where different experiments are installed.

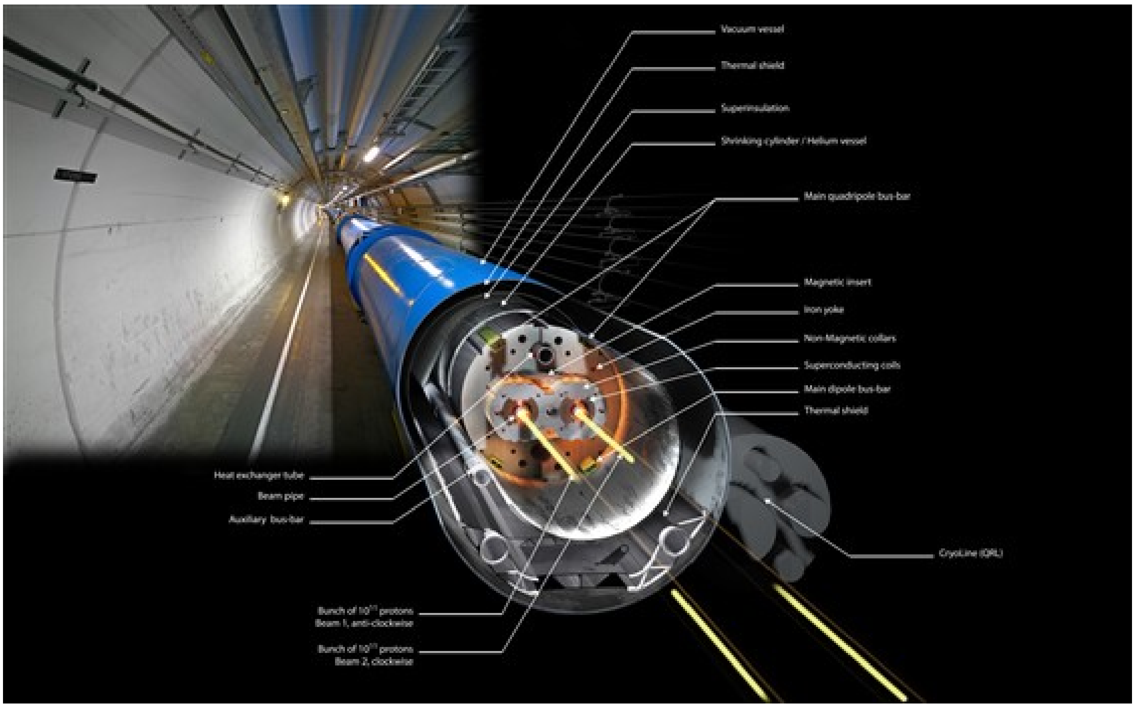


Figure 3.2: Cross section of the LHC beam pipe showing the superconducting dipole magnets, beam cavities, and the cooling and support structures. The figure is taken from Reference [44].

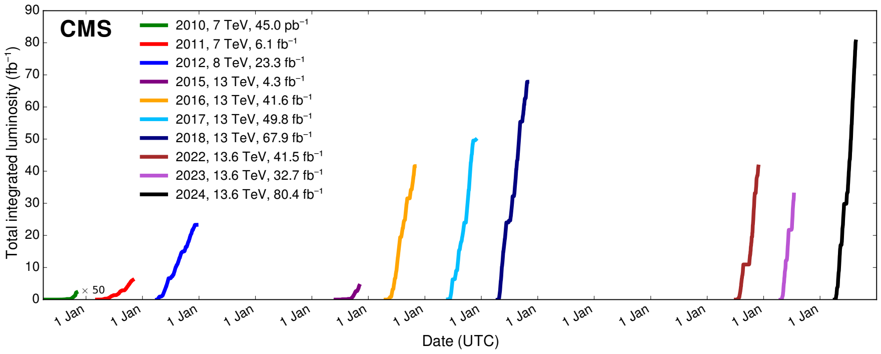
The main task of the LHC is to collide as many proton bunches as possible. To measure the rate of proton-proton collisions per unit time, we define a quantity called instantaneous luminosity as:

$$L_{\text{inst}} = \frac{N_1 N_2 f N_b}{4\pi\sigma_x\sigma_y}, \tag{3.1}$$

where N_1 and N_2 are the number of particles per bunches, N_b the number of bunches, σ_x and σ_y are beam sizes in the horizontal and vertical directions, and f is the revolution frequency. Other effects such as the reduction factor coming from the non-zero crossing angle and possible non-Gaussian shape of the bunches are not considered in the above equation. The number of events for a specific physics process of interest can be written as the product of the process cross section times the time-integrated luminosity:

$$N_p = \sigma_p \times \int L_{\text{inst}}\, dt \tag{3.2}$$

It is therefore desired to maximize the integrated luminosity to have enough events of the rare physics process we are interested in. Figure 3.3 shows the integrated luminosity delivered by the LHC to the CMS detector in each year of its operation, and the cumulative integrated luminosity for all years. The analysis presented in this thesis utilizes the data collected by the CMS experiment in the Run 2 period i.e. in the years 2016, 2017, and 2018.



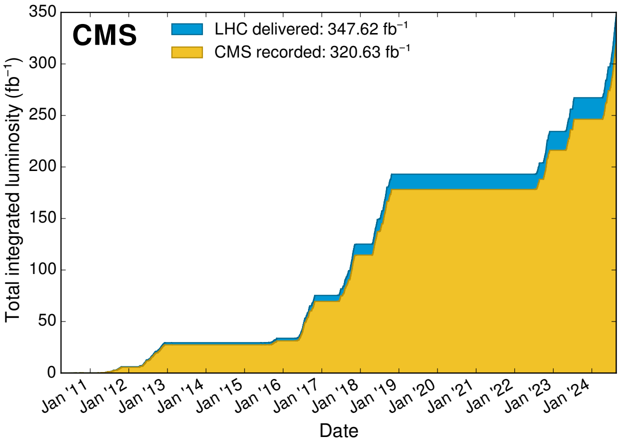


Figure 3.3: The integrated luminosity delivered by the LHC to the CMS detector during proton-proton collisions. The luminosity for each year is shown on the top and the overall cumulative luminosity for all years in the bottom. These figures are taken from Reference [45].

3.1.2. Experiments at the LHC

The beams circulate for several hours inside the LHC beam cavities before they are brought together for collision at four different collision points around the ring. At the four collision points, four different types of detectors are installed, namely ALICE (A Large Ion Collider Experiment), ATLAS (A Toroidal LHC Apparatus), CMS (Compact Muon Solenoid), and LHCb (Large Hadron Collider beauty). Each of these detectors are meant to look for the collision events to gather information of the scattered and newly produced particles. The ATLAS and CMS detectors are the general-purpose detectors, while ALICE and LHCb have specific purposes from a physics perspective. In addition, there are several minor experiments in the surroundings of the major detectors mentioned above.

The ATLAS [46] is the largest detector at the LHC, in terms of volume. It is 46 m long, 25 m high and 25 m wide, with a weight of 7000 tonnes. It sits in the cavern 100 m underground at the CERN main site in Meyrin, Switzerland. It investigates a wide range of physics, from the SM to beyond standard model scenarios, that could possibly explain the existence of dark matter and find evidence for extra dimensions and supersymmetry. When the LHC beams collide, the newly created particles scatter in all possible directions. The ATLAS detector has six different detecting subsystems arranged in layers to detect these particles, record their paths, and measure their momentum and energy. It consists of a toroidal magnetic field, which bend the paths of charged particles to measure their momenta.

The CMS [47] is also a multipurpose detector, like the ATLAS detector, with the same physics goals but different technologies to detect and identify particles. This ensures reproducibility of the results produced by one experiment. With 14000 tonnes, the CMS is the heaviest detector at the LHC. A more detailed description is given in the Section 3.2.

The ALICE [48] is a specific purpose detector that is designed to study heavy ion collisions and quark-gluon plasma. Each year, a part of the LHC collisions are made using lead-ions, which recreates the conditions similar to those just after the Big bang in a laboratory environment. Under extreme conditions, the protons and neutrons melt creating a miniscule fireball, freeing the quarks from their bond with the gluons. This creates a quark-gluon plasma, in which the quarks and gluons are only weakly interacting and are free to move on their own. This phase of matter is particularly interesting to better understand the quark confinement, which is a key issue in the theory of quantum chromodynamics. The ALICE experiment studies the evolution of quark-gluon plasma and observes how it gives rise to the baryonic matter around us. The ALICE detector is 26 m long, 16 m high and 16 m wide, weighing 10,000 tonnes, and sits 56 m underground at one of the LHC collision site.

The LHCb experiment [49] is another special purpose experiment, whose aim is to understand the matter-antimatter asymmetry and to gain insight into CP violation by studying the beauty quarks or b quarks. Unlike the ATLAS and CMS detectors, who cover the entire collision point with the detector material, the LHCb experiment uses a series of subdetectors to detect mainly the particles which are thrown forward by the collision in one direction. The first subdetector is placed close to the collision point, the others are mounted in a sequence one after the other over a length of 20 m. The LHCb detector, 21 m long, 10 m high and 13 m wide, is made up of an asymmetric forward spectrometer and planar detectors, and has a weight of 5600 tonnes.

In addition to the above mentioned experiments, the other experiments at the LHC include: the LHCf (Large Hadron Collider forward), TOTEM (TOTal Elastic and diffractive cross section Measurement), MoEDAL (Monopole and Exotics Detector at the LHC), FASER (Forward Search Experiment), and SND@LHC (Scattering and Neutrino Detector at the LHC). TOTEM [50] and LHCf [51], are the smallest experiments at the LHC, dedicated to study "forward particles"—protons or heavy ions that skim past each other instead of colliding directly. TOTEM has detectors on either side of the CMS interaction point, while LHCf includes two detectors positioned 140 meters apart along the LHC beamline near the ATLAS collision point. MoEDAL [52], located near LHCb, is focused on detecting a theoretical particle known as the magnetic monopole. The newest LHC experiments, FASER [53] and SND@LHC [54], are set up close to the ATLAS collision point to search for light new particles and to explore neutrinos.

3.2. The CMS experiment

The Compact Muon Solenoid experiment is a multipurpose detector system designed to detect and measure particles resulting from collisions at the LHC. It is located 100 m underground near Cessy on the border of France and Switzerland. It has a length of 21.6 m, a diameter of 14.6 m, and a total weight of 14000 tonnes [47]. A full description of the CMS detector is given in Reference [55]. It has a cylindrical design and consists of several subdetectors arranged in layers around the interaction point.

A schematic showing different layers of the CMS detector is shown in Figure 3.4. The first subdetector system surrounding the interaction point is the silicon based tracking system, whose purpose is to measure the trajectories of all charged particles that emerge from the collisions. Surrounding the tracker are the calorimeters that measure the energies of electromagnetically and hadronically interacting particles. The tracker and the calorimeters are contained within the superconducting solenoid magnet, that generates a magnetic field of 3.8 T. This magnetic field curve the trajectories of the charged particles, which is then used for particle identification and momentum measurement. The outer portion of the CMS detector is composed of muon detector systems, which includes drift tubes, cathode strip chambers, and resistive plate chambers. The superconducting solenoid magnet requires a “return yoke”, to control the magnetic field outside of the solenoid, which has a field strength of 2 T. The return yoke is made from steel and acts as a skeleton of the CMS detector as well as a muon filter. It guides the magnetic field and filter all particles except muons and neutrinos. Finally, the muon detectors are placed in between the spaces of the return yoke.

The particles produced during collisions interact with the CMS detector material. Their detection is based on the type of interaction with the detector. The emerging particles are either stable over the lengths of the detector or decay into other

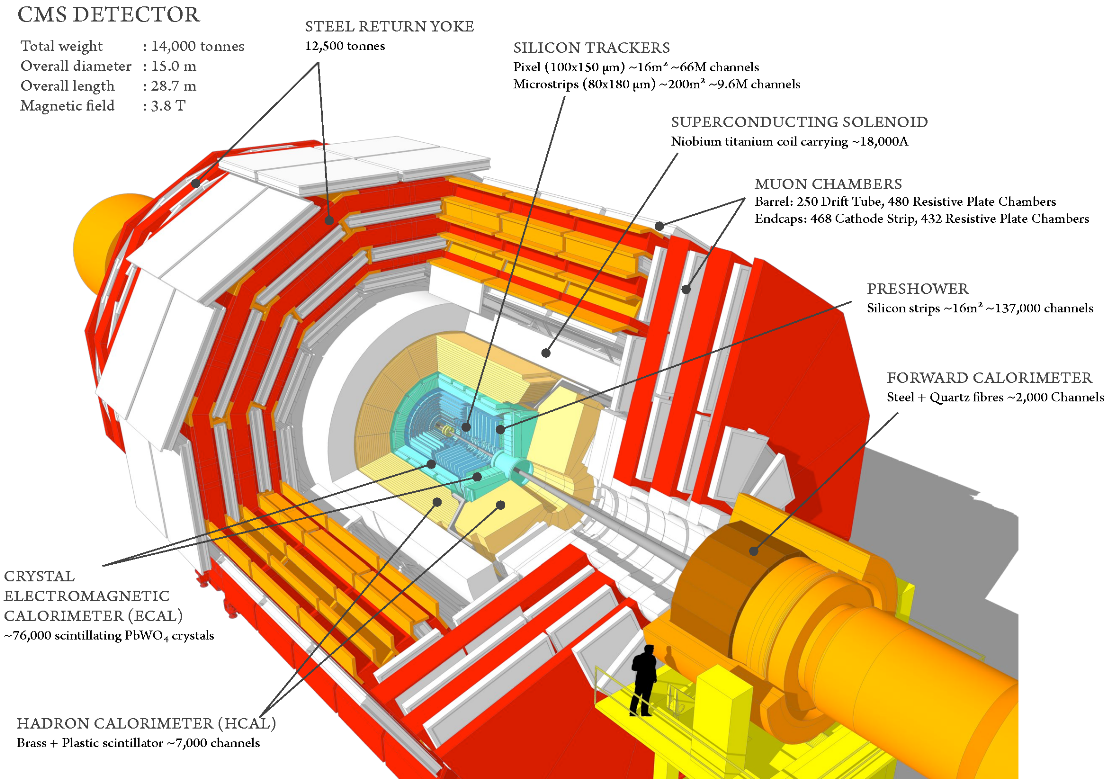

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 3.4: A cutaway view of the CMS detector showing its subdetector systems. Closest to the interaction point are the silicon pixel tracking detectors, followed by the silicon strip tracking detectors. Surrounding these are the electromagnetic and hadron calorimeters. These elements are contained within a solenoid magnet that produces a magnetic field of 3.8 T. Beyond this region, muon detectors are predominantly located. The figure is taken from Reference [56].

particles. The key concept is that charged particles bend when they travel through a magnetic field, with the degree of bending depending on their momentum and charge. This bending, combined with the energy measured in the calorimeters, helps reconstruct the particles and determine their origin or decay paths. Electrons leave a detectable track in the tracker because of their charge and deposit their energy in the Electromagnetic Calorimeter (ECAL). In contrast, photons do not leave tracks and only deposit energy in the ECAL. Charged hadrons also leave tracks and deposit energy in the Hadron Calorimeter (HCAL), while neutral hadrons deposit energy in the HCAL but do not leave any tracks. Muons, which travel the farthest, bend in one direction inside the solenoid magnet and in the opposite direction outside it. Neutrinos are not detected directly by the CMS detector; their presence is inferred from the conservation of momentum.

Different parts of the CMS detector are explained in detail in the following sections.

3.2.1. The coordinate system

To describe the location inside the CMS detector, a right-handed coordinate system is used, with the x-axis pointing towards the centre of the LHC, and the y-axis upwards. The z-axis is fixed in the direction of the beam, pointing in the counter-clockwise direction. The CMS detector is cylindrical, so it is natural to use polar and azimuthal angles. The azimuthal angle ϕ goes from 0 to 2π. The collision products are expected to be symmetric in the azimuthal angle ϕ, measured in the $\mathrm{x}-\mathrm{y}$ plane, starting from the x-axis. The polar angle θ is measured from the z-axis. Instead of polar angle θ, a useful quantity pseudorapidity is defined as:

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right), \tag{3.3}$$

where η is zero in the transverse direction, and diverges as θ approaches zero. Pseudorapidity can also be written as a function of three momentum of a particle:

$$\eta = \frac{1}{2}\ln\left(\frac{|\vec{p}| + p_z}{|\vec{p}| - p_z}\right), \tag{3.4}$$

which under the relativistic limit changes to usual rapidity:

$$y = \frac{1}{2}\ln\left(\frac{E + p_z}{E - p_z}\right). \tag{3.5}$$

The angular separation between two particles is often defined as the distance in the $\eta - \phi$ plane:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{3.6}$$

3.2.2. Silicon tracker

The innermost subdetector system of the CMS detector is the tracker, whose purpose is to measure the trajectories of the electromagnetically charged particles. The basic working principle is that moving charged particles bend in the presence of magnetic field. Knowing the trajectory of a particle, momentum and charge of the particle can be deduced. The tracker is placed in the centre of the CMS detector, where it experiences 3.8 T magnetic field strength from the solenoid magnet. To measure trajectories of charged particles, the tracker records hits at different distances from the centre. These hits are precisely fitted to estimate the trajectories of the particles. The precise measurement of a particle's trajectory is crucial to accurately measure the particle's momenta, and the point of its origin. Therefore, the tracking system is constructed using multiple layers of finely segmented silicon sensors. The choice of silicon based sensor technology was motivated by their light weight, fast response, and good spacial resolution.

The tracker is made up of two types of silicon sensors: the pixels, at the very core of the detector, and the silicon microstrip detector that surrounds it. The silicon pixel detector, being closest to the collision point, receives the most flux of particles. Due to this, a dedicated cooling system is deployed that maintains the temperature of the tracking system at $-20°$C, in order to mitigate the radiation damage. With the increased lifetime, radiation damage is still a challenge for the tracking system. In light of this, pixel detectors have been upgraded during the technical stop in the years 2016-2017, under the so-called Phase 1 upgrade [57]. During the upgrade, additional pixel detectors were installed in both barrel and endcap regions, which now provides an extended coverage in the η plane.

The tracker operation is based on the principle of $\mathrm{p-n}$ junction diode. A basic module is created using positively and negatively doped silicon, to which a reverse-bias voltage is applied, to create a depletion region. When charged particles transverse through this region, they create electron-hole pairs, and induce an electric current, which is then amplified and detected. The electric signals are converted into infrared pulses, which are then transmitted through a 100 m fibre optic cable for further analysis in a radiation free environment. The tracker utilizes 40,000 fibre optic links, offering an efficient, low-power, and lightweight method for signal transmission.

A sketch of one quarter of the CMS tracking system is shown in Figure 3.5. There are a total of 124 million silicon pixels, each with a size of 100 μm by 150 μm and a separate read-out channel, which ensure accurate measurement of particle tracks. The pixel detector gives a coverage up to $|\eta| = 2.5$, and a spatial resolution of $15 - 20$ μm. The strip detectors are placed in the outer part of the tracker, surrounding the pixels, and are composed of 10 million detector strips, read by 72,000 microelectronic chips. The strip detectors are arranged in modules in the inner and outer barrel region, and inner and outer endcap regions. The spatial resolution of the strip detectors is considerably worse than the pixel detectors, and can vary from 20 μm to 200 μm. More details on the pixel and strip detectors can be found in References [57, 58].

During the 2016 data-taking period, the silicon strip tracker experienced an issue with the APV25 readout chip pre-amplifier. This issue led to a decrease in the signal-to-noise ratio and an associated loss of hits from charged particles [60]. The underlying cause was saturation effects in the APV pre-amplifier. Approximately 20 fb^{-1} of data collected in 2016 was affected by this problem. The issue was later resolved, resulting in differences in detector conditions before and after the fix in

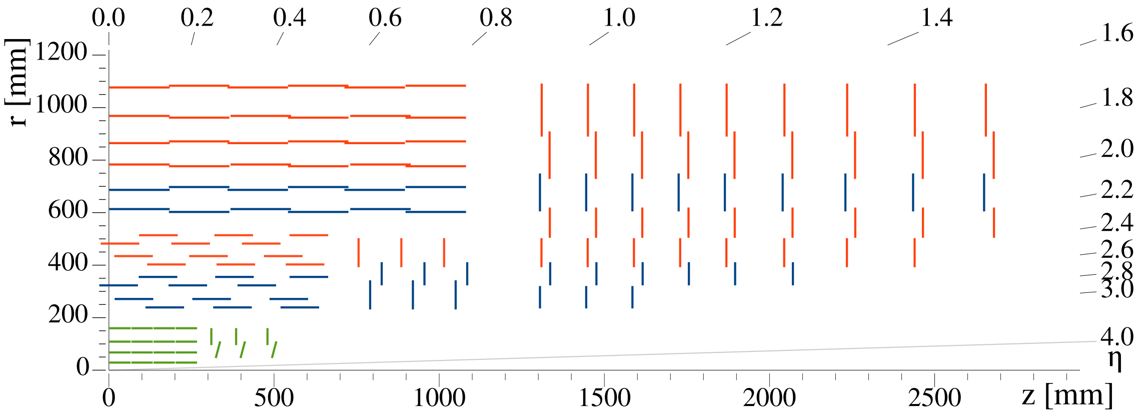


Figure 3.5: The sketch of one quarter of the CMS tracking system after the Phase 1 upgrade is shown in r − z view. The interaction point is located in the lower left corner. Closest to this interaction point are the silicon pixel detectors, which are displayed in green. Surrounding these are the silicon strip detectors, with single-sided detectors shown in red and double-sided detectors depicted in blue. The figure is taken from Reference [59].

the 2016 dataset. To account for these differences, separate simulation production chains were introduced. For the early part of 2016, when the APV issue was present, a dedicated re-reconstruction production chain (2016 preVFP) was created to accurately model the detector conditions. For the later part of 2016, after the issue was fixed, a separate production chain (2016 postVFP) was introduced. This split is crucial, as the detector performance, calibrations, efficiencies, and systematic uncertainties differ between these two periods. To ensure a consistent and precise analysis of the full Run 2 dataset collected by CMS, all simulated samples used in this thesis are categorized into four reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018.

3.2.3. Electromagnetic calorimeter

The electromagnetic calorimeter [61, 62] is the second subdetector system surrounding the tracker, which is designed to measure the energy of electromagnetic particles, such as electrons and photons. It does this by absorbing the particles and measuring the resultant electromagnetic showers. These showers are cascades of secondary particles created when the primary particle interacts with the material of the calorimeter. The electrons and photons primarily interact with the ECAL through bremsstrahlung and pair production, respectively. Electrons and positrons, in the presence of atomic nuclei and with energies greater than a certain critical value, generate photons via bremsstrahlung. These photons can undergo pair production, among other interactions, forming showers in the ECAL. Similarly, incident photons with sufficiently high energy can initiate pair production, resulting in an electron-positron pair that can produce bremsstrahlung photons. This process continues until the energy of the photons in the shower drops below the critical threshold needed for pair production. At this point, the photons are absorbed by the scintillating material, which emits scintillation light. The intensity of the emitted light is proportional to the energy of the initial incident particle, allowing us to deduce the energy of the incident particle.

The CMS ECAL is a homogenous structure made from lead tungstate ($PbWO_4$) crystals. The choice of $PbWO_4$ is well motivated due to its high density $8.28\,g/cm^3$, short radiation length $X_0 = 0.89$ cm, and small Molière radius $R_M = 2.2$ cm, which describes the spread of the shower in the transverse direction. The $PbWO_4$ crystals are optically clear, fast, and radiation hard. In addition, the scintillation decay time of these crystals is of the same order of magnitude as the LHC bunch crossing time, i.e. about 80% of the light is emitted in 25 ns. All these properties of $PbWO_4$ make it a perfect choice for ECAL absorbing and scintillating material.

The ECAL is composed of a central barrel (EB) and forward endcaps (EE) on both sides of the interaction point. The EB covers the pseudorapidity range $|\eta| < 1.479$ and the EE covers the range, $1.479 < |\eta| < 3.0$. Figure 3.6 displays a geometric view of one quarter of the ECAL. The EB contains 61,200 $PbWO_4$ crystals, which are arranged into 36 supermodules. Each supermodule is made up of four modules and contains 1,700 crystals. These crystals have a tapered shape, with an inner area of $22 \times 22\,cm^2$ and outer area of $26 \times 26\,mm^2$. Each crystal is 230 mm in length, equivalent to 25.8 radiation lengths, providing sufficient material to fully develop electromagnetic showers. Altogether, the EB encompasses a crystal volume of 8.14 m^3, and weighs 67.4 tonnes. The scintillation light emitted by the crystals in the EB is detected using avalanche photodiodes (APDs). The EE, on the other hand, is composed of two sections, known as Dees, located on either side of the interaction point. Each Dee contains 3,662 crystals. The EE crystals have a front face area of $28.62 \times 28.62\,mm^2$ and a rear face area of $30 \times 30\,mm^2$. With a length of 220 mm, each crystal corresponds to 24.7 radiation lengths. Altogether, the EE crystals have a total volume of 2.90 m^3 and a combined weight of 24 tonnes. The light signals from the EE are read using vacuum phototriodes (VPTs).

In addition to the EB and the EE, the ECAL has a 20 cm thick preshower (ES) layer in front of the EE, covering the pseudorapidity range $1.653 < |\eta| < 2.6$, as shown in Figure 3.6. The ES is a two-layer sampling calorimeter made from lead layers for absorption and silicon sensors for detection. The purpose of this layer is to reduce the background coming from the decay of a neutral pion decaying into two highly collinear photons in the forward region, which mimics a high-energy photon in ECAL. By looking at the silicon sensor hits, when a high-energy photon seems to appear in ECAL, it is checked whether it's an actual signal from a high-energy photon or a pair of photons from a neutral pion decay.

The energy resolution of the ECAL is given as:

$$\left(\frac{\sigma(E)}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + (C)^2, \tag{3.7}$$

where S is the stochastic term, N the noise term, and C is the constant term. The stochastic term includes the contribution from statistical fluctuations in the width of the electromagnetic shower and the energy deposited in front of the ECAL. The noise term contains contribution from the electronics, digitization, and pileup noise. The constant term originates from the non-uniformity of longitudinal light collection, intercalibration errors, and leakage of energy from the back of the crystal. The numerical values of these three coefficients have been measured using the electron test beams of energy 20 – 250 GeV [63], and found to be 0.028 $\sqrt{\text{GeV}}$, 0.12 GeV, 0.003, for S, N, and C, respectively.

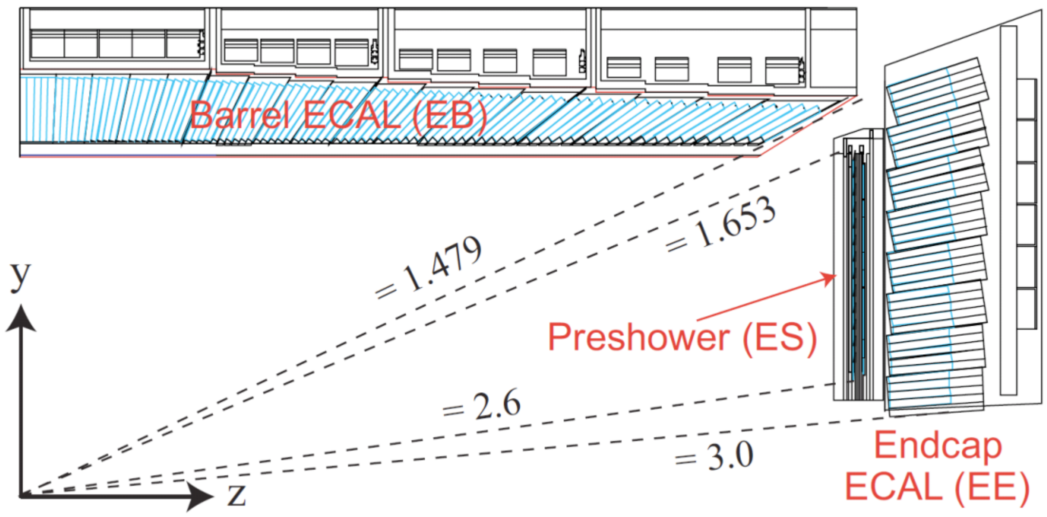


Figure 3.6: The geometric view of one quarter of the CMS ECAL showing the ECAL Barrel (EB), ECAL Endcap (EE), and ECAL preshower (ES) with their respective ranges in η. This figure is taken from Reference [64].

3.2.4. Hadron calorimeter

The next subdetector system of the CMS, after the ECAL, is the Hadron calorimeter [65], which measures the energy of hadrons. The hadrons are not stopped by the ECAL. Therefore, another layer of calorimeter — the HCAL, is required to absorb such particles and measure their energies. Hadrons are the particles made of quarks and gluons, which interact with the HCAL detector material via the strong interaction. When a hadron enters into the HCAL, it produces a cascade of secondary particles through inelastic collisions and produces a hadron shower, similar to the electromagnetic shower in the ECAL. In analogy to the radiation length X_0 in the

ECAL, nuclear interaction length λ_0 can be defined for the hadronic interactions. The nuclear interaction lengths are typically much larger than the radiation lengths. Therefore, in order to contain the full hadron shower within the calorimeter lengths, the HCAL is designed as a sampling calorimeter, with alternating layers of dense absorbing and scintillating materials. The dense absorbing layers are made from brass, which has a density of 8.53 g/cm^3, radiation length of 1.49 cm, and nuclear interaction length of 16.42 cm [55]. The active material chosen for the scintillating layers is Kuraray SCSN81 plastic, because of its long-term stability and moderate radiation hardness.

The structure of the HCAL is divided into four parts: HCAL barrel (HB), HCAL endcaps (HE), HCAL outer (HO) and HCAL forward (HF). The layout of the HCAL system is shown in Figure 3.7. The HB and HE are located inside the solenoid magnet, covering the range $|\eta| < 1.39$ and $1.39 < |\eta| < 3.0$, respectively. They work in conjunction with the ECAL. The combined depth of the HB, and the ECAL crystals in front of the HCAL is not enough to stop all strongly interacting particles, as the depth of HB is constraint by the radius of the solenoid magnet. Due to this reason, an additional layer, the HO, is installed outside the solenoid magnet to fully contain the hadronic showers. In addition, HCAL has another layer, the HF, which covers the forward region $3.0 < |\eta| < 5.0$. Adding HO, provides an absorber length of $11.8\lambda_0$. In the forward region, the particle flux is quite large, therefore, the HF is constructed as a Cerenkov detector, with quartz fibres as an active material to withstand harsh environments.

The energy measurement of charged hadrons in the HCAL provides complementary information to the tracker information. In case of neutral hadrons, the HCAL is the only source of detection, as they do not leave any tracks in the tracking system. The HCAL also plays a very important role in indirect detection of non-interacting particles like neutrinos, which only appear as missing energy after all other particles are detected. Analogous to the ECAL, the energy resolution of the HCAL is measured using pions of energy $20 - 300$ GeV [66] and is given as:

$$\frac{\sigma(E)}{E} = \frac{1.15\sqrt{\text{GeV}}}{\sqrt{E}} \oplus 0.055. \tag{3.8}$$

3.2.5. Superconducting solenoid magnet

The superconducting solenoid magnet of the CMS detector is 12.8 m long, has a diameter of 6.3 m, and weighs 220 tonnes. It can deliver up to 4 T magnetic field strength, but is normally operated at 3.8 T. To generate this strong magnetic field, the number of ampere-turns required are 41.7 MA-turn, therefore, the winding is composed of four layers instead of one layer. The winding is made from Ni-Ti Rutherford cables co-extruded with pure aluminium. To achieve the critical

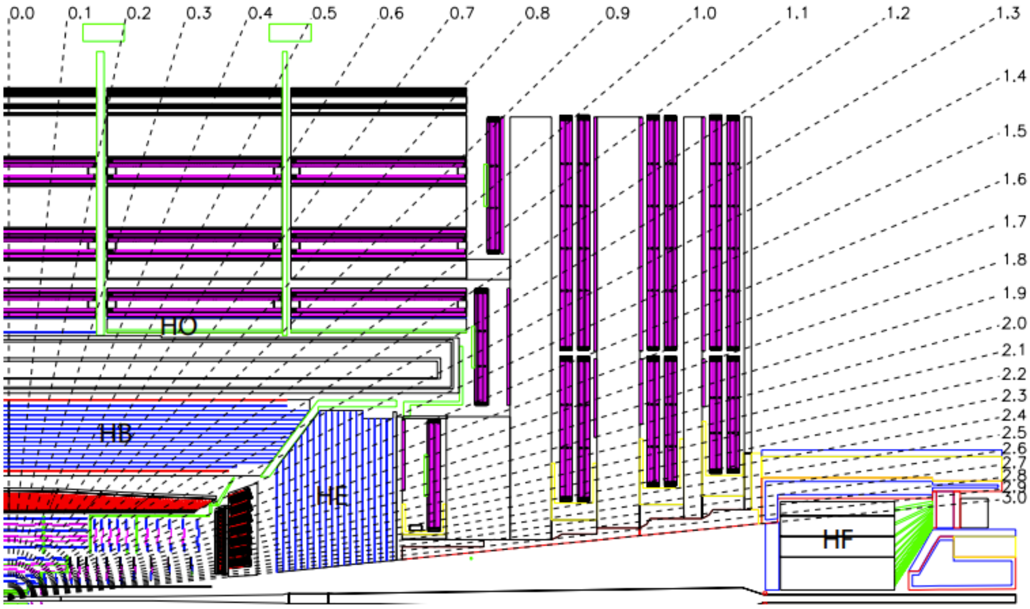


Figure 3.7: Longitudinal view of the CMS HCAL system, showing the HCAL barrel (HB) and HCAL endcap (HE) detectors inside the solenoid magnet, the HCAL outer (HO) detector outside the magnet, and the HCAL forward (HF) detector in the very forward region. This figure is taken from Reference [55].

temperature for the Ni-Ti superconductor, the coil is cooled to 4.5 K using liquid helium. At its standard operating current of 19.14 kA, the coil is capable of storing 2.6 GJ of energy. The field strength generated by the CMS solenoid magnet and predicted field lines are shown in Figure 3.8.

To confine and guide the magnetic field outside the solenoid, a 10,000 tonnes iron return yoke is used. It also acts like a filter, which stops all remaining particles except muons and neutrinos. The return yoke is composed of 11 large elements, 5 barrel wheels, and 6 endcap disks. The modular structure of the yoke allows easy relative movements and facilitates the assembly of the subdetector systems.

3.2.6. Muon system

Almost all the SM particles created in collisions are detected and absorbed by the above mentioned subdetector systems, except muons and neutrinos. Muons are 200 times heavier than electrons, which causes less retardation through bremsstrahlung emissions in the ECAL, so they easily pass through it. In the HCAL, muons do not interact at all because of their leptonic nature. As a consequence, they easily escape all the inner subdetecting systems, making their way outside of the solenoid magnet.

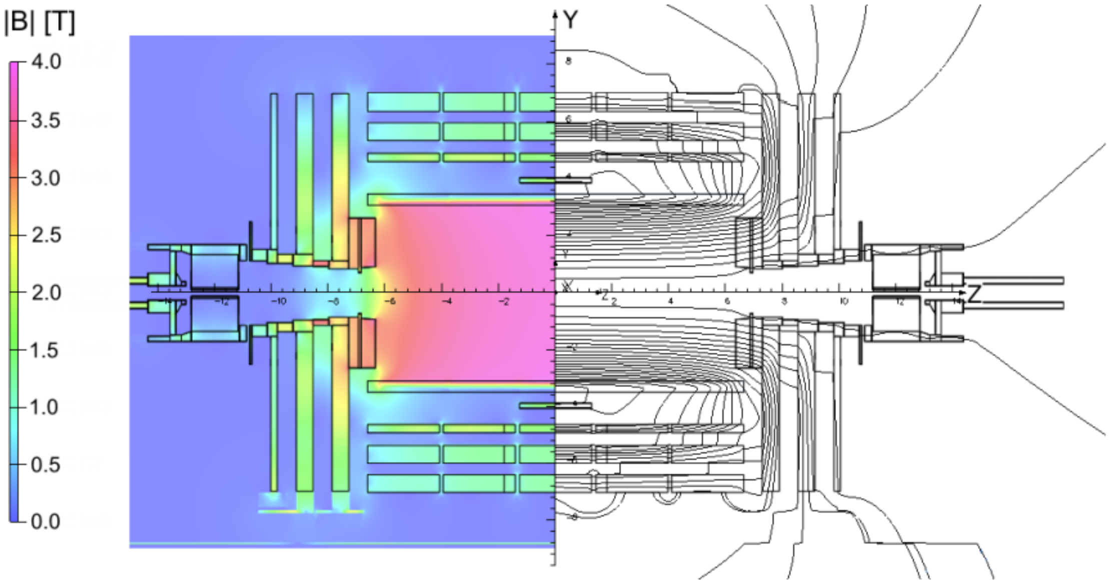


Figure 3.8: A sketch of the CMS solenoid magnet projected onto the longitudinal section of the CMS detector, at the centre magnetic field of 3.8 T. Values of the magnetic field are shown on the left, and field lines are shown on the right. Each field line shows an increment of 6 Wb in the magnetic flux. This figure is taken from Reference [67].

To detect and absorb muons, specialized systems have been installed outside of the solenoid magnet.

The muon detecting system consist of three types of gaseous detectors: drift tube (DT), cathode strip chambers (CSC), and resistive plate chambers (RPC). They are arranged in alternating fashion within the spaces in between the iron return yoke. A schematic overview of muon subdetectors installed in CMS is shown in Figure 3.9. The basic principle of detection is that muons ionize the gas when travelling through the gaseous detector. The charged particles created during ionization are measured to detect muons. The muon system detects muons up to $|\eta| = 2.4$. In the barrel region $|\eta| < 1.2$, the flux of muons is low and the solenoid magnetic field is uniform. The DTs are installed in this region, which provide good time and position resolutions. In the forward region, the rate of muons is high and magnetic field is non-uniform, here the CSCs are used that provide excellent position resolution in the range $0.9 < |\eta| < 2.4$. The RPCs detectors have good time resolution. They are installed in both barrel and forward regions. The details of the CMS muon system are described in Reference [68].

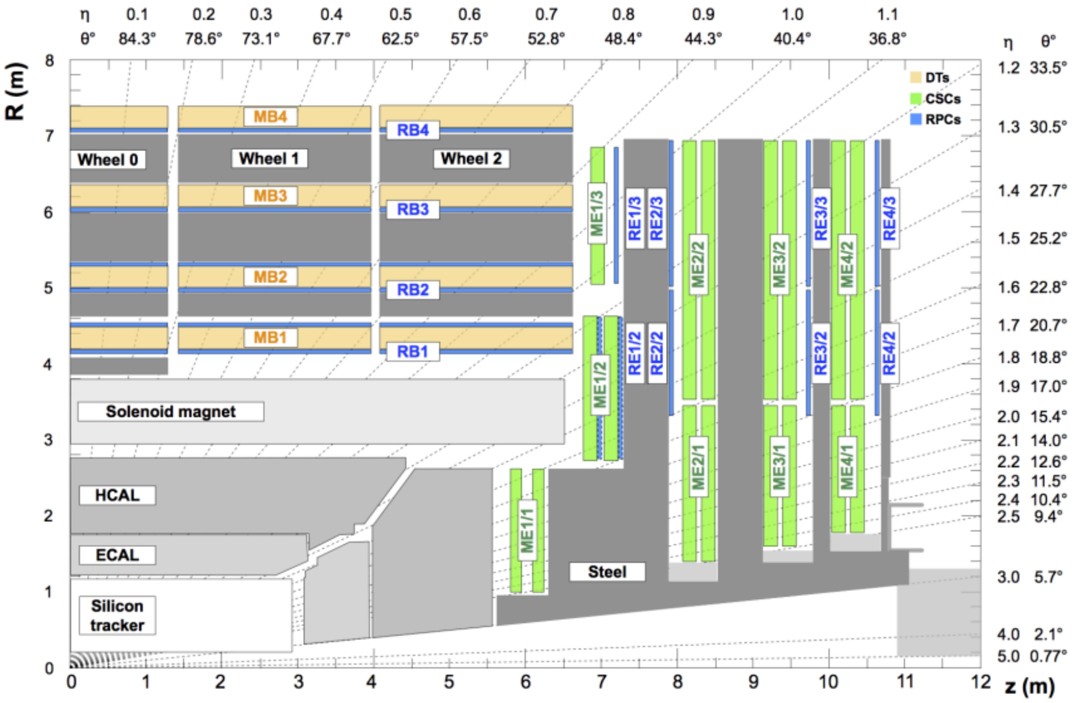


Figure 3.9: A schematic view of the CMS muon detecting system consisting of drift tubes (DTs), cathode strip chambers (CSCs), and resistive plate chambers (RPCs). The muon barrel (MB) are the DTs in the barrel region, and muon endcap (ME) are the CSCs in the endcap region. The RPCs barrel (RB), and RPCs endcap (RE) represents the locations of the RPCs in the barrel and endcap regions, respectively. This figure is taken from Reference [69].

3.2.7. Trigger system and data acquisition

The LHC is designed to create proton-proton collisions at a centre-of-mass energy of 14 TeV, with bunch crossing every 25 ns. At nominal luminosity of $10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$, an average of 27 interactions were recorded per bunch crossing in 2016 [45]. The amount of data created from these collisions is so gigantic that it is physically not possible to store all collision's data due to storage limitations. To reduce the amount of data to store, the CMS detector has a specialized triggering system that selects and stores potentially interesting events while discarding all non-interesting collisions. The CMS trigger is a two-tiered system: Level 1 (L1) trigger and high-level trigger (HLT). A schematic showing the CMS trigger and data acquisition system is shown in Figure 3.10.

The L1 trigger is a hardware based trigger that reduces the original data rate of 40 MHz to 100 kHz. It uses dedicated hardware — field programmable gate arrays (FPGAs) and application specific integrated circuits (ASIC). It consists of a calorimeter trigger and a muon trigger, which utilize the information from the calorimeters

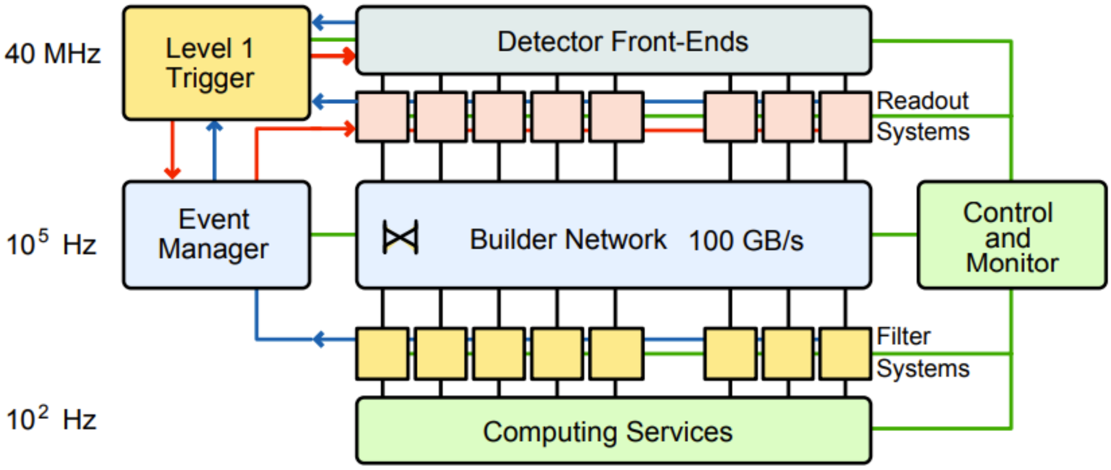


Figure 3.10: Simplified architecture of the CMS data acquisition system. Shown are the key building blocks for a single slice of the system. This figure is taken from Reference [70].

(ECAL and HCAL), and the muon detectors, respectively. The calorimeter trigger has a regional and a global trigger. The regional calorimeter trigger reads out energies from single calorimeter cells of the ECAL and the HCAL along with the quality flags, and pass on the information to the global calorimeter trigger, which further processes this information. The L1 muon trigger collects the information about the tracks of muons from RPC pattern comparator as well as DT and CSC track finders. This information is then fed to the global L1 muon trigger that collects muon tracks and ranks them according to their reconstruction quality. The information from global calorimeter trigger and global muon trigger is combined in a global L1 trigger, which implements a trigger menu — a simplified set of selection criteria to identify candidates (jets, leptons, and photons) of an interesting collision event.

The HLT is a software based trigger system which runs on a farm of computers. It uses events passing the L1 trigger, starting from so-called L1 seeds, and then includes further information for reconstruction, such as from the inner tracker. The HLT has multiple HLT paths, which apply reconstruction and selection algorithms in a specific order to identify physics objects. The HLT trigger further reduces the data rate from 100 kHz to 1 kHz.

To make full use of the available resources, the CMS experiment has devised two new strategies at HLT level called data parking and data scouting [71]. Data parking involves the collection and temporary storage of additional raw data, often with more relaxed trigger requirements. This data is set aside and not immediately processed, but instead is reconstructed later when data-taking runs have ended or when computational resources are otherwise free. On the other hand, data scouting uses a different strategy by recording only a limited subset of information from data that would normally exceed trigger rate limits and therefore not be stored. This lightweight data can then be used for simpler, preliminary analyses. If anything of interest is identified in this reduced dataset, the trigger configurations can be modified in future runs to capture more detailed data on similar events. Both approaches allow for more efficient use of storage and computational resources, ensuring that potentially valuable data is not lost.

3.2.8. Computing infrastructure

The data collected by the data acquisition system, even after reducing it by orders of magnitude using triggers, is still enormous and physically impossible for CERN to store it alone for offline processing. To overcome this problem, a distributed computing infrastructure has been set up, called Worldwide LHC Computing Grid (WLCG) [72], which is common to all experiments of the LHC. It is organized in a tiered system with four levels and is distributed across the globe. A schematic representation of the distributed computing infrastructure used by the CMS experiment as a part of the WLCG is shown in Figure 3.11.

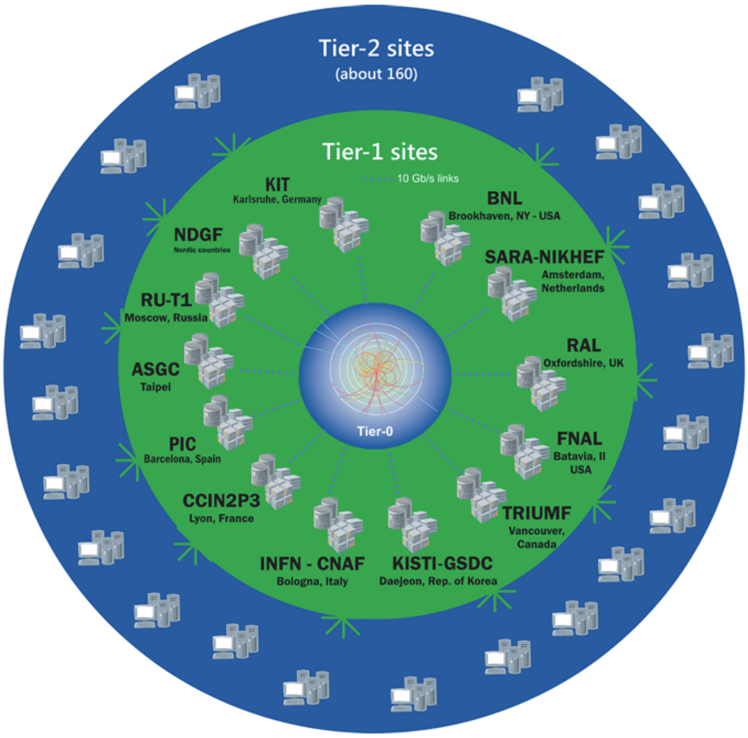


Figure 3.11: A schematic showing distributed computing infrastructure used by the CMS experiment, as a part of the Worldwide LHC Computing Grid (WLCG). The CERN computing site serves as the central Tier-0, while Tier-1 consists of 13 dedicated sites across various countries. Tier-2 includes nearly 160 sites worldwide, with the option to integrate additional local cloud resources as Tier-3 sites. This figure is taken from Reference [73].

Events passing one or more HLT trigger paths are stored in a raw event format on magnetic tapes at the Tier-0 data centre located at CERN. Here, the raw data is split into different datasets according to the HLT trigger paths. Also, prompt calibration and reconstruction takes place and a copy of datasets is distributed among 13 Tier-1 data centres. This distribution of raw and reconstructed data to Tier-1 sites is performed using a dedicated optical fibre network, called LHC Optical Private Network (LHCOPN) [74], which provides a transfer speed of 10 Gb/s.

The function of Tier-1 sites is to store and archive a copy of datasets received from the Tier-0 site, along with the full reconstruction and calibration of the data. The reconstructed data is then distributed from Tier-1 sites to 162 Tier-2 sites, which are generally located at universities and research centres. At Tier-2 sites, the data is available for the analysers, which can be processed for analysis specific purposes. In addition to the above mentioned tiers, additional local and cloud clusters can be added as Tier-3, typically suitable for a small group of users.

4. Object reconstruction and identification

In the previous chapter, the CMS detector and its various subdetector systems were introduced. The data analysed in this thesis was collected by the CMS detector during the Run 2 data-taking period. In order to analyse the recorded data, individual read-out signals from various subdetector systems are systematically combined to reconstruct essential physics objects such as leptons, jets, and neutrinos. The analysis discussed in the subsequent chapters relies on these reconstructed physics objects, which are defined in this chapter. The reconstruction process begins with the identification of tracks and interaction vertices, discussed in Section 4.1. Next, the Particle Flow (PF) algorithm, detailed in Section 4.2, combines information from all detector subsystems to identify final-state particles. Using the particles identified by the PF algorithm, jet clustering algorithms are applied to group sprays of hadrons into jets, as described in Section 4.3. Since this analysis utilizes boosted jets, advanced techniques for boosted object reconstruction, including jet grooming and jet substructure methods, are also detailed in this section. Section 4.4 focuses on the identification of jets originating from b quarks, a critical component for rejecting background processes with b quarks in the final state. A specialized algorithm used in this study for determining jet flavour is introduced here. Lastly, Section 4.5 addresses the reconstruction of missing transverse momentum, accounting for the energy carried by undetected particles like neutrinos.

4.1. Tracks and vertices

The electromagnetically charged particles leave hits in the pixel and strip trackers, when they transverse through the CMS detector. These hits can be combined to form trajectories called tracks. Due to the presence of the magnetic field of the solenoid, these tracks bend and the curvature of these tracks is used to measure the momentum and charge of particles. The track reconstruction uses a combinatorial track finding algorithm [75], which utilizes an extended version of Kálmán filters [76, 77], to iteratively reconstruct tracks. Each iteration begins with the hits in the innermost pixel trackers. Tracks with the largest p_{T} are reconstructed first. In each iteration, hits collection is updated, by discarding hits associated with the successfully reconstructed tracks. Each iteration consist of four steps. In the first step, a track seed is generated by combining neighbouring hits of a track candidate. The second step extrapolates tracks to include more tracking layers using Kálmán filters technique. The third step uses a track-fitting algorithm to fit the track and determine its properties. The uncertainties related to hit positions, energy loss in the detector material are taken into account during fit. In the final step, the quality of the reconstructed tracks is assessed. Tracks that are not associated to a minimum of eight hits and a certain χ^2 value for the fit are removed. A similar track reconstruction method is used to reconstruct muon tracks in muon detector systems.

The points of origin of primary interactions, called primary vertices, are reconstructed by extrapolating the reconstructed tracks to the centre of the detector using a deterministic annealing algorithm [78]. The vertex finding algorithm produces candidates of vertices by clustering tracks based on the z position of their extrapolated origin. The vertex candidates are then fitted using an adaptive vertex filter [79], to determine their position. The vertex with the largest $\sum_i p_{\mathrm{T},i}^2$, where $p_{\mathrm{T},i}$ is the transverse momentum of the i-th track of the reconstructed vertex, is labelled as the primary vertex. All other vertices are labelled as pileup vertices coming from pileup interactions (additional interactions in the same bunch crossing). In addition to the requirement of largest transverse momentum square, the primary vertex should also satisfy geometrical criteria such as: the location of the primary vertex should be within $|z| \leq 24$ cm, and a radius of $r \leq 2$ cm relative to the centre of the detector.

4.2. The Particle Flow algorithm

The CMS Collaboration uses the Particle Flow (PF) algorithm, based on the concept of global event reconstruction [80], to combine data from various subdetectors, aiming to achieve optimal identification of stable particles such as electrons, photons, muons, and charged or neutral hadrons. Figure 4.1 shows a transverse slice of the CMS detector representing the signature of various particles in the detector. In general, a particle gives rise to many PF elements, such as tracks in the tracker, clusters in the ECAL, or tracks in the muon systems. These PF elements are linked into PF blocks recursively, taking into account their compatibility with each other, using a link algorithm. For example, the link algorithm [80] matches tracks to calorimeter clusters when the extrapolated track path aligns with the energy cluster's position, or link bremsstrahlung photons that are tangent to the track extrapolated to ECAL. In each PF block, muons are reconstructed first, followed by combined electron and photon reconstruction. The electron reconstruction also takes into account the energy of the bremsstrahlung photons. After reconstruction of each object, their corresponding PF elements are removed from the PF block. In the end, all remaining PF elements are associated with charged or neutral hadrons.

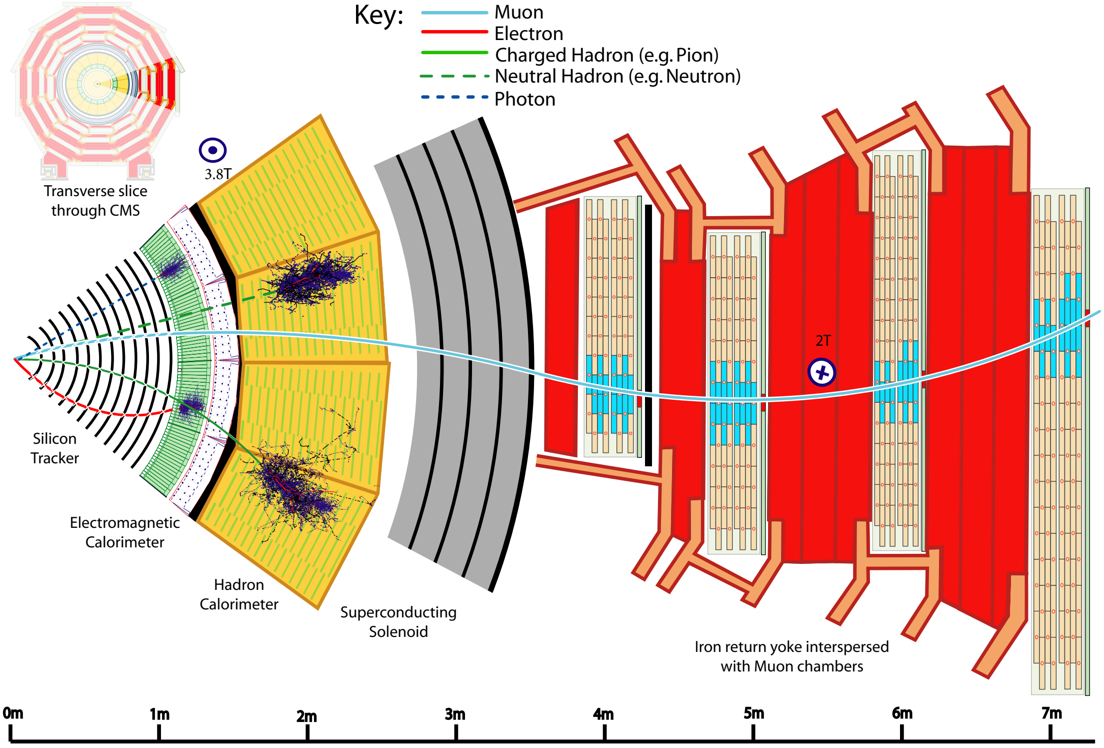

Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
Transverse slice through CMS
3.8T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
0m
1m
2m
3m
4m
5m
6m
7m

Figure 4.1: A sketch of various types of particle interactions in a transverse slice of the CMS detector, starting from the interaction vertex all the way up to the muon detectors. This figure is taken from Reference [80].

4.2.1. Muon reconstruction

Muons are reconstructed first by the PF algorithm, because they have relatively clean tracks compared to electrons. Three types of muons: standalone, tracker, and global, are reconstructed by the PF algorithm, depending on which part of the detector information is used to reconstruct them. Standalone muons are reconstructed using tracks exclusively from the muon system, while tracker muons are identified by taking tracks from the silicon tracker, extending them, and matching them to hits in the muon system. Global muons are formed by combining standalone muons with tracker muons. Most muons are classified as both tracker and global muons. The tracker muon reconstruction works more efficiently at lower momenta, where the silicon tracker offers higher precision. On the other hand, global muons are more effective at higher momenta, as they benefit from multiple hits within the muon detector system.

The PF muons used in the analysis presented in this thesis are required to pass additional requirements to distinguish muons originating from a hard scattering process (like those produced in proton-proton collisions) from cosmic muons or muons resulting from the decay of mesons. Different quality requirements are generally grouped under various IDs of muons provided by CMS Collaboration [81]. Two types of muon ID set: tight and loose, are utilized. The description of various requirements for the tight and loose muon IDs are listed in Table 4.1. In addition, a requirement on the relative muon isolation is applied to reject non-prompt muons coming from semi-leptonic hadron decays in jets. The relative isolation (I_μ) is defined as:

$$I_\mu = \frac{1}{p_{\mathrm{T},\mu}} \left[\sum_{\Delta R<0.4} p_{\mathrm{T,CH}} + \max\left(0, \sum_{\Delta R<0.4} p_{\mathrm{T,NH}} + \sum_{\Delta R<0.4} p_{\mathrm{T},\gamma} - \frac{1}{2} \sum_{\Delta R<0.4} p_{\mathrm{T,PU}} \right)\right] \tag{4.1}$$

It quantifies the isolation of a muon by calculating the transverse momentum of various types of particles (charged hadrons (CH), neutral hadrons (NH), photons (γ), pileup(PU)) in a cone of radius 0.4 around the muon and comparing it to the transverse momentum of the muon itself. The isolation requirement is applied together with the identification criteria. For tight ID, tight isolation is applied, which requires $I_\mu < 0.15$, while for loose ID, loose isolation is applied, which requires $I_\mu < 0.25$. Furthermore, muons are restricted to have $|\eta| \leq 2.4$ to ensure coverage of the full range of the muon system. Muons reconstructed outside this pseudorapidity range exhibit lower efficiencies and accuracies, and are therefore excluded from further analysis.

The efficiency of the final reconstructed muons consists of several factors, including track reconstruction, muon reconstruction and identification, relative isolation, and trigger efficiencies. The tag-and-probe method was used to study the efficiencies related to reconstruction, identification, and isolation [81]. In all cases, the efficiency was found to exceed 95% across the entire range of η and ϕ. The uncertainties are estimated to be at the level of 1% for ID and 0.5% for isolation [81].

4.2.2. Electron reconstruction

Electrons, being lighter than muons, lose a lot of energy through bremsstrahlung radiation. Therefore, they are relatively complicated to reconstruct. Like muons, they leave a track in the tracker system of the CMS detector. The tracker track of an electron is dressed with bremsstrahlung along its curvature. The excessive bremsstrahlung changes the curvature of the tracks, and only a few hits may be recorded in the tracker. A Gaussian sum filter (GSF) [82] is used to re-reconstruct the track candidates, which correctly takes the energy loss into account. In ECAL crystals, electrons produce electromagnetic showers by depositing bremsstrahlung photons. The deposits in individual crystals are clustered together in so-called superclusters (SC). Electron signatures, tracker tracks and the ECAL SC, can be combined in two ways: ECAL-driven approach, and tracker-driven approach. In the ECAL-driven approach, the PF algorithm connects the SC to track candi-

Table 4.1: Muon identification requirements for tight ID and loose ID provided by the CMS Collaboration [81].

Criteria	tight ID	loose ID
Global muon	Yes	–
PF muon	Yes	Yes
Global or tracker muon	–	Yes
χ^2/ndof of the global-muon track fit	< 10	–
No. of hits in muon chambers	> 0	–
No. of segments in muon stations	> 1	–
$\lvert d_{xy}\rvert$	< 2 mm	–
$\lvert d_z\rvert$	< 5 mm	–
No. of pixel hits	> 0	–
No. of tracker layer hits	> 5	–

dates to form an electron. This approach works well for high-p_{T} electrons. In the tracker-driven approach, the tracker tracks are matched to the ECAL SC using a calorimeter-unbiased seed algorithm. This approach works best for low-p_{T} electrons.

For the analysis presented in this thesis, several additional quality requirements are imposed on electron candidates reconstructed by the PF algorithm. Only those electrons which passes tight identification (ID) criteria, as defined by the CMS Collaboration [83], are kept for further analysis. In addition, loose identification criteria is also utilized to veto events with additional loose electrons. Separate sets of ID requirements are set for electrons in the barrel ($|\eta_{SC}| \leq 1.48$) and endcap ($|\eta_{SC}| \geq 1.48$) regions. Details of tight ID and loose ID for electrons are listed in Table 4.2. Similar to muon isolation, a relative isolation of electron (I_e) is defined as:

$$I_e = \frac{1}{p_{\mathrm{T},e}}\left[\sum_{\Delta R<0.3} p_{\mathrm{T,CH}} + \max\left(0, \sum_{\Delta R<0.3} p_{\mathrm{T,NH}} + \sum_{\Delta R<0.3} p_{\mathrm{T},\gamma} - \rho A_{\mathrm{eff}}\right)\right], \qquad (4.2)$$

which quantifies the relative energy content in the vicinity of an electron in a cone of radius 0.3. Like muons, transverse momentum of charged hadrons (CH), photons (γ), and neutral hadrons (NH) are used to calculate isolation, with an additional term ρA_{eff}, which subtract the contribution from pileup to neutral hadrons contribution. The parameter ρ is the average neutral hadron energy density and A_{eff} is the effective area of the electron [83]. Unlike for muons, the isolation criteria of electrons is a part of the identification criteria provided by the CMS Collaboration, as shown in Table 4.2, and is not applied separately. Furthermore, the analysis in this thesis only uses electron having $|\eta| < 2.5$ due to the limited tracker coverage.

Table 4.2: Electron identification criteria for tight and loose electron IDs in two pseudorapidity ranges provided by the CMS Collaboration [83].

Criteria ($|\eta_{SC}| \leq 1.48$)	tight ID	loose ID
$\sigma_{i\eta i\eta}$ (shower shape)	< 0.0104	< 0.0112
$|\Delta\eta(\text{SC, track})|$	< 0.00255	< 0.00377
$|\Delta\phi(\text{SC, track})|$	< 0.022	< 0.0884
Hadronic energy/EM energy	$< 0.026 + 1.15/E_{SC} + 0.0324\rho/E_{SC}$	$< 0.05 + 1.16/E_{SC} + 0.0324\rho/E_{SC}$
I_e	$< 0.0287 + 0.506/p_T$	$< 0.112 + 0.506/p_T$
$|E_{SC}^{-1} - p_{track}^{-1}|$	< 0.159	< 0.193
No. of missing inner hits	≤ 1	≤ 1
Pass conversion veto	Yes	Yes
Criteria ($|\eta_{SC}| > 1.48$)	tight ID	loose ID
$\sigma_{i\eta i\eta}$ (shower shape)	< 0.0353	< 0.0425
$|\Delta\eta(\text{SC, track})|$	< 0.00501	< 0.00674
$|\Delta\phi(\text{SC, track})|$	< 0.0236	< 0.169
Hadronic energy/EM energy	$< 0.0188 + 2.06/E_{SC} + 0.183\rho/E_{SC}$	$< 0.0441 + 2.54/E_{SC} + 0.183\rho/E_{SC}$
I_e	$< 0.0445 + 0.963/p_T$	$< 0.108 + 0.963/p_T$
$|E_{SC}^{-1} - p_{track}^{-1}|$	< 0.0197	< 0.111
No. missing inner hits	≤ 1	≤ 1
Pass conversion veto	Yes	Yes

4.2.3. Photon and hadron reconstruction

Together with electrons, isolated photons are also reconstructed by the PF algorithm. ECAL clusters that cannot be associated with tracks, which are not identified as bremsstrahlung and have a ratio between ECAL and HCAL energy deposits compatible with the expected photon shower, are reconstructed as isolated photons. The remaining ECAL and HCAL clusters without associated tracks are assumed to originate from neutral hadrons (such as neutral kaons or neutrons), or non-prompt photons from decays of neutral pions.

ECAL and HCAL clusters with associated tracks in the tracker are classified as charged hadrons, such as charged pions ($\pi^{\pm}$), charged kaons ($\mathrm{K}^{\pm}$), or protons. HCAL clusters without ECAL energy deposits but with associated tracks are also classified as charged hadrons. In the absence of tracks and for $|\eta| \leq 2.5$, clusters are assumed to correspond to neutral hadrons, as tracks can be reconstructed within this pseudorapidity range. Outside this range ($|\eta| > 2.5$), the lack of reliable track reconstruction means that the PF algorithm cannot distinguish between charged and neutral hadrons, and HCAL clusters without tracks are classified as neutral hadrons.

4.3. Jet reconstruction and boosted object techniques

When two protons collide at very high energies, their constituent quarks and gluons (collectively called partons) can interact in a "hard scatter" process. The energy of this interaction is very large, and the partons are knocked out of the proton with extremely high energy. Quarks and gluons cannot exist freely due to a property called colour confinement — a fundamental principle in quantum chromodynamics (QCD), described in Section 2.1.3.2. The quark or gluon from the collision undergoes a process called hadronization or fragmentation, where it creates a stream of particles by forming colour-neutral hadrons (composite particles like protons, pions, kaons, etc.). The result of hadronization is a spray of hadrons travelling in the same general direction as the original high-energy quark or gluon. This collimated spray of particles is called a jet. To reconstruct a jet, jet algorithms are employed after the PF particle identification and reconstruction to cluster particles together in jet objects. Non-isolated leptons, from the leptonic decays of hadrons, can also be a part of a jet and are also considered in the jet clustering algorithm, described in Section 4.3.1. The properties of the jet are used to infer the properties of the initial hard scattering particle, such as a Higgs boson, or a top quark, or a W boson etc.

When a heavy particle, such as a Higgs boson or a W boson, is produced as a result of a high energy collision, their decay products often get sufficient Lorentz boost. As a result, the jets produced are highly collimated, and are detected as a single large radius jet in the detector rather than multiple separate jets. Such jet objects are called boosted jets and presents a challenge in distinguishing them from ordinary jets produced by lighter particles like quarks and gluons. The analysis presented in this thesis is subject to this challenge, where one of the final state W boson decays into a boosted jet. In order to reconstruct and identify boosted jets, special techniques such as jet grooming and jet substructure are required. They are described in Section 4.3.2. The key idea of these techniques is to analyse the substructure of boosted jets to discriminate between signal and background jets.

4.3.1. Jet clustering

The particles reconstructed by the PF algorithm, as described in Section 4.2, serve as inputs to jet clustering algorithms. There are two major types of jet clustering algorithms: sequential recombination algorithms and cone algorithms. Sequential recombination algorithms: anti-k_T, k_T, Cambridge-Aachen, are most commonly used because they are both infrared and collinear safe. These algorithms iteratively recombine particles by minimizing a distance metric. The key idea behind is that jets are the product of successive parton branchings, and by inverting the process i.e. successively recombining two particles into one, one can mimic the QCD dynamics of the parton shower.

The sequential recombination algorithm used to cluster jets in the analysis presented in this thesis is the anti-k_T algorithm [84]. The choice of this algorithm is standard for all LHC experiments. The primary advantage is that it first clusters the hard particles of a jet and then gradually incorporates soft particles within a distance R from the jet axis, making the algorithm insensitive to soft radiation.

The jets are clustered as follows:

1. Take PF particles as the initial list of jet constituents.

2. From the list, two distance metrics are defined: an inter-particle distance (d_{ij}) and a beam distance ($d_{i\mathrm{B}}$). The definitions are:

$$d_{ij} = \min(p_{\mathrm{T},i}^{2p}, p_{\mathrm{T},j}^{2p})\frac{\Delta_{ij}^2}{R^2}, \tag{4.3}$$

$$d_{i\mathrm{B}} = p_{\mathrm{T},i}^{2p}, \tag{4.4}$$

   where,

$$\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2. \tag{4.5}$$

   Equations 4.3 and 4.4 with $p = -1$ refer to the anti-k_T algorithm. It can also take a value of 1 or 0, which will refer to other types of sequential algorithms, k_T algorithm [85] and Cambridge-Aachen algorithm [86], respectively. In the definitions above, R is the radius parameter, whose value is generally set to 0.4 for resolved jets and 0.8 for boosted jets and parameter p is the characterising feature. Jets reconstructed using anti-k_T algorithm with radius 0.4 and 0.8

are called AK4 jets and AK8 jets, respectively. $p_{\mathrm{T},i}$, y_i, ϕ_i are the transverse momentum, rapidity, and azimuthal angle of the particle i, respectively.

3. Identify iteratively the smallest distance among all the d_{ij} and $d_{i\mathrm{B}}$. If the smallest distance is d_{ij}, two particles i and j are combined to form a pseudo-particle k. The particles i and j are removed from the list and the new pseudo-particle k is added. All new distance measures d_{kl} and $d_{k\mathrm{B}}$ are calculated, and the process is iterated. If the smallest distance happens to be $d_{k\mathrm{B}}$, the particle (or pseudo-particle) k is removed from the list and declared as jet. This process is repeated several times until all PF particles are clustered into jets. An example event clustered with the anti-k_T algorithm is shown in Figure 4.2.

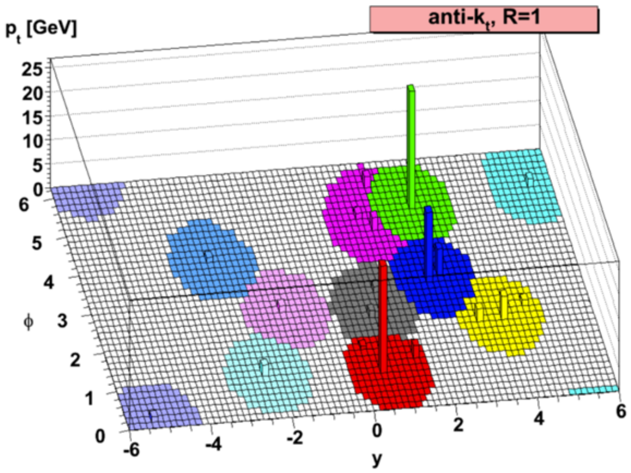


Figure 4.2: An example of simulated events clustered into jets with radius equals to 1 using anti-k_T algorithm. Different colour patches represents different jets clustered by the algorithm. This picture is taken from Reference [84].

The anti-k_T algorithm is implemented using the FastJet package [87]. The analysis presented in this thesis uses both AK4 jets and AK8 jets. The two VBS jets in the final state of the signal process are reconstructed as AK4 jets, while the decay products of a W boson decaying hadronically are reconstructed as a single AK8 jet (boosted jet). In order to optimize the reconstruction and identification of boosted jets from heavy particle decays, further techniques are utilized. They are described in the Section 4.3.2.

An important factor to take into account while jet clustering process is the effect of pileup interactions. Presence of pileup particles degrades the performance of clustering algorithms. Techniques like charge hadron subtraction (CHS) [80] and pileup per particle identification (PUPPI) [88–90] are generally utilized to remove or mitigate the effect of pileup. In the CHS algorithm, charged hadrons are matched to the pileup vertices using the tracking information. All charged hadrons originating from pileup vertices are removed before the application of jet clustering algorithm. The PUPPI algorithm is slightly more sophisticated, which assigns a weight to each particle depending on the probability that the particle is originated from a pileup vertex. These weights are applied to scale the four-momentum of particles, effectively reducing the effect of pileup before clustering jets. In the analysis presented in this thesis, both the CHS algorithm and the PUPPI algorithm are utilized to account for pileup before clustering AK4 jets and AK8 jets, respectively. Therefore, the jets are called AK4 CHS and AK8 PUPPI jets in the rest of this thesis.

Both AK4 CHS and AK8 PUPPI jets are required to pass additional quality criteria for identification, mainly to remove jets originating from calorimetric noise. Requirements are set on the number of constituents clustered into a jet, their neutral and charged energy fractions etc. A detailed list of CMS provided tight identification criteria used in this thesis for AK4 CHS jets and AK8 PUPPI jets are listed in Table 4.3 and 4.4, respectively. Furthermore, to reduce the number of prompt leptons misidentified as jets, additional jet-lepton cleaning procedure is used to remove jets in the vicinity of electrons or muons. The threshold for AK4 jets is $\Delta R(l, \text{jet}) \leq 0.4$ and for AK8 jets is $\Delta R(l, \text{jet}) \leq 0.8$.

4.3.2. Boosted object techniques

The centre-of-mass energy at which the LHC operates is approximately 50 times greater than the electroweak scale of roughly 246 GeV. At this high energy, heavy particles such as W bosons, Z bosons, and Higgs bosons — each with significant hadronic decay branching fractions — typically possess momenta that are much larger than their masses. As a result, their hadronic decay products experience substantial Lorentz boosts, compressing them into a smaller angular region. The relationship between angular distance between the decay products (ΔR), transverse momentum (p_{T}) and mass (M) is roughly described by the following equation:

$$\Delta \mathrm{R} \sim \frac{2\mathrm{M}}{p_{\mathrm{T}}}. \tag{4.6}$$

For a W boson with a mass of 80 GeV and a transverse momentum greater than 200 GeV, the hadronic decay products become so closely packed that they cannot be resolved individually. Instead, they are combined into a single large-radius jet with a radius of 0.8, called boosted jets. Figure 4.3 illustrates the visualization of resolved and boosted decays of a heavy particle. Boosted jets, because of large radius, are more prone to soft and wide angle radiations from pileup and underlying events, therefore they require special treatment for reconstruction, like jet grooming and jet substructure, in order to increase the efficiency of selecting signal-like boosted jets.

Jet grooming

Table 4.3: Tight identification requirements for AK4 CHS jets for different pseudorapidity ranges as provided by CMS collaboration. Values are taken from Reference [91].

2016	$\lvert\eta\rvert \leq 2.4$	$2.4 < \lvert\eta\rvert \leq 2.7$	$2.7 < \lvert\eta\rvert \leq 3.0$	$3.0 < \lvert\eta\rvert \leq 5.0$
Neutral hadron fraction	< 0.90	< 0.90	< 0.90	> 0.2
Neutral EM fraction	< 0.90	< 0.99	> 0 and < 0.99	< 0.9
Number of constituents	> 1	-	-	-
Charged hadron fraction	> 0	-	-	-
Charged multiplicity	> 0	-	-	-
Number of neutral particles	-	-	> 1	> 10
2017-18	$\lvert\eta\rvert \leq 2.4$	$2.4 < \lvert\eta\rvert \leq 2.7$	$2.7 < \lvert\eta\rvert \leq 3.0$	$3.0 < \lvert\eta\rvert \leq 5.0$
Neutral hadron fraction	< 0.90	< 0.90	-	> 0.2
Neutral EM fraction	< 0.90	< 0.99	> 0.01 and < 0.99	< 0.9
Number of constituents	> 1	-	-	-
Charged hadron fraction	> 0	-	-	-
Charged multiplicity	> 0	> 0	-	-
Number of neutral particles	-	-	> 1	> 10

Jet grooming is a widely-used method to mitigate the impact of soft background from additional proton-proton interactions other than the primary hard scatter one (underlying event) and pileup. Typically, the grooming method involves removing soft particles far from the jet axis. Such particles are likely to come from soft contamination rather than the QCD radiation inside the jet. Boosted jets used in the analysis presented in this thesis are groomed with a special method called soft drop [92]. The soft drop algorithm recursively removes soft wide-angle radiations from a jet, thereby highlighting the hard core of the jet associated with the decay of high-momentum particles. It proceeds by reclustering anti-k_{T} jets using Cambridge-Aachen algorithm, to form an angular-ordered pairwise clustering tree. The steps performed during this reclustering are then iterated in reverse order to declustered the jet into two subjets by checking at each step the following soft drop condition:

Table 4.4: Tight identification requirements for AK8 PUPPI jets for different pseudorapidity ranges as provided by the CMS Collaboration. Values are taken from Reference [91].

2016	$\lvert\eta\rvert \leq 2.4$	$2.4 < \lvert\eta\rvert \leq 2.7$	$2.7 < \lvert\eta\rvert \leq 3.0$	$3.0 < \lvert\eta\rvert \leq 5.0$
Neutral hadron fraction	< 0.90	< 0.98	-	-
Neutral EM fraction	< 0.90	< 0.99	-	< 0.9
Number of constituents	> 1	-	-	-
Charged hadron fraction	> 0	-	-	-
Charged multiplicity	> 0	-	-	-
Number of neutral particles	-	-	$>= 1$	> 2
2017-18	$\lvert\eta\rvert \leq 2.4$	$2.4 < \lvert\eta\rvert \leq 2.7$	$2.7 < \lvert\eta\rvert \leq 3.0$	$3.0 < \lvert\eta\rvert \leq 5.0$
Neutral hadron fraction	< 0.90	< 0.99	< 0.9999	-
Neutral EM fraction	< 0.90	< 0.99	-	< 0.9
Number of constituents	> 1	-	-	-
Charged hadron fraction	> 0	-	-	-
Charged multiplicity	> 0	-	-	-
Number of neutral particles	-	-	-	> 2

$$\frac{\min(p_{\mathrm{T},i}, p_{\mathrm{T},j})}{p_{\mathrm{T},i} + p_{\mathrm{T},j}} > z_{\mathrm{cut}} \left(\frac{\Delta R_{ij}}{R}\right)^{\beta}, \tag{4.7}$$

where R is the radius of the jet, and z_{cut} and β are parameters whose values are set to 0.1 and 0, respectively, in the analysis presented in this thesis. If the soft drop condition is met, the jet is labelled as a soft drop jet and the iteration is stop. Otherwise, the iteration continues by dropping the softer subjet and keeping the subjet with larger p_{T}. This jet grooming procedure substantially improves the tagging efficiency of W boson initiated jets by improving jet mass resolution. Removal of soft particles from boosted jets reshapes the mass distribution, and helps differentiate it from QCD initiated boosted jets.

Jet substructure

Jet substructure is a method to study the internal kinematic properties of a boosted

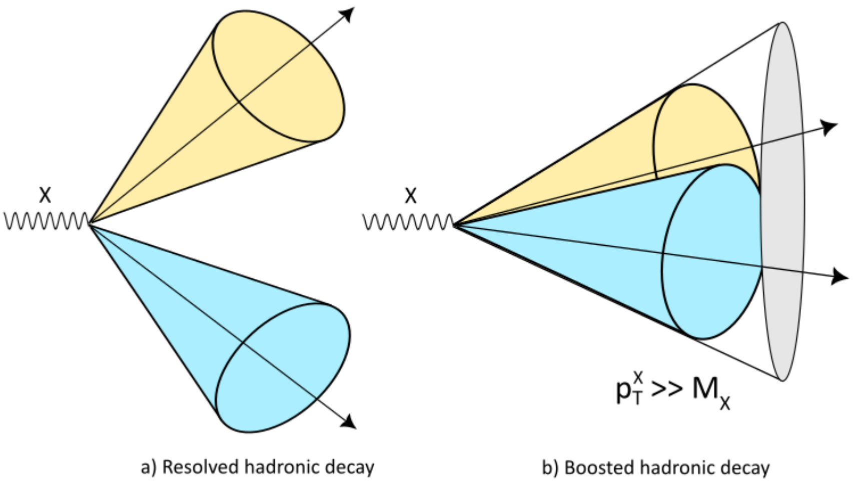


Figure 4.3: An illustration of angular separation between the hadronic decay products of a heavy particle X. Figure a) represents the case with large angular separation, where the two jets can be resolved separately. Figure b) represents the case when the transverse momentum of particle X is much larger than its mass, resulting in smaller separation between the decay products. In this case, the two jets are reconstructed as a single large radius jet.

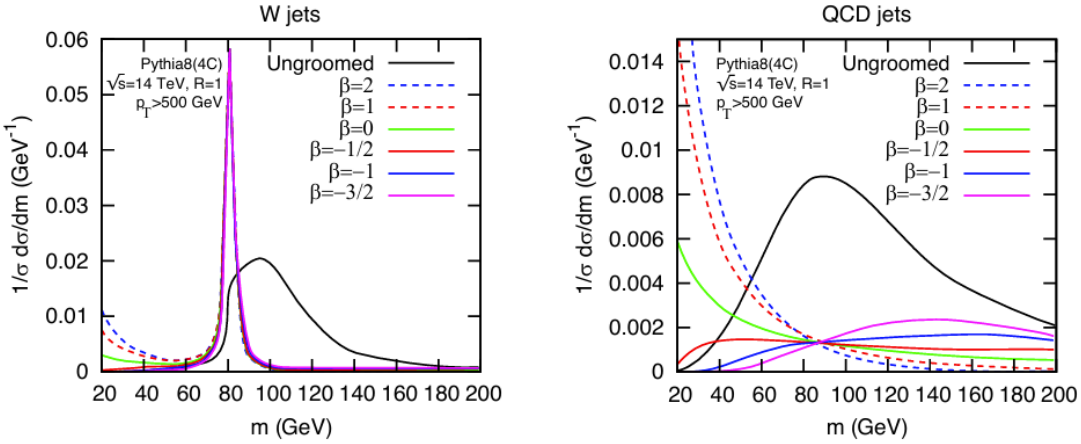


Figure 4.4: Impact of the soft drop algorithm on mass distributions of boosted W boson initiated jets (left) and boosted QCD initiated jets (right). Removal of soft particles corrects the jet mass for both QCD and W boson initiated jets, making it a distinctive feature. This figure is taken from Reference [92] with slight modifications in the legends.

jet in order to distinguish whether it is more likely to be a signal or a background jet. Boosted objects generate jets that can no longer be resolved, but their substructure encodes the information of the parent particle such as the jet mass, the number of decay axis, energy correlations etc. Nowadays, there are more advanced methods powered by machine learning techniques to decode substructure information for signal jets identification. Popular algorithms are Particle Net [93] and Particle Transformer [94], which directly learn from particle-level data of a jet rather than pre-processed variables, to exploit substructure.

In this thesis, we have utilized a traditional substructure variable N-subjettiness [95]. As the name suggests, this variable aims to discriminate jets according to the number N of subjets they are made of. It is defined as:

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min \left(\Delta R_{1,k}, \Delta R_{2,k}, \ldots, \Delta R_{N,k} \right), \tag{4.8}$$

where, the summation is over all particles in a jet. $\Delta R_{\mathrm{i},k}$ is the angular distance between the particle and the subjet candidate, d_0 is the normalization factor given by $d_0 = \sum_k p_{\mathrm{T},k} R_0$ with R_0 as the original jet radius. The N-subjettiness variable quantifies how well a jet's particle distribution aligns with the N-prong hypothesis, where lower values indicate a stronger compatibility with having N subjets. For W, Z, or Higgs bosons, the boosted jet typically have a two-prong structure because these bosons decay into two quarks. In contrast, top quarks produce a three-prong structure, as they decay into a bottom quark and a W boson, which subsequently decays into two quarks. Therefore, boosted jets originating from W, Z, or Higgs bosons, should have lower values of τ_2 and higher values of τ_1. And for top quarks initiated jets, we expect τ_1, τ_2 to be large and τ_3 to be small.

It is typically more beneficial to use ratios of τ_N with different values of N [95]. For the analysis presented in this thesis, the ratio of τ_2, and τ_1,

$$\tau_{21} = \frac{\tau_2}{\tau_1}, \tag{4.9}$$

is used to discriminate between boosted jets originating for W or Z bosons from those originating from light quarks (QCD jets), which typically have one-prong structure and larger values of τ_{21}. Figure 4.5 shows the distribution of τ_{21} variable for W jets and QCD jets. Application of a selection cut on this variable provides separation between the two types.

To sum up, boosted jets with radius 0.8 are corrected for pileup effects using PUPPI algorithm and soft drop grooming is applied to remove soft-wide angle radiations from jets. As a result, jet mass and τ_{21} observable of the corrected and groomed jets form a strong discriminant between signal and background jets.

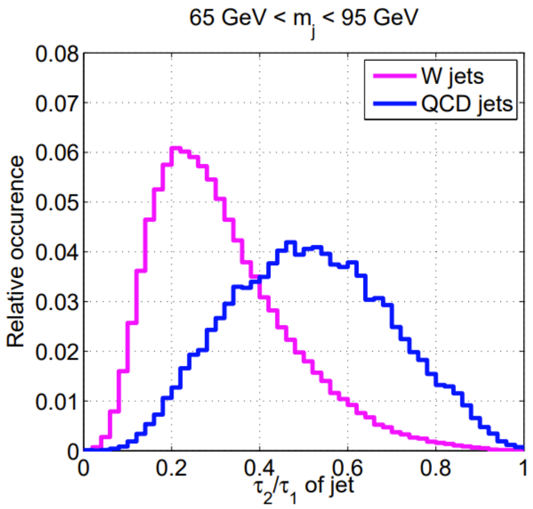


Figure 4.5: Distribution of τ_{21} substructure variable for boosted W and QCD jets, demonstrating the distinguishing power of the variable. This figure is taken from Reference [95].

4.4. Identification of b quark initiated jets

Identification of b quark initiated jets (b jets) is a critical task to suppress backgrounds that contain bottom quarks. For the analysis presented in this thesis, top-antitop pair production is one of the major background which can be significantly suppressed by identifying the presence of b jets. This procedure of identification is called b tagging. The b quarks produced in collisions hadronise to form b hadrons, like the B mesons. They have longer lifetimes, which means that they do not decay immediately but travel a certain distance before decay. As a result, they form a secondary vertex inside the jet which is displaced from the primary vertex of the interaction. Also, the tracks originating from the secondary vertex have large impact parameter values. These distinctive features indicate the presence of a bottom quark initiated jet. The secondary vertices are reconstructed using the same vertex finding algorithms as introduced in Section 4.1. The information of these vertices and tracks associated with them are used to designed b jet identification algorithm. In addition, modern b tagging algorithms make use of full kinematics of jets and their PF constituents to identify the flavour.

The b tagging algorithm used in this thesis is DeepJet [96], which is a neural network based architecture model. The algorithm takes 25 leading neutral and charged PF candidates, and four leading secondary vertices as inputs. The output of this algorithm provides the probability of a jet to be a b quark initiated jet. Different working points (WPs) are defined for the DeepJet algorithm, based on its performance on a set of simulated events. WPs are defined at constant light-jet misidentification rates on that data set. Two WPs defined as 1% (medium) and 10% (loose) light-jet misidentification rates are utilized to select and reject b tagged jets in this thesis. The term light jet refers to jets originating from u, d, s quarks, as well as gluons. The DeepJet discriminator output cut corresponding to these WPs together with their efficiencies are summarized in Table 4.5.

Table 4.5: Summary of DeepJet working points (WPs) and their corresponding efficiencies used in this thesis. Values are taken from Reference [97].

Data-taking era	medium WP	$\epsilon_{\text{medium WP}}$ (%)	loose WP	$\epsilon_{\text{loose WP}}$(%)
2016 preVFP	0.2598	73.3	0.0508	87.3
2016 postVFP	0.2489	71.4	0.0480	86.3
2017	0.3040	79.1	0.0532	91.0
2018	0.2783	80.7	0.0490	91.5

4.5. Missing transverse momentum

The CMS detector measures nearly all stable particles produced in collisions, except neutrinos in the SM or hypothetical dark matter candidate particles in BSM scenarios, which escapes the detector undetected. The presence of such particles can be inferred as missing transverse momentum ($p_{\mathrm{T}}^{\mathrm{miss}}$), calculated as the negative of the vectorial sum of the transverse momenta of all visible particles. The mathematical expression is given as:

$$\vec{p}_{\mathrm{T}}^{\;\mathrm{miss}} = - \sum_{\text{vis. particles}} \vec{p}_{\mathrm{T},i} \; . \tag{4.10}$$

The fact that the initial transverse momentum in collision is either zero or negligible in the transverse $\eta - \phi$ plane is utilized here, which means that the sum of transverse momentum of all final state particles should be zero. In the longitudinal plane, this argument is not valid, as the longitudinal momentum fraction of partons is unknown and hence the sum of initial longitudinal momentum is also unknown.

The calculation of $p_{\mathrm{T}}{}^{\mathrm{miss}}$ is important for the analysis presented in this thesis, as the final state of semi-leptonic WW VBS contains one neutrino from leptonic decay of a W boson. Full reconstruction of the signal process requires reconstructing $p_{\mathrm{T}}{}^{\mathrm{miss}}$ to account for undetected neutrino.

5. Event simulation

The collision data collected by the CMS experiment consists of final-state particles produced during hard scattering events, which are reconstructed using detector information. The exact nature of a single hard scattering event is not directly observable; instead, only the stable particles detected by the experiment contribute to the available information. To understand the underlying processes of hard scattering events, proton-proton collisions are simulated for various physical processes. These simulations aim to reproduce the number of events and kinematic distributions observed in real data. To enable a direct comparison, the simulated events undergo the same reconstruction procedures as the data collected by the CMS detector. The simulation process begins with modelling the hard scattering process, as described in Section 5.1. This is followed by interfacing with parton distribution functions (PDFs), which are explained in Section 5.2. The partons produced in the collision then undergo parton showering and hadronization, which are critical components of the simulation. These steps are detailed in Sections 5.3 and 5.4, respectively. To accurately reflect real data-taking conditions, additional effects such as the underlying event and pileup are included in the simulation. These are described in Sections 5.5 and 5.6. Finally, the simulated events are processed to emulate the CMS detector response, as outlined in Section 5.7. A summary of the generator tools used to produce the simulations relevant to this thesis is provided in Section 5.8.

5.1. Hard scattering

The hard scattering refers to a partonic subprocess, where partons from protons collide and interact to produce new particles. It is typically characterized by the event with the highest momentum transfer between the initial and final states. The probability of incoming partons colliding to produce a specific final state is described by the cross section ($\hat{\sigma}$) of that process. The $\hat{\sigma}$ is calculated by considering all relevant Feynman diagrams at a given order in the coupling strength of the underlying theory. The Feynman rules are applied to compute the matrix element for the process. Finally, the matrix element is integrated over the phase space to obtain the cross section for the desired final state. Mathematically, for a 2 to N particle scattering, the partonic cross section is given by:

$$\hat{\sigma}_{12\to\mathrm{N}} = \frac{1}{4\hat{s}} \int |\overline{\mathcal{M}}|^2 \, (2\pi)^4 \delta^4 \left(p_1 + p_2 - \sum_{i=1}^{\mathrm{N}} k_i \right) \prod_{i=1}^{\mathrm{N}} \frac{d^3 k_i}{(2\pi)^3 2E_i} \tag{5.1}$$

where, $\hat{s}$ is the squared centre-of-mass energy of the initial particle, $|\overline{\mathcal{M}}|^2$ is the squared and spin-averaged matrix element, which encodes the probability amplitude for the scattering process, δ^4 is a four-dimensional delta function ensuring conservation of energy and momentum during the scattering process, and E_i denotes the energy of the ith particle.

The partonic cross section of a process can be parametrized as a perturbative series of QCD coupling strength α_s, whose value is small at high energies. The series is given by:

$$\hat{\sigma} = \sigma_{\mathrm{LO}} \cdot \left(1 + \alpha_s \sigma_{\mathrm{NLO}} + \alpha_s^2 \sigma_{\mathrm{NNLO}} + \dots \right). \tag{5.2}$$

The first term, known as the leading-order (LO) cross section, is calculated using tree-level Feynman diagrams, which involve the minimum number of strong interaction vertices. The second term, referred to as the next-to-leading order (NLO) correction, is derived from one-loop Feynman diagrams. These diagrams include effects such as the radiation and reabsorption of gluons, or the real emission of an additional particle in the initial or final state. The NLO contribution is suppressed by a factor of α_s. The third term corresponds to the next-to-next-to-leading order (NNLO) correction. This includes contributions from two-loop Feynman diagrams, two real emissions, or a combination of one real emission and one-loop diagrams. NNLO corrections are further suppressed by a factor of α_s^2, reflecting the progressively smaller impact of higher-order contributions. For the exact calculation of a cross section, all possible Feynman diagrams need to be taken into account during the matrix element calculation. In principle, one can make an infinite number of Feynman diagrams for a given process by adding more and more radiations. Hence, incorporating all diagrams is not possible. Therefore, we truncate the perturbative series at some order of correction to calculate the cross section of a process.

5.2. Parton distribution functions

At the LHC, we collide protons, not partons, as individual partons do not exist in nature. The four momentum of a parton is an unknown fraction of the total proton momentum and indeterministic to predict. We can however assess the probability density of finding a quark or a gluon with a given momentum fraction x of a proton. These probability density functions are called parton distribution functions (PDFs), which describe the probability density of valence quarks, sea quarks and gluons present in a proton.

The PDFs can not be obtained from first principles but are evaluated from fits to data of various deep inelastic scattering experiments, and some SM processes. The PDFs are obtained at some energy scale, called factorization scale, which are then evolved using Dokshitzer-Gribov Lipatov-Altarelli-Parisi (DGLAP) [98–101] evolution equations at the required energy scale of the experiment. The PDF sets used in the generation of simulated events used in this thesis uses a neural network based formalism called NNPDF3.1 set [102, 103]. Figure 5.1 shows the NNLO PDF sets derived for various quarks and gluons at two different energy scales, 10 GeV^2 and 10^{4}GeV^2. These PDFs sets are used in all simulations used in this thesis.

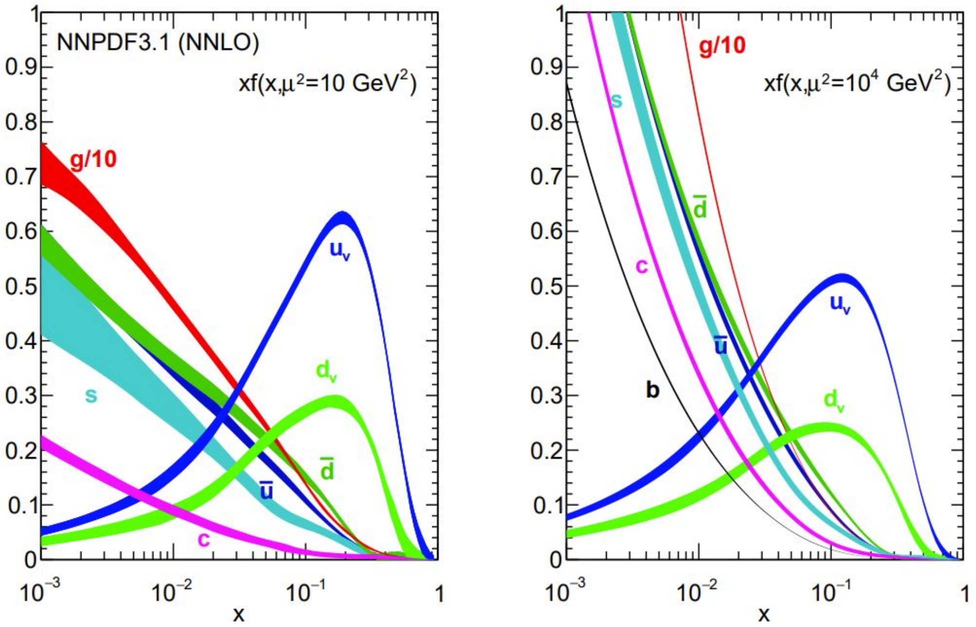


Figure 5.1: The NNLO PDF set (NNPDF3.1) evaluated at energy $\mu^2 = 10$ GeV^2 (left) and $\mu^2 = 10^4$GeV^2 (right). This picture is taken from Reference [103].

$$\sigma(pp \to X) = \sum_{a,b} \int dx_1\, dx_2\, p_a(x_1, \mu_{\mathrm{F}}^2)\, p_b(x_2, \mu_{\mathrm{F}}^2)\, \hat{\sigma}_X(x_1, x_2, \mu_{\mathrm{F}}^2, \mu_{\mathrm{R}}^2). \tag{5.3}$$

Here, p_a and p_b are the PDFs for the partons a and b, respectively. The partonic cross section (hard scattering cross section) $\hat{\sigma}_X$ is a function of two energy scales, the factorization scale μ_{F}, at which the PDFs are evaluated and the renormalization scale μ_{R} that describes the scale dependence of the strong coupling constant. For event generation, along with the total cross section for a particular process calculated from theory, these two energy scales μ_{R} and μ_{F} are explicitly chosen, as they cannot be physically measured.

5.3. Parton showers

Initial- and final-state partons involved in hard scattering can emit gluons, which typically carry lower energy than the originating partons. These emissions are referred to as initial-state radiation (ISR) and final-state radiation (FSR), respectively. The emitted gluons can themselves radiate additional gluons or quark-antiquark pairs, continuing the process and leading to a cascade of partons, each progressively lower in energy. This chain reaction persists until the energy of the partons drops to approximately 1 GeV, at which point non-perturbative effects dominate, and the partons combine to form colour-neutral hadrons.

The evolution of this parton shower can be described using different approaches depending on the energy scale. At high energies, a matrix element approach can effectively describe the dynamics of the process. However, as the energy scale decreases and the strong coupling constant becomes large, perturbative techniques lose their validity. In this non-perturbative regime, Sudakov form factors [104] are employed to simulate the development of the shower.

Importantly, the parton shower calculation can be factorized from the matrix element evaluation, allowing for independent computation of each step. However, care must be taken to avoid double counting when combining the shower evolution with higher-order matrix element evaluations. For this purpose, merging and matching techniques, such as the MLM and FxFx [105, 106] algorithms, are employed. These algorithms ensure a consistent treatment of parton emissions, properly accounting for the overlap between matrix element and parton shower descriptions of radiation, thereby preserving the accuracy and reliability of the simulation.

5.4. Hadronization

As the energy of partons decreases during shower evolution, the partons begin to cluster into colour-neutral hadrons. The resulting cascade of hadrons, usually collimated within a cone around the direction of the originating parton, is termed a jet. To model this hadronization process, phenomenological approaches are employed. The most widely used hadronization models are the Lund string model [107, 108] and the cluster model [109].

The Lund string model is based on the assumption that the potential between two colour-charged objects increases linearly with their separation, leading to the formation of a "string" connecting the partons. When the energy of the string exceeds a certain threshold, it breaks, resulting in the creation of a quark-antiquark pair. These newly created partons combine with the existing ones to produce colour-neutral hadrons. This string-breaking process continues until all colour-charged partons in the system are neutralized, resulting in the formation of a variety of hadrons.

The cluster model, on the other hand, assumes that gluons emitted during the parton shower inevitably decay into quark-antiquark pairs. These quarks and antiquarks, which are colour-connected due to the shower evolution, combine into colourless clusters with a predictable mass distribution. These clusters subsequently decay into hadrons, forming the final state.

In both models, the hadrons produced include unstable particles, which decay further into stable hadrons that interact with the detector. To achieve consistency with observed data, the models include a set of tunable parameters that can be adjusted to reproduce the hadronization patterns measured in experiments, such as those recorded by the CMS detector.

5.5. Underlying event

A key challenge of conducting experiments at hadron colliders arises from the composite nature of hadrons, which are made up of quarks and gluons. As a result, in addition to the primary hard scattering interaction, secondary interactions occur between other partons within the colliding hadrons. These secondary interactions contribute to what is known as the underlying event (UE). The underlying event significantly increases the number of particles produced in the final state, as these secondary interactions generate additional partons that undergo hadronization alongside the particles from the hard scattering. The simulation of the underlying event is achieved using phenomenological models, which incorporate various processes such as multiparton interactions, beam remnants, and soft parton-parton scatterings. These models include a range of tunable parameters that can be optimized to reproduce the features observed in experimental data. Different parameters are set to specific values during event generations, to include the effect of UE, summarized in event tunes. The event tune used in the simulations used in this thesis is the CP5 tune [110].

5.6. Pileup

Another significant factor to consider in hadron collider experiments is the contribution of additional interactions occurring within the same bunch crossing, referred to as pileup. Pileup introduces a large number of extraneous particles into the event, complicating the reconstruction of the hard scattering process and the identification of the physics signal.

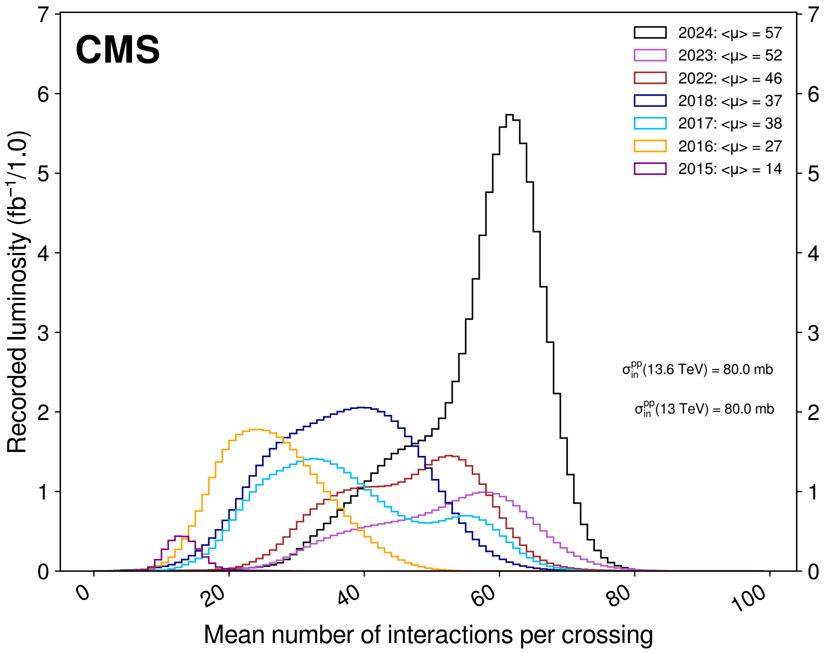


Figure 5.2: Shown are the pileup interactions per bunch crossing for Run 2 (2016-18) and ongoing Run 3 (2022-24). The average interaction values and the minimum bias cross sections are also shown. This figure is taken from Reference [45].

To incorporate pileup effects into simulations, a pileup profile is assumed prior to simulation. This profile is typically modelled using a Poisson distribution, with the mean number of expected pileup interactions per bunch crossing ($<\mu>$) determined by the beam conditions and the instantaneous luminosity of the collider. Figure 5.2 shows the mean number of pileup interactions recorded in proton-proton collision for Run 2 and ongoing Run 3.

5.7. Detector simulation

Once all particles have been simulated, the final step is to simulate the interaction of these particles with the detector material — similar to the interaction of real particles produced during collisions, as described in Chapter 4. This requires simulation of the CMS detector. The GEANT4 framework [111, 112] is used for this purpose, which provides a highly detailed simulation of each subdetector's response and the associated electronic readout, allowing for an accurate representation of the CMS experimental environment. For certain studies, particularly those exploring alternative detector configurations or future upgrades like the High-Luminosity LHC, a faster, lightweight approach can be used. The DELPHES framework [113] offers a parameterized detector simulation, where reconstructed quantities are modified using scaling and smearing factors to approximate detector effects without the computational burden of a full simulation.

For all simulations used in this thesis, GEANT4 has been utilized to simulate the full detector response.

5.8. Generator tools

To carry out the complex simulation tasks described above, a variety of Monte Carlo event generators are employed. Some of these generators are versatile and capable of performing multiple functions independently. However, most simulation workflows rely on a combination of specialized generators, each handling different aspects of the process.

Typically, the hard scattering process, including the matrix element calculation, is performed at next-to-leading order precision using dedicated frameworks. The results from these calculations are then interfaced with other tools that handle subsequent stages, such as parton showering, hadronization, and modelling of the underlying event.

To ensure smooth integration between different generators, a standardized protocol known as the Les Houches Accord [114] has been established. This agreement provides guidelines for interfacing and defines a universal format for sharing event data, referred to as the Les Houches Event (LHE) format. The LHE format streamlines the exchange of information between generators, enabling efficient and consistent simulation workflows across different tools and frameworks.

A summary of the event generators used to produce simulations used in this thesis is provided below:

- **MADGRAPH5_AMC@NLO:** [115] is one of the most commonly used event generators to compute matrix elements, at LO as well as NLO precision. It is a combination of MADGRAPH5 [116] and AMC@NLO [117] event generators. The VBS processes, namely same-sign WW, opposite-sign WW, and WZ VBS, are simulated using MADGRAPH5_AMC@NLO v2.6.5 at leading order. The intermediate vector boson pair is produced using the narrow width approximation. Spin correlations in the decay of intermediate vector boson pair is taken into account by interfacing with a module called MADSPIN [118].

- **POWHEG:** [119, 120] is another commonly used event generator, typically used to simulate events at NLO precision in QCD. In this thesis, POWHEG (positive weight hardest emission generator), is used to simulate events from $t\bar{t}$ and single top quark background processes.

- **Pythia:** [121] is a multifunctional event generator which can not only perform matrix element calculation (at LO) but also parton showering, hadronization, and underlying event simulation. For parton showering, it uses p_{T} ordering and relies on the Lund string model [107, 108] for hadronization. Pythia v8.226 has been used for parton showering, hadronization, and the simulation of the underlying events for all simulations used in this thesis.

5.9. Luminosity reweighting

The number of simulated events does not necessarily match what is expected from the data. To make a fair comparison between data and simulation, each simulated event must be reweighted using a normalization weight, defined as:

$$w_{\text{norm}} = \frac{\sigma \int L \, \mathrm{d}t}{N} \tag{5.4}$$

where σ is a cross section of a process, $\int L \ \mathrm{d}t$ is the integrated luminosity of the dataset, and N is the number of simulated events of that process. The integrated luminosity per reconstruction year, which contributes to the full Run 2 data, is shown in Table 5.1.

Table 5.1: Integrated luminosity collected by the CMS experiment for each reconstruction year and combined Run 2 dataset.

Reconstruction year	Luminosity (fb^{-1})
2016 preVFP	19.5
2016 postVFP	16.8
2017	41.5
2018	59.8
Run 2	137.6

6. Post-simulation corrections

The simulated events from event generators are intended to serve as a reference for comparison with real collision events observed in data collected by the CMS detector. These simulations include all essential steps: hard scattering, parton showering, hadronization, and detector geometry and resolution effects. However, even with these detailed simulations, residual differences often remain when comparing data to simulation. These residual differences can arise from various sources, such as trigger inefficiencies, discrepancies in lepton identification performance, or differences in the behaviour of jet algorithms between data and simulation. To account for these differences and ensure reliable analyses, corrections must be applied to the simulations. These corrections are typically implemented using scale factors derived from auxiliary measurements. Most of these scale factors are determined through a collaborative effort within the CMS Collaboration. The scale factors are applied as additional weights to simulated events, reweighting their kinematic distributions to better align with those observed in the data. This chapter summarizes the necessary corrections applied to simulations for the analysis presented in this thesis.

6.1. Pileup reweighting

The number of pileup interactions simulated can differ from the actual pileup interactions observed in collisions, as simulations are typically generated well in advance and may not perfectly reflect the real conditions. To address these differences, weights are applied to simulated events to ensure that their pileup distribution matches that of the data.

In all simulated samples, additional inelastic proton-proton collisions are accounted for using a reweighting procedure. This procedure involves comparing the pileup distributions in minimum bias events with the cross section for minimum bias events assumed to be 69.2 mb [122]. The reweighting accounts for both "in-time pileup", arising from collisions within the same bunch crossing, and "out-of-time pileup", resulting from interactions in previous or subsequent bunch crossings.

6.2. L1 prefiring

The CMS detector uses a 2-tier trigger system: L1 and HLT, to select and store only interesting collision events, detailed in Section 3.2.7. The L1 trigger, partly installed inside the CMS detector, uses a coarse readout from the detector electronics, called trigger primitives, to make a decision whether to keep the event or not. It has two independent triggering systems, one from the muon system and the other from the calorimeter systems. In the 2016 and 2017 data taking period, the gradual timing shift of the ECAL was not properly propagated to the L1 trigger primitives. It was observed that a large fraction of trigger primitives of the ECAL were associated with the wrong bunch crossing. As two consecutive bunch crossings are not allowed to be triggered, as per rules of the L1 trigger system, this effectively caused self-veto of events. This issue was more pronounced in the ECAL region $2 \leq |\eta| \leq 3$. A similar issue was observed due to the limited timing resolution of the muon systems in all years, which was more pronounced in 2016 data.

The unpleasant effects of L1 prefiring are not reflected in our simulations. To account for this, dedicated correction factors are calculated and applied as prefiring weights to our simulated samples. These weights are derived by calculating prefiring probabilities for all jets, muons, and photons as:

$$w_{\text{prefiring}} = \prod_{\gamma,\, \mu,\, \text{jets}} (1 - \epsilon_{\text{prefire}}) \tag{6.1}$$

where $\epsilon_{\text{prefire}}$ are the observed prefire efficiencies, measured as a function of p_{T} and η, more details in Reference [123].

6.3. Lepton efficiencies

Charged leptons (electrons and muons) are reconstructed using various parts of the CMS detector, as described in Chapter 4. Their reconstruction efficiencies can be different in data and simulation. The CMS Collaboration provides centrally derived scale factors that correct for mismatches stemming from reconstruction and identification inefficiencies [81, 83]. The total lepton efficiency for both muons and electrons can be written as:

$$\begin{aligned} \epsilon^{\mu} &= \epsilon^{\mu}_{\text{ID}} \cdot \epsilon^{\mu}_{\text{iso|ID}} \cdot \epsilon^{\mu}_{\text{reco|iso}} \cdot \epsilon^{\mu}_{\text{trigger|reco}}\ , \text{and} \\ \epsilon^{e} &= \epsilon^{e}_{\text{ID}} \cdot \epsilon^{e}_{\text{reco|ID}} \cdot \epsilon^{e}_{\text{trigger|reco}}\ , \end{aligned} \tag{6.2}$$

respectively. The total lepton efficiency gets contributions from various parts of reconstruction and identification, starting from identification (ID), isolation (iso), reconstruction (reco), and trigger efficiencies. For electrons, isolation is applied with identification, therefore they are not considered as separate, like in the case for muons. All efficiencies are dependent on each other. Each subsequent efficiency factor uses the corrected lepton from the previous step, indicated by the subscripts, e.g. iso|ID, reco|ID etc. These efficiency factors are measured in bins of p_{T} and η using auxiliary measurements in $Z \to \mu^+\mu^-$ and $Z \to e^+e^-$ events, for muons and electrons, respectively.

6.4. Jet energy corrections

Jets reconstructed by the jet clustering algorithms, described in Section 4.3.1, are not a perfect representation of the parent parton. Due to pileup interactions and non-linear detector effect, reconstructed jet energy and resolution can be very different from simulations. The jets reconstructed in both data and simulations are calibrated before they can be used in any analysis. The CMS experiment performs a well-defined series of corrections to correct the energy and resolution of the reconstructed jets [124, 125]. The various steps performed on data and simulation to apply jet energy corrections are shown in Figure 6.1.

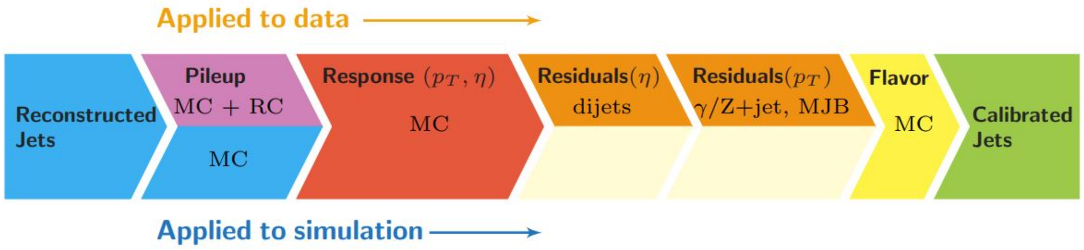


Figure 6.1: Stages of Jet Energy Corrections (JEC) applied sequentially to both data and Monte Carlo (MC) simulations. Corrections labelled as MC are obtained from simulation studies, while RC denotes random cone corrections, and MJB refers to corrections derived from multijet event analysis. The picture is taken from Reference [125].

The CMS collaboration uses a factorized approach to apply jet energy corrections. In the first stage, the residual effects of pileup are mitigated using corrections derived by comparing simulations with and without pileup overlay. Next, the detector response is calibrated to account for its dependence on the transverse momentum and pseudorapidity of the jets. These corrections, determined from simulations, adjust the jet p_{T} such that the ratio of generated to reconstructed jets averages to one. Finally, residual discrepancies in data are addressed by comparing data to QCD dijet simulations for η corrections and to Z/γ+jets events for p_{T} corrections. More details on the procedure can be found in Reference [124].

In addition to jet energy scale corrections, jet energy resolution corrections are also applied to simulations because it has been observed that the resolution of jet energy is quite worse in real data compared to simulations. We use a "hybrid" method, to apply the resolution corrections with the recommended scale factors [126]. In this method, if the jet is reconstructed in the vicinity ($\Delta \mathrm{R} < 0.4$) of a particle-level jet, the jet momentum is scaled by a factor c, given by:

$$c = 1 + (s_{\mathrm{JER}} - 1)\frac{p_{\mathrm{T}} - p_{\mathrm{T}}^{\mathrm{gen}}}{p_{\mathrm{T}}}, \tag{6.3}$$

where, p_{T} is the transverse momentum of the jet, $p_{\mathrm{T}}^{\mathrm{gen}}$ is the transverse momentum of the corresponding jet clustered from generator-level particles, and s_{JER} is the data-to-simulation resolution scale factor. If no particle-level jet is matched to the reconstructed jet, the jet momentum is adjusted by applying the resolution scale factor, smeared using a random value sampled from a standard Gaussian distribution. In both scenarios, c is restricted to be non-negative, ensuring that the resolution in the simulation can only be worsened.

6.5. b tagging efficiencies

The efficiency of tagging b-quark-initiated jets in data and simulation is observed to be slightly different. In the analysis presented in this thesis, the `DeepJet` algorithm [127] is employed to tag b quark initiated jets using a fixed working point. Corrections to simulation are applied in the form of scale factors for the fixed working points used in this thesis, i.e. medium and loose, introduced in Section 4.4. The scale factors are independent for b quark and light quark initiated jets. These scale factors are centrally derived by the CMS Collaboration [128]. For b quark initiated jets, five different methods are used to derive the scale factors and then combined afterwards for smaller uncertainties.

To apply corrections to simulated events, we use a reweighting procedure, recommended in Refernce[129], based on scale factors and tagging efficiencies in simulation. A weight $w_{\text{btag eff.}}$ is constructed, which is the ratio of b jet tagging probability in data and simulation:

$$w_{\text{btag eff.}} = \frac{P(\mathrm{DATA})}{P(\mathrm{MC})}, \tag{6.4}$$

where probability P of a given configuration of jets in data and MC simulation is defined as:

$$P(\mathrm{DATA}) = \prod_{i=\mathrm{tagged}} \mathrm{SF}_i \epsilon_i \prod_{j=\mathrm{not\ tagged}} (1 - \mathrm{SF}_j \epsilon_j), \tag{6.5}$$

and

$$P(\mathrm{MC}) = \prod_{i=\mathrm{tagged}} \epsilon_i \prod_{j=\mathrm{not\ tagged}} (1 - \epsilon_j). \tag{6.6}$$

Here, ϵ_i are the Monte Carlo simulation efficiencies and SF_i are the data-to-simulation tagging efficiency scale factors. Both are functions of jet flavour, p_T, and η. While the scale factors are centrally provided by the CMS Collaboration, the b tagging efficiencies in simulation must be estimated within the analysis phase space. The 2D efficiency maps for b tagging, and mistagging efficiencies of c jets and light quark jets are measured in the VBS signal enriched phase space and are shown in Figure 6.2. The formal definition of VBS signal region will follow in the later chapters. The efficiency maps are calculated separately for all phase spaces used in the analysis. More figures can be found in Appendix A.

6.6. Identification efficiencies of boosted vector boson

To identify jets originating from boosted W or Z bosons and from light quark or gluon initiated jets, the N-subjettiness variable is utilized, which was introduced in Section 4.3.2. The efficiency of this variable is different in data and simulation, therefore corrections are applied to simulation to account for the residual differences. Scale factors for different working points of this variable are centrally calculated and provided by the CMS Collaboration [130]. They are applied as weights to the simulated events. A summary of different working points used, along with the data-to-simulation scale factors, are shown in Table 6.1.

6.7. HCAL Endcap Minus (HEM) 15/16 issue in 2018

The power supply to the HEM15 and HEM16 sector of the HCAL modules was inoperative in the middle of 2018 data taking period. Due to this, the HCAL deposits were not recorded in the region $-1.57 < \phi < -0.87$ and $-3.2 < \eta < -1.3$ resulting in a lower jet energy response in this region, as well as a higher probability

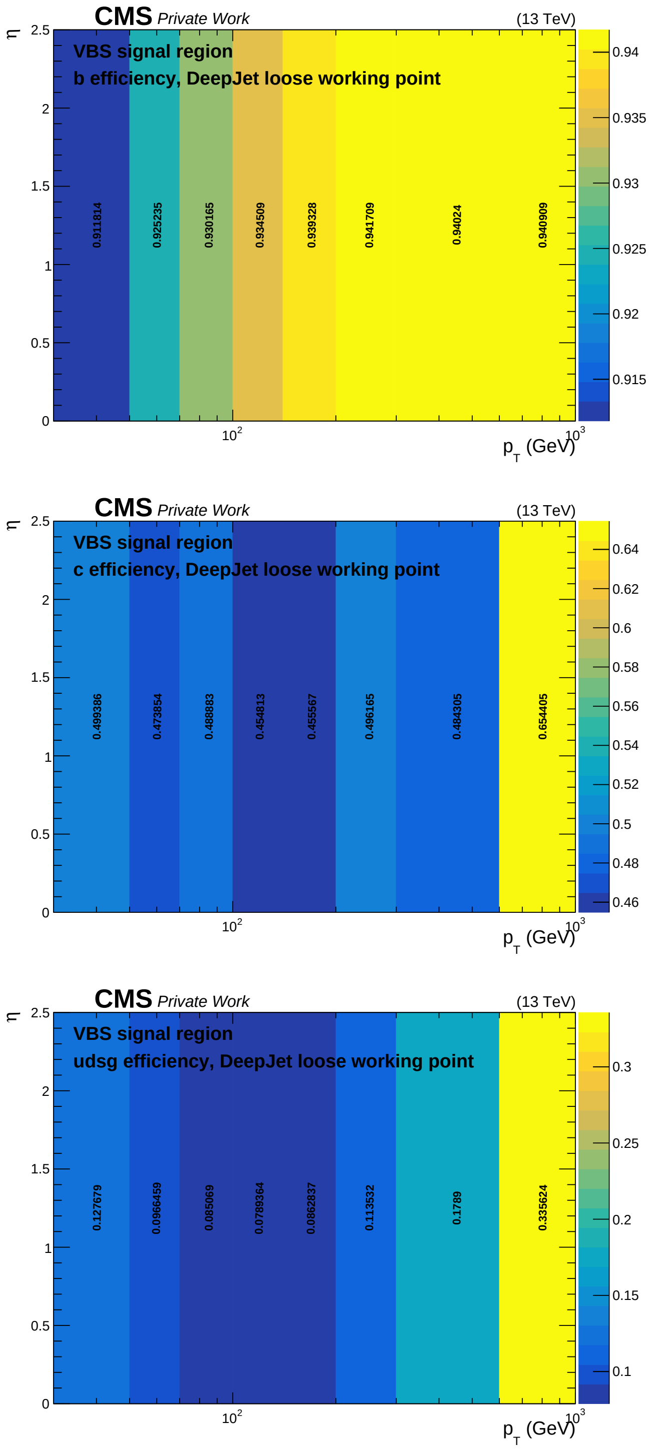


Figure 6.2: Monte Carlo simulation flavour tagging efficiencies as a function of p_{T} and η in 2017. Efficiencies for b (top), c (middle), and light (bottom) jets for the loose working point of the DeepJet tagger are shown in the VBS signal region.

Table 6.1: Summary of N-subjettiness (τ_{21}) working points and scale factors for different data taking eras. Values are taken from Reference [130].

Data-taking era	Working point	Scale factor
2016 preVFP	$\tau_{21} < 0.55$	1.03 ± 0.14
2016 postVFP	$\tau_{21} < 0.55$	1.03 ± 0.14
2017	$\tau_{21} < 0.45$	0.97 ± 0.06
2018	$\tau_{21} < 0.45$	0.980 ± 0.027

of jets being identified as electrons. This effected the data collected in the last certified run of 2018B (run >= 319077) and all runs of 2018C+2018D.

For the analysis presented in this thesis, this issue caused a bump in the η and ϕ distributions of the reconstructed electrons in 2018 data. To correct for this issue, events from 2018 data in which electrons or jets fall in the HEM affected η-ϕ region are vetoed. To account for this correction in simulation, all simulated events falling in this $\eta - \phi$ region are scaled down by a weight that accounts for the affected luminosity fraction. The weight is calculated as following:

$$\text{Affected lumi. fraction} = \frac{\text{RunB(17.37008/pb)} + \text{RunC} + \text{RunD}}{\text{Total integrated lumi. 2018}} = 0.64845 \quad (6.7)$$

After applying the correction to the 2018 data and simulation, the fake electrons arising due to HEM issue were effectively removed. The lepton distributions before and after the corrections are shown in Figure 6.3. The overall effect of the HEM correction on the expected signal selection efficiency is observed to be very small ($< 1\%$).

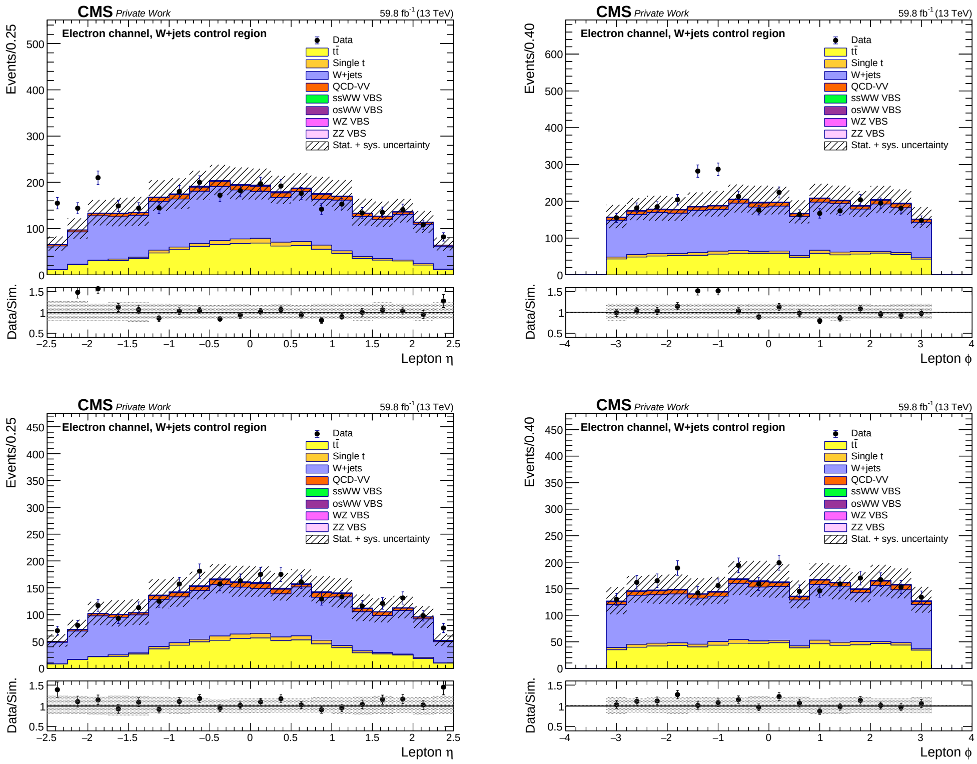


Figure 6.3: Shown are the distributions for the lepton η (left) and lepton ϕ (right), before (top) and after (bottom) vetoing events containing electrons or jets in the HEM affected $\eta - \phi$ region. Weights are applied to simulated events to account for the affected luminosity fraction.

7. Jet charge tagging

Determining the charge of the hadronically decaying W boson is crucial for distinguishing same-sign WW scattering from opposite-sign WW and WZ scattering in the semi-leptonic decay channel. To address this challenge, a dedicated tool — the jet charge tagger — was developed as part of this thesis, and is described in this chapter.

The work presented in this chapter is self-contained. The techniques and results discussed here are broadly applicable to any CMS physics analysis that requires identifying the electric charge of a jet, in particular a large radius jet. All results presented have been published by the CMS Collaboration in a Detector Performance Summary note [131]. A summary of the results was also presented in the form of a poster, shown in Appendix C, which was showcased at the 16th International Workshop on Boosted Object Phenomenology, Reconstruction, Measurements, and Searches at Colliders, held in 2024 in Genoa, Italy.

This chapter is structured as follows: Section 7.1 introduces the motivation behind the need to identify the electric charge of a jet. Section 7.2 provides an overview of the strategy employed for jet charge tagging studies. The data and simulated samples used in the study of jet charge and in the development of the machine learning-based jet charge tagger are detailed in Section 7.3. The object and event selection criteria are described in Section 7.4. Before introducing the jet charge tagger, we present traditional cut-based methods that can be used to discriminate between boosted W^+, W^-, and Z jets. The results of these methods, serving as a baseline, are presented in Section 7.5. Section 7.6 describes the jet charge tagger in detail, including its architecture, input features, training methods, and classification performance in differentiating between various jet charge categories. The systematic uncertainties affecting the results are discussed in Section 7.7. Finally, potential applications of the jet charge tagger are outlined in Section 7.8.

7.1. Motivation

Discriminating the origin of jets in the final state is a pivotal aspect of many analyses at the LHC. In searches where the charge of the particle giving rise to the jet serves as a key distinguishing feature between signal and background, accurately reconstructing the jet charge becomes an essential challenge. This task is straightforward when the originating particle decays into leptons, as the charge of the lepton(s) can be efficiently reconstructed to infer the charge of the parent particle. However, when the decay products are hadrons that form jets, the charge information of the originating particle is lost. An example of such a scenario is the search presented in this thesis, where one of the two vector bosons decays hadronically. In this case, the hadronic decays of W^+, W^-, or Z bosons appear indistinguishable, as they produce only jets in the detector. At the LHC, jets are generally not categorized as positive, negative, or neutral.

To differentiate between processes such as same-sign $\mathrm{W}^\pm\mathrm{W}^\pm$, opposite-sign $\mathrm{W}^\pm\mathrm{W}^\mp$, and $\mathrm{W}^\pm\mathrm{Z}$ VBS in semi-leptonic or fully hadronic final states, jet charge categorization becomes essential. Assigning a charge to jets offers a promising approach to distinguishing between these processes, paving the way for novel research opportunities in vector boson scattering. Such final states are studied collectively in the semi-leptonic decay channel [32], as there is no way to study them separately. Notably, the $\mathrm{W}^\pm\mathrm{W}^\pm$ VBS process is of particular importance compared to the other VBS processes, both for exploring physics beyond the standard model and for deepening our understanding of the electroweak sector, as discussed in Section 2.3.3.

7.2. Strategy

In this study, we utilize large-radius jets (boosted jets), which were introduced in Section 4.3.2. These jets are reconstructed using the anti-k_T algorithm with a radius parameter of 0.8 and are groomed using the soft drop algorithm, as also described in Section 4.3.2. To mitigate the impact of pileup interactions during jet reconstruction, we employ the PUPPI algorithm, detailed in Section 4.3.1. PUPPI weights are incorporated into the kinematic variables of the jets.

The primary objective of this chapter is to distinguish boosted jets originating from the hadronic decays of W^+, W^-, and Z bosons, assuming that jets from light quarks have already been separated. To achieve this, we use the semi-leptonic decay channel of the top quark-antiquark pair ($\mathrm{t\bar{t}}$) production to obtain a pure sample of boosted jets from W^+ and W^- bosons. This choice of process offers several advantages. Firstly, in semi-leptonic decays, the leptonically decaying top quark preserves the charge information of the W boson in the final state. This allows us to infer the true charge of the hadronically decaying W boson, which always has an opposite electric charge to the lepton from the leptonic top quark decay, see Figure 7.1. This logic holds for both simulation and real data, enabling us to validate the performance of jet charge tagging. The second advantage is that in the $\mathrm{t\bar{t}}$ production, W^+ and W^- bosons are produced in approximately equal proportions, resulting in a balanced sample of positive and negative jets from W bosons.

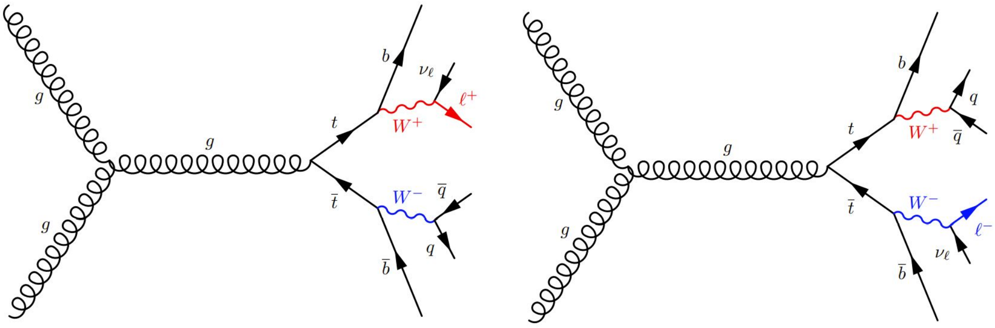


Figure 7.1: Shown are the Feynman diagrams for gluon initiated $\mathrm{t\bar{t}}$ production in the semi-leptonic decay channel. The left diagram depicts the leptonic decay of W^+, while the right shows the leptonic decay of W^-. The charge of the final-state lepton reflects the charge of its parent W boson, with coloured fermion lines emphasizing this correspondence. The charge of the lepton can be used to infer the charge of the jet produced by the other W boson, which will always have the opposite charge to the lepton.

To study boosted jets coming from Z boson decays, we use Z+jets Monte Carlo simulation. The drawback of using this process is that we have two jets in the final state. Jets coming from quarks or gluons (light-quark jets) can be easily misreconstructed as Z jets. To minimize this misidentification, we match reconstructed jets to particle level jets using the truth information of simulated events. We use the distance parameter $\Delta\mathrm{R}$, defined as:

$$\Delta\mathrm{R} = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{7.1}$$

to match particle level jets with reconstructed jets. Only those jets that are reconstructed within $\Delta\mathrm{R} < 0.8$ from the particle level jets are used further in this study.

As a first step, we use the definition of jet charge from the literature to examine the distribution of the electric charge of boosted jets from W^+ and W^- bosons (also called as W^+ and W^- jets) in a $\mathrm{t\bar{t}}$-enriched phase space. Subsequently, we incorporate boosted jets from Z bosons (Z jets) into the analysis. Since W and Z bosons have a mass difference of approximately 10 GeV, jet mass becomes an additional discriminating variable for distinguishing between W and Z jets, alongside jet charge. These studies, detailed in Section 7.5 will serve as a baseline. We then demonstrate that instead of relying solely on jet charge or jet mass as discriminating variables, jet substructure information can be directly exploited using machine learning techniques, described in Section 7.6. This approach allows for more efficient identification of the charge of boosted jets compared to the traditional method based on jet charge and/or jet mass based discrimination.

7.3. Data and simulated samples

The data and Monte Carlo simulation samples used to study jet charge and later on for training the jet charge tagger correspond to proton-proton collisions at a centre-of-mass energy of 13 TeV.

7.3.1. Data samples

The Run 2 dataset collected by the CMS detector during 2018 has been used in this study, which corresponds to a total integrated luminosity of 59.8 fb^{-1}. Centrally produced MINIAODv2 datasets have been used. EGamma and single muon datasets from run periods A to D are utilized, comprising only those runs that pass the quality assessment from the data quality monitoring. The high level trigger paths are used to select single electron and single muon events from the datasets, as shown in Table 7.1.
The global tag used for 2018 data is `106X_dataRun2_v37`. The names of the datasets are listed in Table 7.2.

Table 7.1: Trigger paths for 2018 datasets

Dataset name	Run range	HLT path
Single Muon	A – D	`HLT_IsoMu24`
EGamma	A – D	`HLT_Ele32_WPTight_Gsf`

7.3.2. Simulated samples

Simulated samples are utilized to obtain boosted jets from the hadronic decays of W^{+}, W^{-}, and Z bosons. For validation in data, we use a $\mathrm{t\bar{t}}$-enriched phase space in data, therefore additional simulations are required to estimate backgrounds that mimic the $\mathrm{t\bar{t}}$ pair production-like final state. The simulated samples are generated for the following processes: $\mathrm{t\bar{t}}$ production, Z+jets, W+jets, single top quark production and QCD multijet.

The full names of the simulated samples along with their respective cross section values are listed in Table 7.3. All simulated samples were produced during the `Summer20UL18MiniAODv2` Monte Carlo production campaign.
The global tag `106X_upgrade2018_realistic_v16_L1v1` was used to analyse these samples.

Table 7.2: Datasets used in the study of jet charge and for validating the jet charge tagger performance in data, along with the corresponding luminosity and number of events, are shown. The datasets were collected by the CMS experiment during 2018 and correspond to the total integrated luminosity of 59.8 fb^{-1}.

Dataset	Luminosity (fb^{-1})	Number of events
/SingleMuon/Run2018A-UL2018_MiniAODv2_GT36-v1/MINIAOD	14.03	241,591,525
/SingleMuon/Run2018B-UL2018_MiniAODv2_GT36-v1/MINIAOD	7.06	119,918,017
/SingleMuon/Run2018C-UL2018_MiniAODv2_GT36-v2/MINIAOD	6.90	110,032,072
/SingleMuon/Run2018D-UL2018_MiniAODv2_GT36-v1/MINIAOD	31.74	513,884,680
/EGamma/Run2018A-UL2018_MiniAODv2_GT36-v1/MINIAOD	14.03	339,013,231
/EGamma/Run2018B-UL2018_MiniAODv2_GT36-v1/MINIAOD	7.06	153,822,427
/EGamma/Run2018C-UL2018_MiniAODv2_GT36-v1/MINIAOD	6.90	147,827,904
/EGamma/Run2018D-UL2018_MiniAODv2-v2/MINIAOD	31.74	752,497,815

The majority of the cross sections utilized for normalizing the simulated events have been computed with precision up to the next-to-next-to-leading order. For W+jets, the NLO cross section is used with a k-factor of 1.21 applied [132]. The k-factor has been calculated as ratio of inclusive W+jets cross section at NNLO accuracy to inclusive NLO cross section. The $\mathrm{t\bar{t}}$ cross section was calculated at NNLO+NNLL precision with TOP++V2.0 [133]. Single top cross sections were calculated with the HATHOR V2.1 program [134]. For QCD multijet and Z+jets processes, LO cross section values are used.

7.4. Object and event selections

To select events with a $\mathrm{t\bar{t}}$-like final state in semi-leptonic decays, we apply various physics object selections in addition to the general identification and reconstruction methods discussed in Chapter 4. These selections isolate a pure sample of W^+ and W^- jets.

We define a $\mathrm{t\bar{t}}$-enriched control region by requiring exactly one tight isolated lepton (electron or muon) with $p_\mathrm{T} > 30$ GeV and $|\eta| < 2.4(2.5)$ for muons(electrons). Lepton identification follows the cut-based criteria recommended by the CMS Collaboration, as detailed in Sections 4.2.1 and 4.2.2. Events containing additional leptons that satisfy the loose identification and isolation criteria with $p_\mathrm{T} > 10$ GeV are rejected. This ensures that the selected events contain exactly one isolated lepton from the leptonic top quark decay. To account for the presence of an undetected neutrino from the leptonic decay, we impose a missing transverse momentum

Table 7.3: Simulated samples used to study the electric charge of boosted jets from hadronic decays of W^+, W^-, and Z bosons, along with other simulations used for estimating background processes in data.

Process	Dataset name	Cross section (pb)
TTTosemi-leptonic	TTTosemi-leptonic⋆⋆⋆	370.62
SingleTop s-channel	ST_s-channel_4f_leptonDecays⋆⋆⋆⋆	10.32
SingleTop t-channel (top)	ST_t-channel_top_4f_InclusiveDecays⋆⋆	136.02
SingleTop t-channel (antitop)	ST_t-channel_antitop_4f_InclusiveDecays⋆⋆	80.95
SingleTop tW-channel (top)	ST_tW_top_5f_inclusiveDecays⋆⋆⋆	35.85
SingleTop tW-channel (antitop)	ST_tW_antitop_5f_inclusiveDecays⋆⋆⋆	35.85
W+Jets HT70-100	WJetsToLNu_HT-70To100⋆	1529.44
W+Jets HT100-200	WJetsToLNu_HT-100To200⋆	1519.76
W+Jets HT200-400	WJetsToLNu_HT-200To400⋆	405.96
W+Jets HT400-600	WJetsToLNu_HT-400To600⋆	54.75
W+Jets HT600-800	WJetsToLNu_HT-600To800⋆	13.27
W+Jets HT800-1200	WJetsToLNu_HT-800To1200⋆	5.97
W+Jets HT1200-2500	WJetsToLNu_HT-1200To2500⋆	1.40
W+Jets HT2500-Inf	WJetsToLNu_HT-2500ToInf⋆	0.0317
W+Jets Inclusive	WJetsToLNu⋆	61526.7
Z+Jets HT200-400	ZJetsToQQ_HT-200to400⋆	1010.20
Z+Jets HT400-600	ZJetsToQQ_HT-400to600⋆	114.208
Z+Jets HT600-800	ZJetsToQQ_HT-600to800⋆	25.35
Z+Jets HT800-Inf	ZJetsToQQ_HT-800toInf⋆	12.915
QCD Pt170to300	QCD_Pt_170to300⋆⋆⋆⋆⋆	117276
QCD Pt300to470	QCD_Pt_300to470⋆⋆⋆⋆⋆	7823
QCD Pt470to600	QCD_Pt_470to600⋆⋆⋆⋆⋆	648.2
QCD Pt600to800	QCD_Pt_600to800⋆⋆⋆⋆⋆	186.9
QCD Pt800to1000	QCD_Pt_800to1000⋆⋆⋆⋆⋆	32.29
QCD Pt1000to1400	QCD_Pt_1000to1400⋆⋆⋆⋆⋆	9.418
QCD Pt1400to1800	QCD_Pt_1400to1800⋆⋆⋆⋆⋆	0.8426
QCD Pt1800to2400	QCD_Pt_1800to2400⋆⋆⋆⋆⋆	0.1149
QCD Pt2400to3200	QCD_Pt_2400to3200⋆⋆⋆⋆⋆	0.0068
QCD Pt3200toInf	QCD_Pt_3200toInf⋆⋆⋆⋆⋆	0.000165

⋆ _TuneCP5_13TeV-madgraphMLM-pythia8

⋆⋆ _TuneCP5_13TeV-powheg-madspin-pythia8

⋆⋆⋆ _TuneCP5_13TeV-powheg-pythia8

⋆⋆⋆⋆ _TuneCP5_13TeV-amcatnlo-pythia

⋆⋆⋆⋆⋆ _TuneCP5_13TeV_pythia8

requirement of $p_{\mathrm{T}}^{\mathrm{miss}} > 40$ GeV.

To identify boosted hadronic W boson decays, we require the presence of exactly one AK8 PUPPI jet with $p_{\mathrm{T}} > 200$ GeV and $|\eta| < 2.4$. Further requirements, such as N-subjettiness ($\tau_{21} < 0.55$) and soft drop mass of the boosted jet within the window 65 GeV $< m_{\mathrm{SD}} <$ 105 GeV, enhance the selection probability for jets originating from W boson decays. Since both the top quark and antiquark decay into a W boson and a b quark, we require events to have at least two b-tagged AK4 CHS jets. The b tagging is performed using the DeepJet algorithm medium working point, as described in Section 4.4. The b tagged jets must have a $p_{\mathrm{T}} > 30$ GeV and $|\eta| < 2.4$.

To avoid double counting, we remove events with overlapping objects. In the event selection process, leptons are selected first. All AK8 PUPPI and AK4 CHS jets that are overlapping with the selected lepton are removed from their respective collection. Next, we identify the boosted AK8 PUPPI jet and all AK4 CHS jets overlapping with the boosted jet are removed. This ensures a clean selection of reconstructed objects. A summary of event selections used to define the $\mathrm{t\bar{t}}$-enriched control region is shown in Table 7.4.

To reconstruct Z jets from Z+jets simulations, we require the same kinematic properties of the boosted jet i.e. p_{T}, η, soft-drop mass and N-subjettiness, as required for the reconstruction of W boson initiated jets. The other jets (originating from quarks and gluons) present in this process can be misreconstructed as Z jets. To avoid this, we match the reconstructed jet with the generator level Z boson jet using the distance parameter R defined in Eq. 7.1. The full set of object and event selections applied on the Z+jets simulated events are listed in Table 7.5.

Table 7.4: Event selections for $\mathrm{t\bar{t}}$ control region.

Object	Criteria
Tight lepton (e/μ)	Exactly one, $p_{\mathrm{T}} > 30(30)$ GeV, $\lvert\eta\rvert < 2.4(2.5)$
	cutBasedIDs for both e/μ
Additional lepton veto	$p_{\mathrm{T}} > 10(10)$ GeV, $\lvert\eta\rvert < 2.4(2.5)$
Missing transverse momentum	$p_{\mathrm{T}}^{\mathrm{miss}} > 40$ GeV
AK8 PUPPI jet	Exactly one jet, $p_{\mathrm{T}} > 200$ GeV, $\lvert\eta\rvert < 2.4$
ΔR(AK8 PUPPI jet, lepton)	> 0.8
AK4 CHS jets	At least two, b tagged (DeepJet medium working point)
	$p_{\mathrm{T}} > 30$ GeV, $\lvert\eta\rvert < 2.4$
ΔR(AK4 CHS jet, lepton)	> 0.4
ΔR(AK8 PUPPI jet, AK4 CHS jet)	> 0.8
N-subjettiness (τ_{21}) of AK8 PUPPI jet	< 0.55
AK8 PUPPI jet mass (m_{SD})	65 GeV $< m_{\mathrm{SD}} <$ 105 GeV

Table 7.5: Event selections for Z+jets Monte Carlo simulation.

Object	Criteria
AK8 PUPPI jet	Exactly one jet, $p_{\mathrm{T}} > 200$ GeV, $\eta < 2.4$
Generator-level (Gen) Z boson	pdgID = 23
Gen AK8 PUPPI jet	min(ΔR(Gen AK8 PUPPI jet, Gen Z boson))
ΔR(Gen AK8 PUPPI jet, Gen Z boson)	< 0.8
AK8 PUPPI jet	min(ΔR(AK8 PUPPI jet, Gen AK8 PUPPI jet))
ΔR(AK8 PUPPI jet, Gen AK8 PUPPI jet)	< 0.8
AK4 CHS jets	At least one, $p_{\mathrm{T}} > 30$ GeV
ΔR(AK8 PUPPI jet, AK4 CHS jet)	> 0.8
N-subjettiness (τ_{21}) of AK8 PUPPI jet	< 0.55
AK8 PUPPI jet mass (m_{SD})	65 GeV $< m_{\mathrm{SD}} <$ 105 GeV

7.5. Jet charge based discrimination

Measuring an electric charge of a particle which originates a jet is not straightforward, as a jet is a complex object consisting of many particles. The energy deposits and tracks associated with a jet can come from a wide range of charged particles in the p_{T} spectrum. The charged particles with low p_{T} can originate from pileup interactions or an underlying event. If a charged particle has a very low $p_{\mathrm{T}} \ll 1$ GeV, it can even curl up in the magnetic field of the detector and might not be measured in the tracker or the calorimeter. Hence, it is very important to define jet charge as a p_{T}-weighted sum of the charge of all particle constituents of a jet [135]. Mathematically, the jet charge is defined as:

$$Q^{\kappa} = \frac{\sum_i q_i (p_{\mathrm{T}}^i)^{\kappa}}{(p_{\mathrm{T}}^{\mathrm{jet}})^{\kappa}}, \tag{7.2}$$

where i runs over all particle/tracks inside the jet. q_i is the measured charge and p_{T}^i is the transverse momentum of the i-th track associated with the reconstructed particle. The parameter κ is a free regularization parameter, which can be tuned to get additional discrimination between different types of jets [135]. Typical choice of κ is 1. The way in which jet charge is defined makes it less sensitive to pileup interactions and detector resolution effects. The charged hadrons used to build the jet charge are identified and reconstructed using tracker information, where those from pileup vertices are discarded. More details about object reconstruction and identification methods are mentioned in Chapter 4.

To visualize the jet charge distribution of the boosted W^+ and W^- jets, we use the $\mathrm{t\bar{t}}$ control region, which gets the dominant contribution from the $\mathrm{t\bar{t}}$ production and little contamination from other processes like W+jets and multijet background. Each reconstructed event in this region contains one boosted jet and one isolated tight lepton. The jet charge for jets in events is calculated using Eq. 7.2. By utilizing the charge of the lepton, we can infer the true charge of the jet. Using this information, the jet charge distribution can be split into positive and negative contributions. The same reconstruction technique can also be applied to real data in the control region. Figure 7.2 shows the jet charge distributions for W^+ and W^- jets in the decay channels consisting of a muon or an electron in the final state. It is worth noting that the charge distributions of W^+ jets are peaking on the positive charge axis, while the distributions for W^- have a peak on the negative charge axis, giving a nice discrimination between the two peaks around zero. Also, the simulation agrees quite well with the data within the uncertainties.

The hyperparameter κ in the definition of jet charge can be tuned by testing different values between 0 and 1. We have observed that decreasing the value of κ smear the charge distribution, shown in Figure 7.3. To find the best value of κ which provides the most discrimination between W^+ and W^- jets, we have made efficiency curves for various values, shown in Figure 7.4. The best value of κ providing the most discrimination between the two jet categories is found to be 0.5. The jet charge distributions with $\kappa = 0.5$ are also shown in Figure 7.5.

$\mathbf{W}^+$ vs $\mathbf{W}^-$ vs **Z** jets:

For discrimination between W^+, W^-, and Z jets, we can use the jet charge variable in the similar manner as we did for W^+ and W^- jets. The only difference would be that we can only compare the three categories using simulation, as it is not possible to construct a control sample in data for Z jets. The W^+ and W^- jets comes from $\mathrm{t\bar{t}}$ simulation and Z jets comes from Z+jets simulation, after doing the generator matching described in Section 7.2. The normalized jet charge distribution for the three jet categories is shown in Figure 7.6.

For this classification case, one can also use the mass difference between the W and Z boson as an additional discriminating variable in combination with the jet charge. A two-dimensional plot using jet charge and jet mass is shown in Figure 7.7. It is evident that the resolution of charge and mass is not sufficient to define boundaries between the three categories for efficient discrimination. This motivates us to design an algorithm based on machine learning which can discriminate the three categories better than just charge and mass of the jet. Using machine learning methods, we can use various low-level information of the jet and its particle constituents, and train the network to learn and identify the charge of the boosted jet. This method can potentially enhance the discrimination between W^+, W^-, and Z jets.

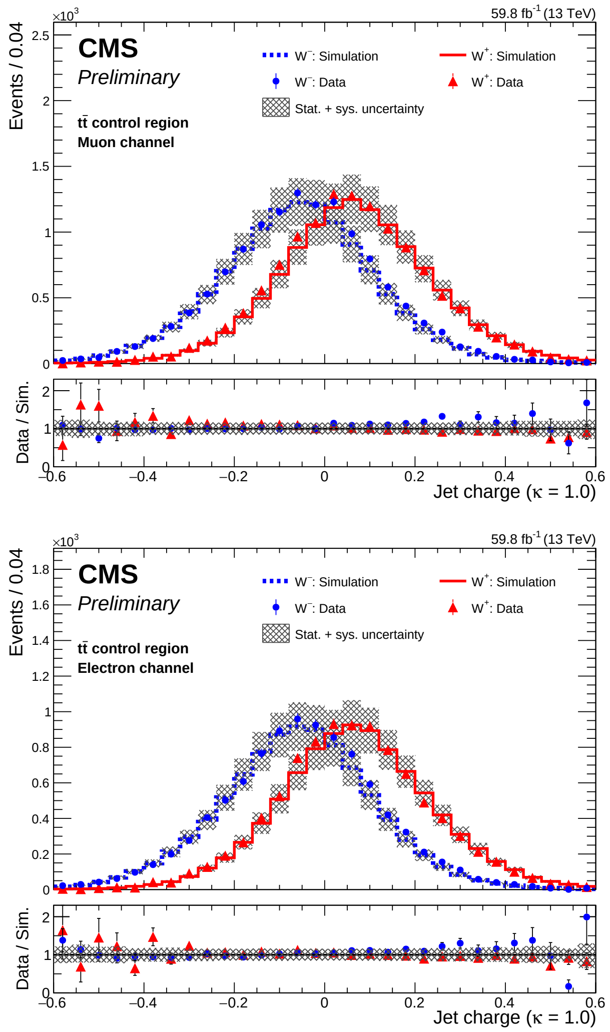


Figure 7.2: The distribution of the jet charge ($\kappa = 1.0$) in the $\mathrm{t\bar{t}}$ control region in data and in simulation for W^+ and W^- boosted jets are shown. The distributions are split by requiring negatively and positively charged muons (top) and electrons (bottom) for W^+ and W^- jets, respectively. This figure has been published in Reference [131].

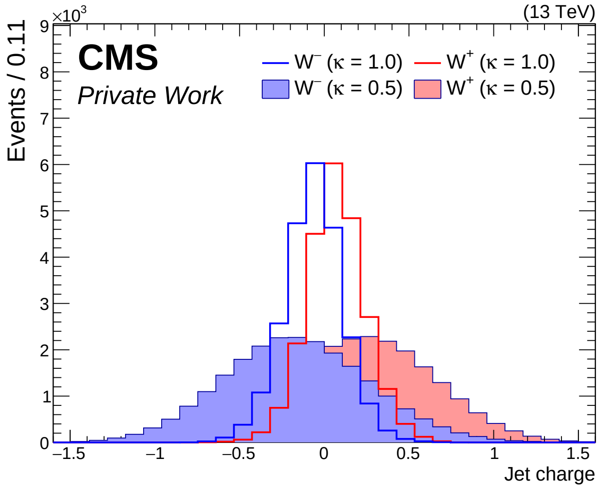


Figure 7.3: The effect of varying hyperparameter κ on the jet charge distributions is demonstrated using the $t\bar{t}$ simulation. Small values of κ smears the charge distributions.

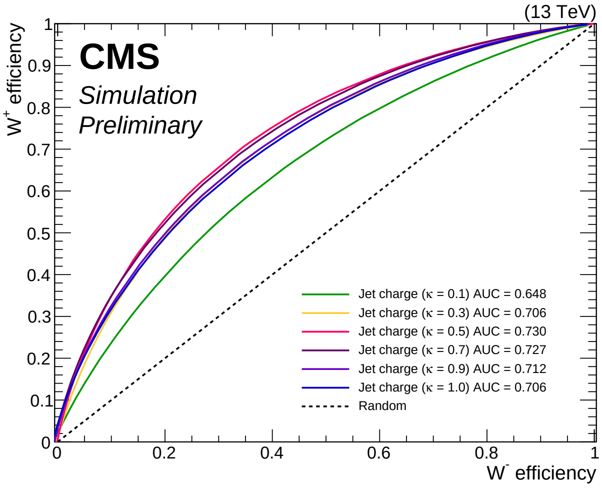


Figure 7.4: Efficiency curves for various values of κ from 0 to 1 are shown. The best performance as suggested by the Area Under the Curve (AUC) score is for κ equals to 0.5., providing the most discrimination between W^+ and W^-. This figure has been published in Reference [136].

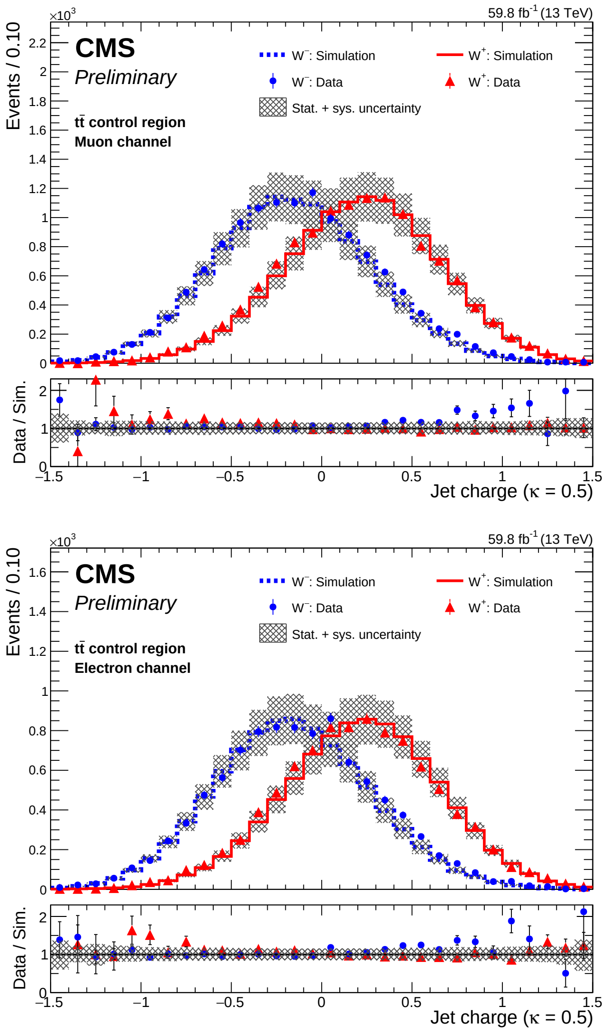


Figure 7.5: The distribution of the jet charge ($\kappa = 0.5$) in the $t\bar{t}$ control region in data and in simulation for W^+ and W^- boosted jets are shown. The distributions are split by requiring negatively and positively charged muons (top) and electrons (bottom) for W^+ and W^- jets, respectively. With $\kappa = 0.5$, jet charge provides better separation compared to $\kappa = 1.0$. This figure has been published in Reference [131].

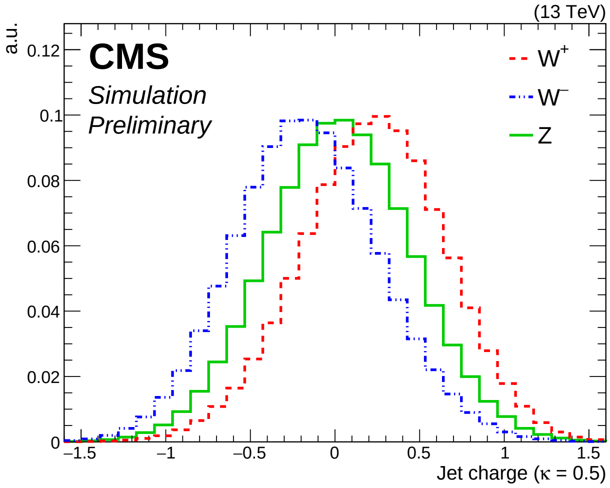


Figure 7.6: Discrimination between positive, negative, and neutral boosted jets coming from the hadronic decay of W^+, W^-, and Z bosons, respectively, using the jet charge variable is shown. The distributions are normalized to unity. This figure has been published in Reference [131].

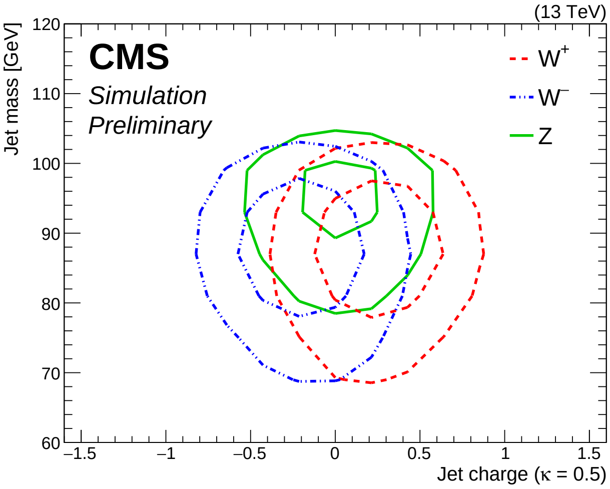


Figure 7.7: Shown is a two-dimensional plot with jet charge plotted on the x-axis and jet mass on the y-axis for boosted W^+, W^-, and Z jets. The discrimination power of the two variables is not sufficient to define boundaries to discriminate the three jet categories. This figure has been published in Reference [131].

7.6. Jet charge tagger

It was demonstrated in the previous section that the jet mass and jet charge resolution is not sufficient to discriminate between boosted jets originating from W^+, W^-, and Z bosons. This section describes an advanced algorithm referred as "jet charge tagger", which gives us more efficient and reliable discrimination between positive, negative, and neutral jets.

7.6.1. Architecture

The jet charge tagger uses the architecture of a Dynamic Graph Convolutional Neural Network model called ParticleNet [93], which has been successfully used for various jet classification tasks in the CMS [136] and outperforms previous models for tagging purposes. The jets are represented as "particle clouds", in which the particles are treated as unordered, permutational invariant sets. This representation is quite close to the natural definition of a jet and allows flexibility to include arbitrary features of particles, compared to sequences or trees based representation of jets.

A typical convolutional operation cannot be applied to particle clouds, because the particles are irregularly distributed. Instead, a special convolutional operation called the Edge Convolutional operation (EdgeConv) is used, which represents particle clouds as graphs, whose vertices are particles themselves, and the edges are constructed between each particle and its k nearest neighbouring particles, more details in Reference [137]. The EdgeConv operation provides a map from one particle cloud to another and hence multiple such operations can be stacked together to form a deep network, which learns the features of the cloud hierarchically.

The jet charge tagger architecture, uses a stack of two EdgeConv blocks, in contrast to ParticleNet, which uses three EdgeConv blocks. This simplified architecture is chosen to reduce the complexity of the network for charge tagging. The first EdgeConv block uses the coordinates η and ϕ of particles inside the jet to construct a graph. Each particle is connected with its k nearest neighbours and distances are computed using these coordinates. Then the features of the particle are used to construct the edge features with the help of indices of the k neighbours. The Edge convolutional operation itself is a three layer feed-forward neural network, consisting of a linear transformation, batch normalization, and a rectified linear unit (ReLU). The number of k nearest neighbours, and number of neurons in each linear transformation unit are the hyperparameters of a EdgeConv block. The values of these hyperparameters are chosen to be the same as used in the ParticleNet-Lite architecture [93].

The output from one EdgeConv block goes as an input to the other block. After the

EdgeConv blocks, a global average pooling operation is performed to aggregate the learned features over all particles in the cloud. Following the average pooling, two fully connected layers are included in the design. The output is generated using a softmax function, which gives the probabilities of a jet to belong to each category. A schematic representation of full architecture and a detailed schematic of EdgeConv block is shown in Figure 7.8.

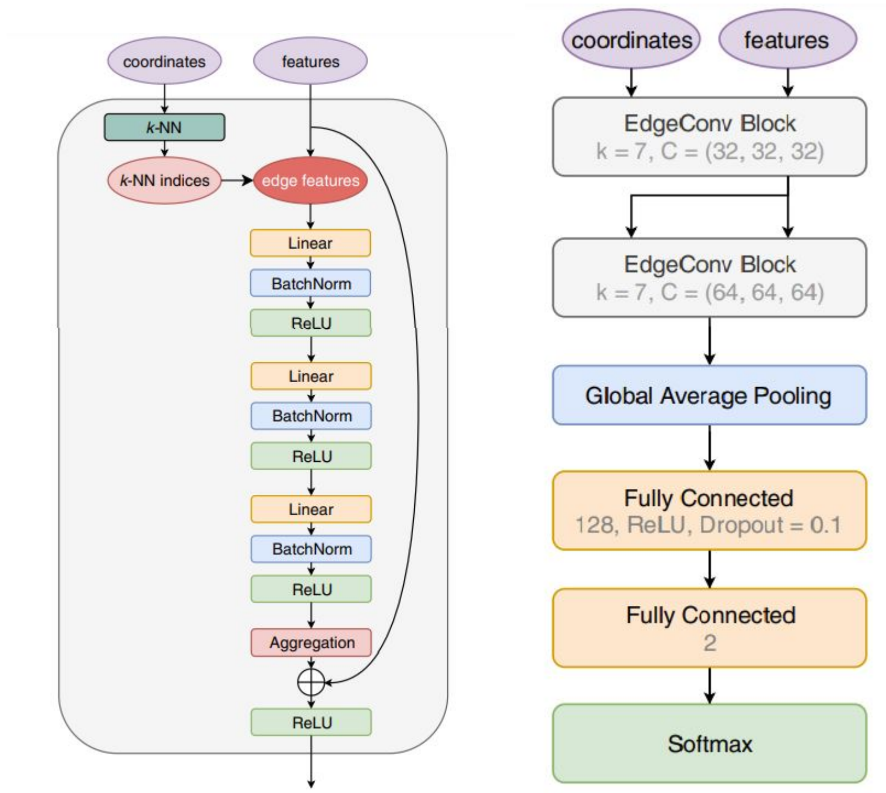


Figure 7.8: Schematic representation of jet charge tagger architecture, which uses the same architecture as ParticleNet-Lite. To the left, the Edge Convolutional block with its various components is shown. The full architecture consisting of two Edge Convolutional blocks and various other layers is shown on the right. The figures are taken from Reference [93].

7.6.2. Inputs

Various input features of the particles inside the boosted jet are used as input to the jet charge tagger. These input variables are categorized as coordinates and features. The pseudorapidity η and azimuthal angle ϕ of each particle belongs to coordinates, which are used by the tagger to construct the EdgeConv graphs, and various other kinematic variables like transverse momentum, energy etc. are categorized among the features of the particles. For some features of particles, such as the transverse momentum and energy, their logarithmic values are used in inputs for better handling of wide ranges. A complete list of input variables used for training is shown in Table 7.6.

Table 7.6: Description of input variables used for jet charge tagging.

Variable	Definition
Coordinates:	
$\Delta\eta$	difference in the pseudorapidity between the particle and the jet axis
$\Delta\phi$	difference in the azimuthal angle between the particle and the jet axis
Features:	
$\log\, p_{\mathrm{T}}$	logarithm of the particle's p_{T}
$\log\, \mathrm{E}$	logarithm of the particle's energy
$\log\, \frac{p_{\mathrm{T}}}{p_{\mathrm{T}}(\mathrm{jet})}$	logarithm of the particle's p_{T} relative to the jet p_{T}
$\log\, \frac{\mathrm{E}}{\mathrm{E(jet)}}$	logarithm of the particle's energy relative to the jet energy
$\Delta\mathrm{R}$	angular seperation between the particle and the jet axis ($\sqrt{(\Delta\eta)^2+(\Delta\phi)^2}$)
q	electric charge of the particle.

It is essential to check the modelling of the input variables before they can be used for a reliable training. We can use the $\mathrm{t\bar{t}}$ region to validate the modelling of input variables using the real data. All input variables were checked, and they show a good agreement between data and simulation. The input distributions in the muon decay channel are shown in Figure 7.9. Distributions for the electron decay channel can be found in Appendix B.

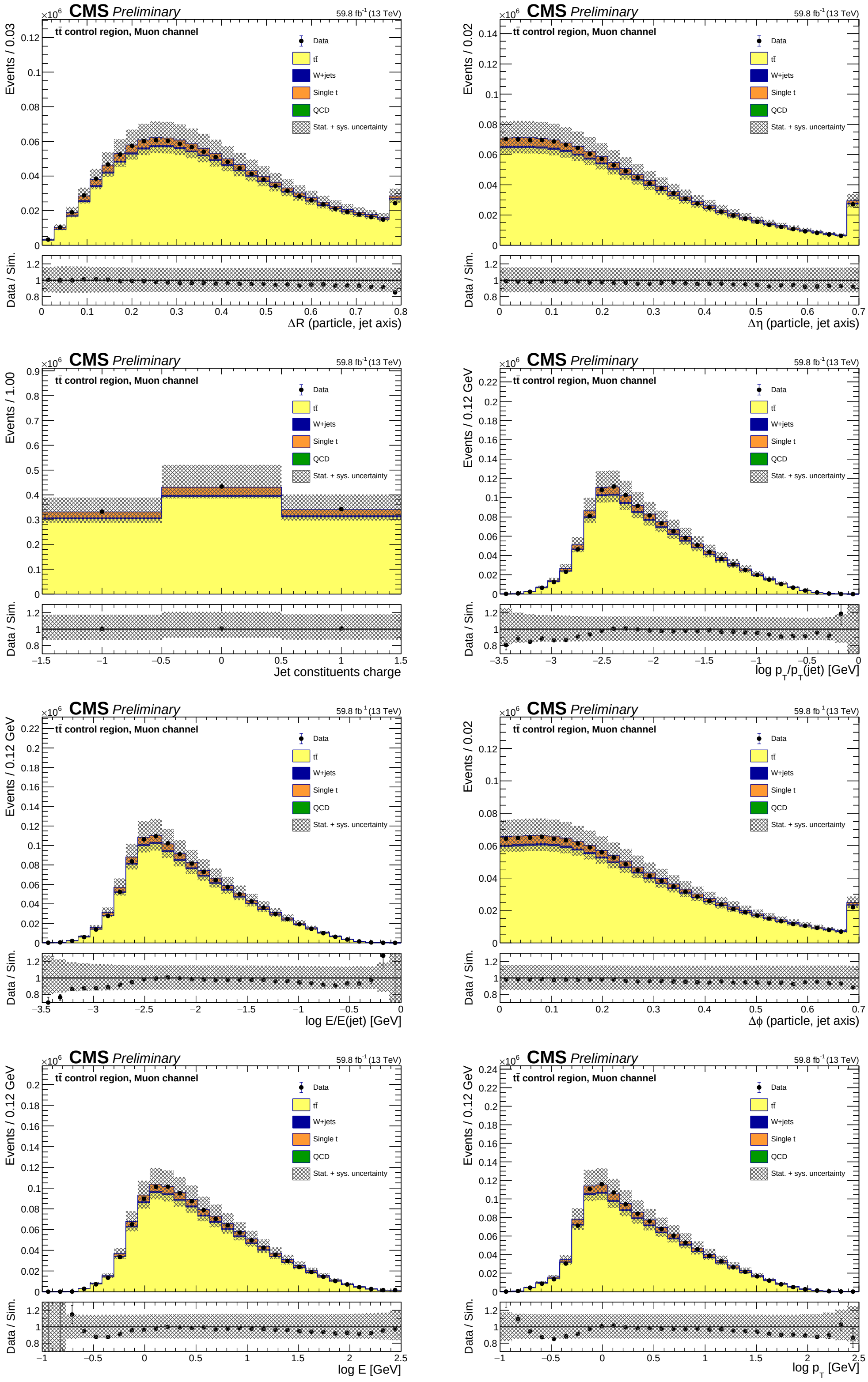


Figure 7.9: Distributions of input variables of the jet charge tagger in data and in simulation using the $t\bar{t}$ control region are shown in the muon decay channel. All inputs are well-modelled in simulation and are within the expected uncertainties. Similar distributions for the electron decay channel can be found in Appendix B.

7.6.3. Training

The jet charge tagger is implemented using the TensorFlow/Keras library. The training is performed using a single GPU. A batch size of 1024 is used due to memory constraints. The tagger is trained for two types of tasks: binary classification (W^+ vs W^- jets), and ternary classification (W^+ vs W^- vs Z jets).

For binary classification, the binary cross-entropy loss function is minimized and for ternary classification, the categorical cross-entropy function has been used. In both cases, the Adam optimizer is used with a learning rate scheduler. The initial learning rate is chosen to be 1×10^{-4}, after every 10 epochs the rate is reduced by 0.1. The training is performed for 30 epochs. The parameters of the training are chosen after optimization using a few initial trial trainings.

The loss and accuracy metric on the validation dataset is monitored during the training. The snapshot of the model is saved after each epoch, if the validation accuracy is improved. To avoid overtraining, early stopping method is employed with a patience level of five epochs. In addition, the network also uses a dropout method with dropout rate of 0.1, applied in one of the fully connected layers to avoid overtraining. For the final evaluation, the model showing the best performance on the validation dataset is chosen.

The training is performed using simulated events, each containing exactly one boosted jet. For each jet, a true label was assigned for supervised training based on its true expected electric charge. For binary classification, the boosted jets are taken from $t\bar{t}$ simulation. The whole sample is divided into three parts: training, validation, and testing sets using the ratio 60:15:25. The same splitting strategy has been used for ternary classification. The events containing exactly one boosted Z jet are taken from Z+jets simulation. This size of the training samples were set in a way to have a balanced training set containing equal number of jets from each category. The training is performed on the combined muon and electron decay channels, to have increased number of events during training.

7.6.4. Binary classification

The jet charge tagger is first trained for a binary classification task to categorize boosted jets as positive or negative jets or, in other words, W^+ or W^- jets. The main aim of the binary classification is to compare the tagger performance with our baseline study done with the jet charge variable, detailed in Section 7.5. The trained network model is used to evaluate on the unseen dataset (25% of the total), reserved for testing. The jet charge tagger output scores for the muon and the electron decay channels are shown in Figure 7.10. The distributions are split based on the charge of the lepton in the event, which gives the information of the true charge of the boosted jet. The tagger output score is consistent in data within the uncertainties of the simulation. A significant enhancement in the separation is observed in the tagger output score between the two jet categories.

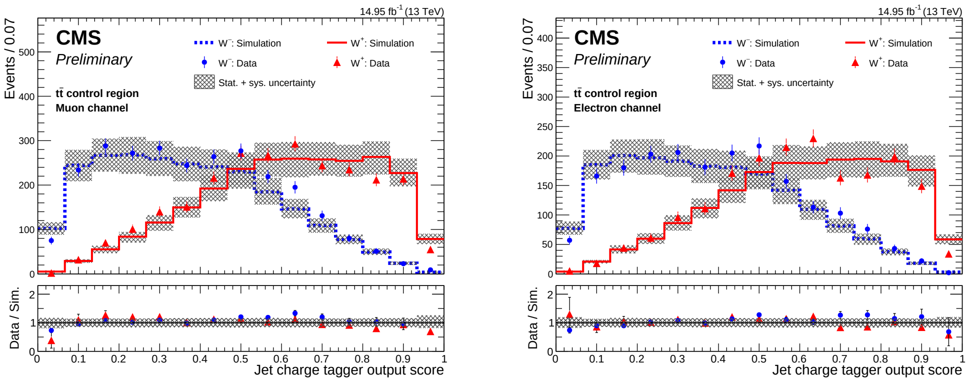


Figure 7.10: Jet charge tagger output score in the binary classification task to distinguish positively and negatively charged boosted jets is shown in the muon decay channel (left) and the electron decay channel (right). The distributions are split based on the charge of lepton. The jet charge tagger output score shows a significant increase in the separation between W^+ and W^- jets as compared to the jet charge variable, shown in Figure 7.5. This figure has been published in Reference [131].

To compare the performance of the jet charge tagger with the jet charge variable, efficiency curves are shown in Figure 7.11. The jet charge tagger outperforms the jet charge based discrimination.

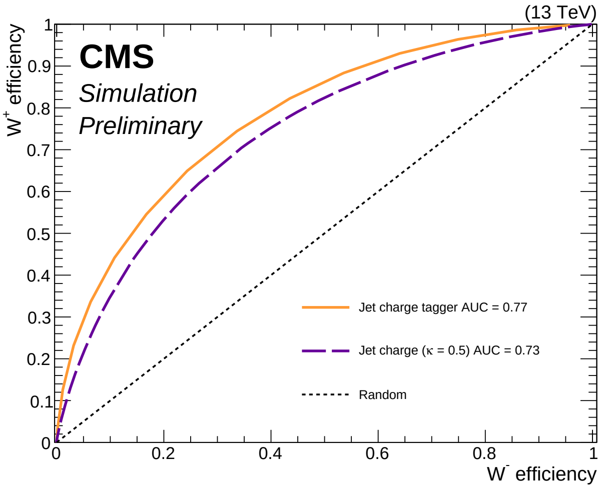
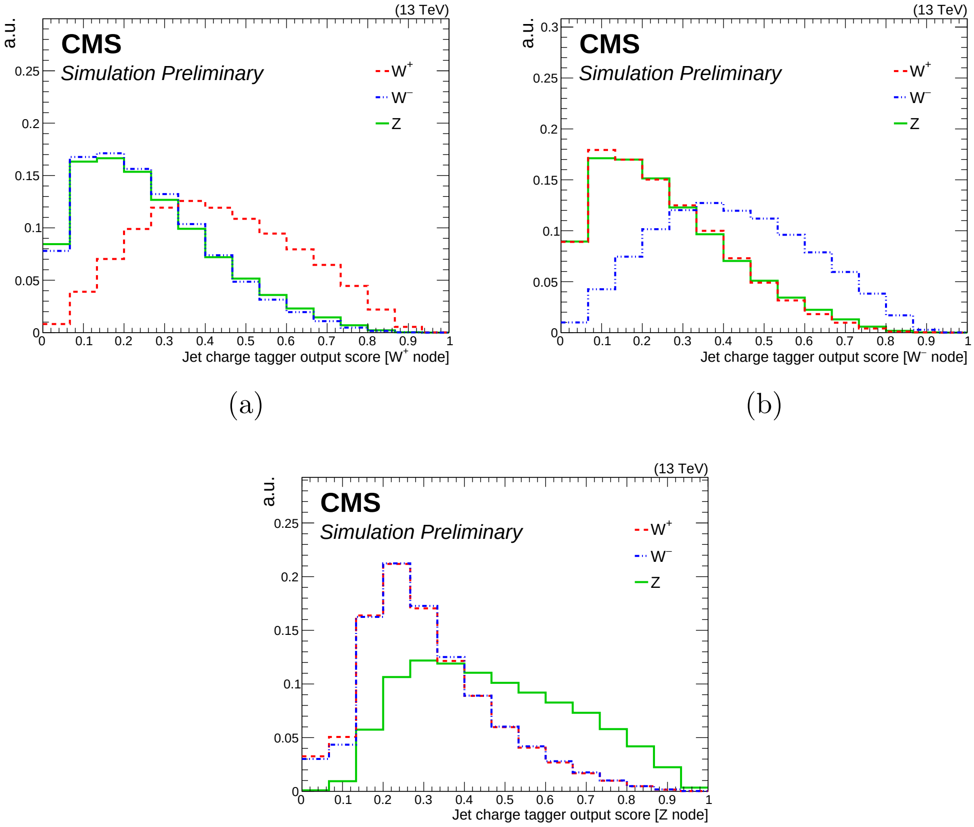


Figure 7.11: Comparison of the jet charge tagger and the traditional jet charge variable, with the best performing κ value for distinguishing between W^+ and W^- jets. The jet charge tagger demonstrates superior performance, significantly enhancing the efficiency of W^+ and W^- jet identification. This figure has been published in Reference [131].

7.6.5. Ternary classification

For the ultimate classification task between positive, negative, and neutral boosted jets, the tagger is trained on a combined sample consisting of boosted jets originated from W^+, W^-, and Z bosons. The tagger in the ternary classification gives three output probabilities for each jet to be classified as W^+, W^-, or Z -like jet. The tagger output scores for each of the output categories is shown in Figure 7.12. For all the output categories, the jet charge tagger performs very well to correctly predict the true jet class.



(a)

(b)

(c)

Figure 7.12: Jet charge tagger output score in the ternary classification task on each output category, W^+ shown in (a), W^- shown in (b), and Z shown in (c). The distributions are split based on the true electric charge of the jet. For each output category, the tagger is performing very well to predict the true charge. This figure has been published in Reference [131].

In order to quantify the tagger performance in the ternary classification task using efficiency curves, we need to reduce the task into binary classification. This can be done in two ways: One-vs-All scheme or One-vs-One scheme. In the One-vs-All scheme, one category is treated as a signal and the remaining two categories are treated as backgrounds. While in the One-vs-One scheme, we make unique pairs of all categories and compare them one-to-one. We have used both of these schemes to calculate the efficiency curves of the tagger, shown in Figure 7.13. The jet charge tagger performs equally well to classify the three categories of jets. The best performance score is for discrimination between W^+ and W^- jets.

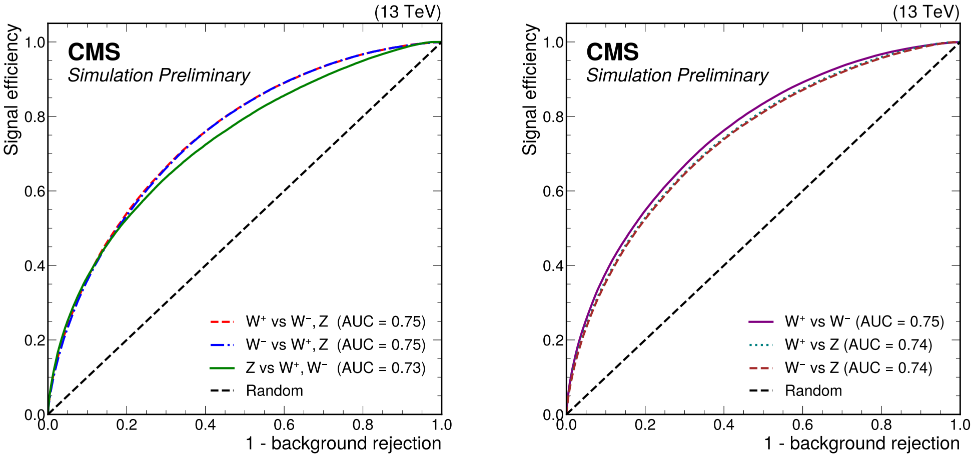
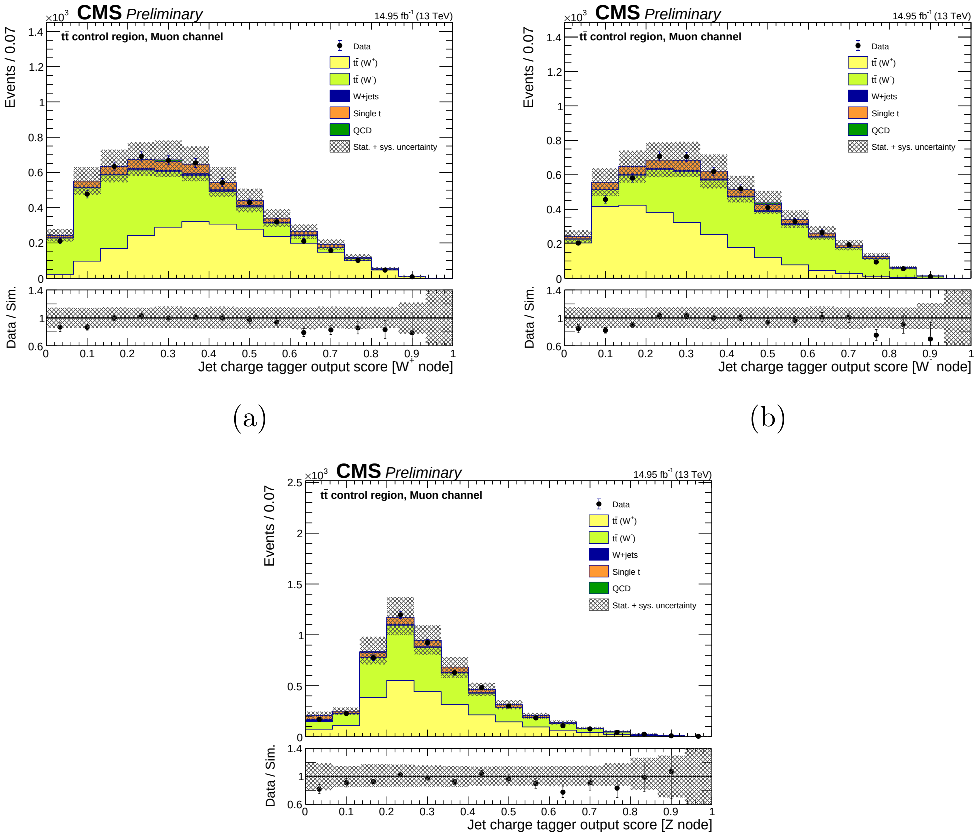


Figure 7.13: Performance of the jet charge tagger in classifying W^+, W^-, and Z jets using the One-vs-All scheme (left) and One-vs-One scheme (right). The legend represents signal vs background for each of the curves in the plots.

Validation in data:

To validate the jet charge tagger performance for ternary classification, we utilize the $t\bar{t}$ control region. One can check the tagger output score in simulation and in data for all three output categories. Since, we do not expect boosted Z jets in this control region, the tagger output score for the Z node must reject most of the events in this region. This will serve as a validation of the tagger performance, as a ternary classifier using the real data. Figure 7.14 shows the validation plots for each of the tagger output score, comparing its performance in data and simulation.



(a)

(b)

(c)

Figure 7.14: Jet charge tagger output score in the ternary classification task on each output category, W^+ shown in (a), W^- shown in (b), and Z shown in (c), in the $t\bar{t}$ control region. Various contributions from simulated backgrounds are shown. The $t\bar{t}$ simulation is split based on the true electric charge of the jet (which is inferred by the charge of the lepton). The performance of the tagger is similar in data and in simulation. The tagger output score at the Z node, shown in (c), is particularly interesting. Most of the events are pushed towards the lower values of the Z node output score, since in this region we do not have boosted Z jets. This figure has been published in Reference [131].

7.7. Systematic uncertainties

The jet charge tagger is trained using simulated events, which are affected by various theoretical and experimental uncertainty sources. In principle, the systematic sources can change both the event yields and shapes of kinematic observables. The uncertainty with the largest impact comes from the choice of factorization scale and renormalization scale used for simulating events. Simulated events are reweighted by alternative event weights, where renormalization and factorization scales are varied up and down by a factor 2, to estimate its impact on the jet charge and jet charge tagger outputs. A total of 6 scale variations are considered, the opposite variations are not included. The final estimate of the scale uncertainty is made by employing the envelope method. This uncertainty only has a normalization effect.

The uncertainty source with the second highest impact comes from the jet energy scale and resolution variations. Separate simulated events are produced by varying the boosted jet energy and resolution up and down by one standard deviation. The jet charge tagger, trained using nominal jet energy scale and resolution corrections, is evaluated on the varied samples to assess the effect relative to the nominal ones on the output.

The uncertainty related to the parton shower modelling is evaluated by means of event reweighting by weights calculated by PYTHIA8. Reweighting factors are applied at the per-event level, which correspond to variations in the ISR and FSR scales, where the scales are varied up and down by a factor of 2 relative to their nominal values. This uncertainty source is the third leading source of systematic uncertainty relevant for charge tagging. The relative impact of the systematic uncertainty sources in percentage are summarized in Table 7.7.

Uncertainty source	Relative percentage
QCD factorization and renormalization scale	13
Jet energy corrections	0.8
Parton shower	0.7

Table 7.7: The relative impact of the systematic uncertainty sources considered in the charge tagging study are shown in percentage.

7.8. Applications of jet charge tagger

The jet charge tagger is a powerful tool that can be utilized in various CMS analyses. It is the first machine-learning based algorithm that distinguishes between the jets originating from the same kind of particle but with opposite electric charge. One possible application of the jet charge tagger is under discussion in this thesis, i.e. in the search of same-sign WW scattering in the semi-leptonic decay channel. The jet charge tagger output together with the charge of the lepton can be used to efficiently separate different VBS processes, such as same-sign WW, opposite-sign WW, and WZ, which are otherwise indistinguishable in this final state.

One exciting application of the jet charge tagger lies in the search for a doubly charged Higgs boson, predicted by the Georgi-Machacek model [24], in vector boson scattering, shown in Figure 7.15. At present, this search is predominantly limited to fully leptonic final states (final state with two same-sign leptons), as identifying the charge of vector bosons in other decay channels is currently not possible. The jet charge tagger could overcome this limitation, enabling the inclusion of additional final states, such as semi-leptonic or fully hadronic.

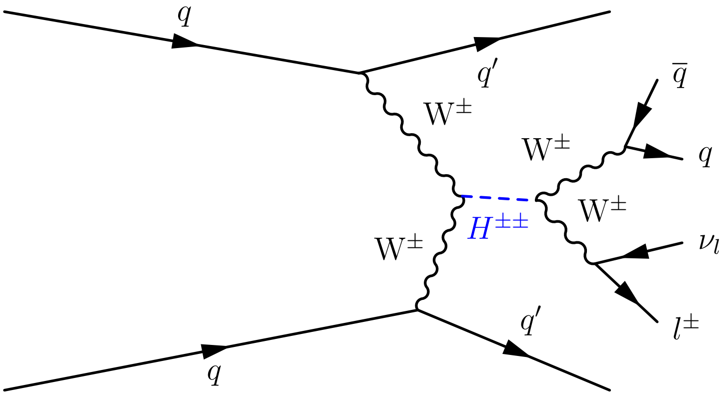


Figure 7.15: Feynman diagram for VBS mediated by a doubly charged Higgs boson, which decays 100% of the times into two pairs of same-sign W bosons. The existence of doubly charged Higgs boson can significantly alter the cross section of same-sign WW scattering. The jet charge tagger can be employed to select the same-sign WW final state in semi-leptonic final state, which provides larger branching fractions compared to the fully leptonic final state that is typically accessible.

8. A search for same-sign WW scattering in the semi-leptonic decay channel

The analysis presented in this chapter is the first search for same-sign WW scattering events in the semi-leptonic decay channel. The primary goal of the analysis is to measure the cross section of the same-sign WW VBS process. The physics motivation for this search, along with the distinctive experimental signature of the signal process, has been discussed in Section 2.3.

This chapter summarizes the overall analysis strategy, including the estimation of major background contributions and the techniques employed to isolate the signal. Finally, the results of the statistical analysis, detailing both the expected and observed signal significances, are presented. Upper limits are set on the cross section times the branching fraction of the same-sign WW VBS process.

8.1. Analysis strategy

We define a phase space enriched in the VBS signal process through specific object selection criteria. The VBS signal region receives significant contributions from background processes, primarily W+jets and $\mathrm{t\bar{t}}$, even after applying optimal signal selection criteria. This is largely due to the substantially higher cross sections of these background processes compared to the VBS signal. With an integrated luminosity of 138 fb^{-1} collected by the CMS experiment during Run 2, the total expected yield of same-sign WW VBS events is approximately 17,000. In contrast, W+jets and $\mathrm{t\bar{t}}$ production, are about five orders of magnitude more abundant, posing a significant challenge for signal extraction. To estimate the contributions from these backgrounds, we use dedicated control regions that are orthogonal to the signal region and are enriched in a specific background process.

To enhance the discrimination of the VBS signal from overwhelming background processes, a multivariate analysis is performed using several kinematic observables. A feed-forward Deep Neural Network, referred to as the “WV-DNN”, is developed to perform binary classification between VBS processes and major standard model backgrounds, including W+jets, $t\bar{t}$ production, non-resonant diboson production (QCD-VV), and single top quark production. The WV-DNN effectively suppresses the bulk of background events, enabling the extraction of a phase space enriched in VBS processes, designated as the high-purity VBS signal region.

The high-purity VBS signal region is enriched in all three VBS processes, namely same-sign WW, opposite-sign WW, and WZ VBS, together called as “WV VBS” signal, where “V” can be a W boson or a Z boson decaying into hadrons. These VBS processes are generally indistinguishable, as the charge of a boson decaying hadronically is not trivial to reconstruct. To effectively distinguish them, we employ the jet charge tagger, described in Chapter 7, on jet level in the high-purity VBS signal region. Based on the jet charge tagger output and the charge of the lepton, the VBS signal events are categorized into same-sign WW, opposite-sign WW, and WZ VBS signal regions, for the first time in this decay channel, making this analysis unique. A schematic representing the application of the jet charge tagger in the event selection chain is shown in Figure 8.1.

Finally, a template-based statistical analysis is conducted to evaluate the significance of the electroweak same-sign WW VBS signal. Upper limits are set on the product of the scattering cross section and the branching fraction of same-sign WW VBS in the semi-leptonic decay channel.

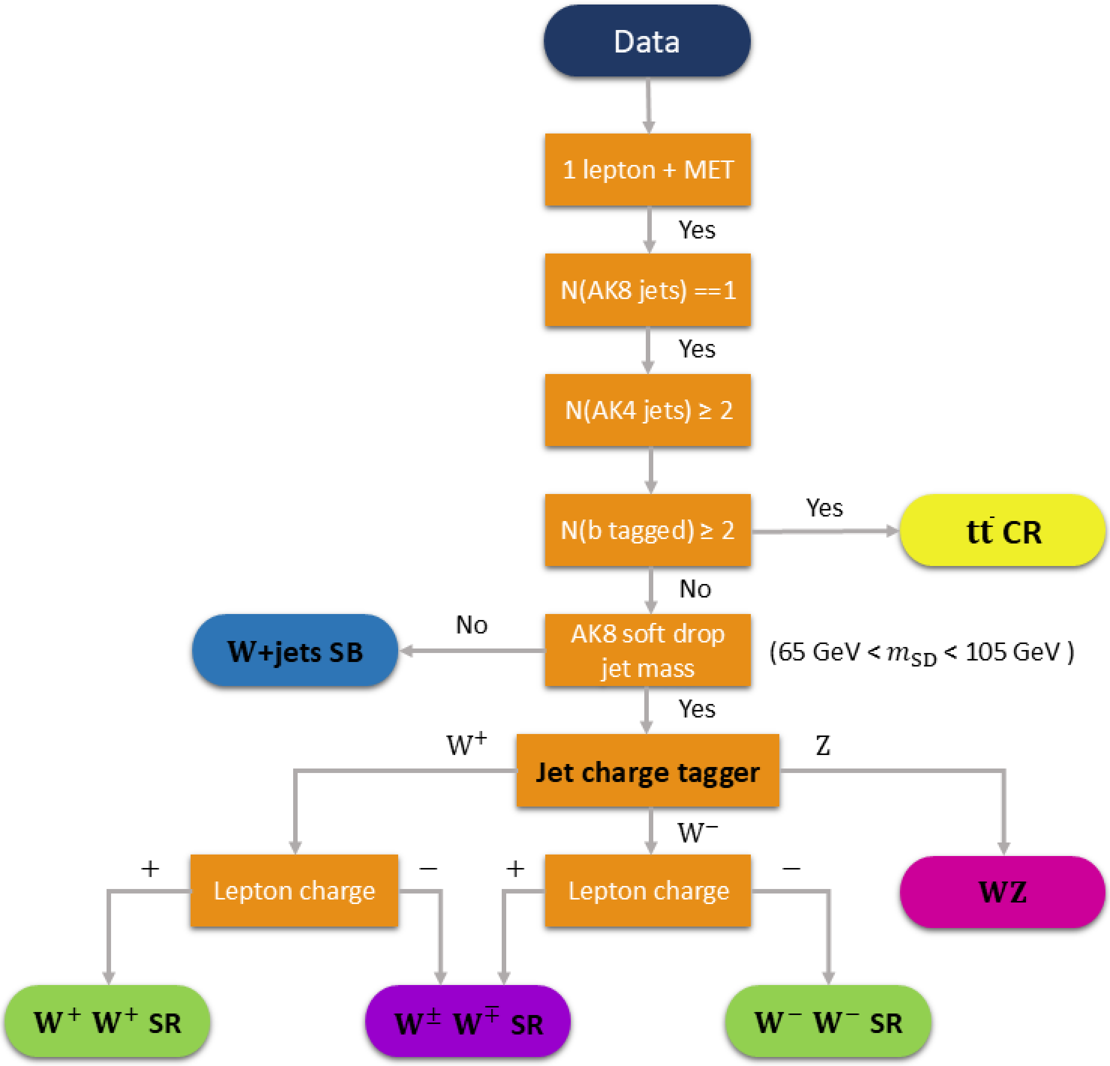


Figure 8.1: A schematic representation showing deployment of the jet charge tagger in the analysis chain to categorize the VBS signal region into same-sign WW ($W^{\pm}W^{\pm}$), opposite-sign WW ($W^{\pm}W^{\mp}$), and WZ VBS.

8.2. Data and simulated samples

8.2.1. Data samples

In this analysis, the data collected from proton-proton collisions during 2016–2018 data taking period with the CMS detector at a centre-of-mass energy of 13 TeV has been analysed. Only data that passed the quality certification by all detector subsystems is used, using the following golden JSON files:

- `Cert_271036-284044_13TeV_Legacy2016_Collisions16_JSON.txt`
- `Cert_294927-306462_13TeV_UL2017_Collisions17_GoldenJSON.txt`
- `Cert_314472-325175_13TeV_Legacy2018_Collisions18_JSON.txt`

The global tag used for the data is `106X_dataRun2_v37`. The analysis aims to reconstruct single lepton events, therefore datasets corresponding to the single lepton triggers have been utilized. The HLT paths used to select single lepton events for each year are listed in Tables 8.1, 8.2 and 8.3. The full names of the Run 2 datasets are listed in Table D.1.

Table 8.1: Trigger paths for 2016 datasets.

Dataset name	Run range	HLT path
Single Muon	B – H	`HLT_IsoMu2` OR `HLT_IsoTkMu24`
Single Electron	B – H	`HLT_Ele27_WPTight_Gsf`
		`HLT_Ele25_eta2p1_WPTight_Gsf`

Table 8.2: Trigger paths for 2017 datasets.

Dataset name	Run range	HLT path
Single Muon	B – F	`HLT_IsoMu27`
Single Electron	B – F	`HLT_Ele35_WPTight_Gsf`

Table 8.3: Trigger paths for 2018 datasets.

Dataset name	Run range	HLT path
Single Muon	A – D	HLT_IsoMu24
EGamma	A – D	HLT_Ele32_WPTight_Gsf

8.2.2. Simulated samples

Several event generators, introduced in Section 5.8, are used to simulate the signal and background processes. For all processes, the detector response is simulated using a detailed description of the CMS detector, based on the GEANT4 package [111, 112], and the event reconstruction is performed with the same algorithms as used for data. Proton-proton interactions occurring in the same or adjacent bunch crossings (pileup) are included in the simulation samples. Minimum bias events simulated with PYTHIA are used to reweight events so that the pileup distribution matched the data, with an average pileup of $\sim$ 23 for 2016 and $\sim$ 32 for 2017–2018 data.

The VBS EW processes are simulated at leading order (LO) with the MADGRAPH5_aMC@NLO generator, requiring two vector bosons and two quarks in the final state. One vector boson decays hadronically and the other decays leptonically using MADSPIN. For the VBS signal simulation, the dipole shower configuration of PYTHIA8 was turned ON, as it is known to model more accurately the additional hadronic emissions in the VBS events [25]. Separate simulations are produced for $W^{\pm}W^{\pm}$, $W^{\pm}W^{\mp}$, $W^{\pm}Z$, and ZZ VBS, with a generator level selection on jets requiring their p_{T} greater than 10 GeV and dijet mass greater than 100 GeV. The list of all signal samples used with their respective cross section values are listed in Table 8.4.

The W+jets background is modelled using MADGRAPH and PYTHIA8 in bins of HT at LO to increase the number of simulated events. For lower HT values, i.e. less than 70 GeV, the W+jets inclusive sample is used with a selection on parton level HT to cover this region of phase space. The cross sections of the W+jets samples are corrected for the next-to-leading order (NLO) corrections using a k-factor of 1.21.

The $t\bar{t}$ process with semi-leptonic decays is simulated at NLO using POWHEG2.0 and PYTHIA8 generators. It has been shown that the simulation at NLO is not accurate enough to describe the measured spectrum in data. To improve the agreement of the simulated $t\bar{t}$ events with data, events are reweighted using the generator level transverse momentum of the top quark and antitop quark, called as top-p_{T} reweighting method [138], based on predictions at NNLO. For the single top quark production, the s-channel, t-channel, and tW-channel are simulated separately, and their contributions are collectively analysed as single top quark background.

Table 8.4: VBS Monte Carlo simulations used to model signal processes, along with their cross sections. The cross section values are taken at LO from the generator.

Dataset name	Cross section (pb)
WplusToLNuWplusTo2JJJ_dipoleRecoil_EWK_LO_SM_⋆	0.09121
WminusToLNuWminusTo2JJJ_dipoleRecoil_EWK_LO_SM_⋆	0.03353
WplusTo2JWminusToLNuJJ_dipoleRecoil_EWK_LO_SM_⋆	0.9435
WplusToLNuWminusTo2JJJ_dipoleRecoil_EWK_LO_SM_⋆	0.9441
WminusToLNuZTo2JJJ_dipoleRecoil_EWK_LO_SM_⋆	0.1029
WminusTo2JZTo2LJJ_dipoleRecoil_EWK_LO_SM_⋆	0.03058
WplusToLNuZTo2JJJ_dipoleRecoil_EWK_LO_SM_⋆	0.1887
WplusTo2JZTo2LJJ_dipoleRecoil_EWK_LO_SM_⋆	0.05613
ZTo2LZTo2JJJ_dipoleRecoil_EWK_LO_SM_⋆	0.01638
⋆ MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	
2016 preVFP: RunIISummer20UL16MiniAODAPV-106X_mcRun2_asymptotic_preVFP_v8-v1/MINIAODSIM	
2016 postVFP: RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v3/MINIAODSIM	
2017: RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v3/MINIAODSIM	
2018: RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v3/MINIAODSIM	

The QCD initiated production of two gauge bosons with two final state quarks and at least one QCD vertex (which we refer to as QCD-VV) is considered as irreducible background. All EW VBS signal samples are also generated with a QCD vertex at LO to model this background using the MadGraph5_aMC@NLO generator. The interference between the EW and QCD process was reported to be negligible [32] and therefore is not considered in this analysis. All used background samples are listed in Appendix D.

8.3. Event reconstruction

We define a VBS signal region based on a set of object and event selection criteria to reconstruct VBS processes in the semi-leptonic final state. The event reconstruction begins by requiring exactly one tight, isolated muon (electron) with transverse momentum greater than 30 (35) GeV and an absolute pseudorapidity less than 2.4 (2.5). Events containing additional loosely identified muons or electrons with transverse momentum greater than 10 GeV are vetoed. The lepton identification and isolation criteria for both tight and loose requirements are described in Chapter 4.

To account for an undetected neutrino from the leptonic decay of the W boson, a moderate missing transverse momentum ($p_{\mathrm{T}}^{\mathrm{miss}}$) requirement is applied. If the reconstructed lepton is a muon, $p_{\mathrm{T}}^{\mathrm{miss}}$ is required to be greater than 40 GeV. For electrons, a stricter threshold of 90 GeV is imposed to suppress multijet background contamination. This choice is motivated to reject most of the multijet background in the signal region. Additionally, the transverse mass of the leptonically decaying W boson is required to be less than 185 GeV, where the transverse mass is defined as:

$$\mathrm{M}_{\mathrm{T}}^{\mathrm{W}} = \sqrt{2\, p_{\mathrm{T}}(\ell)\, p_{\mathrm{T}}^{\mathrm{miss}}\, (1 - \cos(\Delta\phi(\ell, p_{\mathrm{T}}^{\mathrm{miss}})))}. \tag{8.1}$$

After reconstructing the W boson decay products, jets that overlap with the selected lepton (within a cone of radius equal to the jet radius) are removed to ensure clean event reconstruction and suppress the QCD background.

To identify the hadronically decaying vector boson, the event must contain exactly one AK8 PUPPI jet, defined in Section 4.3.1, with transverse momentum greater than 200 GeV, $|\eta|$ smaller than 2.4, and a softdrop mass between 65 and 105 GeV. To efficiently select boosted jets coming from a W/Z boson, the τ_{21} tagger with its high-purity working point is employed, described in Section 4.3.2. Events with zero or more than one reconstructed AK8 PUPPI jet are vetoed. Furthermore, any overlapping AK4 CHS jets within ΔR $<$ 0.8 of the AK8 PUPPI jet are removed before proceeding to the next selections.

The reconstruction of the two VBS jets, which originate from quarks radiating off vector bosons in the initial state, is performed using the AK4 CHS jet collection, defined in Section 4.3.1. The event must contain at least two AK4 CHS jets with transverse momentum greater than 30 GeV. Since the VBS jets are expected to be highly forward, the pseudorapidity requirement is relaxed to cover the range -5.0 to 5.0. Additionally, a b jet veto is implemented: any AK4 CHS jet with transverse momentum greater than 30 GeV and absolute pseudorapidity less than 2.4 that is b tagged according to the loose working point of the DeepJet algorithm, described in Section 4.4, leads to the rejection of the event. This veto effectively suppresses backgrounds such as top quark antiquark pair production where b jets are expected, whereas in the signal process b jets are not anticipated.

Further, in the event reconstruction process, the VBS topology is targeted by requesting the AK4 CHS jets in the event to have a large invariant mass (greater than 500 GeV) and large pseudorapidity separation (greater than 2.5). If more than two AK4 CHS jets are present in the event, all pairs are tested for the invariant mass requirement, and the pair with the largest invariant mass is selected as VBS jets. In addition, the leading VBS jet is required to have a transverse momentum greater than 50 GeV. A summary of event selections for the VBS signal region is provided in Table 8.5.

In addition to the VBS signal region, two control regions are defined, each designed to enrich and study a specific major background process. The same set of event selections are used for these control regions, except a few changes to make the regions independent and orthogonal to each other as well as to the VBS signal region. The changes made in the event and object selection set for these regions compared to

Table 8.5: Event selections for VBS signal region.

Object	Criteria
Tight isolated lepton	Exactly one, $p_{\mathrm{T}} > 30(35)$ μ(e) GeV, $\eta < 2.4(2.5)$ μ(e)
Additional lepton veto	$p_{\mathrm{T}} > 10(10)$ μ(e) GeV, $\eta < 2.4(2.5)$ μ(e)
$p_{\mathrm{T}}^{\mathrm{miss}}$	$> 40(90)$ μ(e) GeV
AK8 PUPPI jet	Exactly one jet, $p_{\mathrm{T}} > 200$ GeV, $\eta < 2.4$
ΔR(AK8 PUPPI jet, lepton)	> 0.8
B-tag veto	Veto events with AK4 jets $p_{\mathrm{T}} > 30$ GeV, $\eta < 2.4$ with loose working point
AK4 CHS jets	Atleast two, $p_{\mathrm{T}} > 30$ GeV, $\eta < 5.0$
ΔR(AK4 CHS jet, lepton)	> 0.4
ΔR(AK8 PUPPI jet, AK4 CHS jet)	> 0.8
N-subjettiness (τ_{21}) of AK8 PUPPI jet	$< 0.45, 0.45, 0.55, 0.55$ (2016 preVFP, 2016 postVFP, 2017, 2018)
AK8 PUPPI jet mass	65 GeV $< m_{\mathrm{SD}} <$ 105 GeV
VBS selections	Jets pair (max m_{jj}), leading VBS jet $p_{\mathrm{T}} > 50$ GeV, $m_{\mathrm{jj}} > 500$ GeV, $\lvert\Delta\eta\rvert > 2.5$
Leptonically decaying W transverse mass	< 185 GeV

the signal region are listed below:

- **W+jets control region:** The requirement on the softdrop mass of the AK8 jet is inverted in this region. The AK8 jet is requested to have a mass 40 GeV $< m_{\mathrm{SD}} <$ 65 GeV or 105 GeV $< m_{\mathrm{SD}} <$ 250 GeV. The W+jets control region is also referred to as sidebands to the signal region.

- **t$\bar{\text{t}}$ control region:** The t$\bar{\text{t}}$ control region is defined in the same mass interval of the AK8 jet as the VBS signal region, i.e. 65 GeV $< m_{\mathrm{SD}} <$ 105 GeV but requires at least the presence of two b-tagged AK4 jets using the medium working point of the DeepJet algorithm.

A schematic representation of the phase space division into the signal and the control regions is shown in Figure 8.2.

8.4. Validation of control regions

The two control regions, t$\bar{\text{t}}$ and W+jets, defined based on specific event selections described in Section 8.3, serve as a tool to assess the accuracy of the simulation. The validation of these regions is performed using both simulated samples and data, ensuring a robust understanding of the modelling of background contributions.

The first check to verify the reliability of the control regions is to check whether

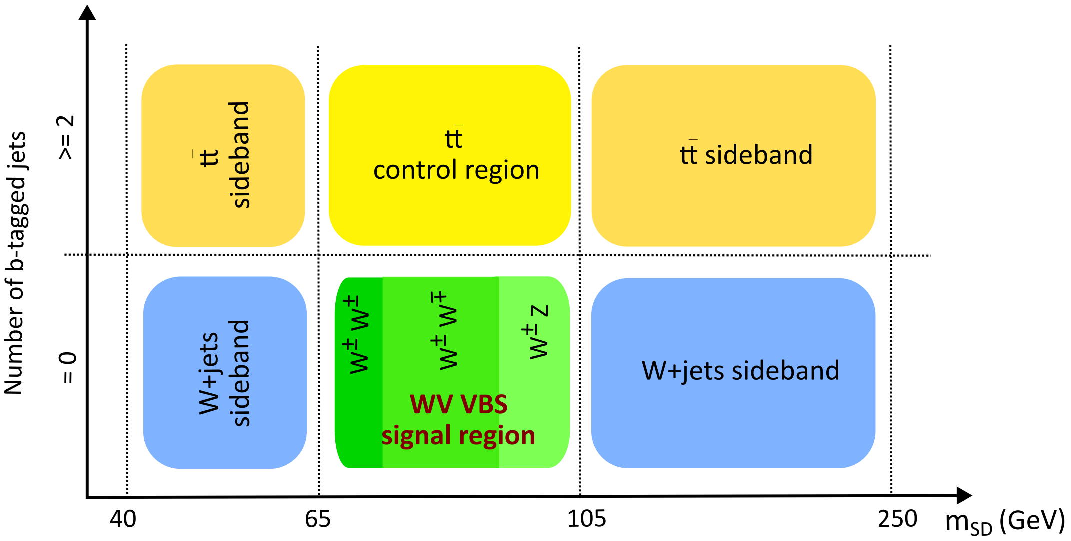


Figure 8.2: Schematic representation of different phase space regions used in the analysis presented in this thesis. The W+jets sidebands (also called together as the W+jets control region) are used to estimate the W+jets process contribution in the signal region. The $t\bar{t}$ control region is used to check the modelling of the $t\bar{t}$ background. While the $t\bar{t}$ sidebands are not used in the analysis.

they are enriched in the specific background and the contribution from the signal process is negligible in the control regions. This can be done by looking at the event yields of different processes predicted using simulations in the control regions. Table 8.6 lists the event yields for various backgrounds and the signal processes in the $t\bar{t}$ control region for different data-taking eras. The yields for the muon and the electron decay mode are quoted separately, and verify that the $t\bar{t}$ control region is enriched in $t\bar{t}$ background, with negligible contribution from the signal. Similarly, the event yields for the W+jets control region are shown in Table 8.7, verifying the enrichment in W+jets events.

In addition to event yields, various kinematic distributions of fundamental observables — such as lepton transverse momentum, missing transverse momentum, jet multiplicity, softdrop mass of the AK8 jet, invariant mass of the two VBS jets etc — are compared between simulation and data in the control regions. These comparisons help evaluate whether the simulations accurately describe the observed distributions and maintain the expected shapes. Significant discrepancies, if present, could indicate potential mismodelling, requiring further investigation or reweighting procedures based on data-driven techniques to improve the agreement. The effect of various systematic uncertainties on overall normalizations, as well as the shapes of these kinematic distributions are assessed as well. Figure 8.3 shows a comparison of softdrop mass of the AK8 jet in the $t\bar{t}$ control region and the W+jets control region for the two decay channels, muon and electron. The shape and normalization of this observable are well modelled in the $t\bar{t}$ control region, showing a good agreement. The same level of agreement has been observed for the other kinematic observables in this region, shown in Appendix E. This establishes our confidence in the use of $\mathrm{t\bar{t}}$ simulation for estimating its contribution in the signal region. However, in the W+jets control region, specially in the muon channel, this variable shows a discrepancy in the lower mass sidebands. The discrepancy is also present in other kinematic distributions in this region, shown in Appendix K. This suggests that the modelling of the W+jets background is not reliable. To improve the agreement in the W+jets control region, special corrections are derived for the W+jets simulation based on a data-driven technique, which is described in Section 8.5.

Table 8.6: Event yields in the $\mathrm{t\bar{t}}$ control region for different processes in different data-taking eras and different decay channel (Muon, Electron) are shown. The quoted uncertainties reflect the statistical limitations of the simulated samples. The dominant process in this control region is the $\mathrm{t\bar{t}}$ production. The contribution from the VBS signal processes is negligible.

Process	2016 preVFP		2016 postVFP		2017		2018	
	Muon	Electron	Muon	Electron	Muon	Electron	Muon	Electron
$\mathrm{t\bar{t}}$	1523.86 ± 8.98	527.66 ± 5.35	1265.84 ± 7.28	438.98 ± 4.32	2884.16 ± 10.32	1030.68 ± 6.17	3632.13 ± 11.82	1316.83 ± 7.08
Single Top	150.55 ± 6.29	52.47 ± 3.79	115.35 ± 4.87	44.35 ± 3.22	280.23 ± 7.59	115.21 ± 4.88	365.33 ± 8.72	143.28 ± 5.57
QCD-VV	9.97 ± 0.88	4.90 ± 0.65	8.61 ± 0.81	2.81 ± 0.43	16.68 ± 1.16	9.33 ± 0.89	26.04 ± 1.70	12.29 ± 1.20
W+jets	25.04 ± 2.03	8.76 ± 1.26	15.66 ± 1.50	7.66 ± 1.16	32.37 ± 2.16	41.80 ± 30.58	43.82 ± 2.71	17.02 ± 1.67
ssWW VBS	0.08 ± 0.01	0.04 ± 0.01	0.04 ± 0.01	0.02 ± 0.00	0.09 ± 0.01	0.05 ± 0.01	0.12 ± 0.02	0.06 ± 0.01
osWW VBS	3.61 ± 0.25	1.53 ± 0.17	3.46 ± 0.25	1.06 ± 0.13	7.66 ± 0.36	2.89 ± 0.22	9.99 ± 0.50	3.68 ± 0.29
WZ VBS	0.85 ± 0.05	0.36 ± 0.03	0.71 ± 0.04	0.26 ± 0.03	1.62 ± 0.07	0.62 ± 0.04	1.96 ± 0.09	0.91 ± 0.06
ZZ VBS	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.03 ± 0.00	0.01 ± 0.00
Data	1792.00	559.00	1607.00	546.00	3234.00	1095.00	4118.00	1471.00
Total MC	1713.98 ± 11.18	595.73 ± 6.71	1409.66 ± 8.93	495.14 ± 5.53	3222.84 ± 13.05	1200.59 ± 31.59	4079.43 ± 15.04	1494.08 ± 9.25

8.5. Data-driven W+jets background estimation

To correct for the disagreements observed in the W+jets control region, described in Section 8.4, we utilize a data-driven method. The method uses data from the W+jets control region to correct for the modelling deficiencies, improving the reliability of the W+jets background estimate. Instead of applying a global correction, the W+jets simulation is split into subcategories using the bins of leptonically decaying W boson transverse momentum and the normalization of the simulation is independently corrected for each bin. The binning scheme used is detailed in Table 8.8. The transverse momentum of the leptonically decaying W boson is calculated as:

$$p_{\mathrm{T}}{}^{\mathrm{W_{lep}}} = \sqrt{(p_{\mathrm{T}}^{\ell} \cos\phi_{\ell} + p_{\mathrm{T}}^{\mathrm{miss}} \cos\phi_{\mathrm{miss}})^2 + (p_{\mathrm{T}}^{\ell} \sin\phi_{\ell} + p_{\mathrm{T}}^{\mathrm{miss}} \sin\phi_{\mathrm{miss}})^2}, \tag{8.2}$$

Table 8.7: Event yields in the W+jets control region for different processes in different data-taking eras and different decay channel (Muon, Electron) are shown. The quoted uncertainties reflect the statistical limitations of the simulated samples. The dominant process in this control region is the W+jets production. The contribution from the VBS signal processes is negligible.

Process	2016 preVFP		2016 postVFP		2017		2018	
	Muon	Electron	Muon	Electron	Muon	Electron	Muon	Electron
$\mathrm{t\bar{t}}$	746.30 ± 6.00	304.08 ± 3.88	743.14 ± 6.16	306.35 ± 3.97	1147.94 ± 6.67	500.87 ± 4.40	1441.38 ± 7.62	626.66 ± 4.99
Single top	99.08 ± 4.50	40.11 ± 3.04	84.17 ± 4.12	35.49 ± 2.80	111.23 ± 4.47	59.94 ± 3.52	166.84 ± 5.55	81.19 ± 4.03
QCD-VV	122.22 ± 2.75	53.04 ± 1.89	127.67 ± 3.29	56.83 ± 2.27	209.28 ± 3.85	92.45 ± 2.63	265.68 ± 5.13	115.29 ± 3.44
W+jets	2757.79 ± 29.73	802.48 ± 13.89	2383.85 ± 28.45	730.11 ± 14.00	4365.18 ± 46.80	1340.59 ± 16.80	5319.27 ± 51.04	1939.05 ± 21.70
ssWW VBS	3.98 ± 0.07	1.68 ± 0.05	3.57 ± 0.07	1.73 ± 0.05	6.51 ± 0.10	2.98 ± 0.06	8.64 ± 0.13	3.96 ± 0.09
osWW VBS	9.91 ± 0.39	4.57 ± 0.27	11.08 ± 0.47	4.88 ± 0.31	18.66 ± 0.56	8.19 ± 0.37	23.67 ± 0.77	11.61 ± 0.54
WZ VBS	4.85 ± 0.11	1.97 ± 0.07	4.58 ± 0.12	1.93 ± 0.08	8.45 ± 0.15	3.78 ± 0.10	10.96 ± 0.20	5.24 ± 0.14
ZZ VBS	0.25 ± 0.01	0.06 ± 0.00	0.21 ± 0.01	0.06 ± 0.00	0.44 ± 0.01	0.12 ± 0.01	0.59 ± 0.02	0.17 ± 0.01
Data	3533.00	1208.00	3123.00	1083.00	5555.00	1921.00	6795.00	2570.00
Total MC	3744.40 ± 30.78	1207.99 ± 14.86	3358.28 ± 29.59	1137.39 ± 14.99	5867.69 ± 47.64	2008.92 ± 17.92	7237.01 ± 52.16	2783.16 ± 22.89

where p_{T}^{ℓ} and ϕ_{ℓ} are the transverse momentum and azimuthal angle of the lepton, and $p_{\mathrm{T}}^{\mathrm{miss}}$ and ϕ_{miss} are the missing transverse momentum and its azimuthal angle. This variable shows an evident trend of discrepancy between data and simulation, specially for the lower momentum values. Figure 8.4 shows the prefit distributions of this variable in the muon and the electron decay channel. We expect the corrections to be different in the two decay channels, as the missing transverse momentum requirement is different depending on the lepton flavour.

Table 8.8: Bins of transverse momentum of the leptonically decaying W boson (${p_{\mathrm{T}}}^{\mathrm{W_{lep}}}$), defined in Eq. 8.2 used to correct the W+jets simulation using the W+jets control region data.

Bin	${p_{\mathrm{T}}}^{\mathrm{W_{lep}}}$ bin
1	${p_{\mathrm{T}}}^{\mathrm{W_{lep}}} < 50$ GeV
2	$50 \leq {p_{\mathrm{T}}}^{\mathrm{W_{lep}}} < 100$ GeV
3	$100 \leq {p_{\mathrm{T}}}^{\mathrm{W_{lep}}} < 150$ GeV
4	$150 \leq {p_{\mathrm{T}}}^{\mathrm{W_{lep}}} < 200$ GeV
5	$200 \leq {p_{\mathrm{T}}}^{\mathrm{W_{lep}}} < 300$ GeV
6	$300 \leq {p_{\mathrm{T}}}^{\mathrm{W_{lep}}} < 400$ GeV
7	${p_{\mathrm{T}}}^{\mathrm{W_{lep}}} \geq 400$ GeV

The leptonically decaying W boson transverse momentum distributions are fit using the CMS COMBINE tool [139]. During the fit, all subcategories of the W+jets

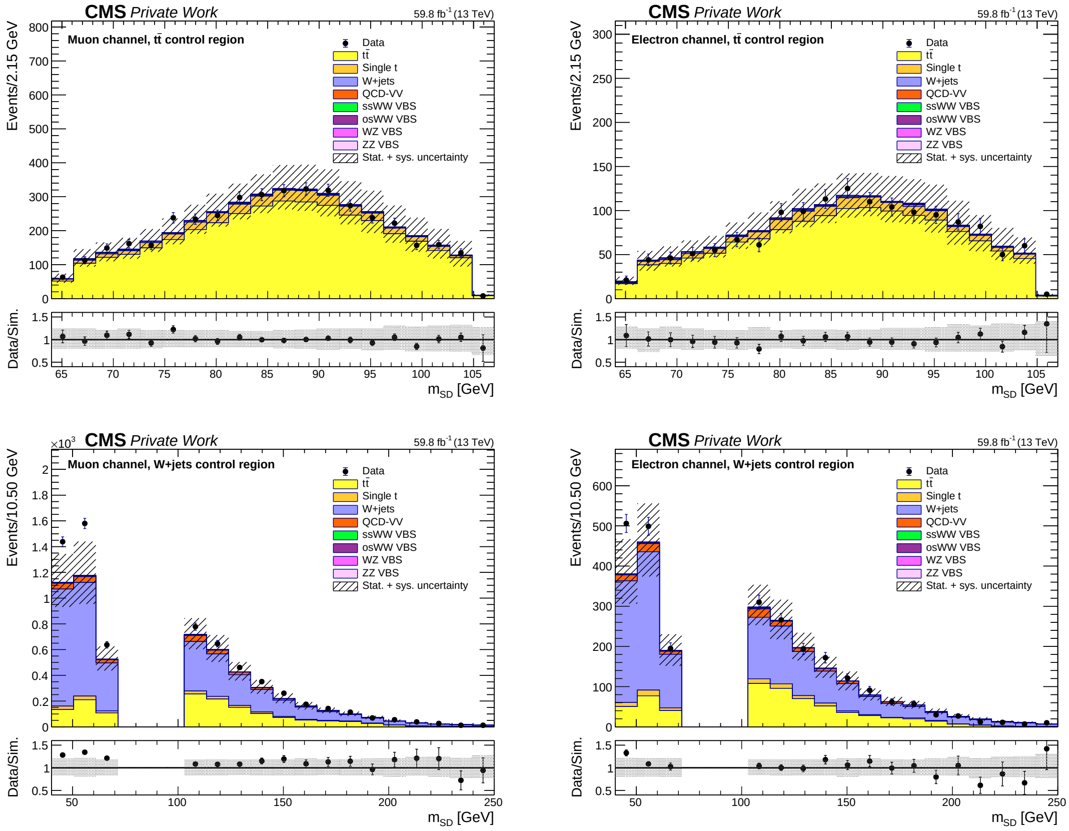


Figure 8.3: Distributions of the softdrop mass of the AK8 jet (m_{SD}) in the $\text{t}\bar{\text{t}}$ control region (top) and the W+jets control region (bottom) for the muon decay channel (left) and the electron decay channel (right) are shown. The $\text{t}\bar{\text{t}}$ control region shows a good agreement between the simulation and data. For the W+jets control region, the lower mass sideband shows discrepancy between the data and simulation, suggesting that the simulation is not well modelled in this region. The distributions are from 2018 reconstruction year. Other years are shown in Appendix E and K.

simulation are scaled to match the data in each bin. After the fit, correction factors are derived as a ratio between postfit over prefit yields for each bin of this variable, in both decay channels. They are shown in Figure 8.5. The corrections factors are applied to the W+jets simulation as additional event weights.

To visualize the effect of the corrections applied as additional event weights for W+jets simulation, different kinematic observables are checked again to compare data and simulation in the W+jets control region. We have observed that the agreement is improved for all kinematic observables in this region after applying these correction weights, see Figure 8.6 for a few of the observables, more are shown in Appendix F. To estimate the W+jets process contribution in the signal region, the same correction strategy is applied. To have a robust estimate, instead of applying these correction weights by hand, we include the W+jets control region directly in the final fit. This ensures that the systematic uncertainties associated

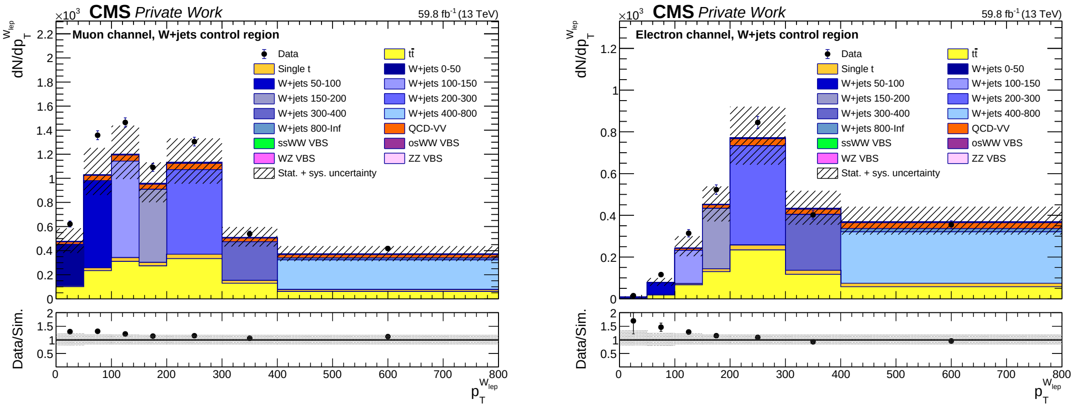


Figure 8.4: Distributions of the leptonically decaying W boson transverse momentum ($p_{\mathrm{T}}{}^{\mathrm{W_{lep}}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region are shown. This kinematic observable shows evident trend of data and simulation discrepancy. The discrepancy is more pronounced in the muon channel, where the missing transverse momentum requirement is much lower than the electron channel. These distributions are from 2018 reconstruction year.

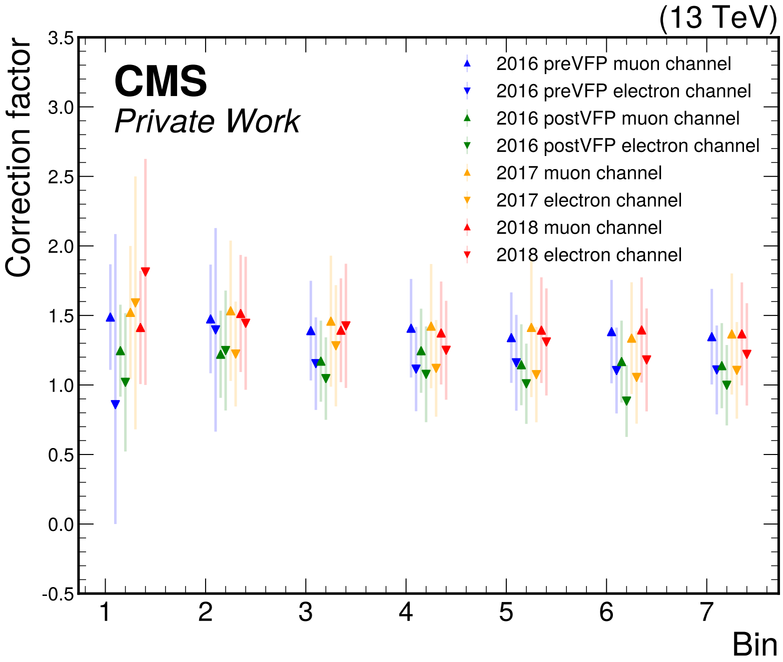


Figure 8.5: Correction factors shown are derived as a ratio between postfit and prefit yields of the leptonic W transverse momentum for the muon and the electron channel for each reconstruction era. These correction factors are applied to visualize and check for the agreement in the W+jets control region and the VBS signal region. In the final fit, the W+jets control region is directly used to incorporate these corrections in a robust way.

with these corrections are propagated to the signal region within the fit, ensuring a robust estimate of the W+jets background contribution.

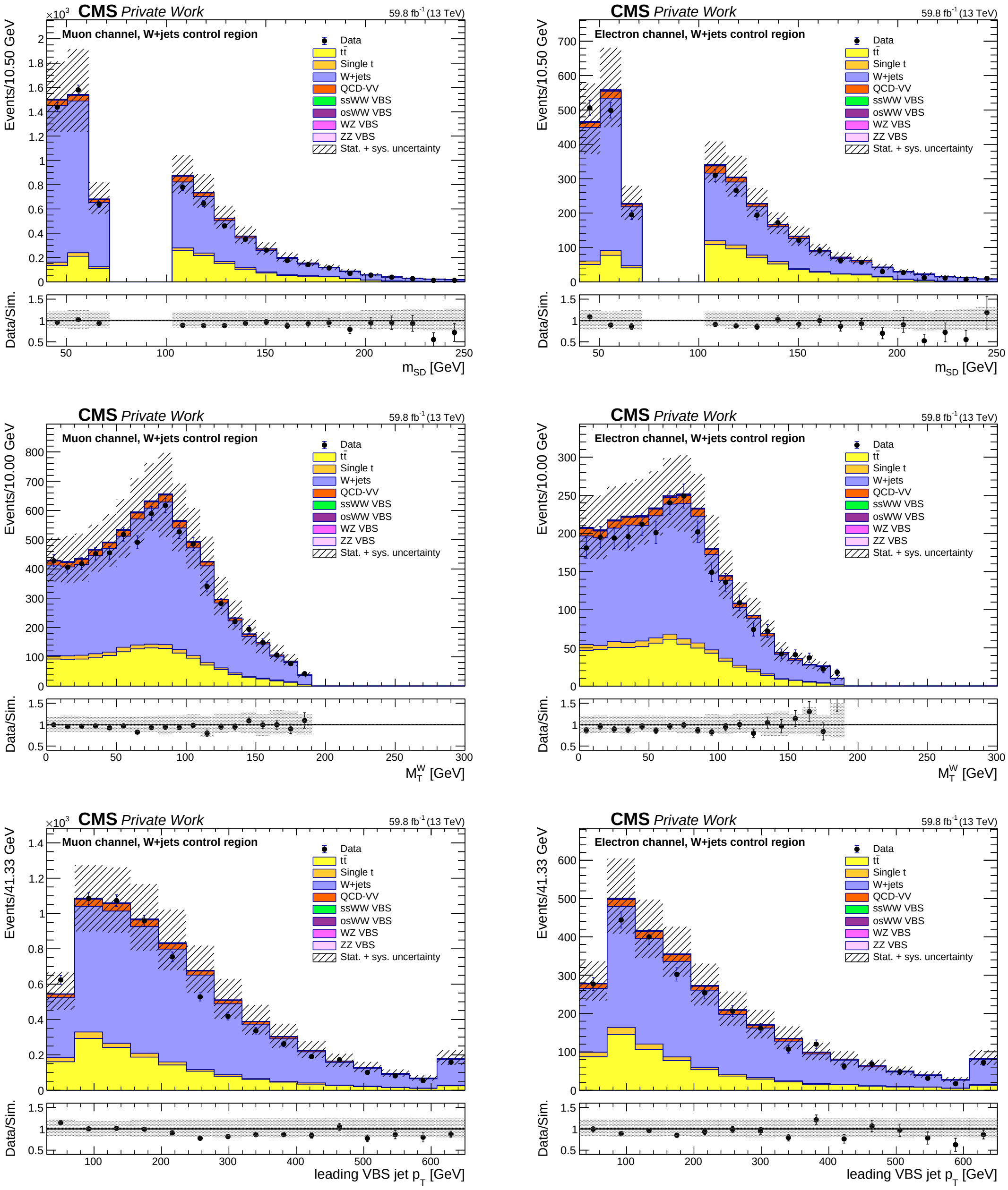


Figure 8.6: Various kinematic distributions in the W+jets control region for the muon decay channel (left) and the electron decay channel (right) are shown after applying the correction factors to reweight events in the W+jets simulation. A nice agreement is observed between the simulation and data following the data-driven corrections. The distributions shown are from 2018 reconstruction year.

8.6. Multivariate analysis

This section outlines the multivariate analysis designed to enhance the discrimination between various background processes: W+jets, $\mathrm{t\bar{t}}$, QCD-VV, and single top quark production, and the VBS signal. A dedicated feed-forward deep neural network, is trained using a set of carefully selected kinematic observables in the VBS signal region. The goal is to optimize the separation between signal and background, thereby enriching the phase space with events characteristic of VBS processes.

8.6.1. Inputs

Several kinematic variables showing distinctive features of the VBS signal with respect to backgrounds have been chosen to train the WV-DNN discriminator, listed in Table 8.9. The Zeppenfeld variable, used as one of the input variables, is defined as follows:

$$\mathrm{Z}^{x} = \frac{\left|\eta^{x} - \frac{\eta_{j_1}+\eta_{j_2}}{2}\right|}{|\eta_{j_1} - \eta_{j_2}|}, \tag{8.3}$$

where x can be a lepton or a AK8 jet, j_1 and j_2 are the two VBS jets. All these variables show differences in their distribution between the VBS signal and backgrounds, but none of them alone can be used as a powerful discriminant. The WV-DNN combines the information from all of them and learns from their correlations to build a discriminator output to differentiate the VBS signal from background processes. The distributions, in particular the shapes of the most important kinematic observables for the signal and background processes, are shown in Figure 8.7.

The correlation of various input variables pairs are estimated using the correlation matrix for both signal and background events, shown in Figure 8.8. Differences in correlation structure between signal and background plays a key role in event classification, which the DNN can exploit.

Table 8.9: List of kinematic observables used for the training of the WV-DNN discriminator.

Variable	Description
No. of AK4 jets	Number of AK4 jets with transverse momentum > 30 GeV
m_{jj}	Invariant mass of the two VBS jets
$\Delta\eta_{\mathrm{jj}}$	Pseudorapidity difference between the two VBS jets
VBS jet 1 p_{T}	Transverse momentum of the leading VBS jet
VBS jet 2 p_{T}	Transverse momentum of the trailing VBS jet
Lepton p_{T}	Transverse momentum of the lepton
Lepton η	Lepton pseudorapidity
AK8 jet p_{T}	Transverse momentum of the AK8 jet
m_{SD}	Soft drop mass of the AK8 jet
$\mathrm{Z}^{\mathrm{lep}}$	Zeppenfeld variable for the lepton
$\mathrm{Z}^{\mathrm{AK8\ jet}}$	Zeppenfeld variable for the AK8 jet

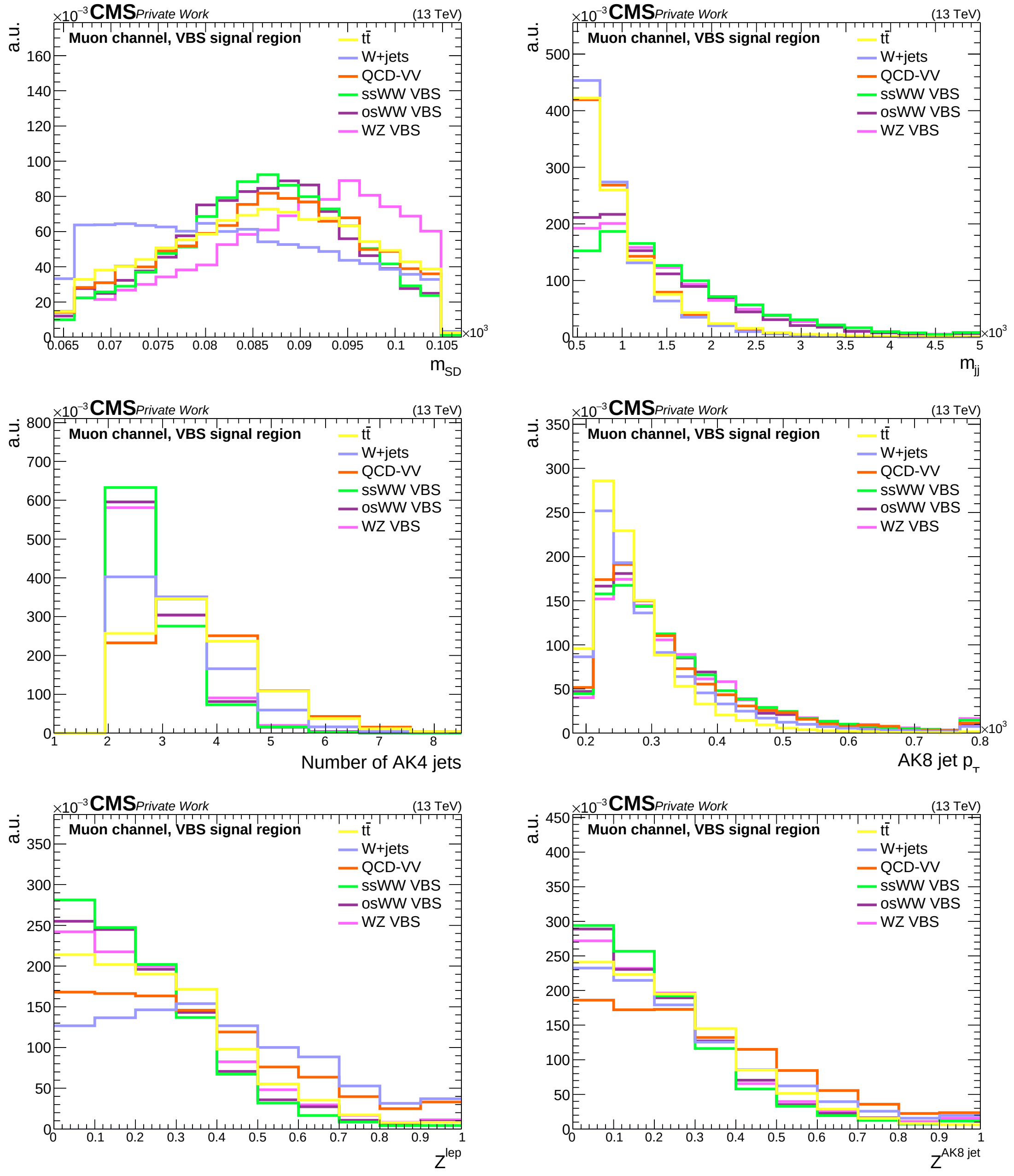


Figure 8.7: The top most discriminating kinematic observables between the VBS signal processes (same-sign WW, opposite-sign WW, and WZ) and major background processes. Only the muon decay channel is shown. The distributions are similar for the electron channel. All distributions are normalized to unity to compare their shapes.

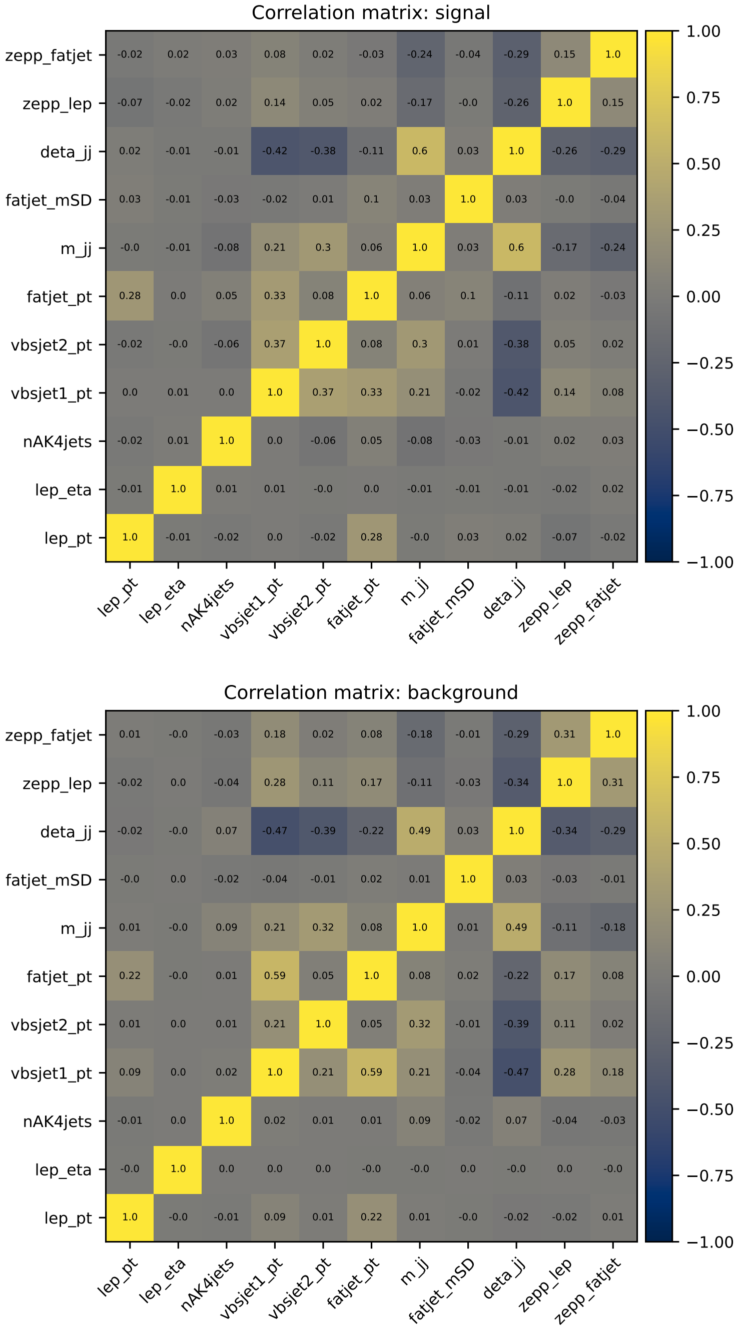


Figure 8.8: The correlation matrices of the different input variable pair for signal events (top) and background events (bottom) are shown. Some pair of variables have different correlation in signal and background, are utilized by the WV-DNN to construct a powerful multivariate discriminant.

8.6.2. Architecture and training strategy

The WV-DNN is constructed using the TensorFlow/Keras library. The architecture of the network is chosen to be as follows:

1. Input: 11 variables, as listed in Table 8.9
2. Layer 1:
   - 64 nodes
   - L2 weights regularization
   - Batch normalization
   - Dropout layer (rate = 0.2)
   - ReLU activation
3. Layer 2:
   - 32 nodes
   - Batch normalization
   - Dropout layer (rate = 0.2)
   - ReLU activation
4. Layer 3:
   - 32 nodes
   - Batch normalization
   - Dropout layer (rate = 0.2)
   - ReLU activation
5. Layer 4:
   - 32 nodes
   - Batch normalization

  - Dropout layer (rate = 0.2)
  - ReLU activation

6. Ouput:
  - 1 node
  - Sigmoid activation

The network is designed for a binary classification task between VBS processes and SM background ($t\bar{t}$, W+jets, single top quark production, QCD-VV), therefore a binary cross-entropy loss function is used to optimize its performance. The architecture employs the ReLU activation function for all hidden layers, while the sigmoid activation function is applied to the output layer to produce a probability score.

To enhance the generalization ability of the network, batch normalization is applied before the activation function in each hidden layer. This technique helps regularize the network and stabilize the training process. Additionally, L2 regularization is implemented on the weights of the first layer, which aids in preventing overfitting and improves the robustness of the model.

A dropout rate of 0.2 is incorporated after every layer, further regularizing the network by reducing the likelihood of neuron co-adaptation. The architecture consists of three hidden layers: the first with 64 nodes, followed by two layers of 32 nodes each.

The network uses the Adam optimizer, which adapts the learning rate for each weight dynamically during training, leading to faster convergence and effective optimization. To ensure stable and efficient learning, the learning rate is gradually decreased during the training process.

To prevent overtraining, early stopping is implemented. Training is terminated if the validation loss does not improve by less than 0.0001 after 10 consecutive epochs. This ensures the model stops training when no significant improvement is observed, balancing performance and generalization.

To train the WV-DNN effectively, both signal ($\sim$ 0.1 million) and background ($\sim$ 0.4 million) samples are combined into a single dataset with the cross section, selection and scale factor coefficients applied so that the relative weight between the backgrounds is correct. This allows the WV-DNN not to learn to discriminate the signal against a minor background, while losing discrimination power against main backgrounds. To avoid any bias in the training due to imbalance present in the signal and background samples, class weights are applied to the signal samples during training, so that the total number of weighted events of the signal dataset matches the background.

The combined dataset is then split into three equal parts, denoted as A, B, and C, which are used for a 3-fold training strategy. This approach improves the robustness and generalization of the model by ensuring it is exposed to all parts of the data during training and validation.

In each fold, two parts of the dataset are used for training and validation, while the remaining part is reserved for testing. The folds are rotated to ensure that each part is used for validation exactly once. Specifically:

- Parts A and B are used for training and validation, and part C is reserved for testing
- Parts B and C are used for training and validation, and part A is reserved for testing
- Parts C and A are used for training and validation, and part B is reserved for testing

Table 8.10: Cross validation folds distribution

Whole MC	A	B	C
fold 0	Train	Val	Test
fold 1	Test	Train	Val
fold 2	Val	Test	Train

This 3-fold training strategy, summarized in Table 8.10, has several advantages. First, it ensures that the model is evaluated on unseen data during each fold, providing an unbiased estimate of its performance. Secondly, by training on different combinations of the data, the network is exposed to a more diverse range of training examples, which helps to reduce overfitting and increase generalization to new data. This approach is particularly beneficial when the training dataset is not too large. It ensures that the model is trained in a balanced and comprehensive manner, leveraging all available data, and avoids the necessity to remove training events in the statistical inference.

The training is monitored during each fold by visualising model loss and accuracy curves. Consistent training is observed without any signs of overtraining in each fold for all years. The training is done separately for each reconstruction era. The model loss curves for each fold and year are shown in Figure 8.9.

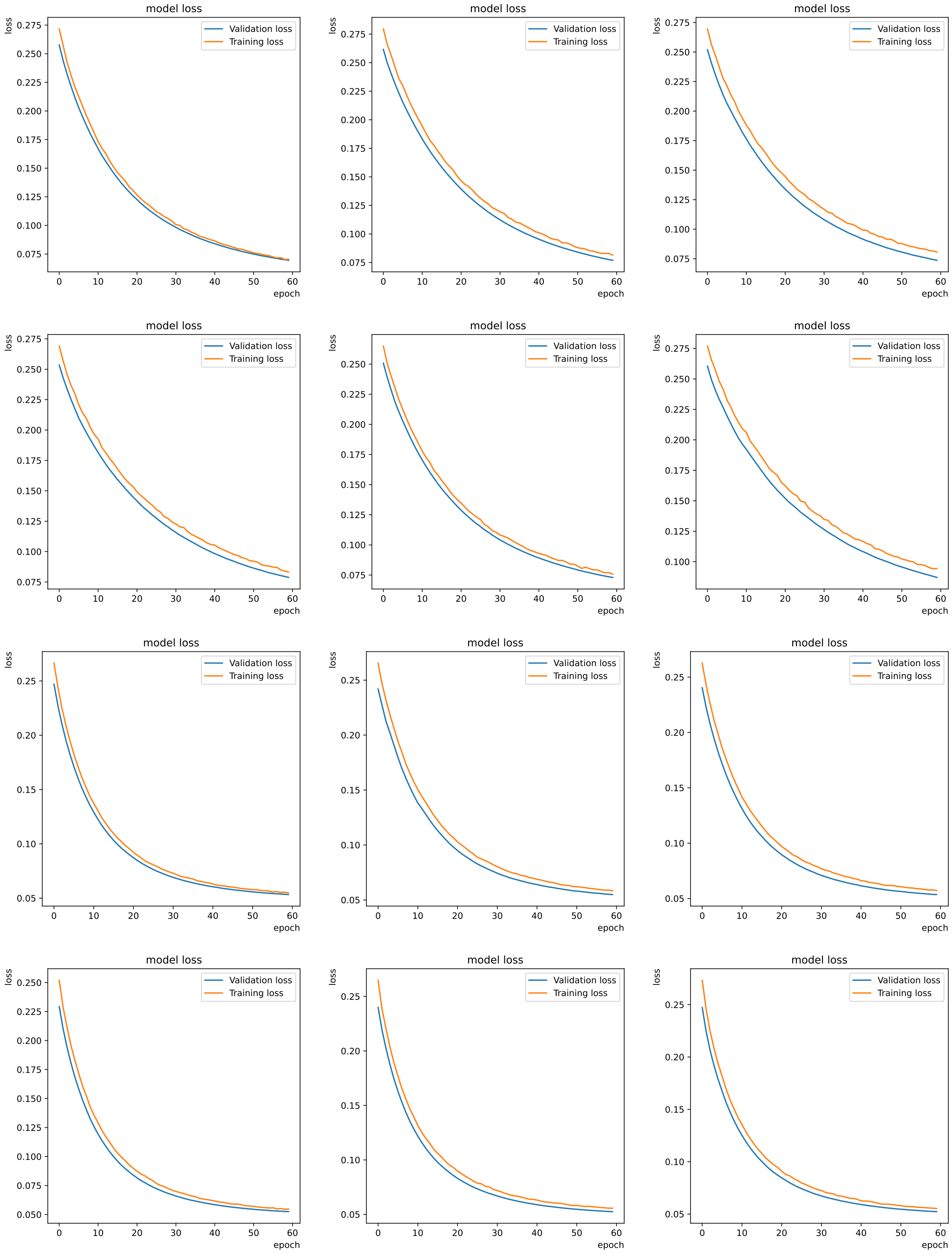


Figure 8.9: WV-DNN model loss curves during training in three folds A, B, and C (from left to right) for the different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 from (top to bottom) are shown.

8.6.3. Performance and validation

The performance of the WV-DNN model is evaluated using an independent test dataset that remains unseen during the training process, ensuring an unbiased assessment. The test dataset for each fold is distinct and corresponds to the portion of the dataset excluded during the training phase of that specific fold. This approach, often referred to as k-fold cross-validation, helps to maximize the utilization of the available data and provides a robust estimate of the model's generalization performance.

When applying the WV-DNN model to variations of simulations and actual data, we use the same splitting strategy and use all three models from each fold training for evaluation. This method ensures that the evaluation is not biased toward any particular fold, while maintaining consistency with the overall training and validation procedure.

To quantify the effectiveness of the model, the performance metric the area under the curve is computed using the training and the test datasets. The ROC curves for different training folds and years are shown in Figure 8.10.

For simulated samples, the agreement between the predicted and true labels is further validated by comparing the model outputs with the known input distributions, ensuring consistency and reliability. The WV-DNN output score is shown in Figure 8.11. The background and signal are split using the true labels.

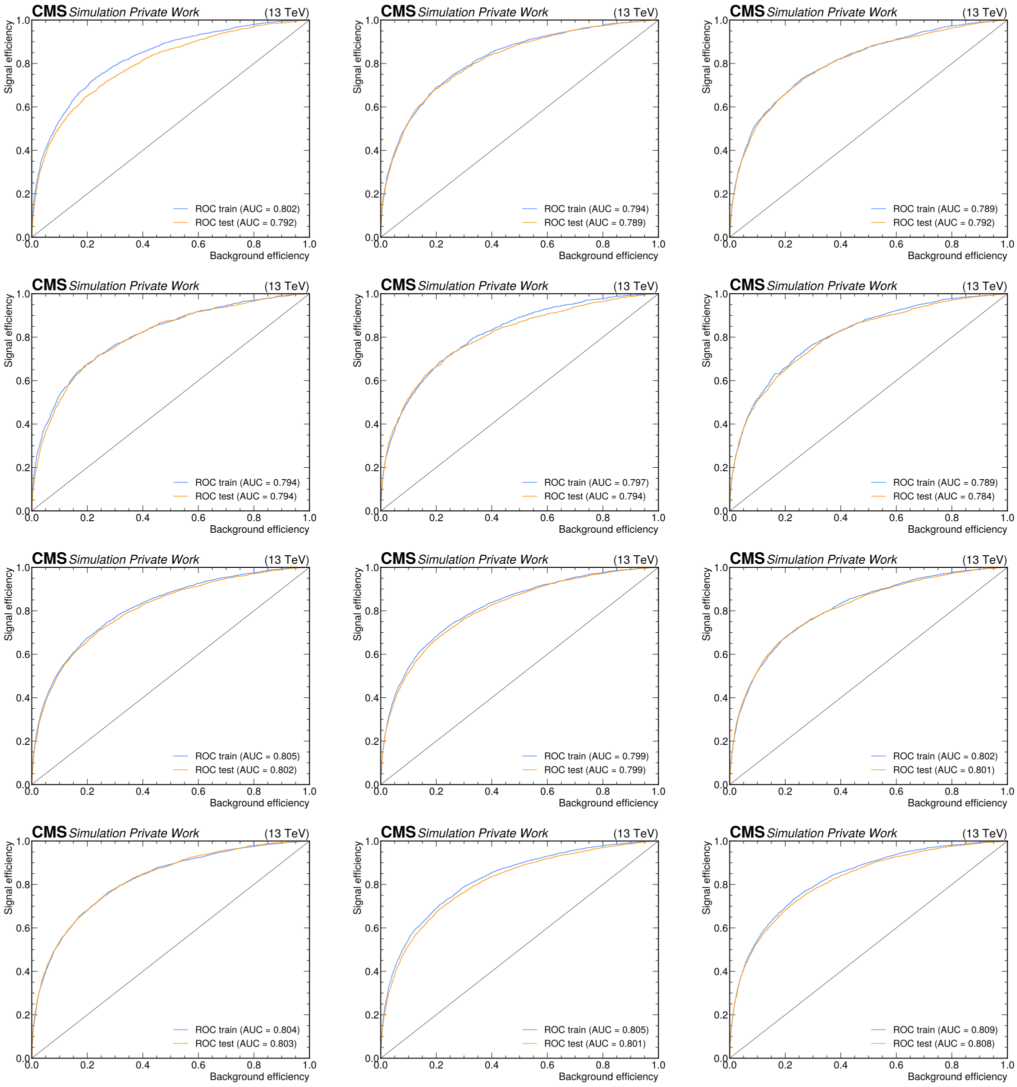


Figure 8.10: The receiver operator characteristic (ROC) curves for the performance evaluation of the three folds A, B, and C (from left to right) for the different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom) are shown.

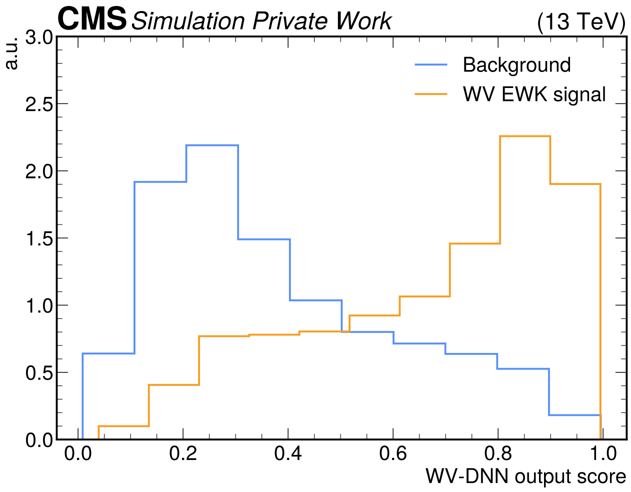


Figure 8.11: The WV-DNN output distribution for 2018, evaluated using the test dataset for fold A is shown. The signal and background contributions are normalized to unity. The signal events are nicely discriminated from background events. Similar results for other reconstruction eras and folds are not shown.

8.6.4. WV-DNN discriminator output

After a successful training, and validation of the WV-DNN discriminator, its output can be evaluated on unseen data. The events are classified as signal-like and background-like at the output score, leaving the signal like events at higher thresholds of the discriminator. In the VBS signal region, the WV-DNN output score is a new distribution, which provides a good discrimination between the VBS signal and all background processes. The WV-DNN discriminator outputs in the signal region for all years are shown in Figure 8.12.

The optimal threshold for the WV-DNN output is determined by maximizing the purity times efficiency metric, which ensures a balance between selecting a signal-enriched region and maintaining sufficient event statistics. As shown in Figure 8.13, this metric peaks around a specific threshold value, indicating the point where the classifier achieves the best trade-off between signal purity and efficiency. Based on this observation, we define the high-purity VBS signal region (WV-DNN output $>$ 0.6) as the electroweak VBS-enriched region, where further separation of different VBS contributions is performed. Meanwhile, the low-purity VBS signal region (WV-DNN output $<$ 0.6) is retained for the final fit to better control the backgrounds. To cross-check whether the events selected as signal-like actually corresponds to the physical expectations, such as large dijet invariant mass, low QCD activity etc, the WV-DNN output greater than a threshold of 0.6 is plotted for various kinematic variables in Figure 8.14. The events classified as signal-like by the DNN indeed have kinematic properties similar to the true signal events.

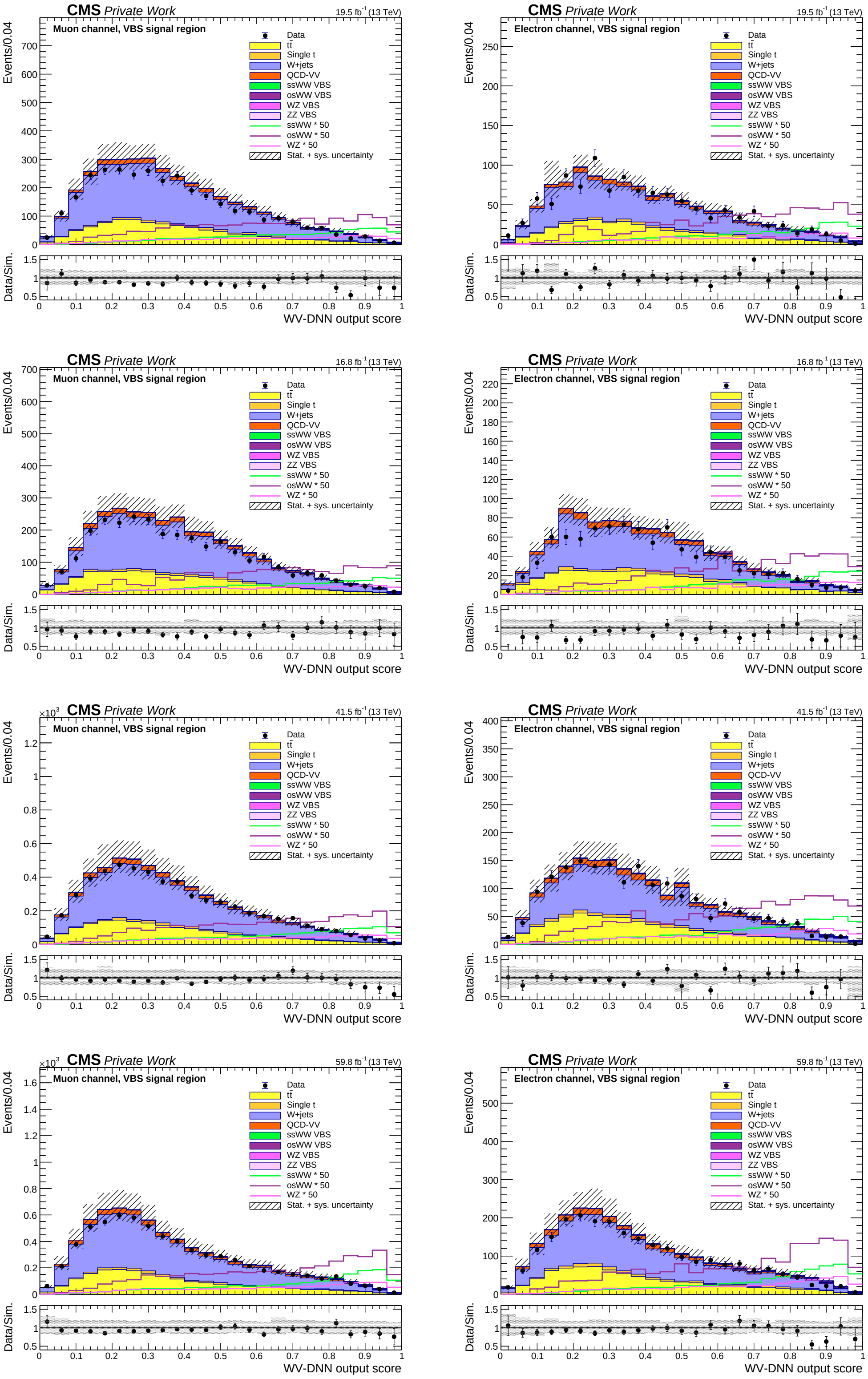


Figure 8.12: WV-DNN output distributions for the muon channel (left) and the electron channel (right) in the VBS signal region for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom). The WV (same-sign WW, opposite-sign WW, WZ) VBS signal processes are also shown after rescaling using solid lines. The output shows a good discrimination between the enormous background and the electroweak signal.

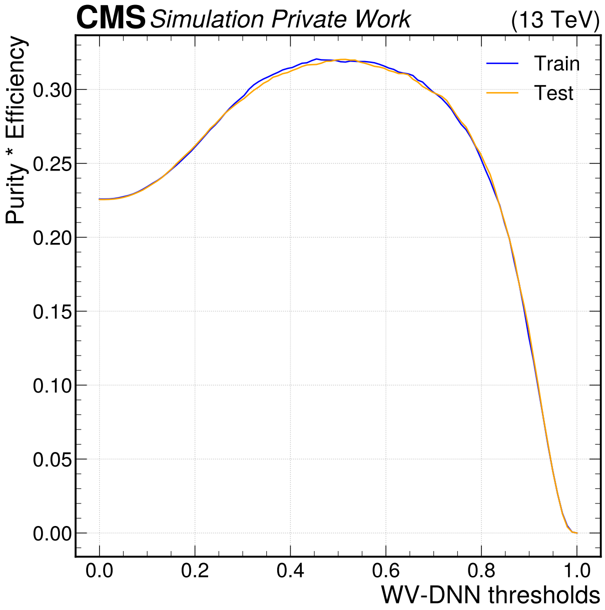


Figure 8.13: The purity times efficiency as a function of the WV-DNN output threshold. The optimal threshold is determined at the peak of the curve, ensuring a balance between signal purity and efficiency. This informs the selection of high-purity VBS signal region (WV-DNN output > 0.6) and low-purity VBS signal region (WV-DNN output < 0.6) to maximize the sensitivity of the analysis.

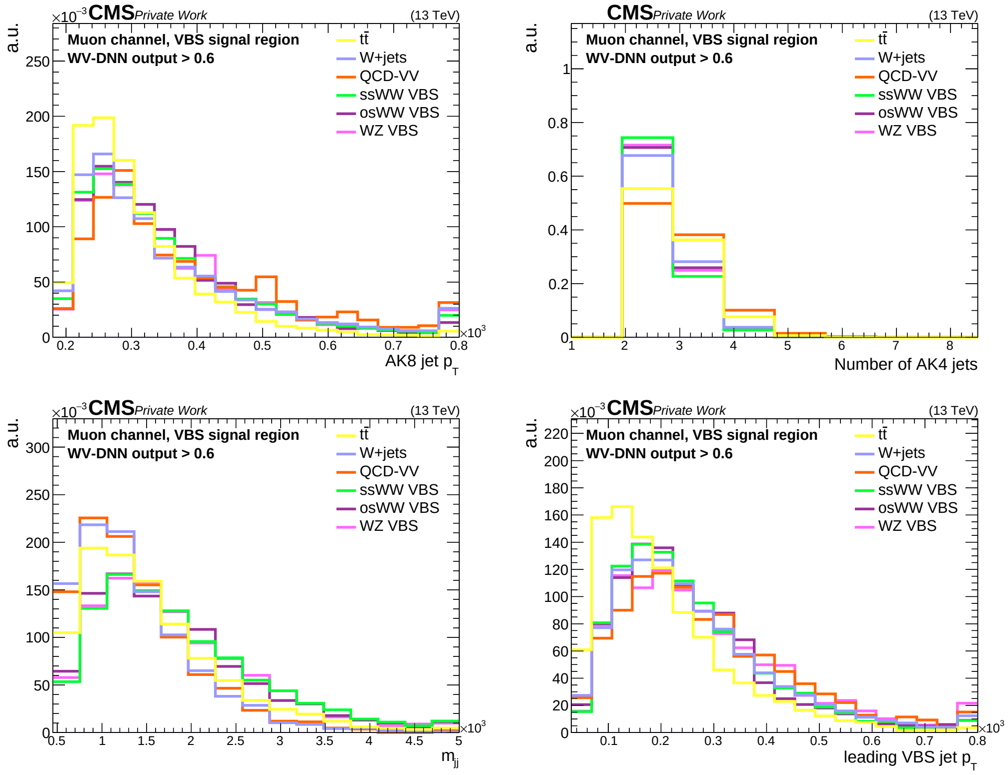


Figure 8.14: WV-DNN kinematic distributions for events with scores greater than 0.6 are shown. Events classified as signal-like by the network exhibit kinematic distributions similar to those of true signal events.

8.7. Classification of VBS events using the jet charge tagger

This section describes the application of the jet charge tagger, detailed in Chapter 7, to distinguish different VBS processes — same-sign WW, opposite-sign WW, and WZ — in the high-purity VBS signal region. This categorization is performed by identifying whether the AK8 jet in the event is W^+-like, W^--like, or Z-like using the jet charge tagger.

Various jet substructure variables, listed in Table 7.6, of the AK8 jets in the high-purity VBS region, are given as inputs to the tagger. The tagger provides output probability scores on its three output nodes (W^+, W^-, and Z) for each AK8 jet present in events in this region. To ensure a unique classification, the jet type corresponding to the highest probability score is assigned to the event.

Finally, the predicted jet type is combined with the charge of the lepton in the event to categorize events into the five signal categories listed in Table 8.11.

Table 8.11: Categorization of the high-purity VBS signal region.

WV VBS signal region				
Jet charge tagger W^+		Jet charge tagger W^-		Jet charge tagger Z
+ive lepton	-ive lepton	+ive lepton	-ive lepton	–
W^+W^+	W^+W^-	W^-W^+	W^-W^-	WZ

For the analysis under discussion, the same-sign WW category is of primary interest. The jet charge tagger distributions for the W^+W^+ and W^-W^- VBS signal categories are shown in Figure 8.15 and 8.16, respectively. The distributions for the opposite-sign WW and WZ VBS categories can be found in Appendix G. The distributions effectively distinguish the same-sign WW VBS process from opposite-sign and WZ VBS, and other SM background processes. The separation is more pronounced for higher values of the jet charge tagger output score. The W^-W^- category contributes little to the total same-sign WW VBS yield because of the W^-W^- VBS process has low cross section compared to W^+W^+ VBS. This feature stems from the differences in proton parton distribution functions. The up quarks, being more abundantly available as valence quarks in the initial state, favour the production of W^+ over W^-. This lead to a higher cross section of W^+W^+ VBS compared to W^-W^- VBS.

The distributions of the jet charge tagger output are used as inputs to perform a statistical analysis. The details of the statistical model used is described in Section 8.8. To maximize the sensitivity of the signal, all five categories listed in Table 8.11 are used in the final fit.

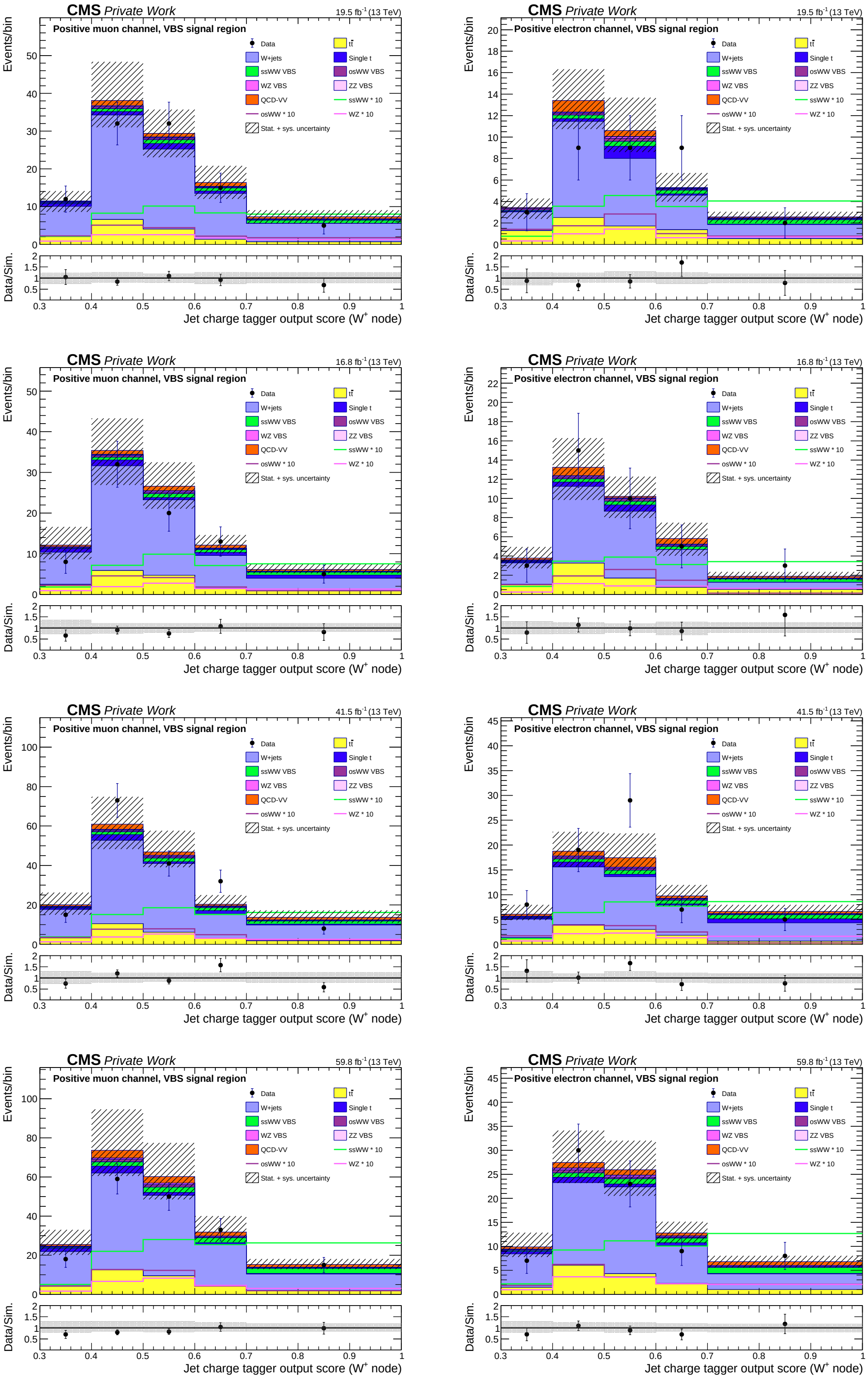


Figure 8.15: Jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **same-sign WW** (W^+W^+) **VBS signal category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom). The VBS signal processes are also shown after rescaling using solid lines. For higher values of the tagger output score, the same-sign WW VBS process is dominant compared to the other VBS processes and QCD-VV.

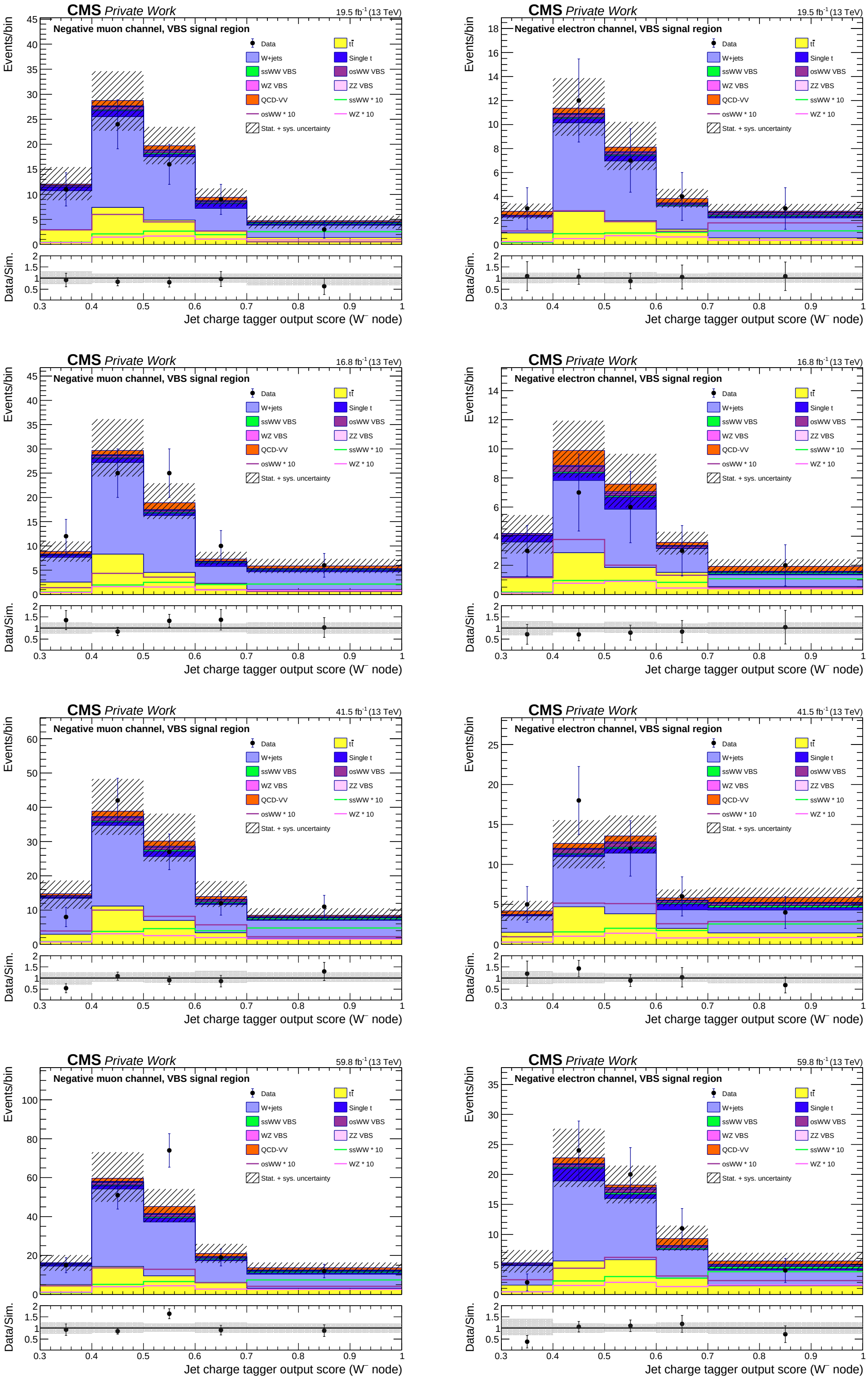


Figure 8.16: Jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **same-sign WW (W^-W^-) VBS signal category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom). The W^-W^- VBS process contribute less to the total same-sign WW VBS signal, because of its low cross section compared to its positive counterpart.

8.8. Statistical model

A statistical model is required to estimate the contribution of the signal process (same-sign WW VBS) in the events observed in data. The data events (N) observed in the experiment follow the Poisson distribution function given as:

$$\mathcal{P}(N|\lambda) = \frac{\lambda^N e^{-\lambda}}{N!}, \tag{8.4}$$

where λ is the total expected number of events. Our interest is to quantify the contribution of the signal process to these N observed data events, given we expect n_S and n_B as number of signal and background events, respectively from our simulations. To allow modifications in the signal yields, we introduce a signal strength parameter r, which when multiplied with the expected signal yield (n_S) results in the observed signal yield (N_S) i.e. $N_\mathrm{S} = rn_\mathrm{S}$. Based on Eq. 8.4, we can re-write the likelihood distribution as:

$$\mathcal{L}(r|N) = \mathcal{P}(N|rn_\mathrm{S} + n_\mathrm{B}) = \frac{(rn_\mathrm{S} + n_\mathrm{B})^N e^{-(rn_\mathrm{S}+n_\mathrm{B})}}{N!}. \tag{8.5}$$

In reality, our knowledge of the expected number of signal and background events is limited by various uncertainties. Theoretical uncertainties arise from cross section calculations, including scale variations and parton distribution functions (PDFs). Experimental uncertainties stem from detector resolution, calibration, and efficiency corrections for reconstructed objects. Some of these systematic uncertainties have an effect on overall event yields, while some alter the shapes of kinematic observables. The details of all relevant systematic uncertainty sources are described in Section 8.9. While performing a statistical analysis, all such sources of uncertainties needs to be taken into account in the likelihood function. To do this, a set of nuisance parameters are introduced, where each nuisance parameter represents a source of systematic uncertainty. The likelihood function in Eq. 8.5 takes the form:

$$\mathcal{L}(r,\boldsymbol{\theta}|N) = \frac{(rn_\mathrm{S}(\boldsymbol{\theta}) + n_\mathrm{B}(\boldsymbol{\theta}))^N e^{-(rn_\mathrm{S}(\boldsymbol{\theta})+n_\mathrm{B}(\boldsymbol{\theta}))}}{N!} \times \prod_{i=1}^{n_{\boldsymbol{\theta}}} \pi_i(\theta_i), \tag{8.6}$$

where $\boldsymbol{\theta}$ is the set of nuisance parameters associated with systematic uncertainties, and $\pi_i(\theta_i)$ are the prior probability density functions for a nuisance parameter, which is typically a log-normal distribution for normalization uncertainties and a Gaussian distribution for shape uncertainties.

To extract the signal strength parameter r from data and to test a certain hypothesis, a maximum likelihood estimate (MLE) is performed. For the analysis presented in this chapter, a binned likelihood function is maximized, which is the product of the likelihood function introduced in Eq. 8.6 in each bin of the distribution of the discriminating observables across the VBS signal regions (low-purity and high-purity) and the W+jets control region.

To estimate the significance of an observed signal with respect to the background-only hypothesis, the profile likelihood ratio is used, which is defined as:

$$q_r = -2\Delta\ln\mathcal{L} = \frac{\mathcal{L}(r, \hat{\boldsymbol{\theta}}_{\boldsymbol{r}})}{\mathcal{L}(\hat{r}, \hat{\boldsymbol{\theta}})}, \tag{8.7}$$

where the numerator represents the value of the likelihood that is maximized with the set of nuisance parameters $\hat{\boldsymbol{\theta}}_{\boldsymbol{r}}$ at a specified value of r, while the denominator is the value of the likelihood that is globally maximized with the set of nuisance parameters $\hat{\boldsymbol{\theta}}$. Large values of q_r correspond to a large disagreement between the observed data and the hypothesis defined by the choice of r. According to Wilk's theorem [140], in the limit of a large data sample, the profile likelihood ratio approaches a chi-squared distribution with the corresponding degrees of freedom. The level of compatibility of the data with a given hypothesis can be quantified in terms of p-value, which represents the probability of obtaining a test statistic q_r that is as extreme as, or more extreme than, the observed value, assuming the null hypothesis is true. The p-value is typically converted into an equivalent significance level, which is determined based on the area in the tail of a standard Gaussian distribution. A significance of 3σ is conventionally considered the threshold to claim evidence for a process. On the other hand, a discovery or observation requires a higher significance of 5σ. The test statistic q_r can also be used to estimate confidence intervals (CL$_s$) defined as:

$$\mathrm{CL}_s = \frac{p_{\mathrm{S+B}}}{1 - p_{\mathrm{B}}}, \tag{8.8}$$

where $p_{\mathrm{S+B}}$ refers to the p-value under signal-plus-background hypothesis, and p_{B} refers to the p-value under background-only hypothesis. The CL$_s$ criterion is designed in a way to avoid exclusion of a signal hypothesis when the signal sensitivity is low, or due to underfluctuations in data. In this analysis, the CL$_s$ method is used to calculate exclusion limits on the same-sign WW VBS signal strength parameter.

8.9. Systematic uncertainties

This section summarizes the systematic uncertainties that are taken into account in the maximum likelihood fit performed to extract the signal strength parameter. We use a binned likelihood function in which each systematic uncertainty is represented by a nuisance parameter that morphs the template shape and/or scales the normalization of the templates. The shape variations are included in the COMBINE tool [139] by providing the histogram templates shifted up and down by one standard deviation.

8.9.1. Theoretical uncertainties

PDF uncertainty: The uncertainty in the choice of parton distribution function is evaluated by using 100 alternate sets of the central NNPDF3.1 densities. The PDF uncertainty in the signal and background processes is evaluated by reweighting simulated events with these alternate sets and the envelope method is used for the final estimate of this uncertainty, as per recommendations of PDF4LHC [141]. For some processes, the uncertainty is treated as correlated, while for the rest uncorrelated. Also, full correlation is used among the years.

Renormalization and factorization scale: Monte Carlo simulated events are reweighted by alternative event weights, where renormalization and factorization scales are varied up and down by a factor 2. A total of 6 variations are considered and the envelope of the varied distributions are taken as one standard deviation variations. The uncertainty is treated as uncorrelated between different processes, but correlated across years

Parton shower modelling: The uncertainty related to the parton shower modelling is evaluated by means of event reweighting by weights obtained from PYTHIA8. Reweighting factors are applied at the per-event level, which correspond to variations in the ISR and FSR scales, where the scales are varied up and down by a factor of 2 relative to their nominal values. The ISR and FSR uncertainties are treated as correlated among processes and years.

8.9.2. Experimental uncertainties

Integrated luminosity: The integrated luminosity uncertainties measured by the CMS collaboration for the 2016, 2017, and 2018 data-taking periods are included following the year-to-year correlation scheme recommended by the CMS Collaboration [45]. The correlated uncertainties are 0.6% for 2016, 0.9% for 2017, and 2.0% for 2018. An additional correlated component, applicable only to 2017 and 2018, contributes 0.6% and 0.2%, respectively. The uncorrelated uncertainties are 1.0% for 2016, 2.0% for 2017, and 1.5% for 2018.

Pileup reweighting: To evaluate the uncertainty introduced by the pileup reweighting procedure, explained in Section 6.1, the true pileup distribution in the simulation is systematically varied by ±4.6%, as recommended in Reference [122]. The resulting impact on event yields and kinematic distributions is used to quantify the systematic uncertainty associated with pileup reweighting. This uncertainty is treated as uncorrelated among years.

L1 prefiring: The L1 prefiring issue, explained in Section 6.2, is mitigated by applying prefiring weights to the simulated events. The uncertainty in this reweighting procedure is estimated by the recipe provided in Reference [123]. The prefiring uncertainty is treated as uncorrelated among years.

Lepton efficiency: The efficiencies for muon and electron triggers, reconstruction, identification, and isolation are calculated using data through the standard tag-and-probe method with Z bosons. Scale factors are then applied to the simulation to correct for differences with the measured data. To evaluate the uncertainties on these scale factors, both statistical uncertainties from the measurements and efficiency variations are taken into account. These uncertainties are considered uncorrelated between years.

B tagging efficiency: B tagging is utilized in this analysis to reject backgrounds mainly $\mathrm{t\bar{t}}$ containing b jets in the final state. Uncertainties in the b tagging efficiencies and mistag rates as function of the jet p_T and η are provided by the CMS Collaboration [128]. The effect of these uncertainties on the event yields is evaluated by adjusting the data-to-simulation correction factors within their respective uncertainty ranges and reanalysing the events. Ten separate uncertainties are considered as recommended for Run 2 analyses in Reference [142].

Top p_T reweighting: Uncertainty related to top quark p_T reweighting, which is applied to improve the $\mathrm{t\bar{t}}$ simulation, is estimated by quantifying the effect on yields with and without applying corrections. More details in Reference [143].

Limited size of MC simulation: The uncertainty related to the limited size of simulated samples is taken into account using the COMBINE tool `autoMCStats`, which uses the Barlow-Beeston method [144]. For each bin in the distribution of the fit template, a Gaussian nuisance parameter is introduced that varies the predicted yields of all processes simultaneously. The uncertainties are treated as uncorrelated among years.

V tagging efficiency: The uncertainty related to the use of the τ_{21} tagger for identification of boosted AK8 jets originating from W/Z bosons is applied according to the recommendations provided by the CMS Collaboration [130]. The scale factors and uncertainties are provided centrally, which are listed in Table 6.1. The uncertainty is treated as uncorrelated between years.

Jet energy scale: The uncertainty related to jet energy scale corrections, explained in Section 6.4, is propagated by varying the jet energy scale (JES) by $\pm 1\sigma$ variation for both AK4 and AK8 jets. The events are reanalysed with the varied set of corrections, including the full analysis chain. The JES variations have an impact on both the rate and shape of the fit templates. The JES varied shapes of the templates are provided during the fit to estimate the impact of the uncertainty relative to the nominal jet energy scale. The JES uncertainties are treated as correlated among processes and uncorrelated across years.

Jet energy resolution: Measurements show that the jet energy resolution (JER) in data is worse than in the simulation and the jets in simulation. Therefore, we smear the jet energy resolution in simulation to match the one in data, details of the procedure are given in Section 6.4. We estimate the effect of JER uncertainty by shifting the scale factors by $\pm 1\sigma$ and repeating the smearing procedure together with the full analysis chain. Like JES, JER variations have an impact on both shapes and rate of the fit templates. The uncertainty is treated as correlated among processes and uncorrelated across years.

A summary of the impacts of systematic uncertainties in the high-purity VBS region is shown in Table 8.12.

Table 8.12: A summary of the systematic uncertainties affecting different process rates in the high-purity VBS signal region is shown. The values are reported as percentages relative to the nominal predictions. For asymmetric uncertainties, the larger value is quoted. Uncertainties that also impact the shape of distributions are marked with ✓.

Uncertainty source	shape	$t\bar{t}$	Single top	W+jet	QCD-VV	ssWW VBS	osWW VBS	WZ VBS
PDF	✓	0.52	3.0	0.93	1.9	0.8	2.2	1.26
QCD scale	✓	21.5	13.3	24.04	31.07	12.04	11.8	12.5
Parton shower (ISR)	✓	3.5	0.6	4.3	-	1.04	1.3	0.4
Parton shower (FSR)	✓	4.6	17.7	9.3	-	2.4	3.4	3.0
Integrated lumiosity		1.6	1.6	1.6	1.6	1.6	1.6	1.6
Pileup		0.8	1.4	0.5	0.2	0.6	2.8	1.6
L1 prefiring	✓	0.09	0.09	0.08	0.07	0.08	0.08	0.08
Lepton efficiency	✓	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Trigger efficiency	✓	0.2	0.2	0.2	0.2	0.2	0.2	0.2
B tagging efficiency	✓	5.6	4.8	0.93	1.15	0.7	1.4	0.7
Top-p_T reweighting		6.61	-	-	-	-	-	-
V-tagging efficiency		15.4	15.4	15.4	15.4	15.4	15.4	15.4
Jet energy scale	✓	10.2	16.0	26.2	9.8	4.1	8.2	6.0
Jet energy resolution	✓	6.2	3.8	11.5	4.8	0.9	1.8	1.3

8.10. Results

For the final analysis and extraction of the EW same-sign WW VBS signal strength (r_{ssWW}), a maximum likelihood fit is performed by combining the information from the VBS signal regions (high-purity and low-purity) and the W+jets control region. In the high-purity VBS signal region, five categories are considered, for each muon and electron channel, and each year. The statistical model is defined using separate datacards for each category, year, and region. For the final statistical inference, various datacards are combined using the Combine tool. The jet charge tagger output score is used to build the templates in the high-purity VBS signal region, where in the low-purity signal region, the WV-DNN output score is used. For the W+jets control region, the transverse momentum of leptonically decaying W boson is used as a fit variable.

The normalization of the W+jets background in the signal region is controlled by including the W+jets control region in the fit. The fit is performed simultaneously in the signal regions and the control region. The W+jets background is divided into various subcategories, as described in Section 8.5. These subcategories are linked together in the signal regions and the control region by using the Combine rate parameters (`rateParam`). These constant parameters are derived during the fit in the control region, and are propagated to the signal region to estimate the correct normalization of the W+jets background. All sources of systematic uncertainties, affecting rates and shapes of these templates, are properly encoded in the datacards.

The post-fit distributions of the high-purity signal region for the full Run 2 dataset are shown in Figure 8.17 and 8.18, for the W^+W^+ and W^-W^- categories, respectively. Other categories distributions are shown in Appendix H.

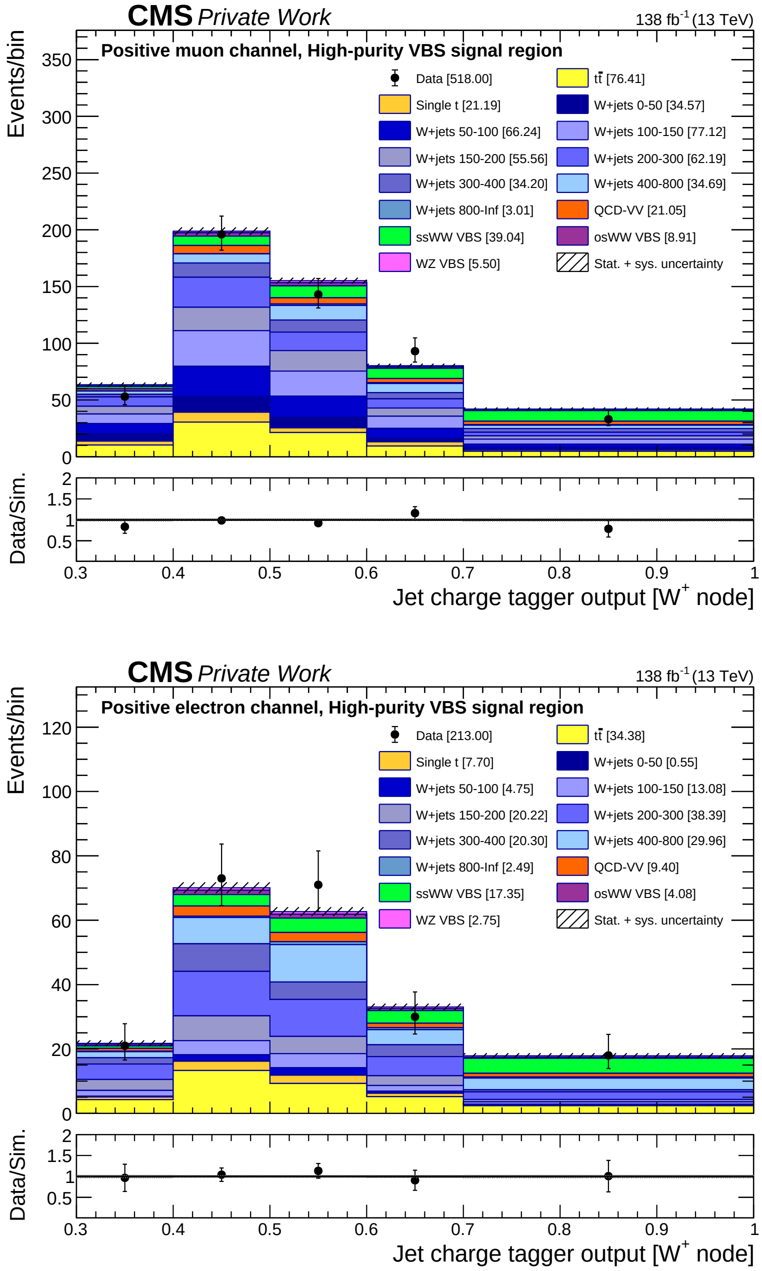


Figure 8.17: Post-fit distributions of the jet charge tagger output in the **same-sign WW** (W^+W^+) **VBS signal category** in the muon channel (top) and the electron channel (bottom) using the full Run 2 dataset. The signal strength of the same-sign WW VBS is left floating, while the other VBS processes and backgrounds are kept fixed to their standard model expectation values.

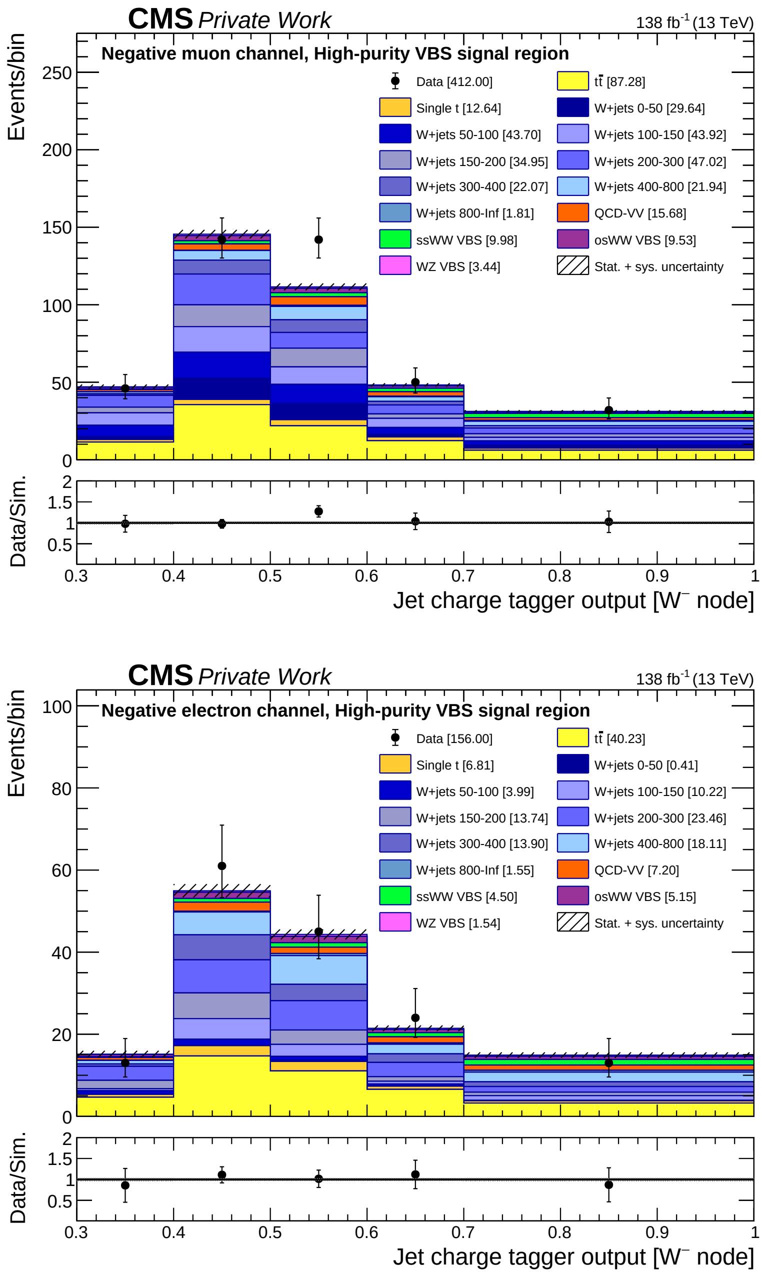


Figure 8.18: Post-fit distributions of the jet charge tagger output in the **same-sign WW** (W^-W^-) **VBS signal category** in the muon channel (top) and the electron channel (bottom) using the full Run 2 data. The signal strength of the same-sign WW VBS is left floating, while the strengths of the other VBS processes and backgrounds are kept fixed to their expected standard model values.

A profile likelihood scan for various values of the signal strength parameter is shown in Figure 8.19, showing both the expected and observed curves. The minimum of the curves gives the best-fit value of the parameter of interest. The observed best-fit value of the signal strength is found to be:

$$\hat{\mathrm{r}}_{\mathrm{ssWW}} = \frac{\sigma^{\mathrm{obs}}}{\sigma^{\mathrm{SM}}} = 1.63\ ^{+0.40}_{-0.32}(\mathrm{syst})^{+0.59}_{-0.57}(\mathrm{stat}) = 1.63\ ^{+0.72}_{-0.66}, \tag{8.9}$$

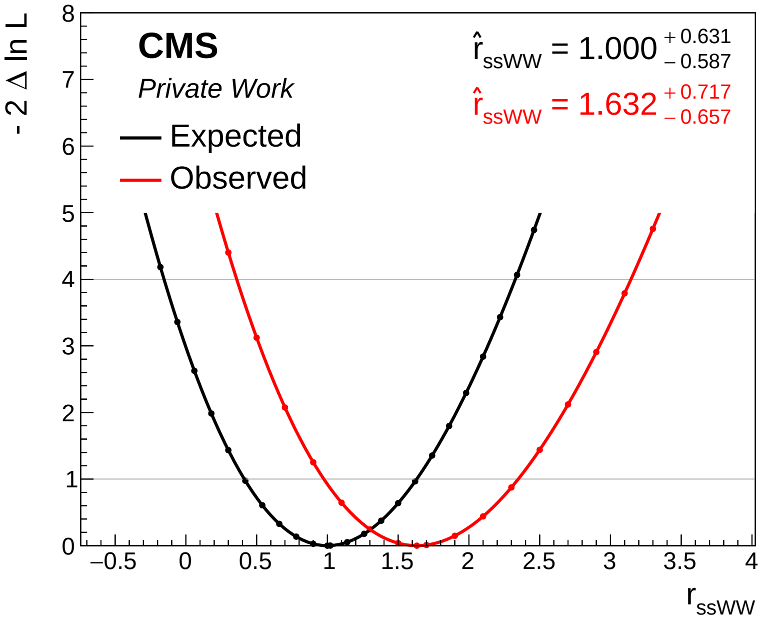


Figure 8.19: Profile likelihood scans are shown for various values of the signal strength parameter of the same-sign WW VBS signal. The expected curve represents the scan assuming the standard model prediction ($\hat{\mathrm{r}}_{\mathrm{ssWW}}$) while the observed curve corresponds to data. The minimum of the observed curve provides the best-fit value of the signal strength parameter. The quoted uncertainties include contributions from both statistical and systematic sources.

with $1.00^{+0.63}_{-0.58}$ expected, where σ^{obs} and σ^{SM} are the observed and predicted cross section, respectively. The measured cross section times the branching fraction is thus found to be 204^{+90}_{-82} fb, with the theory value of 125 fb. The observed significance of the EW same-sign WW VBS signal is 2.8 standard deviations, with 1.8 expected with respect to the background-only hypothesis. The significance from individual years is summarized in Table 8.13. As the significance of the signal is not enough for an evidence or a discovery, we set upper limits on the cross section times the branching fraction using the 95% confidence level intervals (CL_s). The expected and observed upper limits in the asymptotic approximation [145] are shown in Table 8.14. The upper limits for individual years as well as for the combined Run 2 are summarized in a limit plot, shown in Figure 8.20. A slight excess is observed in data that is not significant enough for an evidence of a potential new signal.

This small excess is mainly coming from the electron decay channel in 2017, and muon decay channel in 2018, as shown in Figures H.1 and H.2, respectively. The jet charge tagger scores for these excess events are between 0.5 and 0.6. After explicitly checking the kinematic distributions of these events, we expect them to be multijet events. The jet charge tagger is not trained to handle QCD jets and therefore, it gives a random score for them and the events with QCD jets ends up in the 0.5 - 0.6 bin. The future iterations of this analysis should either utilize an increased missing transverse momentum cut or should estimate the multijet background using a data-driven method to mitigate the issue.

Table 8.13: Expected and observed significance of the EW same-sign WW VBS signal by combining both the muon and electron channels. Significances are shown separately for each reconstruction era as well as for the combined Run 2 dataset.

Year	Expected	Observed
2016 preVFP	0.64	0.08
2016 postVFP	0.61	1.3
2017	0.91	2.04
2018	1.31	1.96
Run 2	**1.8**	**2.8**

Table 8.14: Expected and observed upper limits at 95% confidence intervals for the same-sign WW VBS signal strength parameter.

Quantile	2016 preVFP	2016 postVFP	2017	2018	Run 2
Expected (-2σ)	1.74	1.87	1.19	0.82	0.59
Expected (-1σ)	2.4	2.57	1.62	1.11	0.79
Expected (median)	3.45	3.77	2.31	1.56	1.13
Expected (+1σ)	5.14	5.65	3.4	2.34	1.63
Expected (+2σ)	7.46	8.25	4.8	3.31	2.28
Observed	4.03	6.55	5.07	3.25	2.85

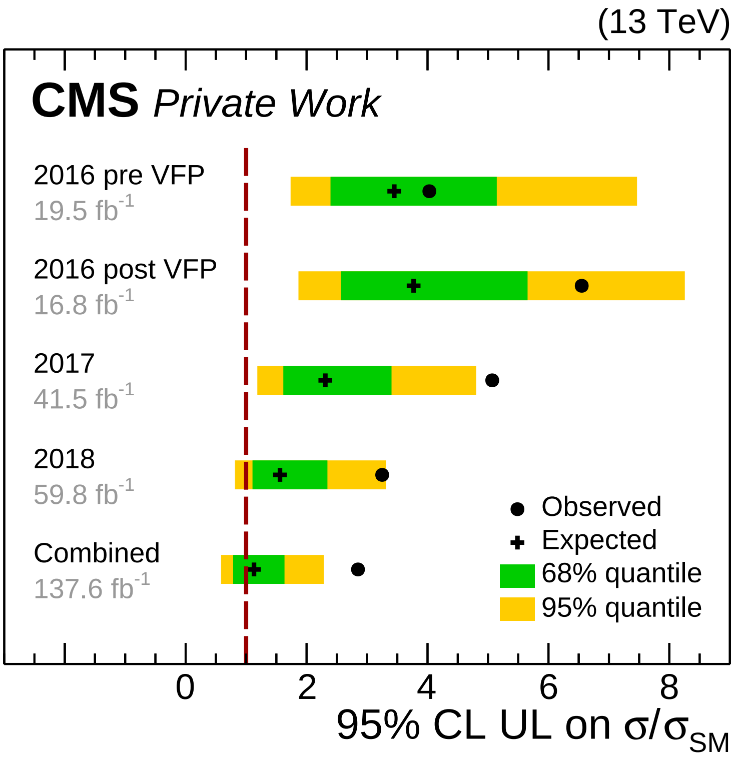


Figure 8.20: A summary of 95% confidence level upper limits on the same-sign WW signal strength, defined as ratio of measured cross section and expected standard model cross section. The upper limits are shown on the x-axis. The reconstruction years and their combination are shown on the y-axis. For each year and their combination, the observed limit is shown with a circle marker, while the median expected limit is shown with a plus marker. 68% and 95% quantiles bands are shown in green and yellow, respectively.

9. Summary and outlook

This thesis presents the first search for same-sign WW vector boson scattering in the semi-leptonic decay channel. The analysis is based on proton-proton collision data recorded by the Compact Muon Solenoid experiment at the CERN Large Hadron Collider during 2016–2018, corresponding to a total integrated luminosity of 138 fb^{-1}. One of the two W bosons is required to decay hadronically, producing a large-radius jet in the detector. To distinguish same-sign WW events from opposite-sign WW and WZ processes, a jet charge tagger is designed and employed to predict the charge of the large-radius jet.

The jet charge tagger, designed and trained using a ParticleNet based Dynamic Convolutional Graph Neural Network model, exploits jet substructure information to classify jets as positively, negatively, or neutrally charged. This innovative algorithm enables an analysis that was previously infeasible due to the inherent difficulty of jet charge reconstruction. The algorithm is standalone and can be utilized by any future CMS analysis that requires jet charge identification, in particular the charge of jets initiated by a vector boson.

A typical VBS topology is targeted by selecting events with two jets of large invariant mass and pseudorapidity separation. Events are reconstructed to contain exactly one tight lepton (electron or muon), one large-radius jet from a boosted vector boson decay, and moderate missing transverse momentum to account for an undetected neutrino. The semi-leptonic decay channel suffers from significant background contributions, primarily from W+jets and top quark antiquark pair production. Various Monte Carlo simulations and data-driven techniques are employed to estimate background contributions in the VBS signal phase space. To suppress background contamination, a dedicated deep neural network is developed and applied, with the jet charge tagger further categorizing events into same-sign WW, opposite-sign WW, and WZ VBS categories. The primary focus of this thesis is the same-sign WW VBS process, which is treated as the signal, while other VBS processes and non-resonant diboson production are considered background in the final statistical inference.

A template-based shape analysis is performed by combining information from the signal and control regions. The normalization of the W+jets background in the signal region is derived from a dedicated W+jets control region during the fit. Various experimental and theoretical uncertainties are incorporated into the model. A binned maximum likelihood estimation is used for the final fit, yielding a 95%

confidence level upper limit on the signal strength, $\mathrm{r_{ssWW}} < 2.85$, using the full Run 2 dataset. The observed and expected significances of the same-sign WW VBS signal are 2.8 and 1.8 standard deviations, respectively. The best-fit value of the signal strength is $1.63^{+0.72}_{-0.66}$, with an expected value of $1.00^{+0.63}_{-0.58}$. The measured cross section times the branching fraction is found to be 204^{+90}_{-82} fb, with the SM theory value of 125 fb. These results represent the first measurement of this final state at the LHC, demonstrating that novel reconstruction techniques, such as the jet charge tagger, can unlock previously inaccessible areas of the physics research program.

The signal process studied in this thesis has a very low cross section, leading to a small number of signal events compared to background processes. Detector inefficiencies and reconstruction challenges further reduce this yield. Consequently, VBS studies are expected to benefit from increased luminosity and larger datasets from future LHC runs, particularly the High-Luminosity LHC (HL-LHC). At the HL-LHC, the study of longitudinally polarized vector boson scattering is one of the main research targets.

The ongoing Run 3, when combined with Run 2 data, could significantly enhance both the statistical precision and signal significance of the same-sign WW process. Future extensions of this analysis could explore polarized scattering, with a particular focus on longitudinally polarized WW scattering, which is closely tied to electroweak symmetry breaking. Additionally, this analysis could be adapted to search for a doubly charged Higgs boson in the same-sign WW semi-leptonic final state. When combined with fully leptonic final states, such searches could improve the discovery potential for exotic signatures.

A. Additional figures for b-tagging efficiencies

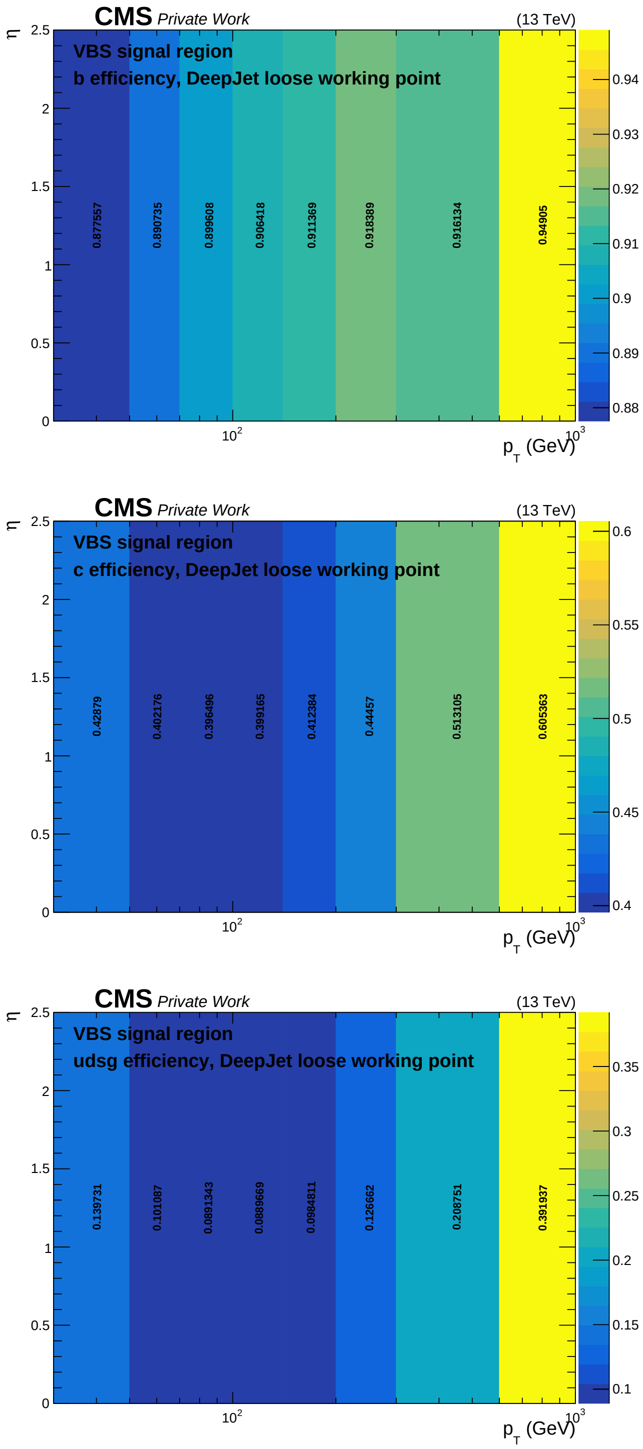


Figure A.1: Monte Carlo simulation flavour tagging efficiencies as a function of p_T and η in 2016 preVFP era. Efficiencies for b (top), c (middle), and light (bottom) jets for the loose working point of the DeepJet tagger are shown in the VBS signal region.

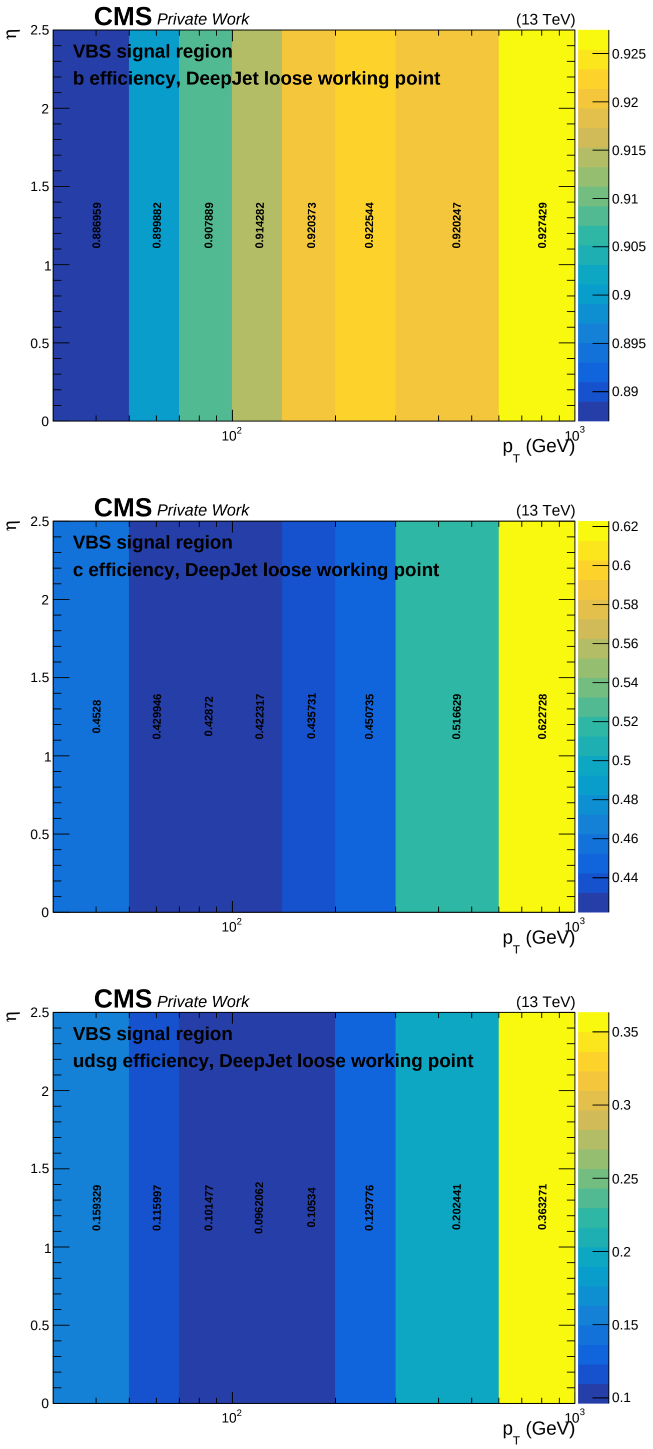


Figure A.2: Monte Carlo simulation flavour tagging efficiencies as a function of p_{T} and η in 2016 postVFP era. Efficiencies for b (top), c (middle), and light (bottom) jets for the loose working point of the DeepJet tagger are shown in the VBS signal region.

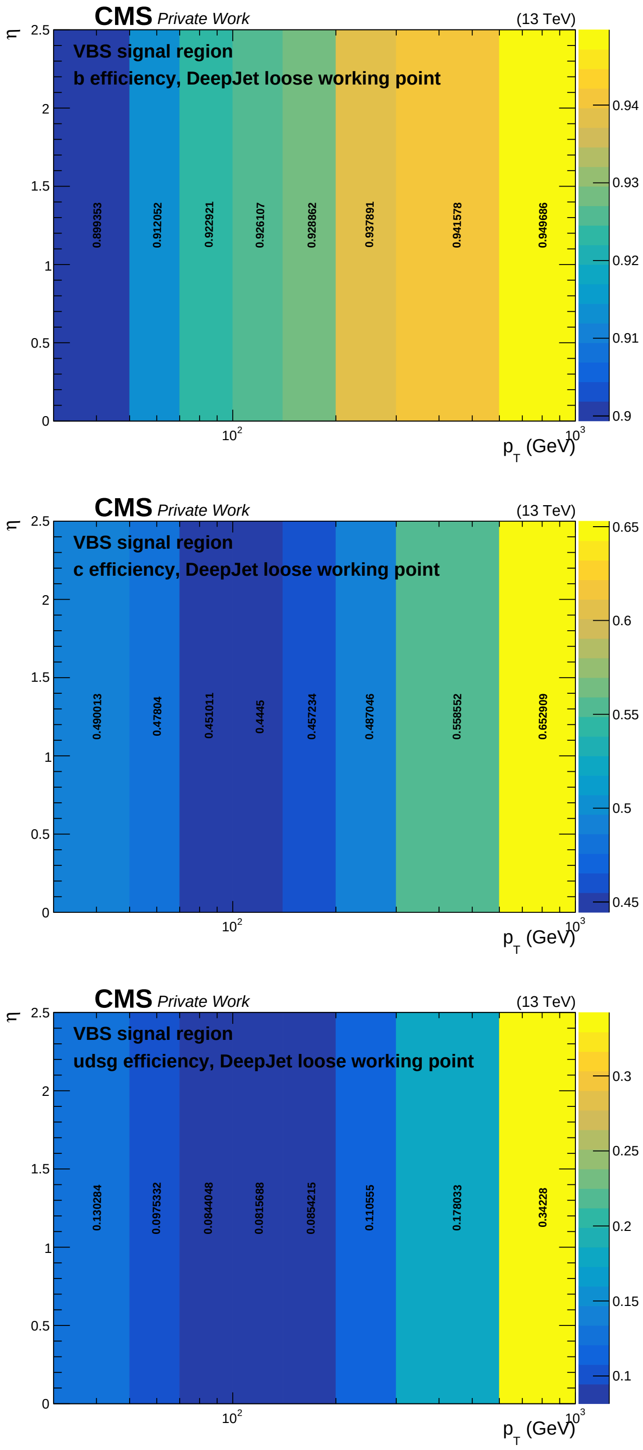


Figure A.3: Monte Carlo simulation flavour tagging efficiencies as a function of p_T and η in 2018. Efficiencies for b (top), c (middle), and light (bottom) jets for the loose working point of the DeepJet tagger are shown in the VBS signal region.

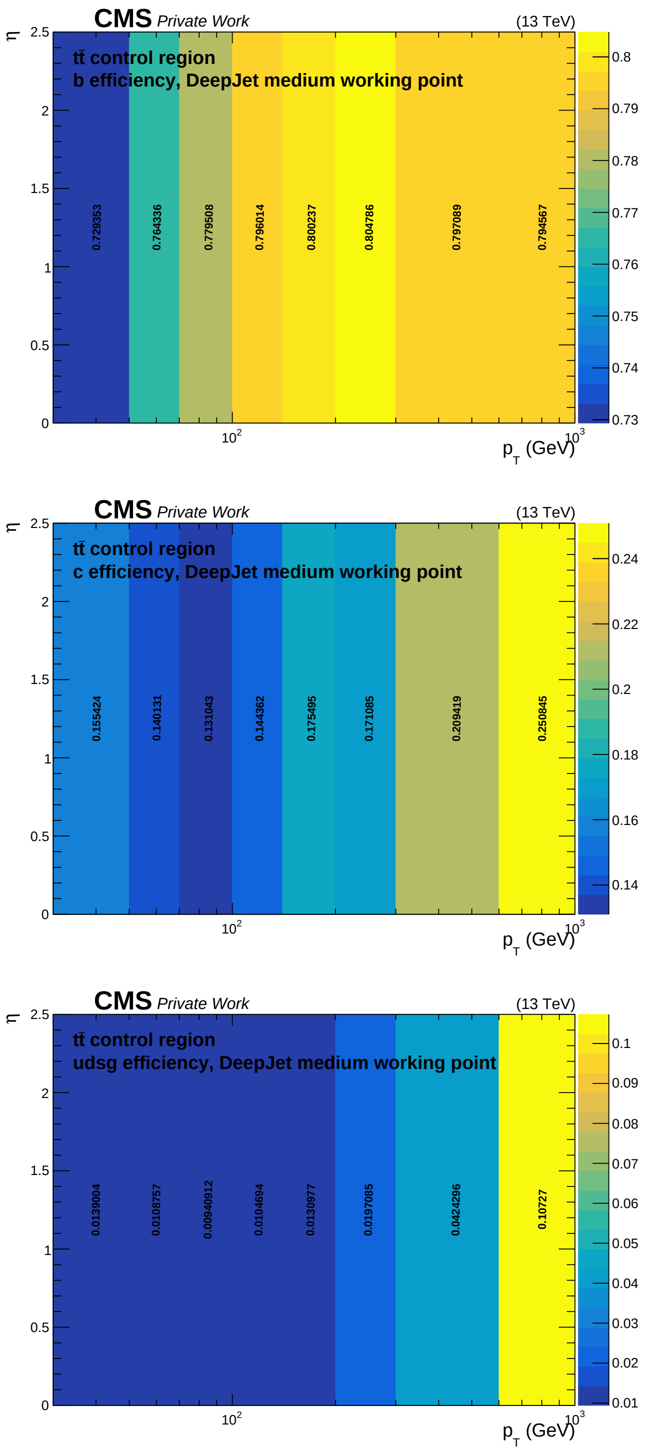


Figure A.4: Monte Carlo simulation flavour tagging efficiencies as a function of p_T and η in 2016 preVFP era. Efficiencies for b (top), c (middle), and light (bottom) jets for the medium working point of the DeepJet tagger are shown in the $\mathrm{t\bar{t}}$ control region.

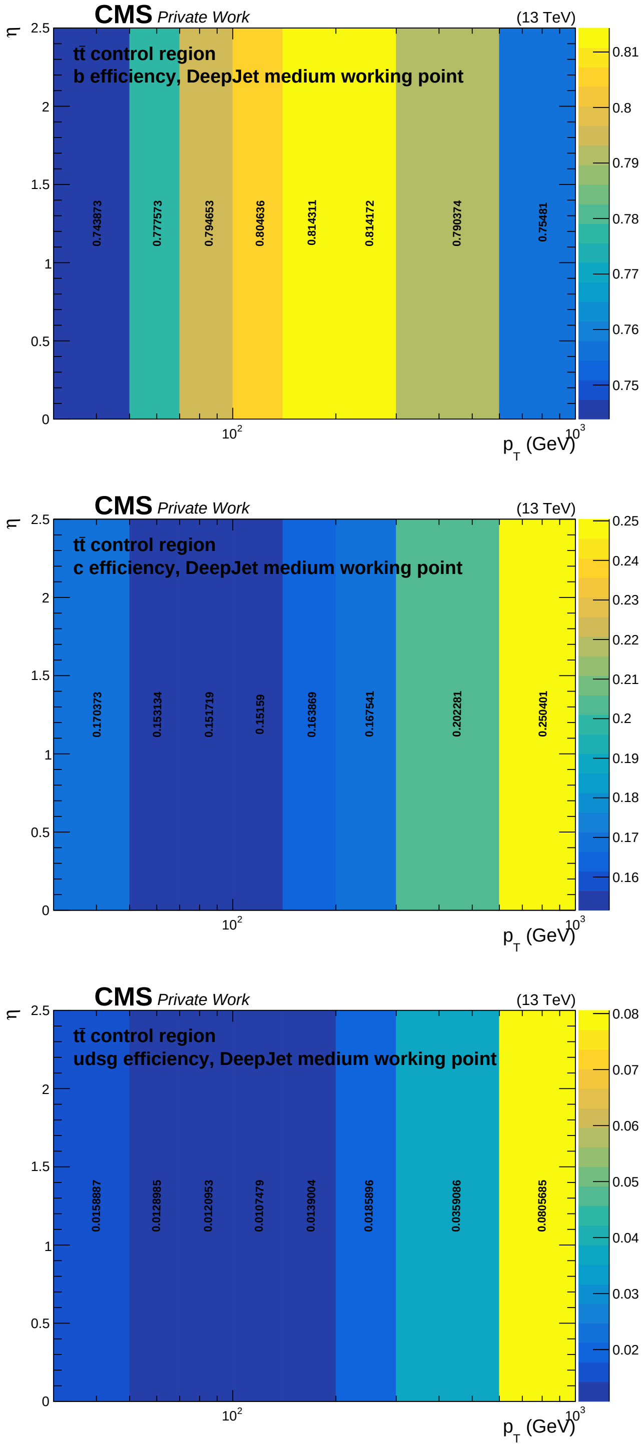


Figure A.5: Monte Carlo simulation flavour tagging efficiencies as a function of p_T and η in 2016 postVFP era. Efficiencies for b (top), c (middle), and light (bottom) jets for the medium working point of the DeepJet tagger are shown in the $\mathrm{t\bar{t}}$ control region.

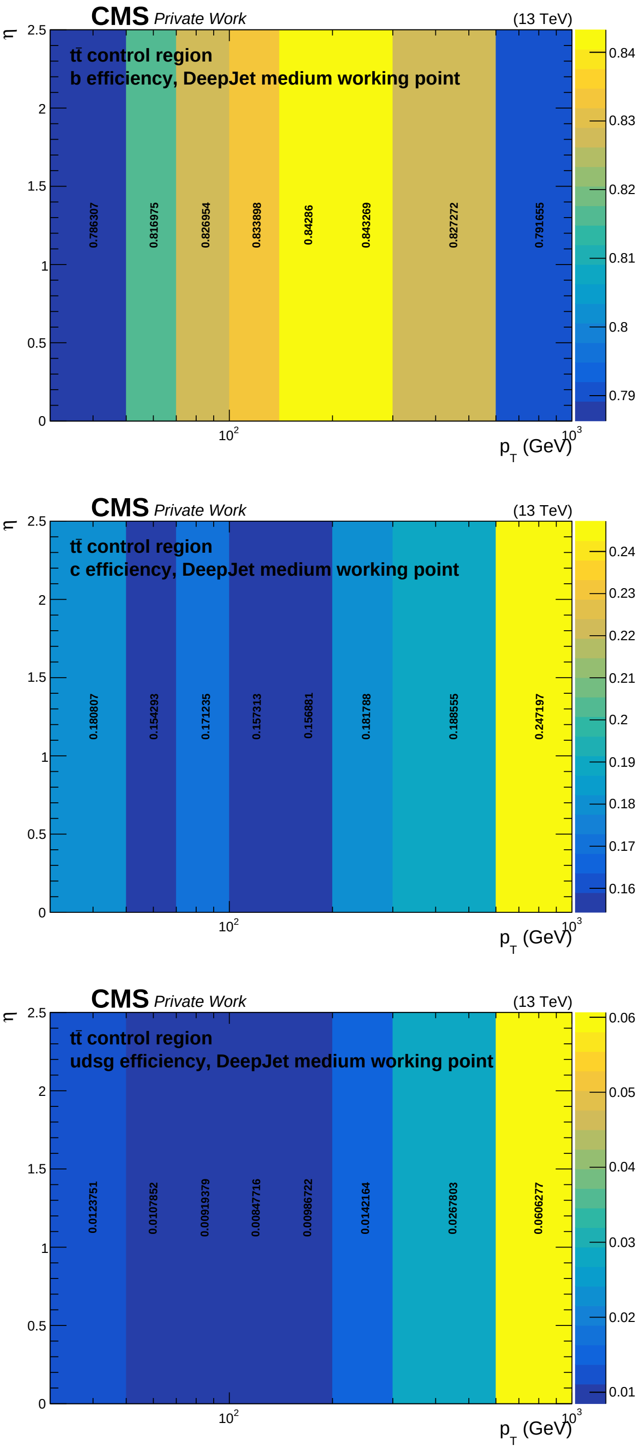


Figure A.6: Monte Carlo simulation flavour tagging efficiencies as a function of p_{T} and η in 2017. Efficiencies for b (top), c (middle), and light (bottom) jets for the medium working point of the DeepJet tagger are shown in the $t\bar{t}$ control region.

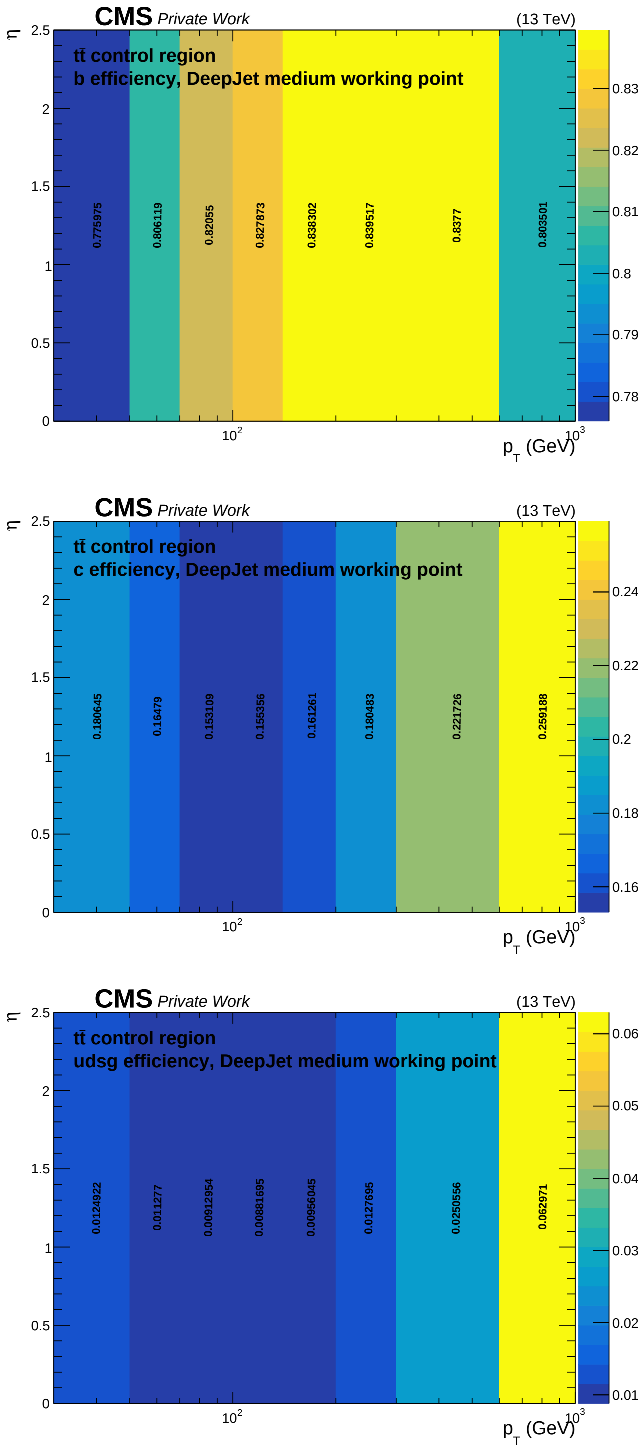


Figure A.7: Monte Carlo simulation flavour tagging efficiencies as a function of p_T and η in 2018. Efficiencies for b (top), c (middle), and light (bottom) jets for the medium working point of the DeepJet tagger are shown in the $\mathrm{t\bar{t}}$ control region.

B. Jet charge tagger input variables

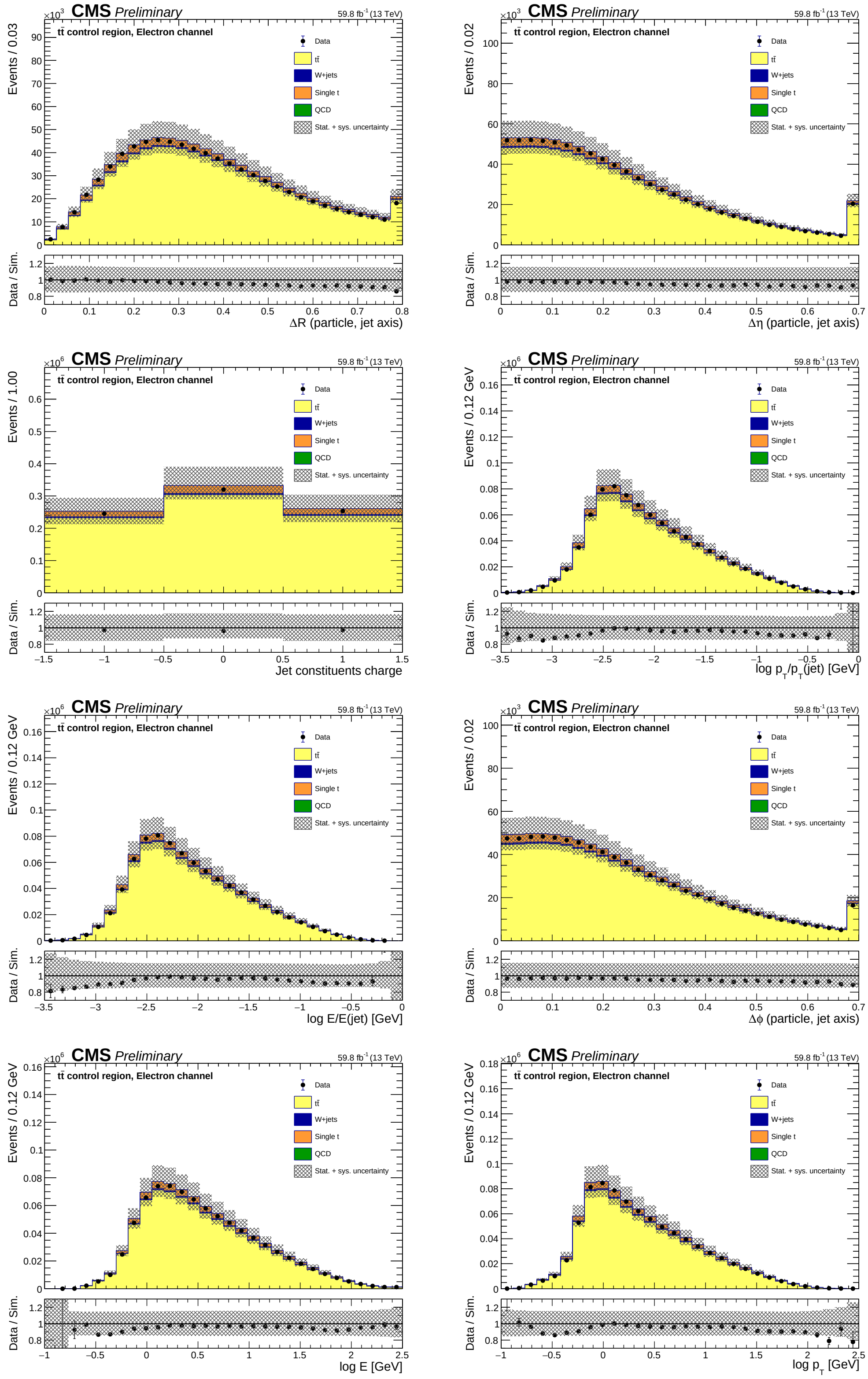


Figure B.1: Input variables of the jet charge tagger in data and in simulation using the $t\bar{t}$ control region in the electron decay channel. All inputs are well-modelled in simulation and are within the expected uncertainties.

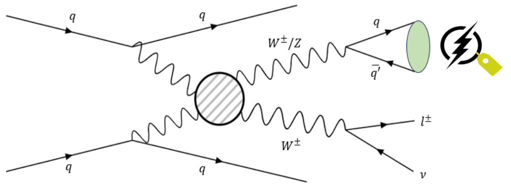
C. Jet charge tagger poster

A novel approach for distinguishing between hadronically decaying W$^+$, W$^-$, and Z bosons in the CMS experiment

CMS DP-2024/044

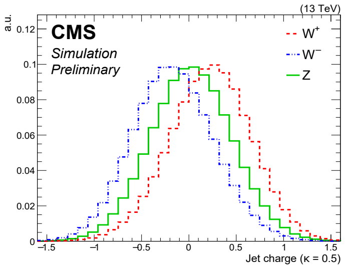
Komal Tauqeer
on behalf of the CMS Collaboration
Institute of Experimental Particle Physics (ETP), Karlsruhe Institute of Technology (KIT), Germany

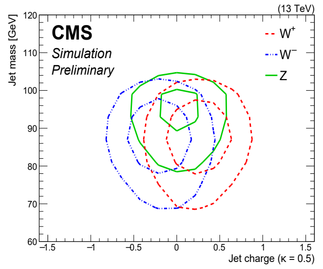
Motivation



- ❑ Same-sign WW Vector Boson Scattering (VBS), opposite-sign WW VBS, and WZ VBS are typically indistinguishable in the hadronic or semi-leptonic decay channels.
- ❑ These processes can be disentangled by identifying the charge of the jet(s).
- ❑ A novel method for identifying jet charge is presented in this poster, which can greatly help whenever the charge of the originating particle is a distinctive feature of a process.

Jet charge and jet mass



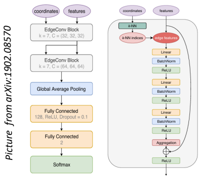


- ❑ Jet charge is defined as the p$_T$-weighted sum of the charge of all particles in the jet:

$$Q_\kappa = \frac{\sum_i q_i (p_T^i)^\kappa}{(p_T{}^{jet})^\kappa}$$

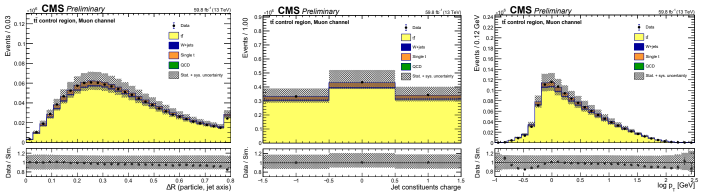
- ❑ The discrimination power of the jet charge and mass observables is weak by itself.

ParticleNet based jet charge tagger



Picture from arXiv:1902.08570

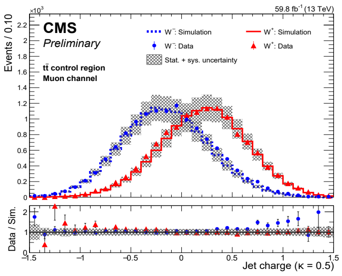
Input variable	Description
ΔR	angular separation between the particle and the jet axis$\sqrt{(\Delta\eta)^2+(\Delta\phi)^2}$
Δη	difference in the pseudorapidity between the particle and the jet axis
Δϕ	difference in the azimuthal angle between the particle and the jet axis
log E	logarithm of the particle’s energy
log p$_T$	logarithm of the particle’sp_T
log E/log E^{jet}	logarithm of the particle’s energy relative to the jet energy
log p$_T$/log p$_T$^{jet}	logarithm of the particle’sp_T relative to the jet p_T
Jet constituents charge	electric charge of the particle

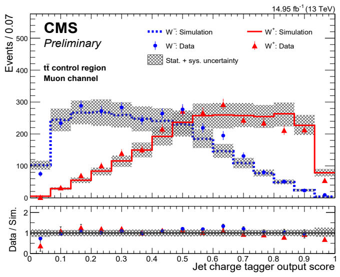


- ❑ The algorithm learns from low-level features of the jets that are well modelled in simulation to predict the charge.

Jet charge tagger as a binary classifier and validation in data

W$^+$ vs W$^-$

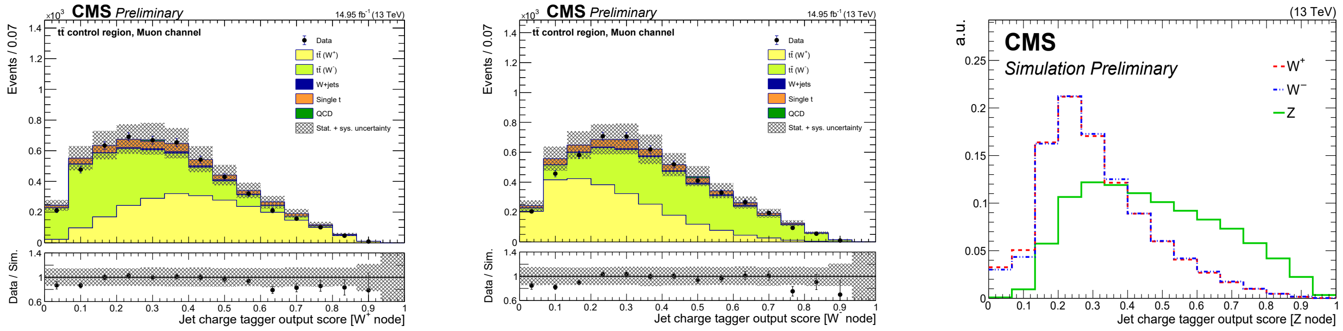




- ❑ W$^+$ and W$^-$ are better separated in the jet charge tagger output score compared to the jet charge variable.
- ❑ The method is well described in data as well.

Jet charge tagger as a multi-classifier and validation in data

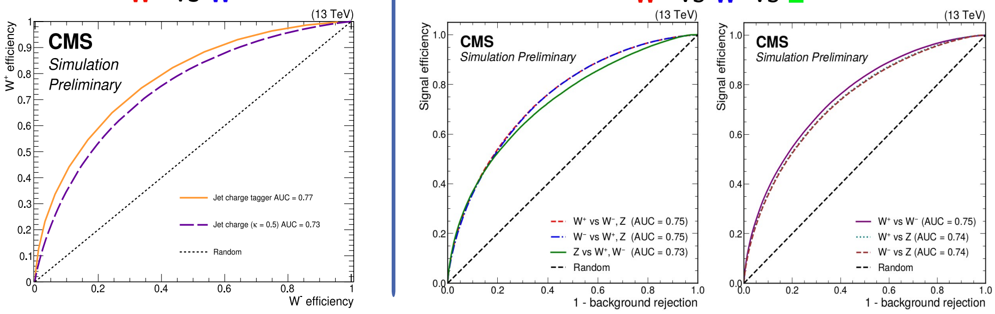
W$^+$ vs W$^-$ vs Z



Jet charge tagger output score separating boosted jets coming from the decay of W$^+$, W$^-$, and Z bosons

Performance evaluation

W$^+$ vs W$^-$ | W$^+$ vs W$^-$ vs Z



ParticleNet based jet charge tagger outperforms traditional cut-based methods

Conclusion

- ❑ The first study in CMS at the center-of-mass energy of 13 TeV to distinguish hadronic decays of W$^+$, W$^-$, and Z bosons.
- ❑ Use of machine learning based algorithm shows substantial improvement compared to the variable-based methods.
- ❑ The jet charge tagger performs equally well to classify all three types of jets.
- ❑ The best performance is for W$^+$ vs W$^-$.
- ❑ Similar performance in data, expect scale factors close to unity.

BOOST, Genova, 2024

scan to read the CMS DP-2024/044

komal.tauqeer@cern.ch

D. Samples used for the same-sign WW VBS search

Table D.1: Full names and the total number of events in datasets used for the search of same-sign WW VBS in the semi-leptonic decay channel.

Dataset	Number of events
2016 preVFP	
/SingleMuon/Run2016B-ver1_HIPM_UL2016_MiniAODv2-v2/MINIAOD	2,789,243
/SingleMuon/Run2016B-ver2_HIPM_UL2016_MiniAODv2-v2/MINIAOD	158,145,722
/SingleMuon/Run2016C-HIPM_UL2016_MiniAODv2-v2/MINIAOD	67,441,308
/SingleMuon/Run2016D-HIPM_UL2016_MiniAODv2-v2/MINIAOD	98,017,996
/SingleMuon/Run2016E-HIPM_UL2016_MiniAODv2-v2/MINIAOD	90,984,718
/SingleMuon/Run2016F-HIPM_UL2016_MiniAODv2-v2/MINIAOD	57,465,359
/SingleElectron/Run2016B-ver1_HIPM_UL2016_MiniAODv2-v2/MINIAOD	1,422,819
/SingleElectron/Run2016B-ver2_HIPM_UL2016_MiniAODv2-v2/MINIAOD	246,440,440
/SingleElectron/Run2016C-HIPM_UL2016_MiniAODv2-v2/MINIAOD	97,259,854
/SingleElectron/Run2016D-HIPM_UL2016_MiniAODv2-v2/MINIAOD	148,167,727
/SingleElectron/Run2016E-HIPM_UL2016_MiniAODv2-v5/MINIAOD	117,269,446
/SingleElectron/Run2016F-HIPM_UL2016_MiniAODv2-v2/MINIAOD	61,735,326
2016 postVFP	
/SingleMuon/Run2016F-UL2016_MiniAODv2-v2/MINIAOD	8,024,195
/SingleMuon/Run2016G-UL2016_MiniAODv2-v2/MINIAOD	149,916,849
/SingleMuon/Run2016H-UL2016_MiniAODv2-v2/MINIAOD	174,035,164
/SingleElectron/Run2016F-UL2016_MiniAODv2-v2/MINIAOD	8858,206
/SingleElectron/Run2016G-UL2016_MiniAODv2-v2/MINIAOD	153,363,109
/SingleElectron/Run2016H-UL2016_MiniAODv2-v2/MINIAOD	129,021,893
2017	
/SingleMuon/Run2017B-UL2017_MiniAODv2-v1/MINIAOD	136,300,266
/SingleMuon/Run2017C-UL2017_MiniAODv2-v1/MINIAOD	165,652,756
/SingleMuon/Run2017D-UL2017_MiniAODv2-v1/MINIAOD	70,361,660
/SingleMuon/Run2017E-UL2017_MiniAODv2-v1/MINIAOD	154,618,774
/SingleMuon/Run2017F-UL2017_MiniAODv2-v1/MINIAOD	242,140,980
/SingleElectron/Run2017B-UL2017_MiniAODv2-v1/MINIAOD	60,537,490
/SingleElectron/Run2017C-UL2017_MiniAODv2-v1/MINIAOD	136,637,888
/SingleElectron/Run2017D-UL2017_MiniAODv2-v1/MINIAOD	51,526,521
/SingleElectron/Run2017E-UL2017_MiniAODv2-v1/MINIAOD	102,122,055
/SingleElectron/Run2017F-UL2017_MiniAODv2-v1/MINIAOD	128,467,223
2018	
/SingleMuon/Run2018A-UL2018_MiniAODv2_GT36-v1/MINIAOD	241,591,525
/SingleMuon/Run2018B-UL2018_MiniAODv2_GT36-v1/MINIAOD	119,918,017
/SingleMuon/Run2018C-UL2018_MiniAODv2_GT36-v2/MINIAOD	110,032,072
/SingleMuon/Run2018D-UL2018_MiniAODv2_GT36-v1/MINIAOD	513,884,680
/EGamma/Run2018A-UL2018_MiniAODv2_GT36-v1/MINIAOD	339,013,231
/EGamma/Run2018B-UL2018_MiniAODv2_GT36-v1/MINIAOD	153,822,427
/EGamma/Run2018C-UL2018_MiniAODv2_GT36-v1/MINIAOD	147,827,904
/EGamma/Run2018D-UL2018_MiniAODv2-v2/MINIAOD	752,497,815

Table D.2: Background Monte Carlo simulations used in the search of same-sign WW VBS.

Process	**Dataset name**	**Cross section (pb)**
Semileptonic $t\bar{t}$	TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8	370.62
Single top s-channel	ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-amcatnlo-pythia8	3.364
Single top t-channel (top)	ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	136.02
Single top t-channel (antitop)	ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	80.95
Single top tW-channel (top)	ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8	35.85
Single top tW-channel (antitop)	ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8	35.85
W+Jets LO HT 70-100	WJetsToLNu_HT-70To100_TuneCP5_13TeV-madgraphMLM-pythia8	1529.44
W+Jets LO HT 100-200	WJetsToLNu_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8	1519.76
W+Jets LO HT 200-400	WJetsToLNu_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8	405.96
W+Jets LO HT 400-600	WJetsToLNu_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8	54.75
W+Jets LO HT 600-800	WJetsToLNu_HT-600To800_TuneCP5_13TeV-madgraphMLM-pythia8	13.27
W+Jets LO HT 800-1200	WJetsToLNu_HT-800To1200_TuneCP5_13TeV-madgraphMLM-pythia8	5.97
W+Jets LO HT 1200-2500	WJetsToLNu_HT-1200To2500_TuneCP5_13TeV-madgraphMLM-pythia8	1.40
W+Jets LO HT 2500-Inf	WJetsToLNu_HT-2500ToInf_TuneCP5_13TeV-madgraphMLM-pythia8	0.0317
W+Jets LO Inclusive	WJetsToLNu_TuneCP5_13TeV-madgraphMLM-pythia8	61526.7
QCD-VV	WplusToLNuWplusTo2JJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	0.07557
QCD-VV	WminusToLNuWminusTo2JJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	0.03276
QCD-VV	WplusTo2JWminusToLNuJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	4.853
QCD-VV	WplusToLNuWminusTo2JJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	4.859
QCD-VV	WminusToLNuZTo2JJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	1.122
QCD-VV	WminusTo2JZTo2LJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	0.3334
QCD-VV	WplusToLNuZTo2JJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	1.868
QCD-VV	WplusTo2JZTo2LJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	0.5558
QCD-VV	ZTo2LZTo2JJJ_QCD_LO_SM_MJJ100PTJ10_TuneCP5_13TeV-madgraph-pythia8	0.328

E. Additional $t\bar{t}$ control plots

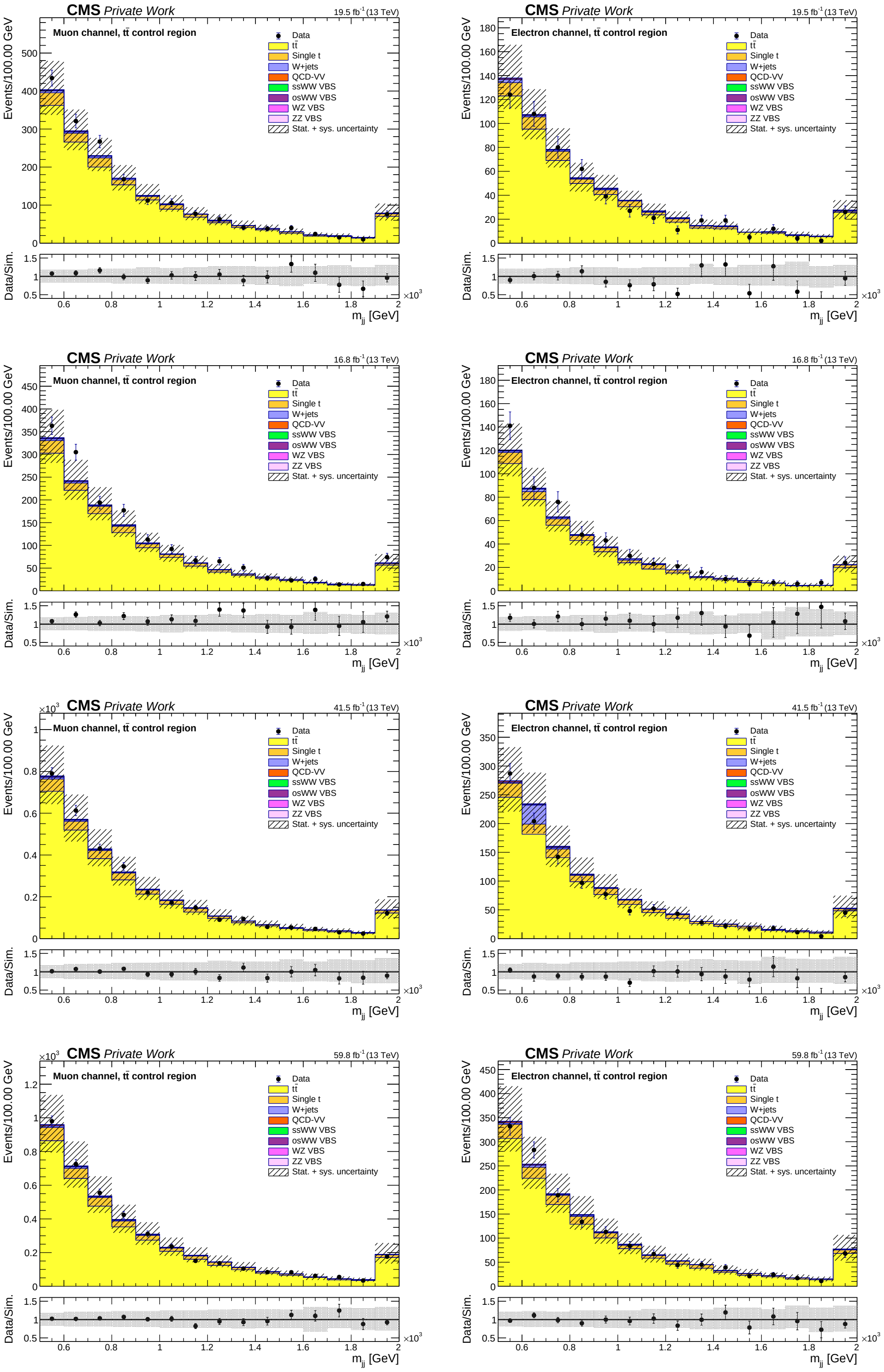


Figure E.1: Data and simulation comparison of **dijet invariant mass** ($\mathbf{m}_{\rm jj}$) of the two VBS jets in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

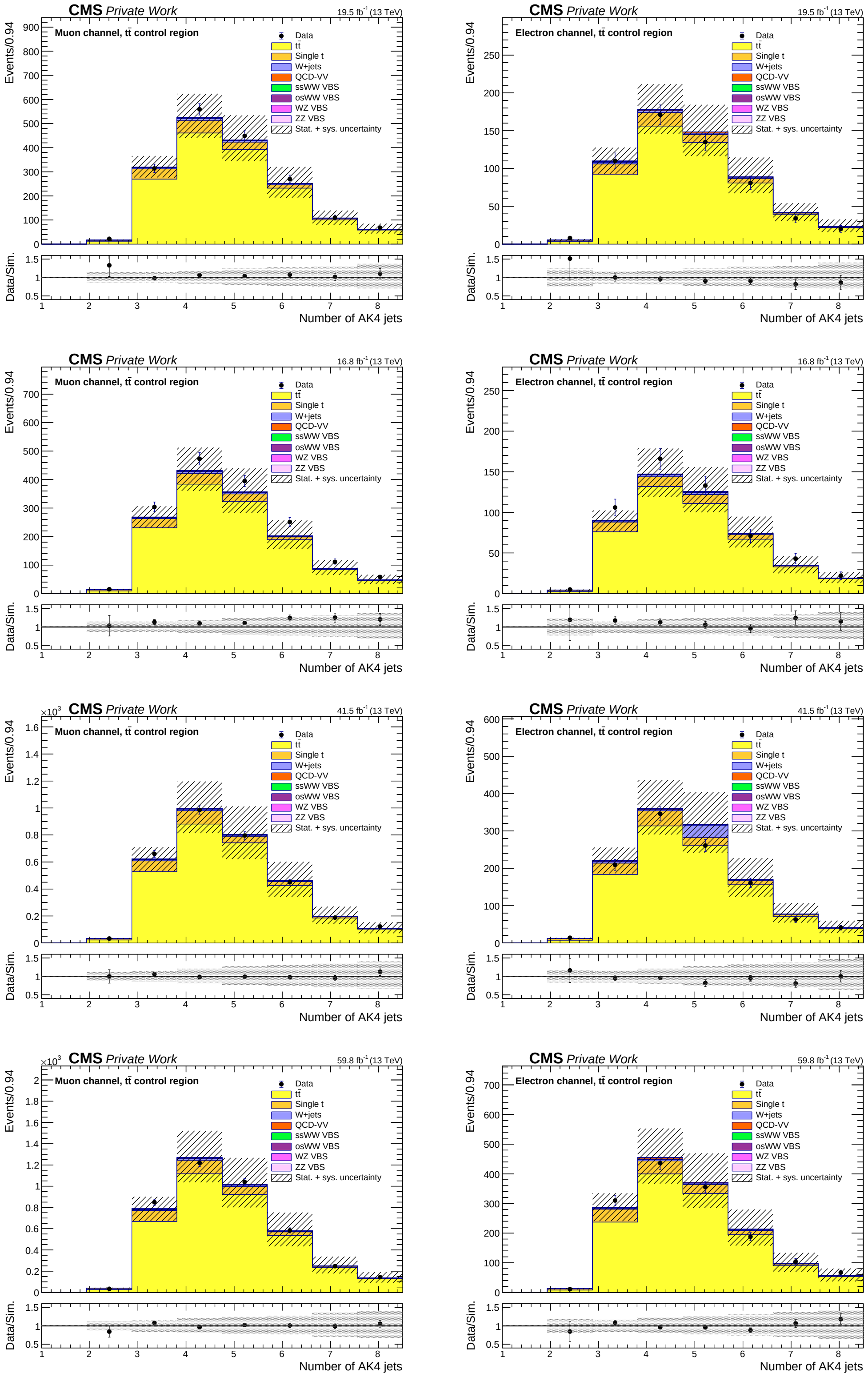


Figure E.2: Data and simulation comparison of **number of AK4 jets** in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

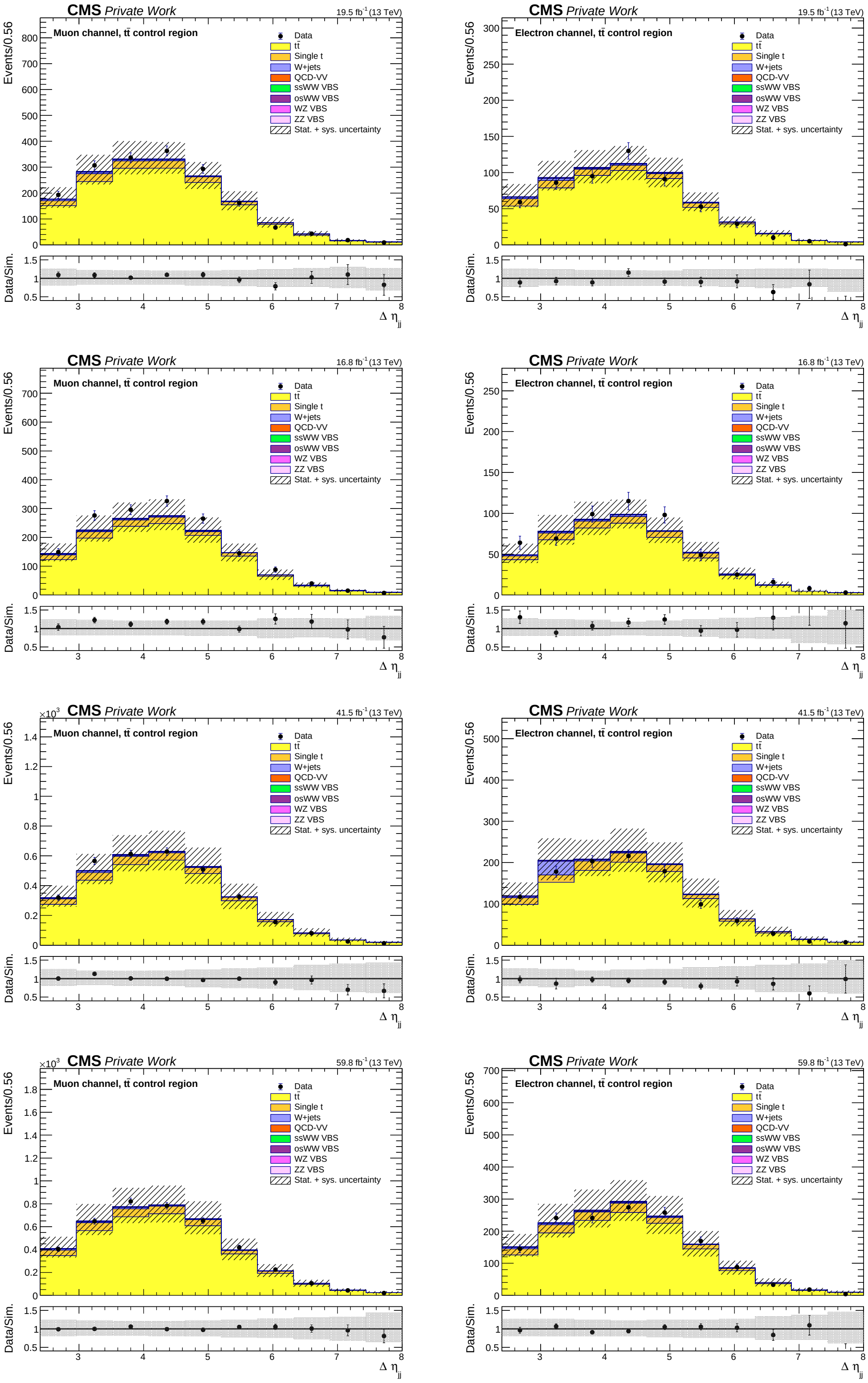


Figure E.3: Data and simulation comparison of **psuedorapidity separation** ($\boldsymbol{\Delta}\eta_{\mathrm{jj}}$) of the two VBS jets in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

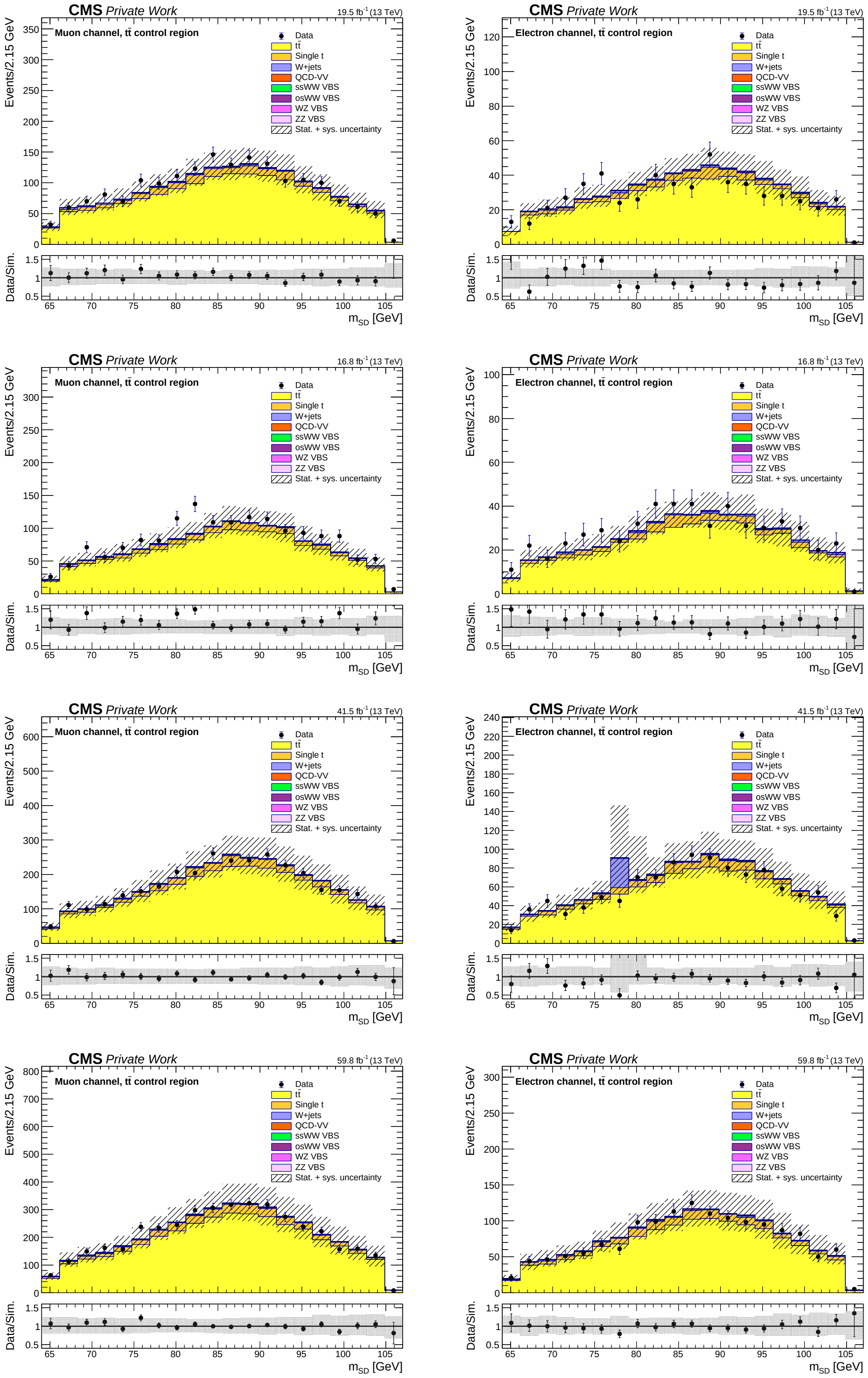


Figure E.4: Data and simulation comparison of the **soft drop mass of the AK8 jet** (**m**$_{\text{SD}}$) in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

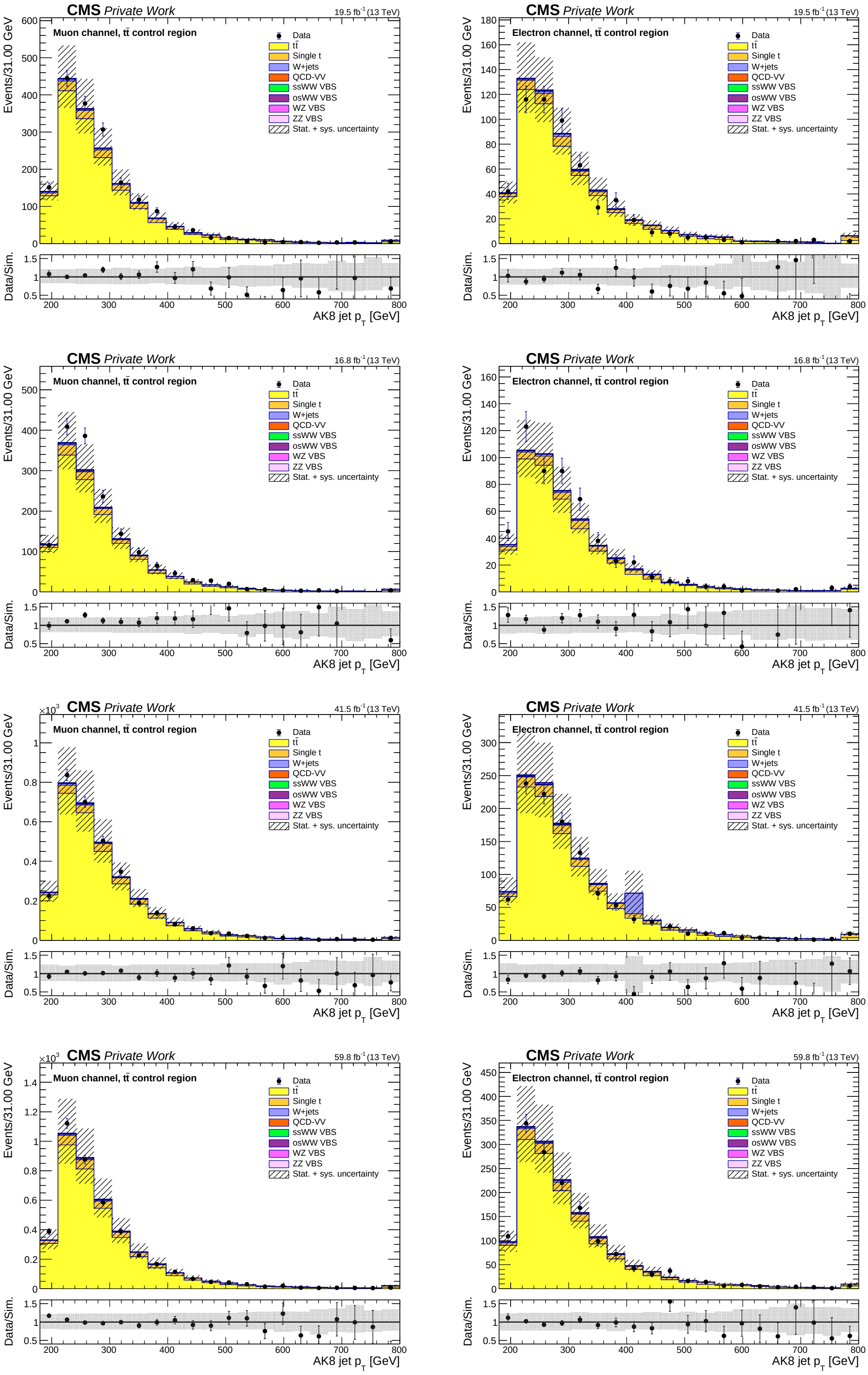


Figure E.5: Data and simulation comparison of the **transverse momentum of the AK8 jet** in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

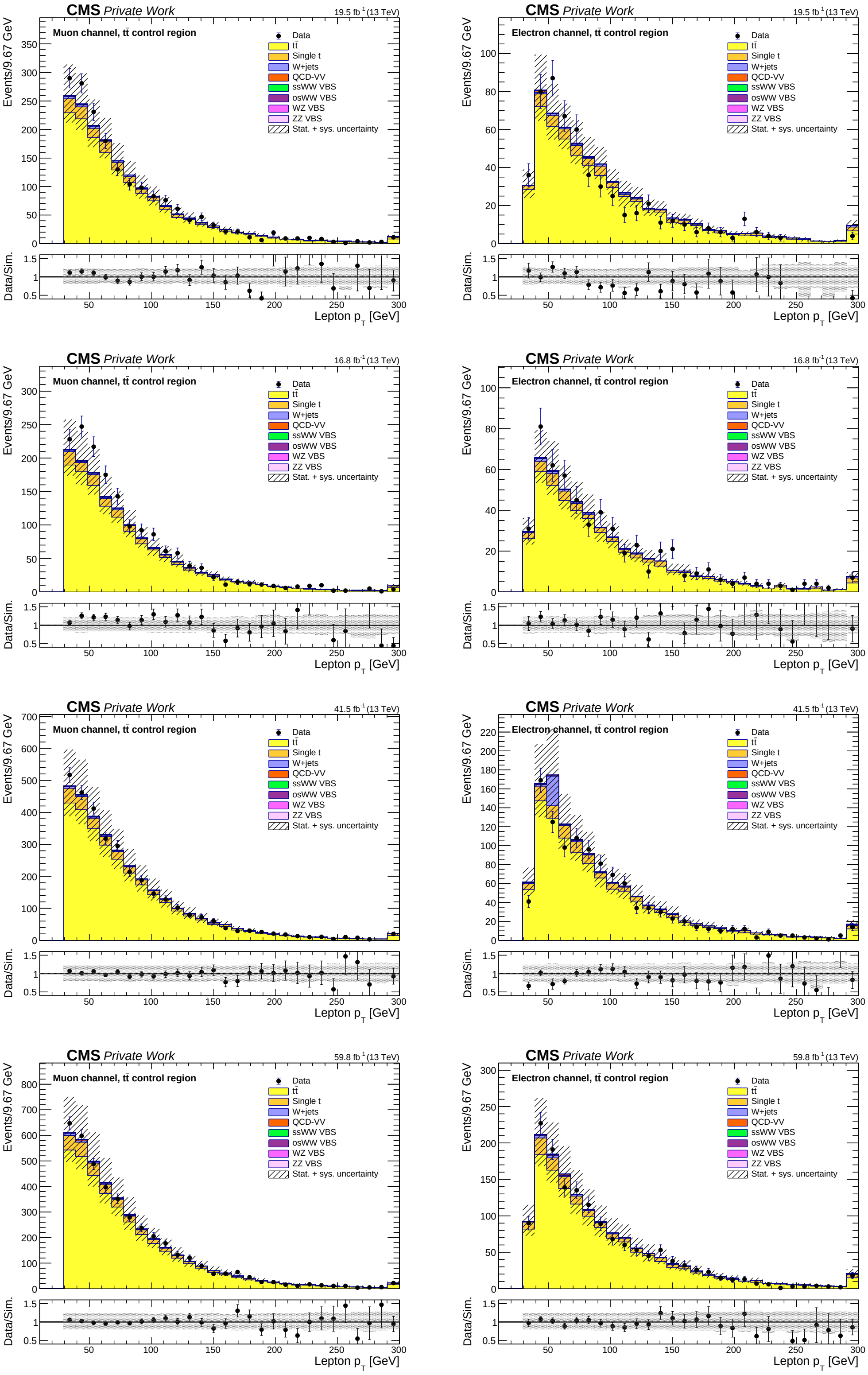


Figure E.6: Data and simulation comparison of the **transverse momentum of the lepton** in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

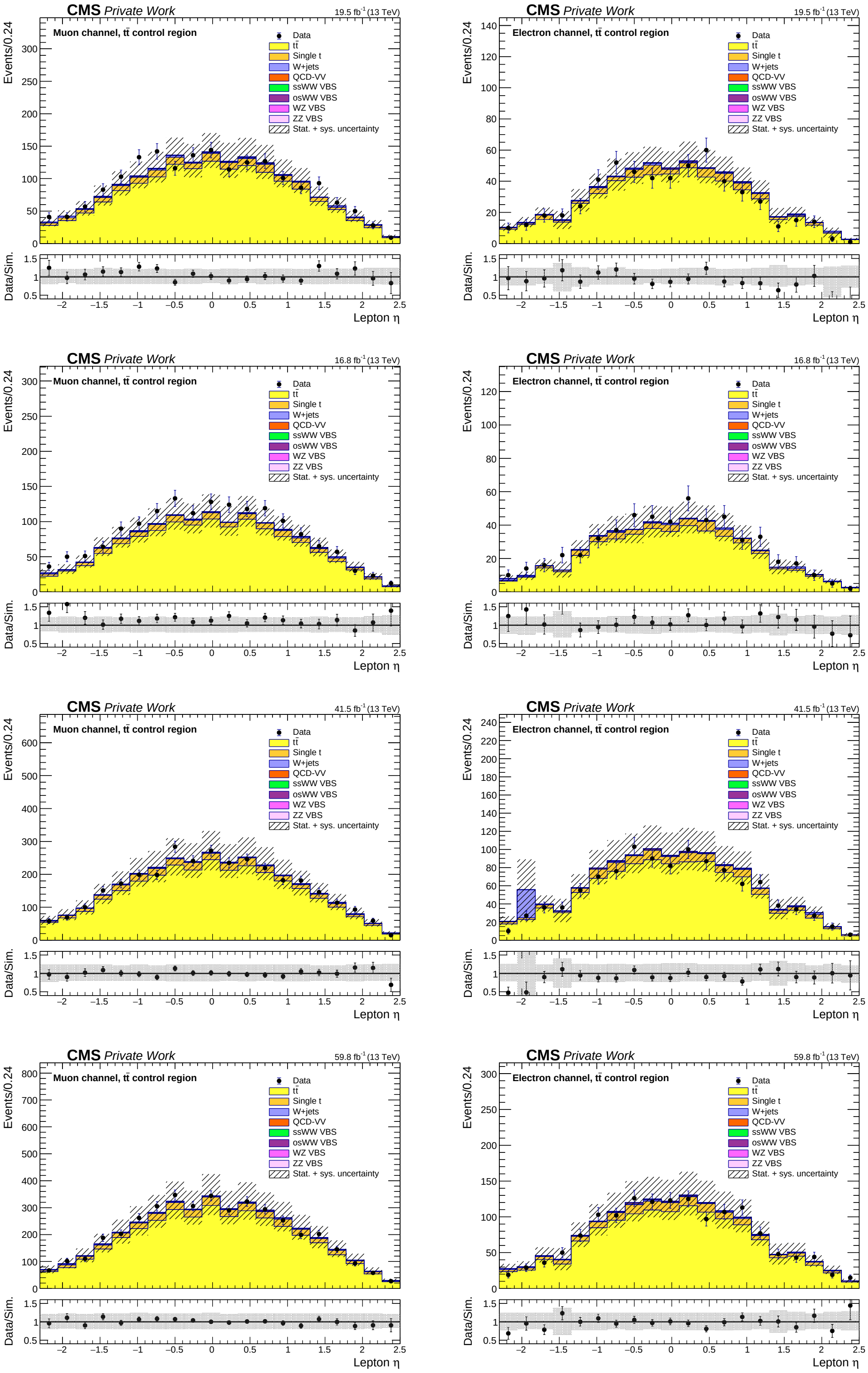


Figure E.7: Data and simulation comparison of the **pseudorapidity of the lepton** in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

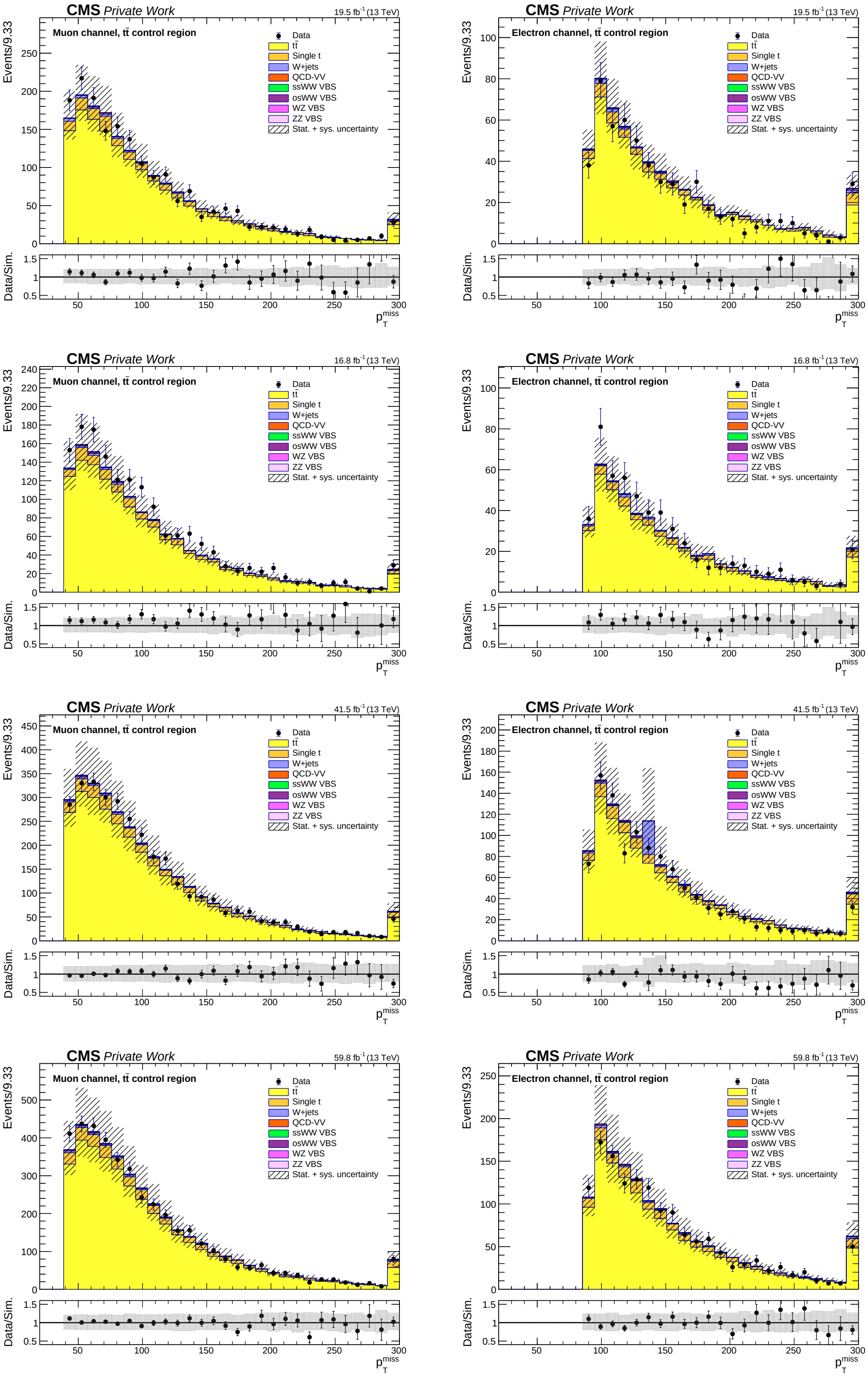


Figure E.8: Data and simulation comparison of the **missing transverse momentum** ($\mathbf{p_T}^{\mathrm{miss}}$) in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

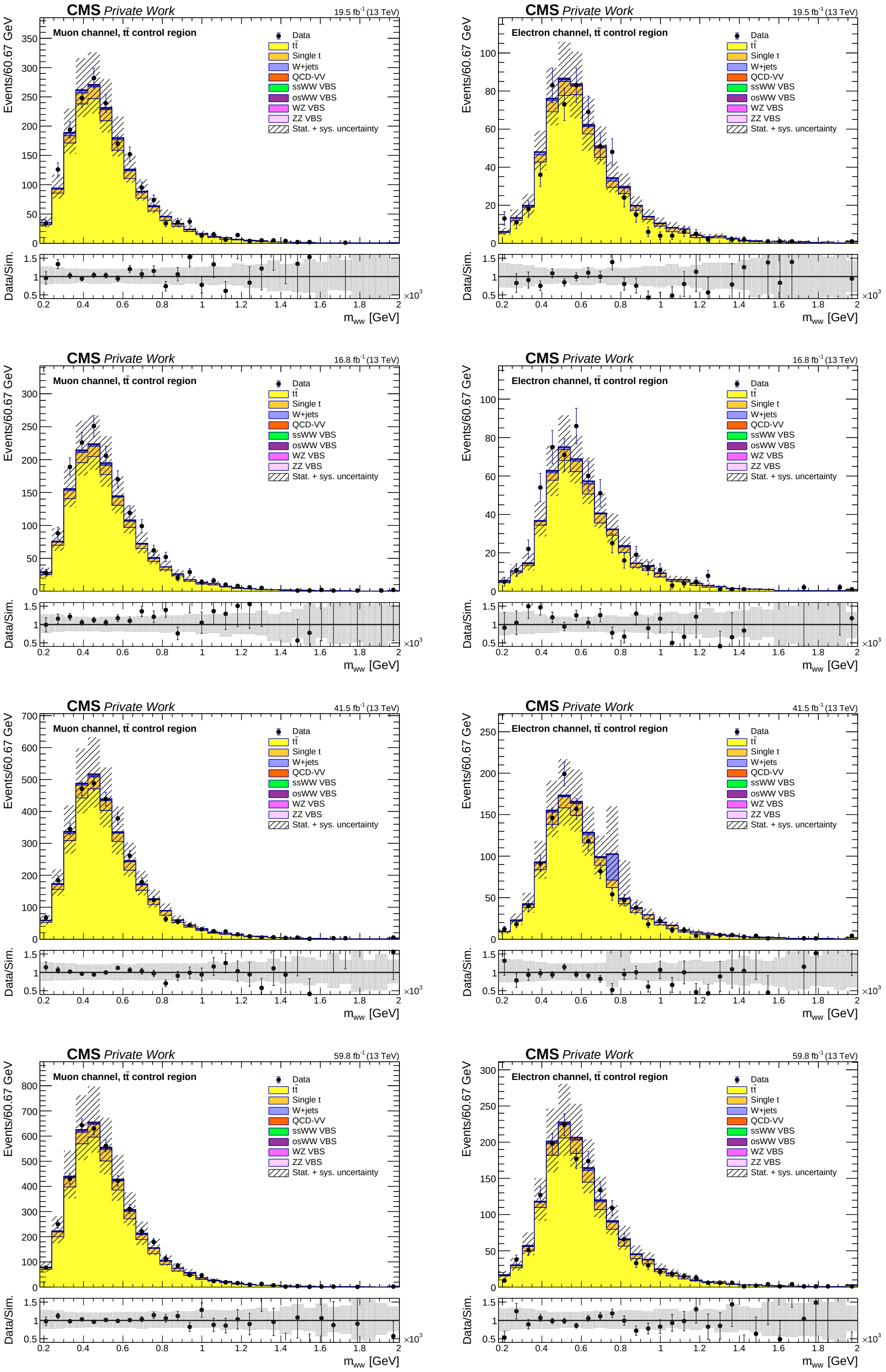


Figure E.9: Data and simulation comparison of the **invariant mass of the two W bosons** ($\mathbf{m}_{\mathrm{ww}}$) in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

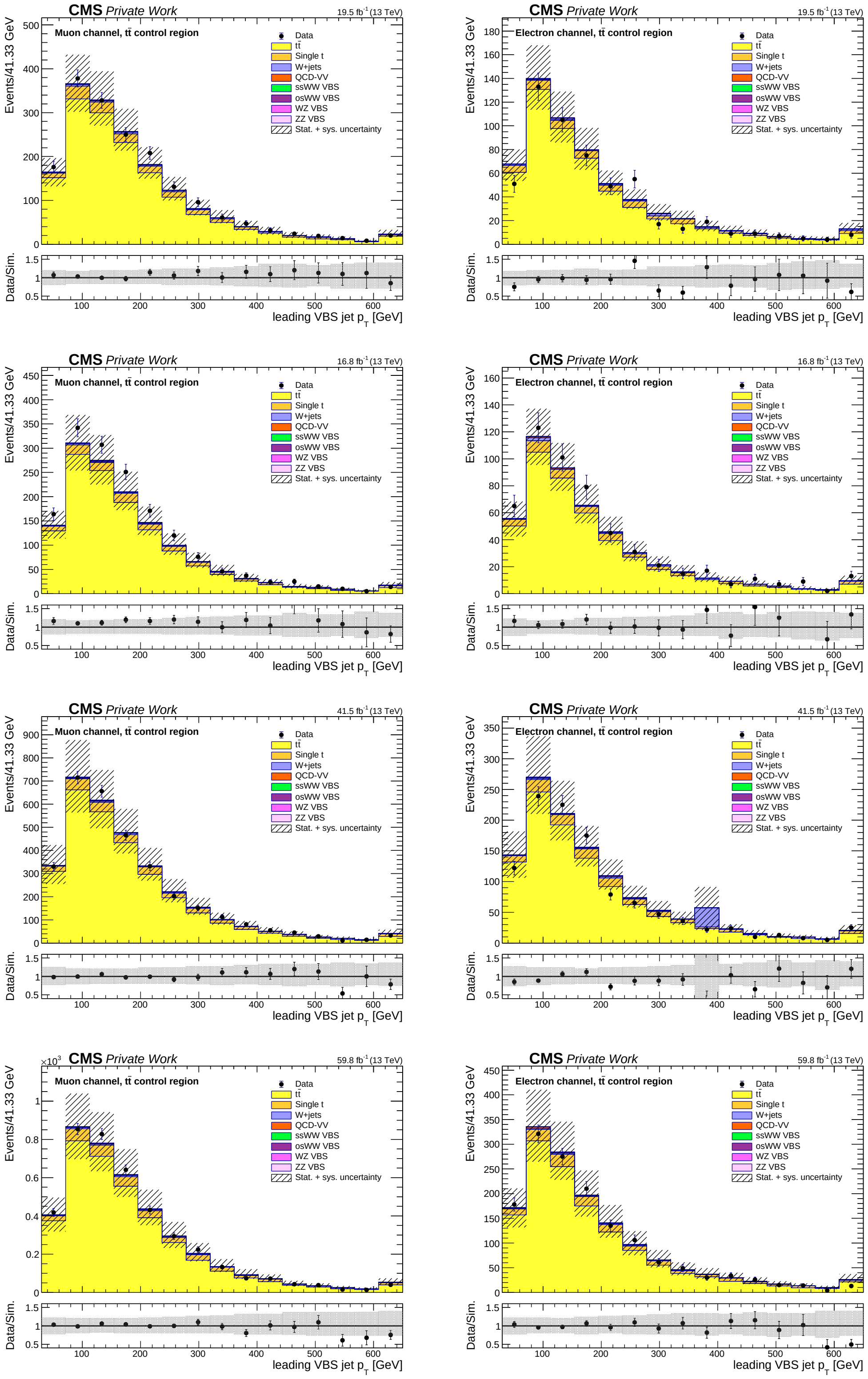


Figure E.10: Data and simulation comparison of the **leading VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

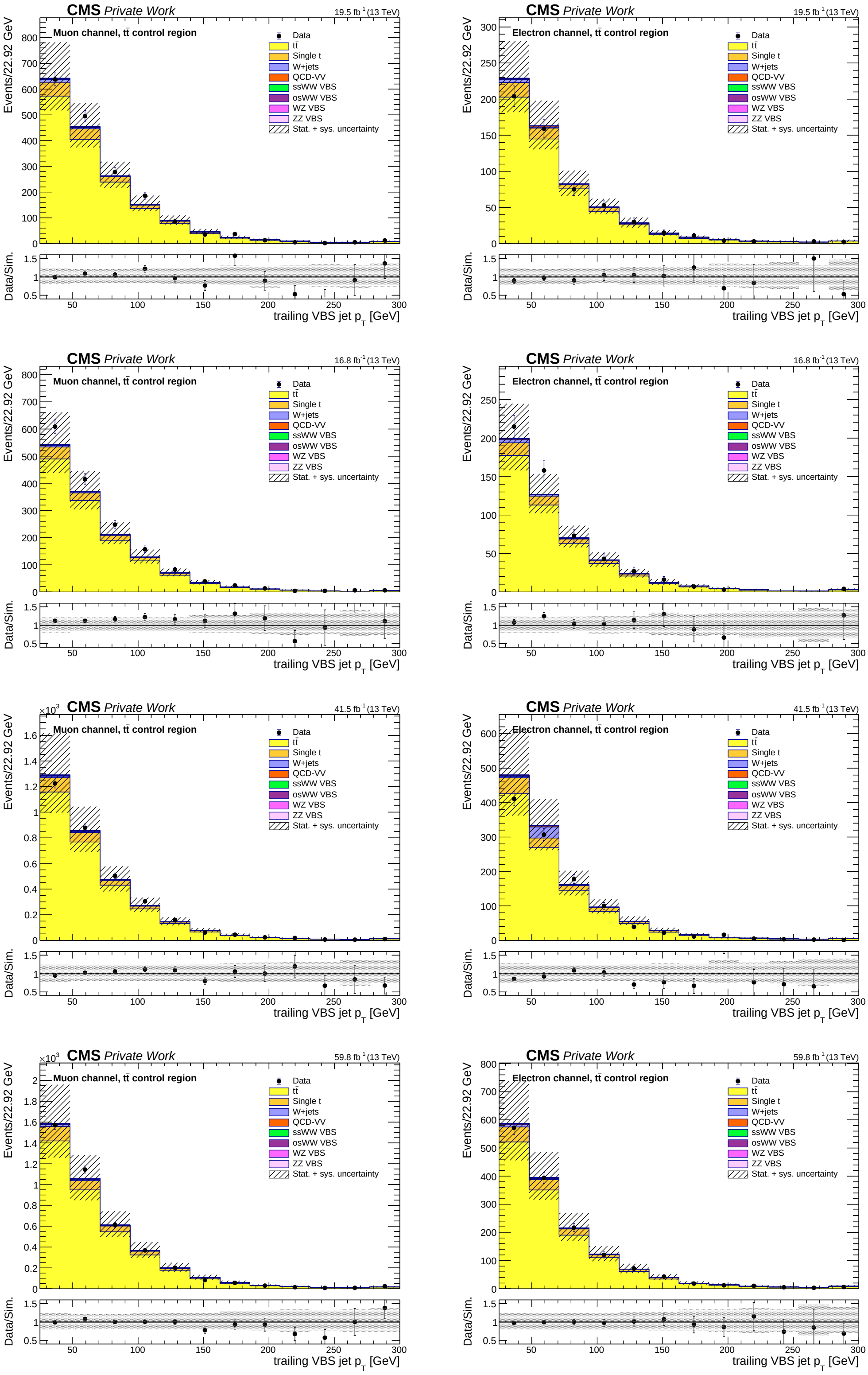


Figure E.11: Data and simulation comparison of the **trailing VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

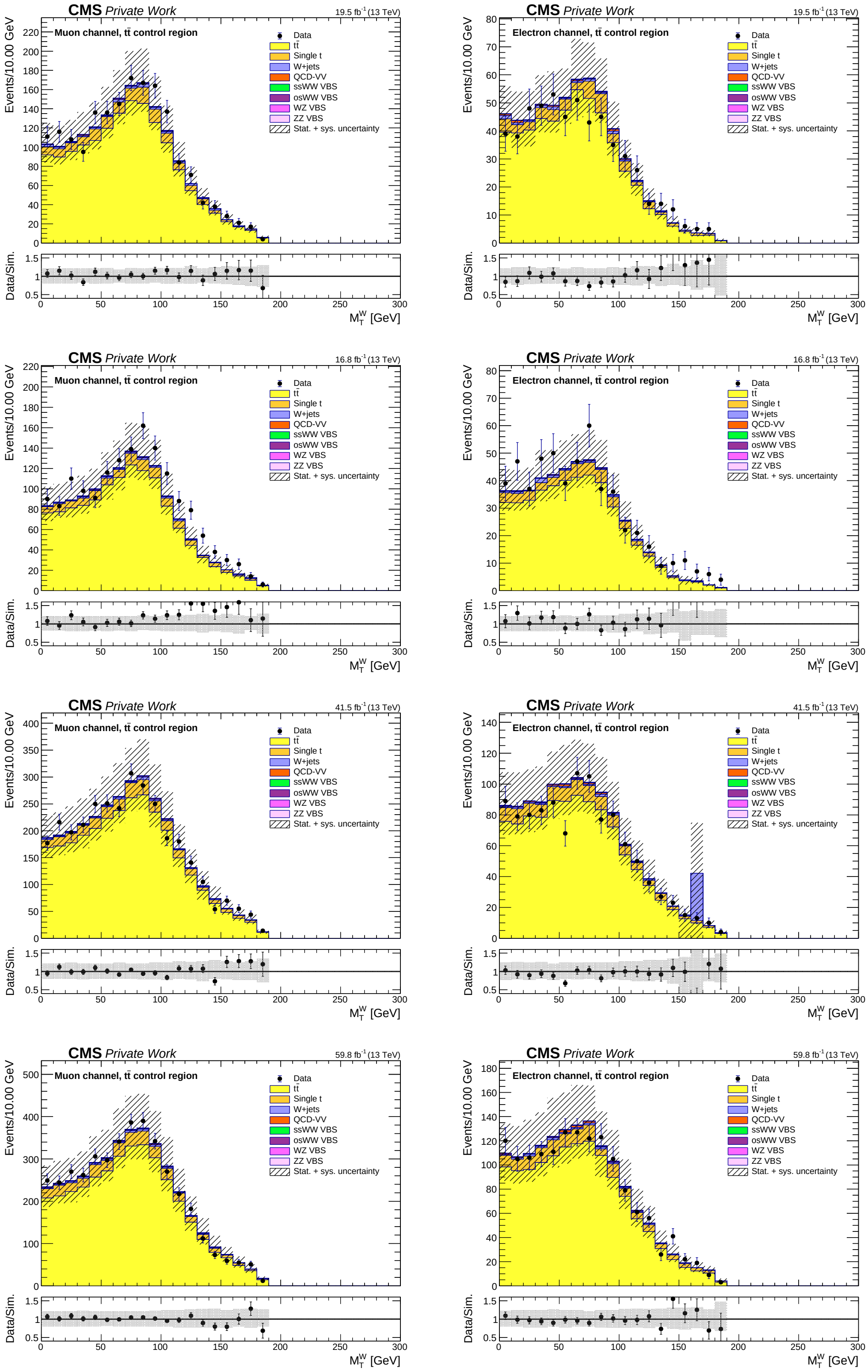


Figure E.12: Data and simulation comparison of the **transverse mass of leptonically decaying W boson** ($\mathrm{M_T}^\mathrm{W}$) in the muon channel (left) and the electron channel (right) in the $\mathrm{t\bar{t}}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

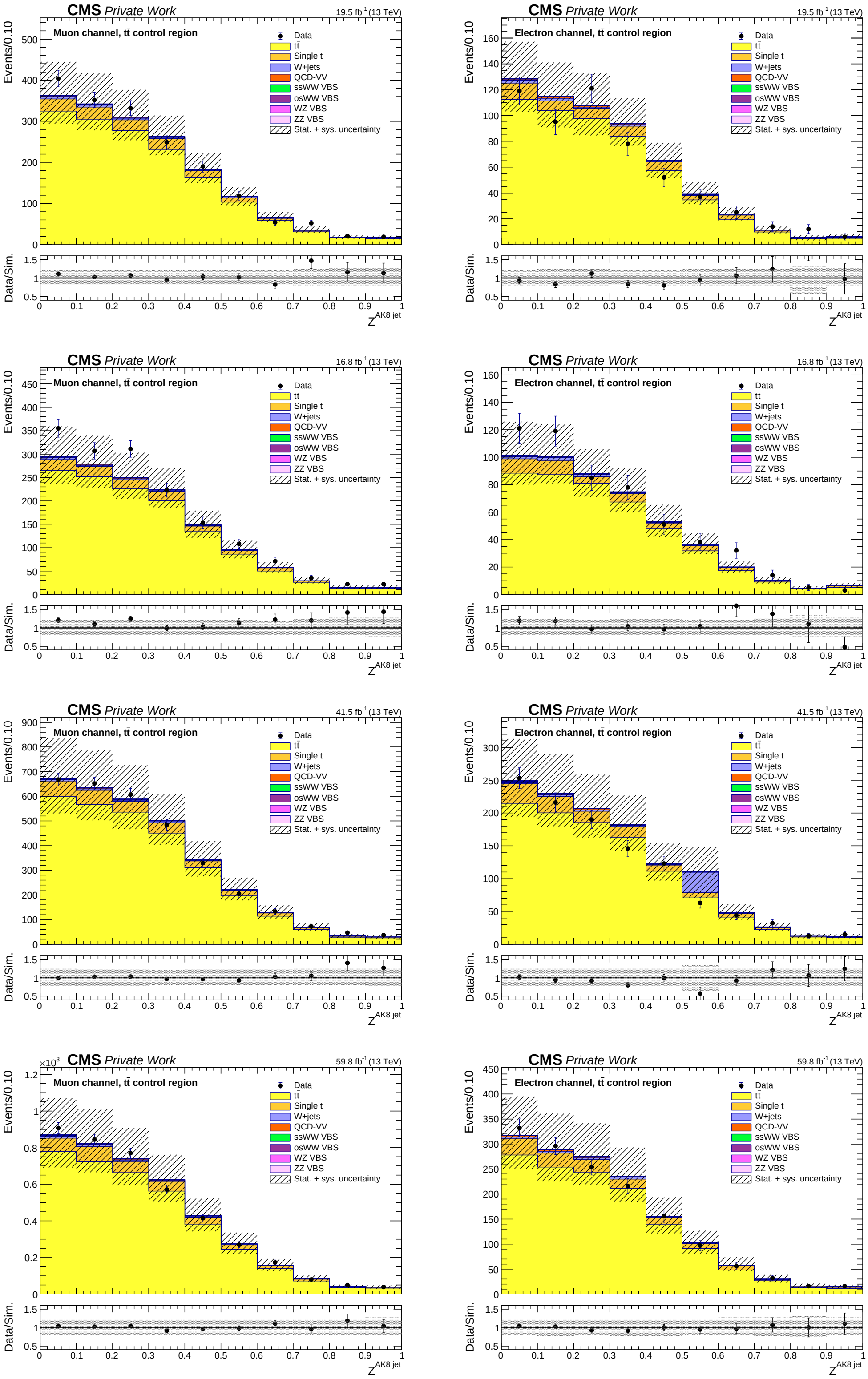


Figure E.13: Data and simulation comparison of **Zeppenfeld variable of the AK8 jet** in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

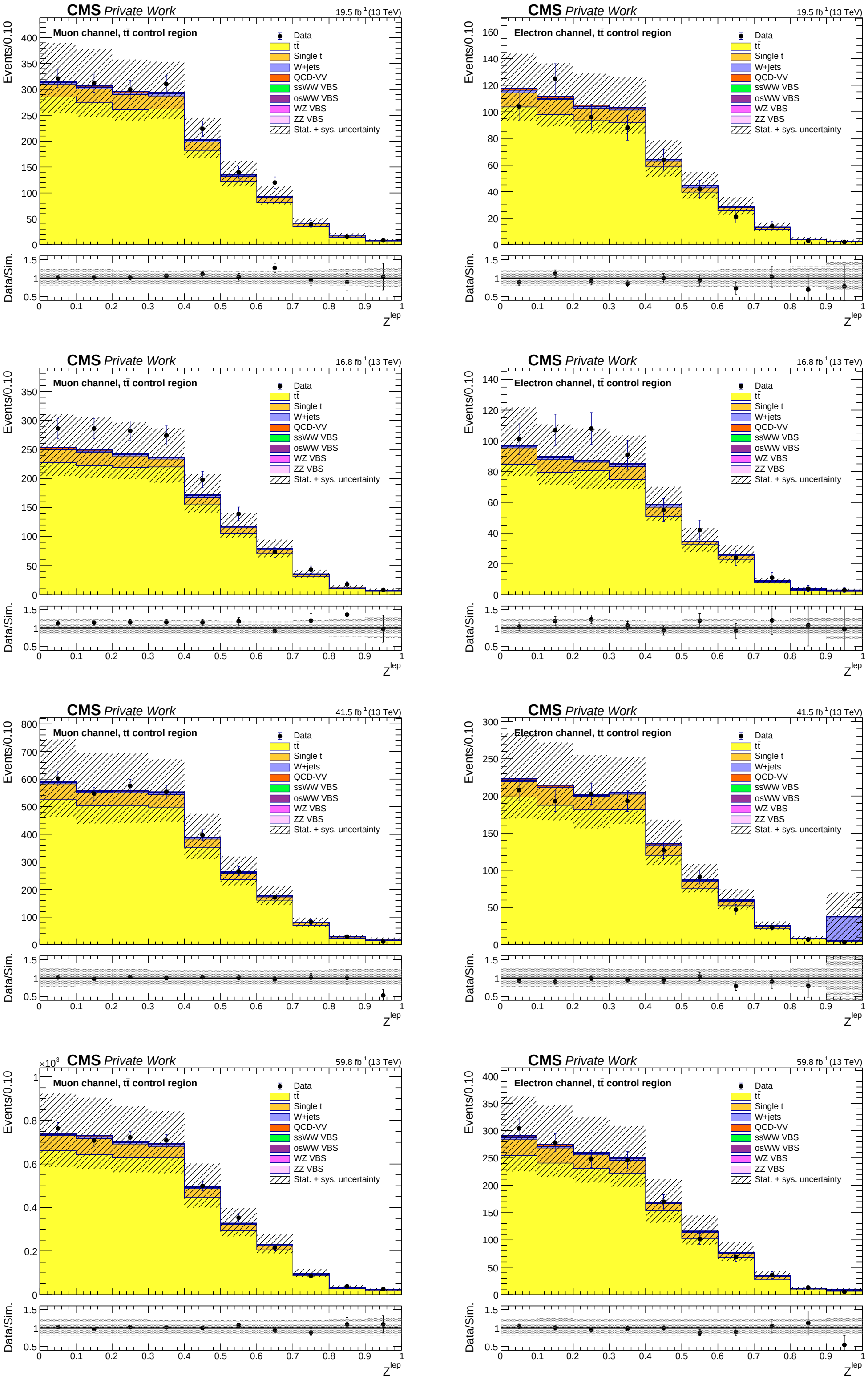


Figure E.14: Data and simulation comparison of **Zeppenfeld variable of the lepton** in the muon channel (left) and the electron channel (right) in the $t\bar{t}$ control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

F. Additional W+jets control plots after corrections

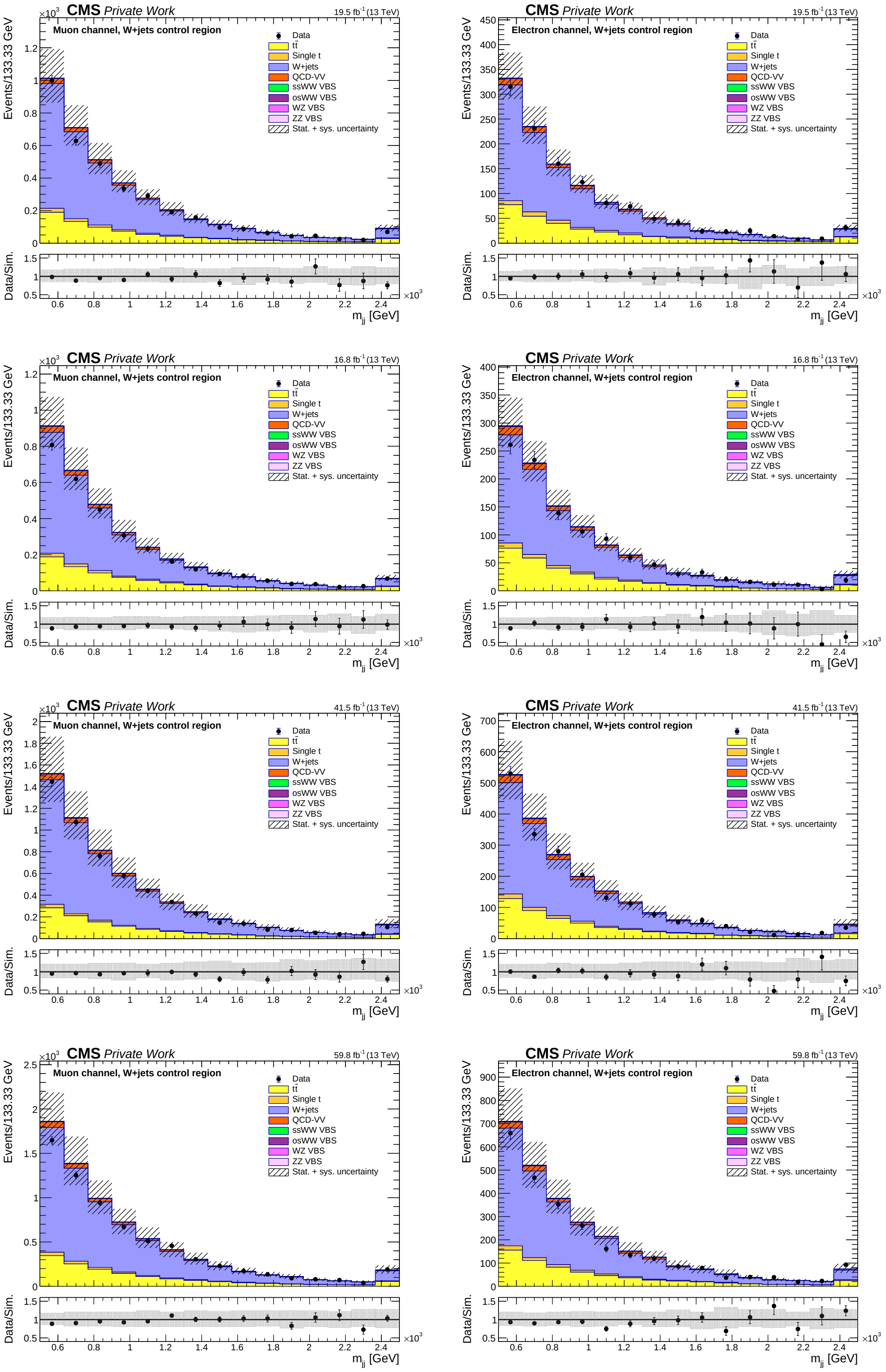


Figure F.1: Data and simulation comparison of **dijet invariant mass** ($\mathbf{m_{jj}}$) of the two VBS jets in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

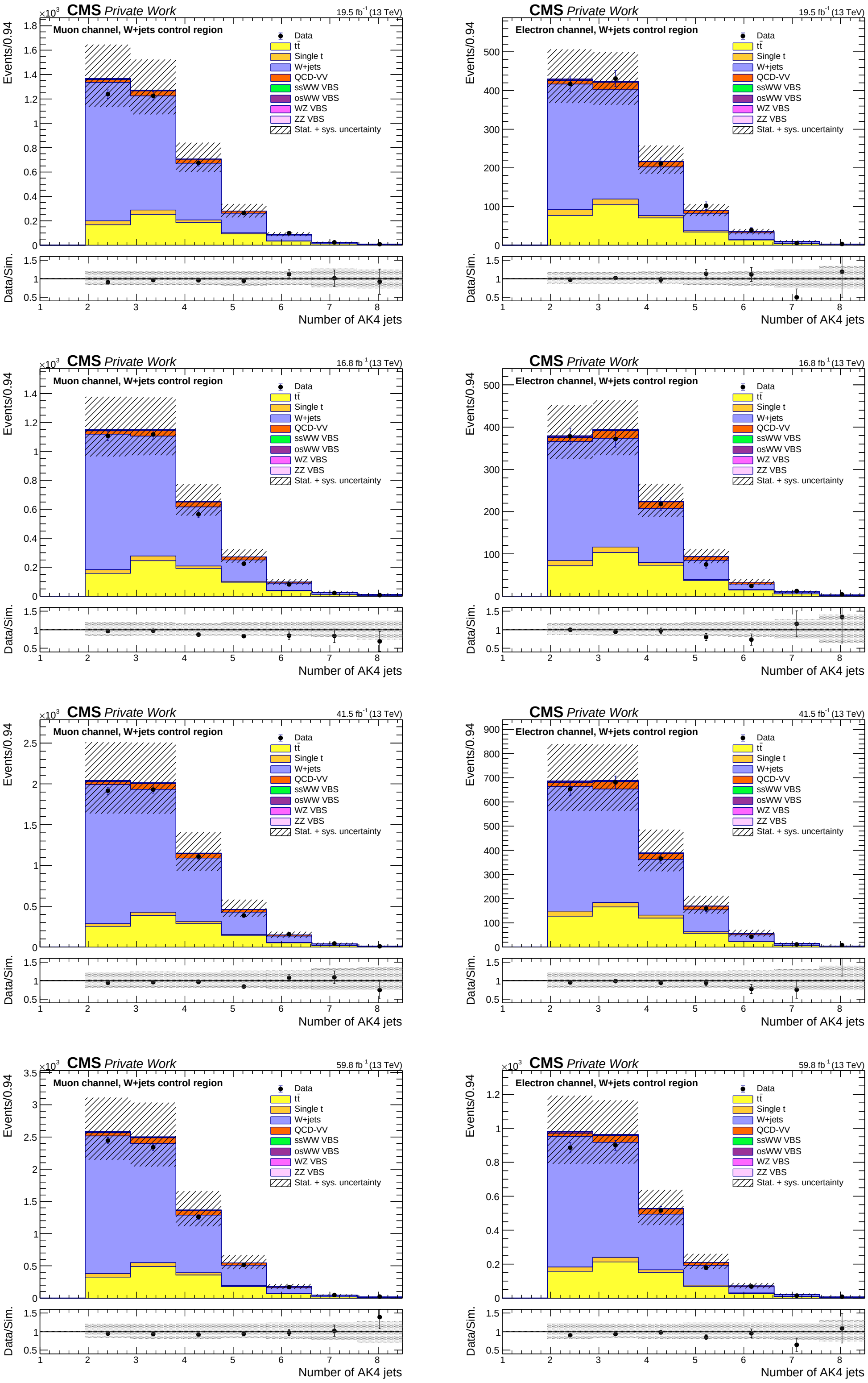


Figure F.2: Data and simulation comparison of **number of AK4 jets** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

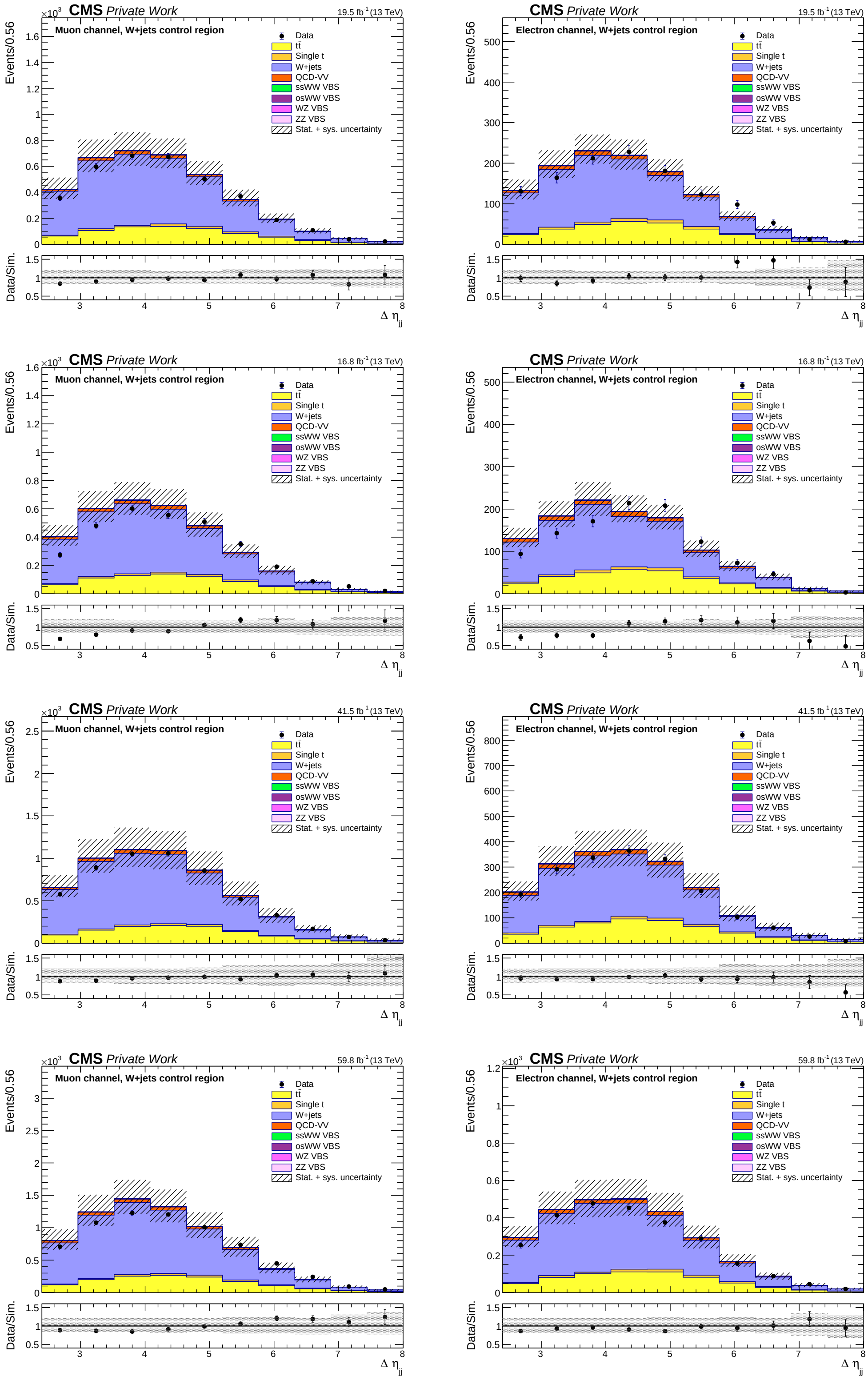


Figure F.3: Data and simulation comparison of **psuedorapidity separation** ($\boldsymbol{\Delta}\eta_{\mathrm{jj}}$) of the two VBS jets in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

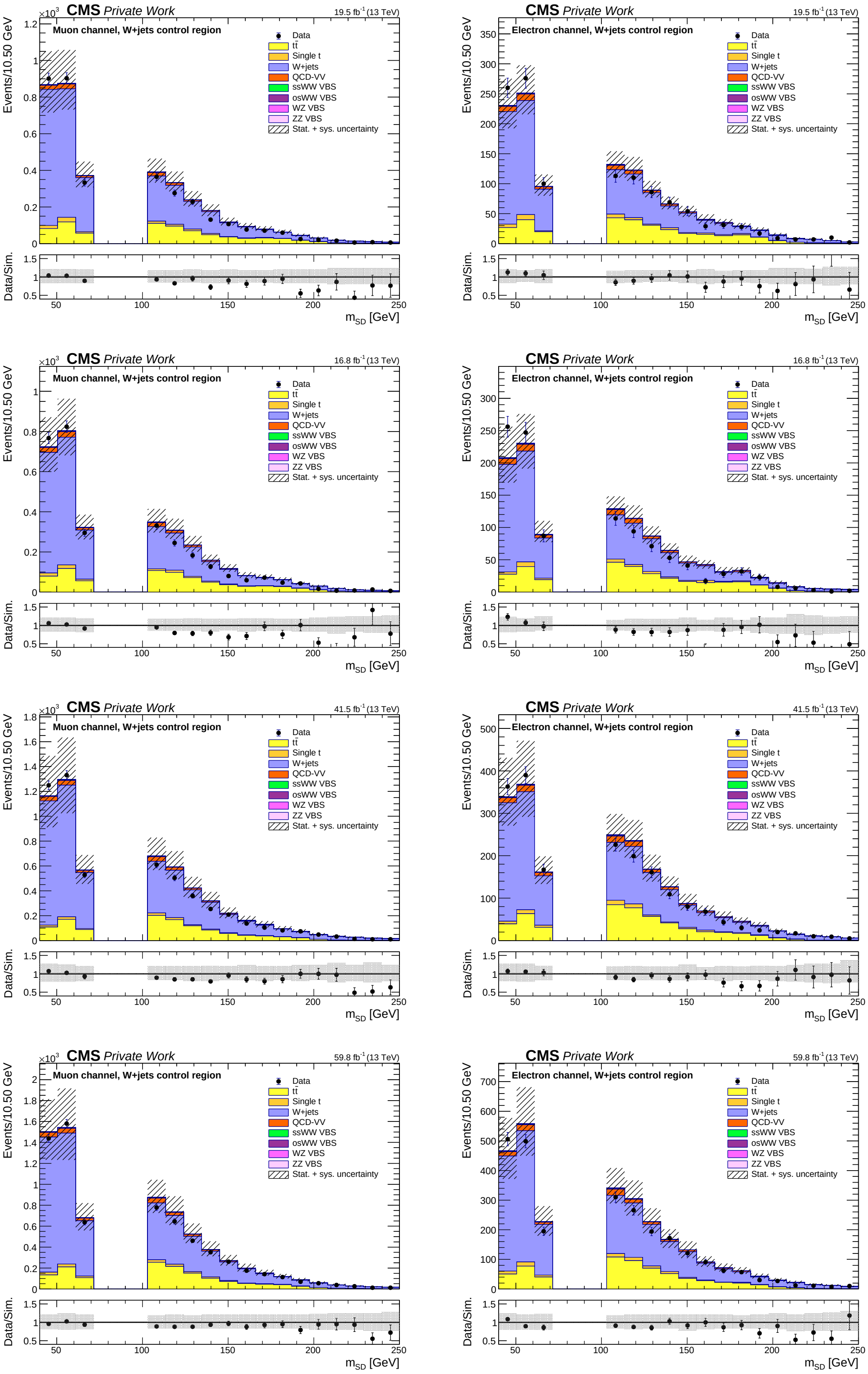


Figure F.4: Data and simulation comparison of the **soft drop mass of the AK8 jet** ($\mathbf{m}_{\text{SD}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

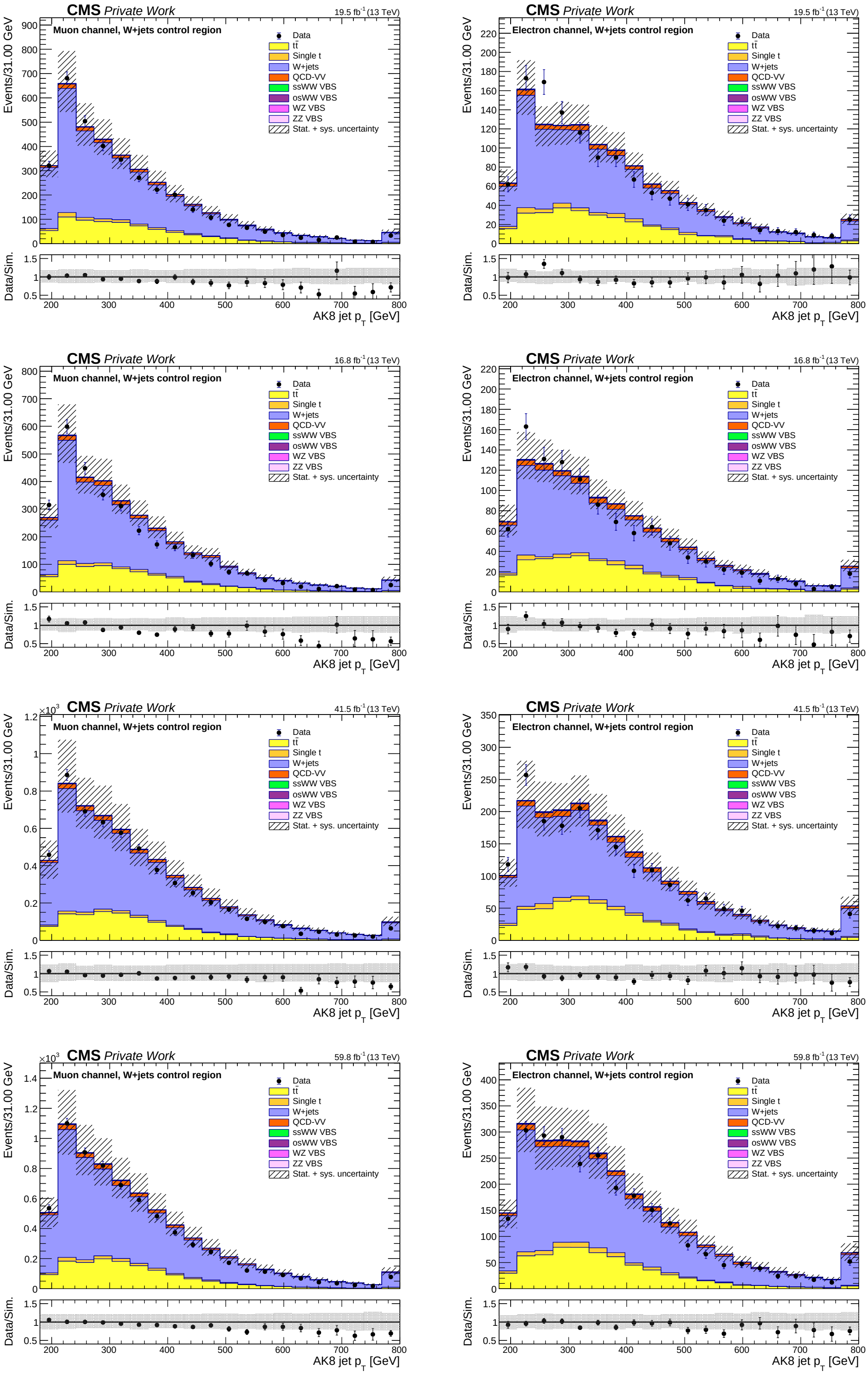


Figure F.5: Data and simulation comparison of the **transverse momentum of the AK8 jet** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

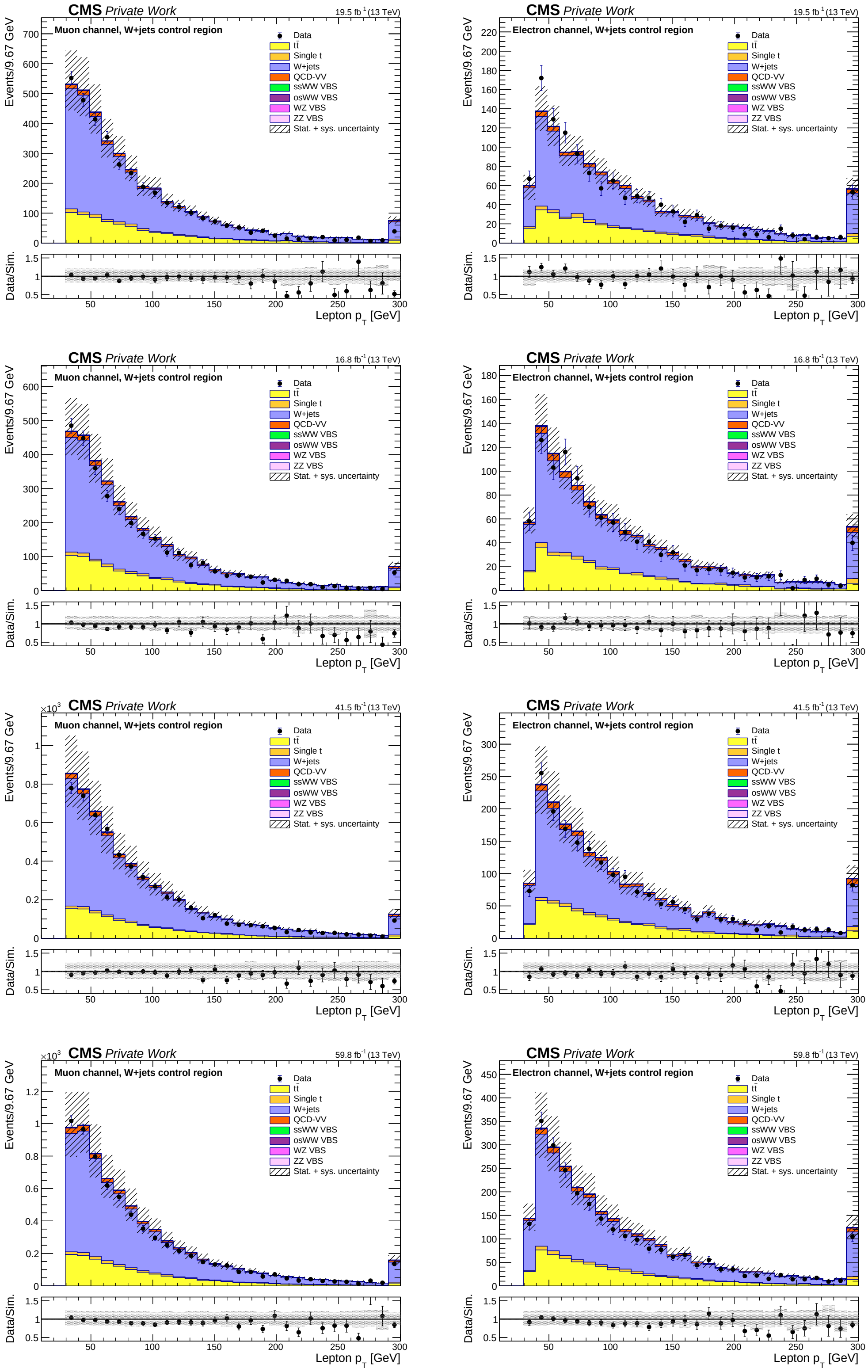


Figure F.6: Data and simulation comparison of the **transverse momentum of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

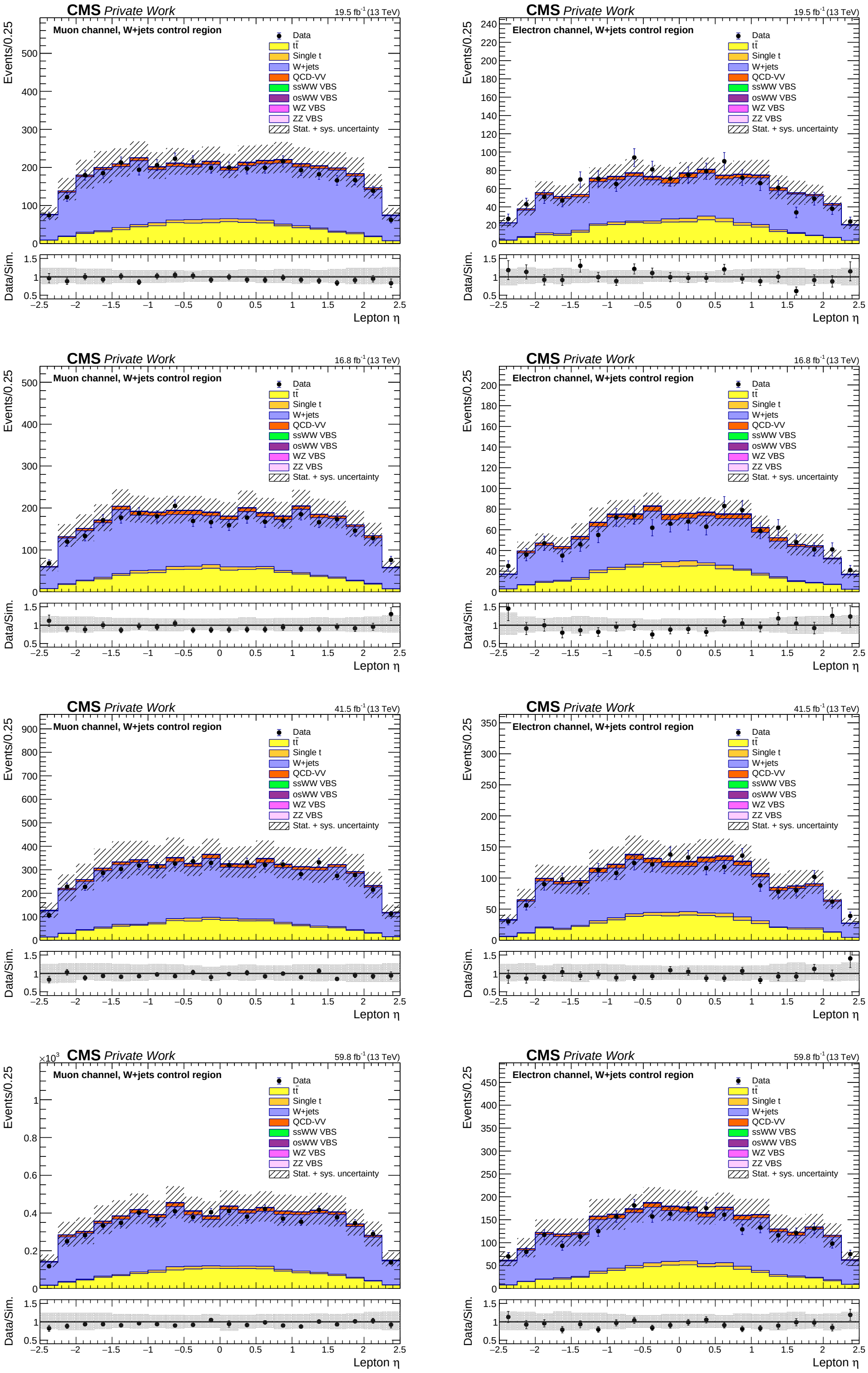


Figure F.7: Data and simulation comparison of the **pseudorapidity of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

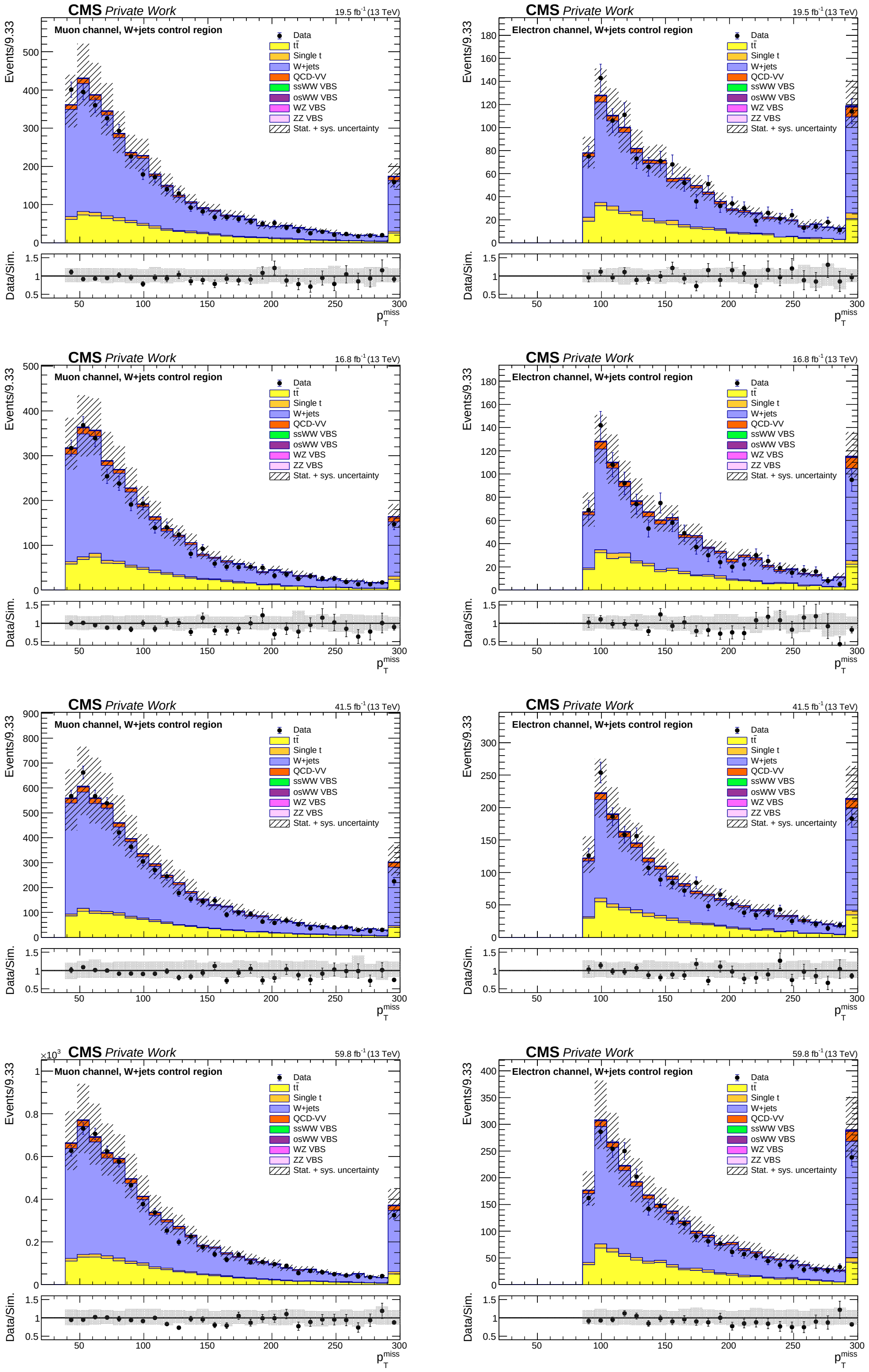


Figure F.8: Data and simulation comparison of the **missing transverse momentum** ($\mathbf{p}_{\mathrm{T}}^{\mathrm{miss}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

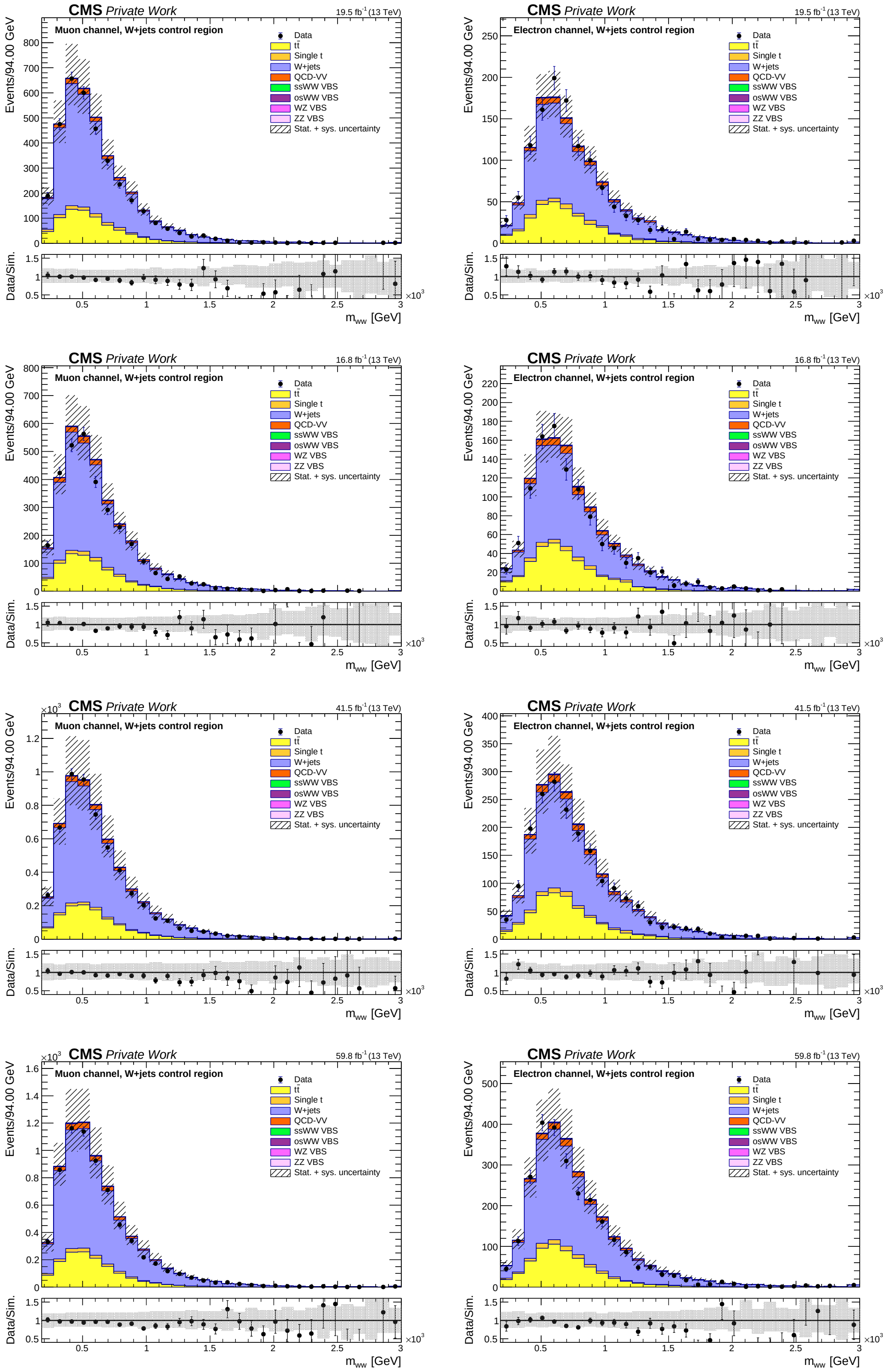


Figure F.9: Data and simulation comparison of the **invariant mass of the two W bosons ($\mathbf{m}_{\mathrm{ww}}$)** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

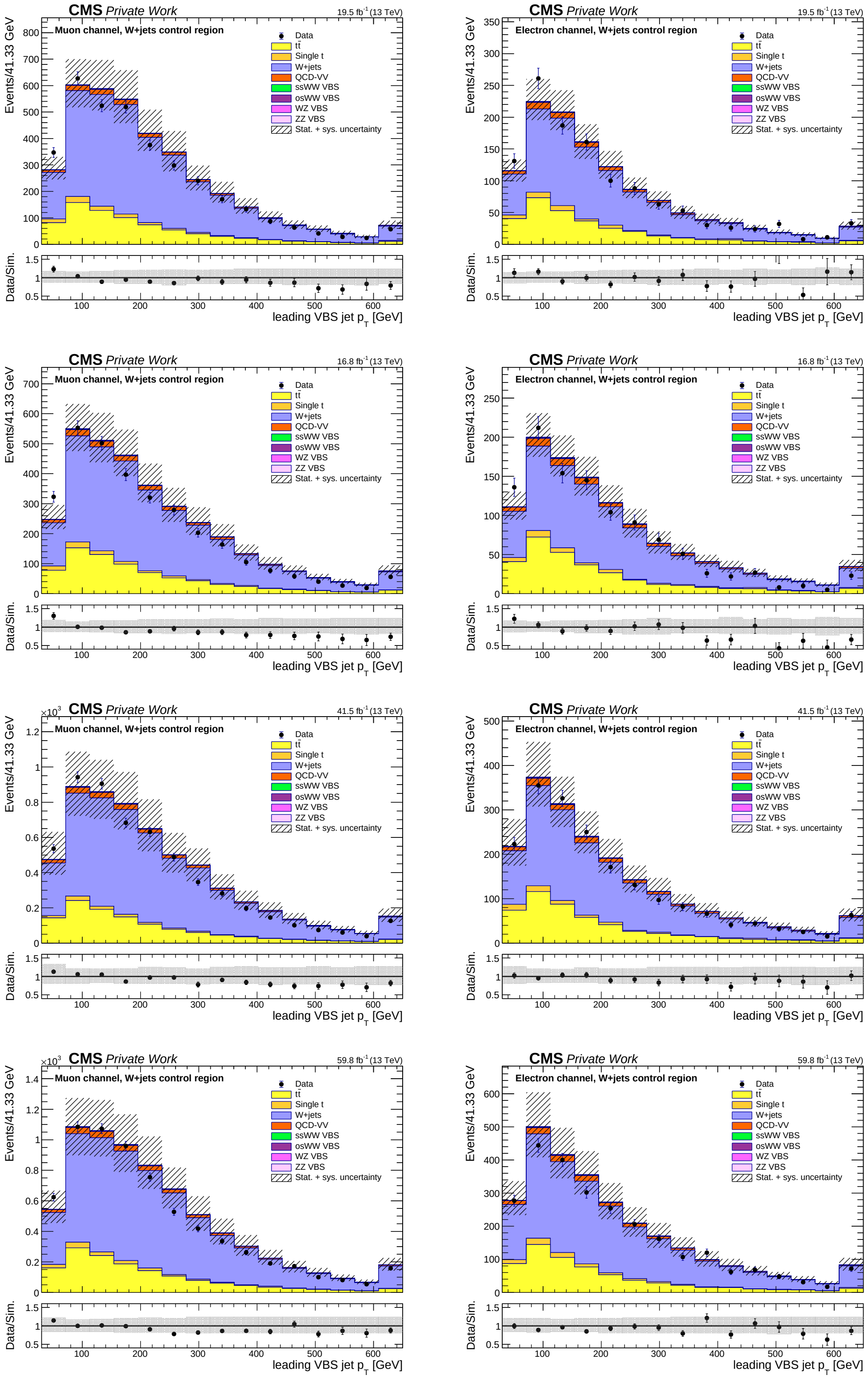


Figure F.10: Data and simulation comparison of the **leading VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

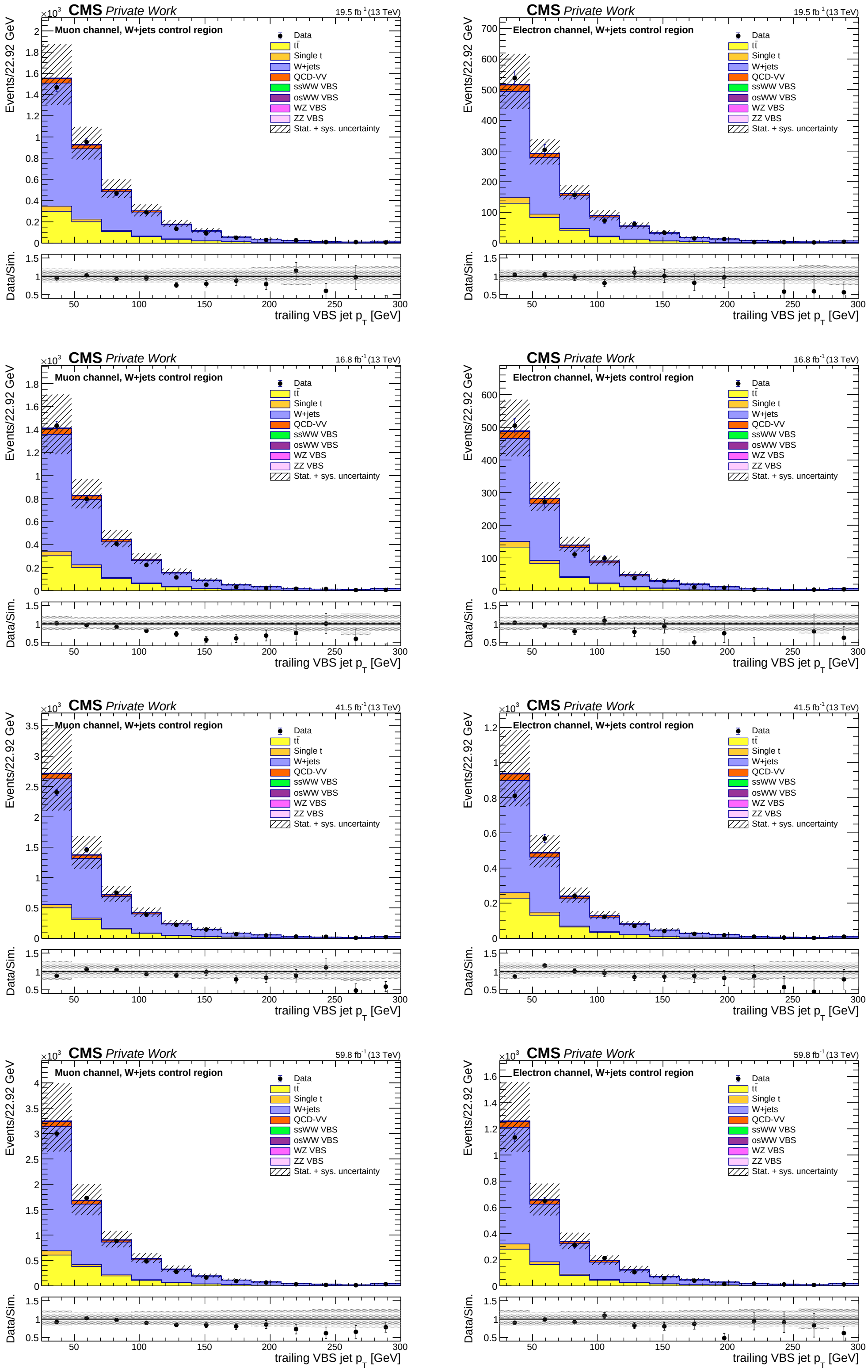


Figure F.11: Data and simulation comparison of the **trailing VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

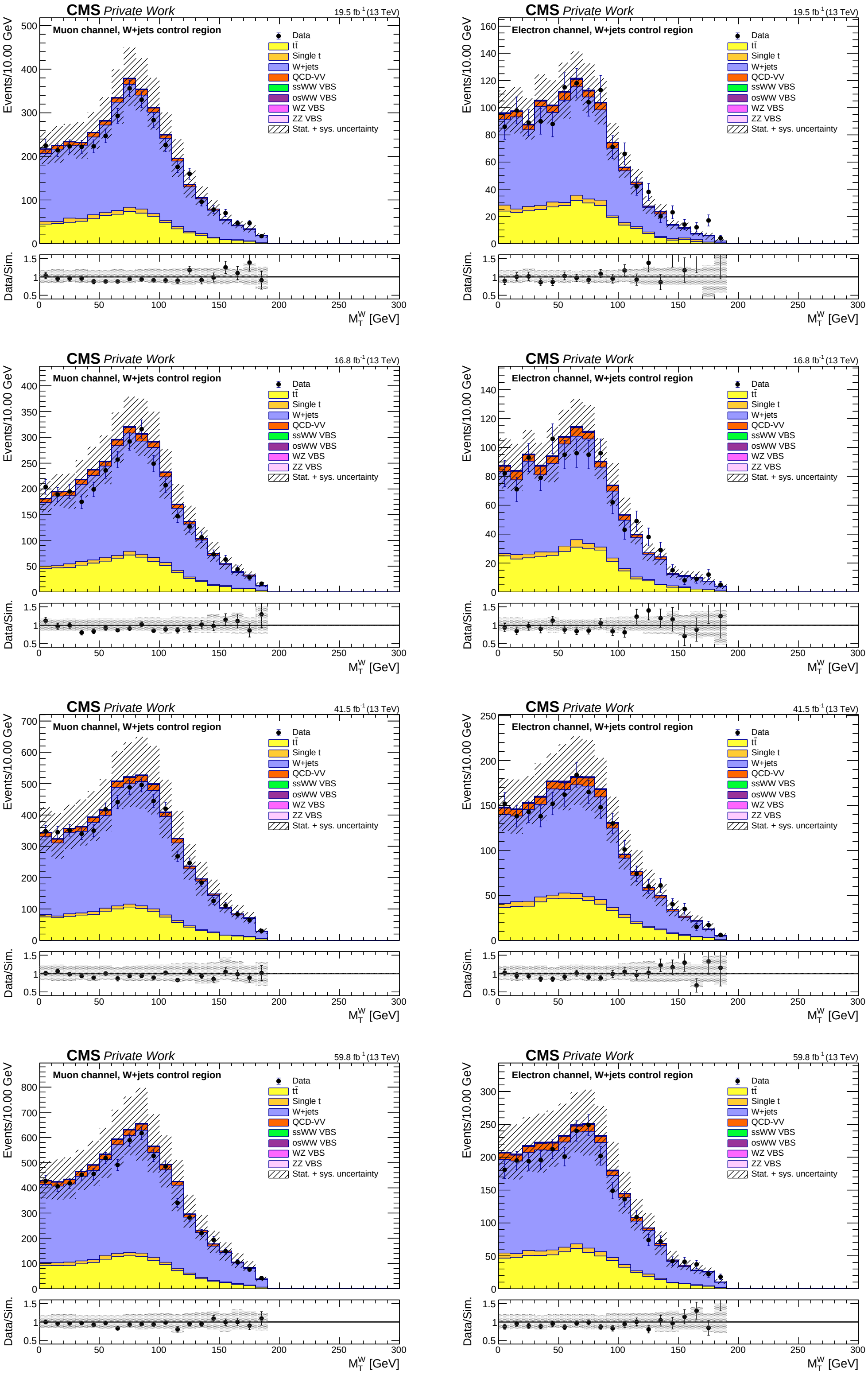


Figure F.12: Data and simulation comparison of the **transverse mass of leptonically decaying W boson** ($M_T{}^W$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

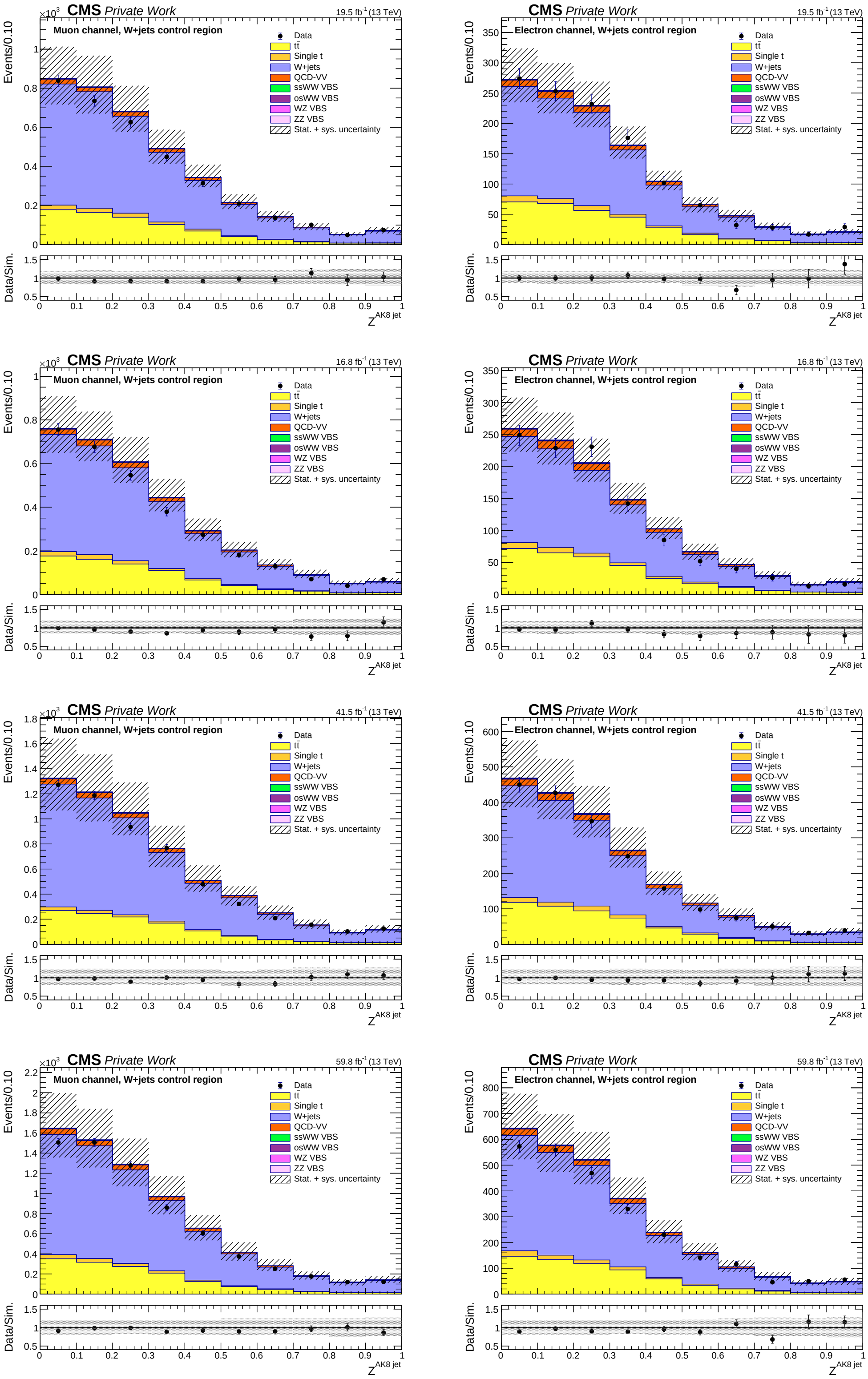


Figure F.13: Data and simulation comparison of **Zeppenfeld variable of the AK8 jet** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

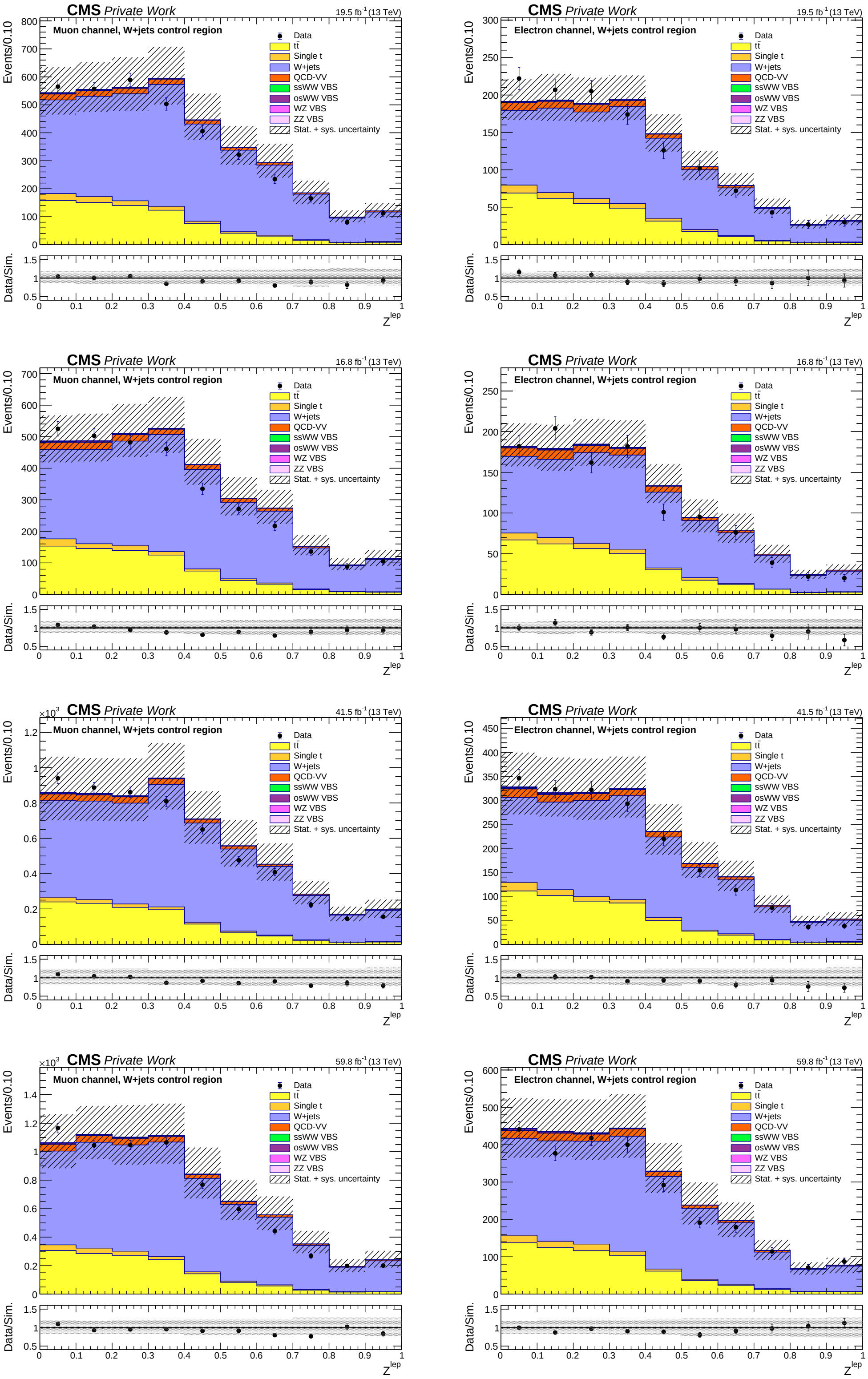


Figure F.14: Data and simulation comparison of **Zeppenfeld variable of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

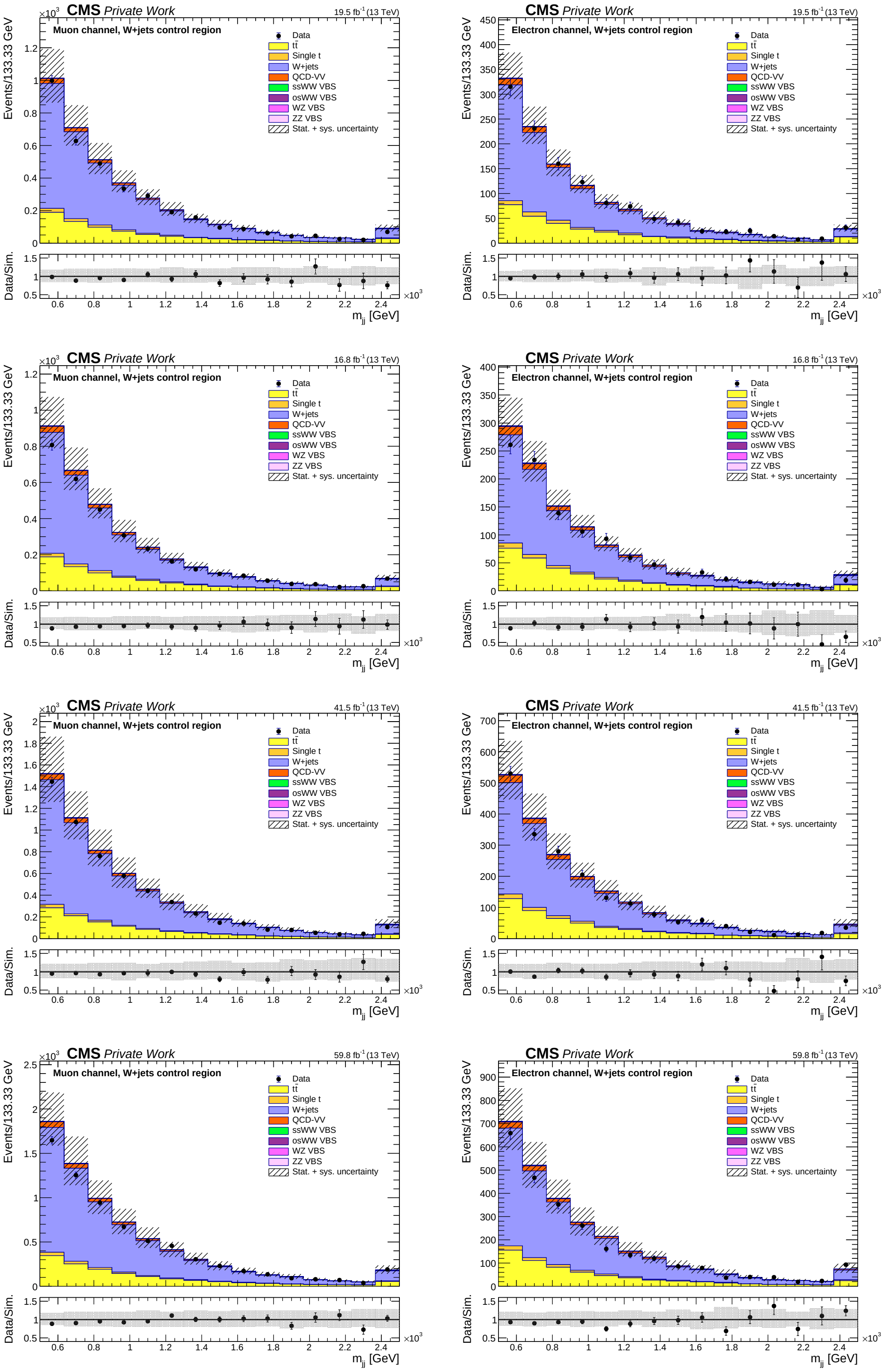


Figure F.15: Data and simulation comparison of **dijet invariant mass ($\mathbf{m_{jj}}$)** of the two VBS jets in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

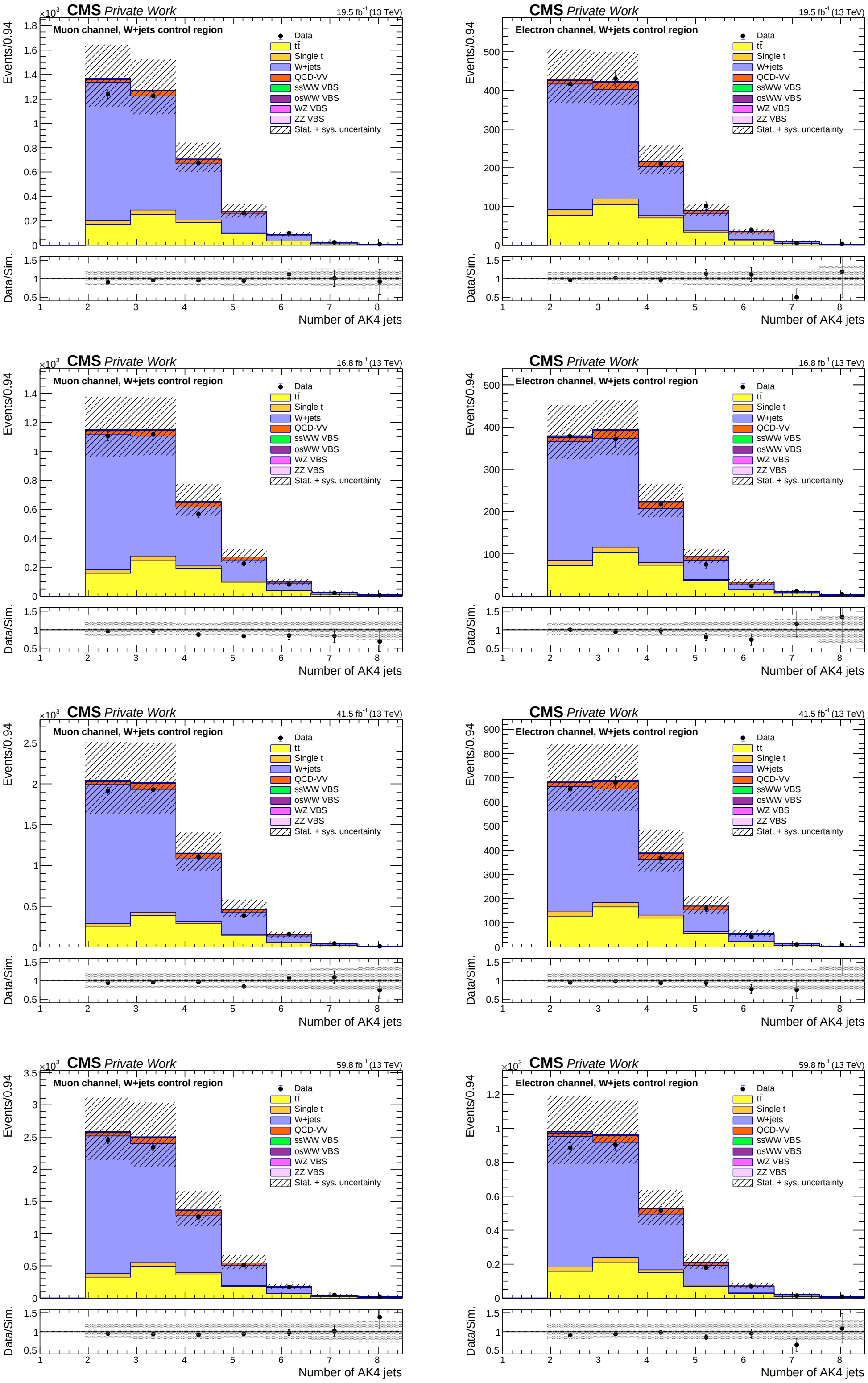


Figure F.16: Data and simulation comparison of **number of AK4 jets** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

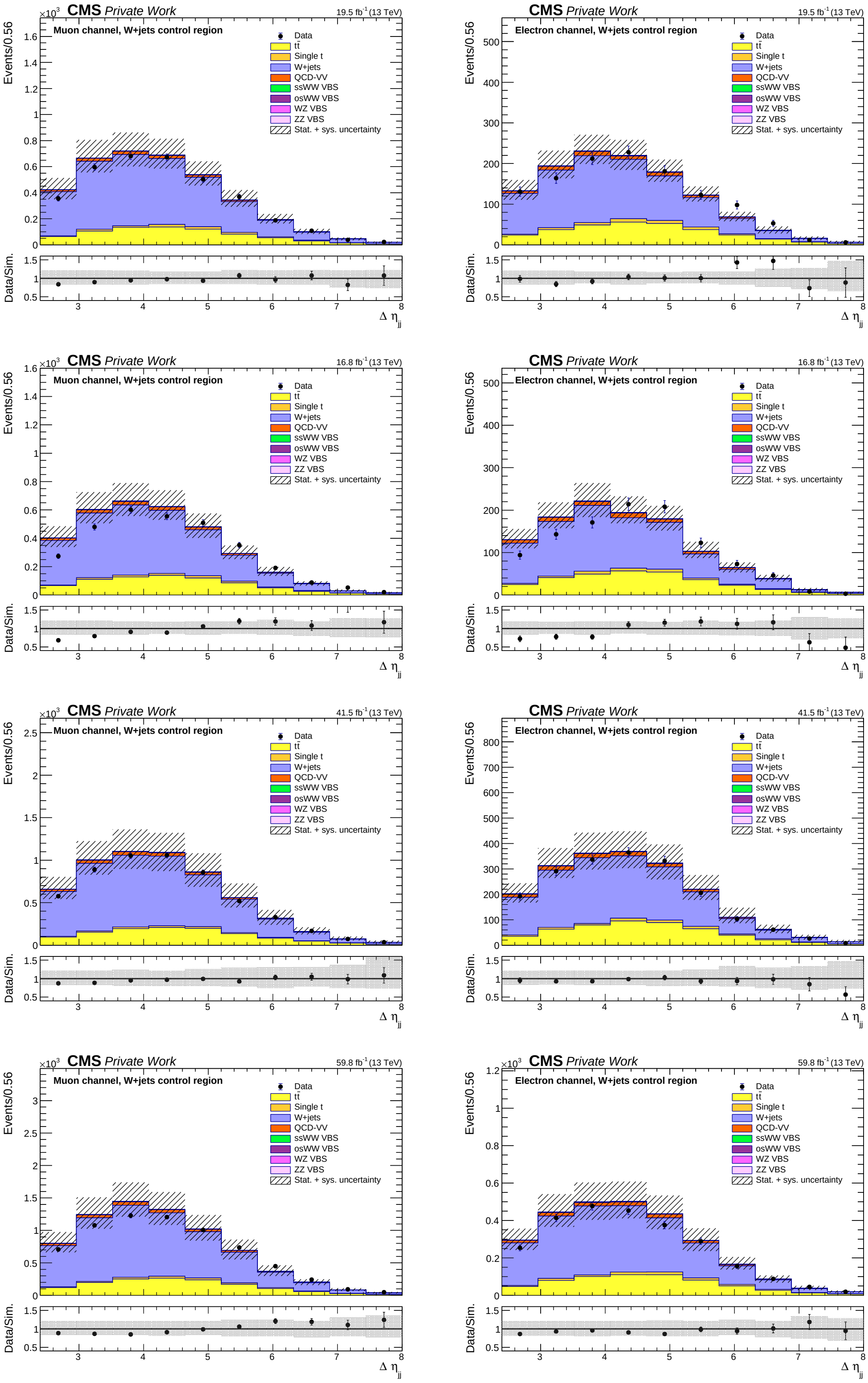


Figure F.17: Data and simulation comparison of **psuedorapidity separation** (**$\Delta\eta_{jj}$**) of the two VBS jets in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

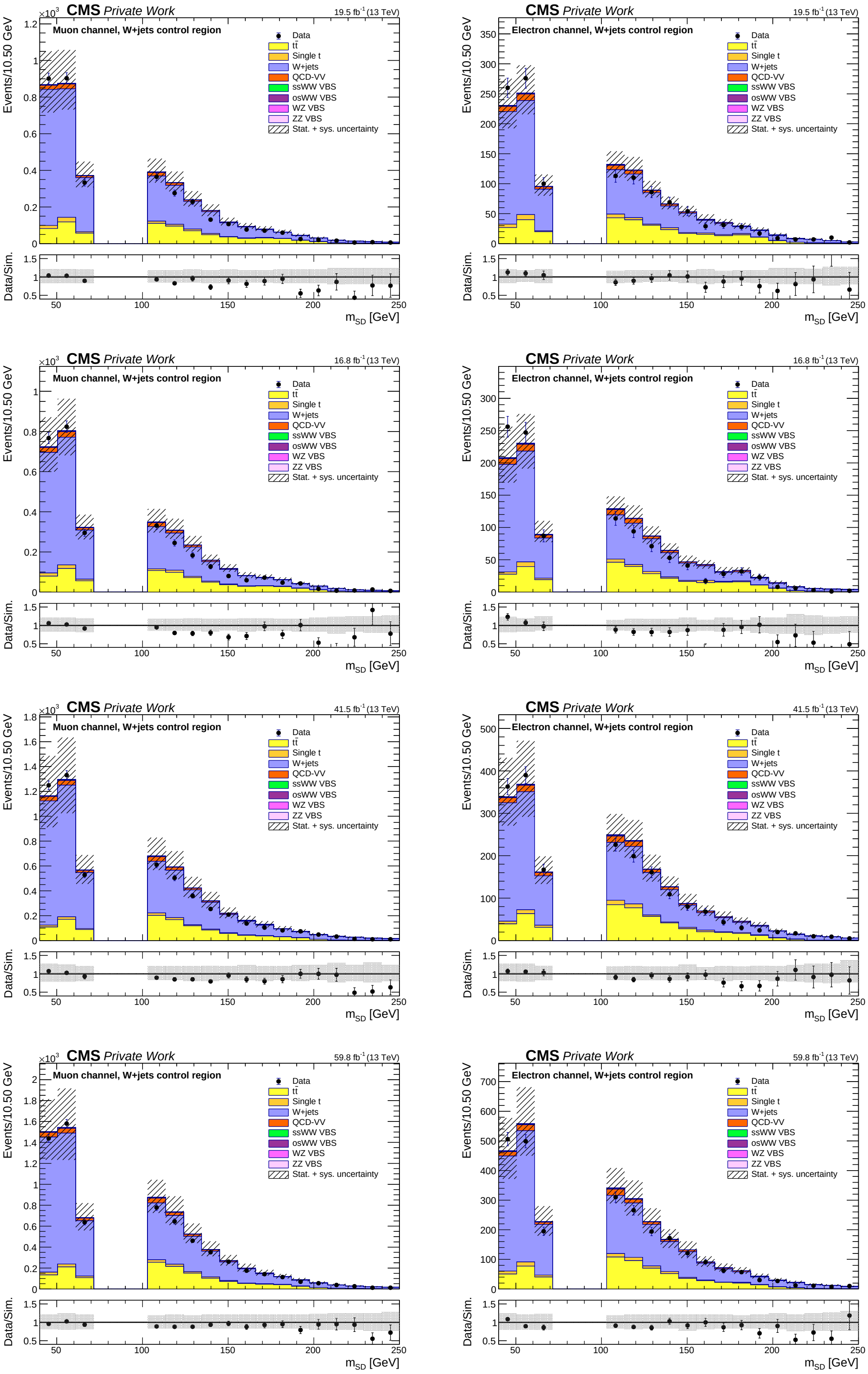


Figure F.18: Data and simulation comparison of the **soft drop mass of the AK8 jet** ($\mathbf{m}_{\mathrm{SD}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

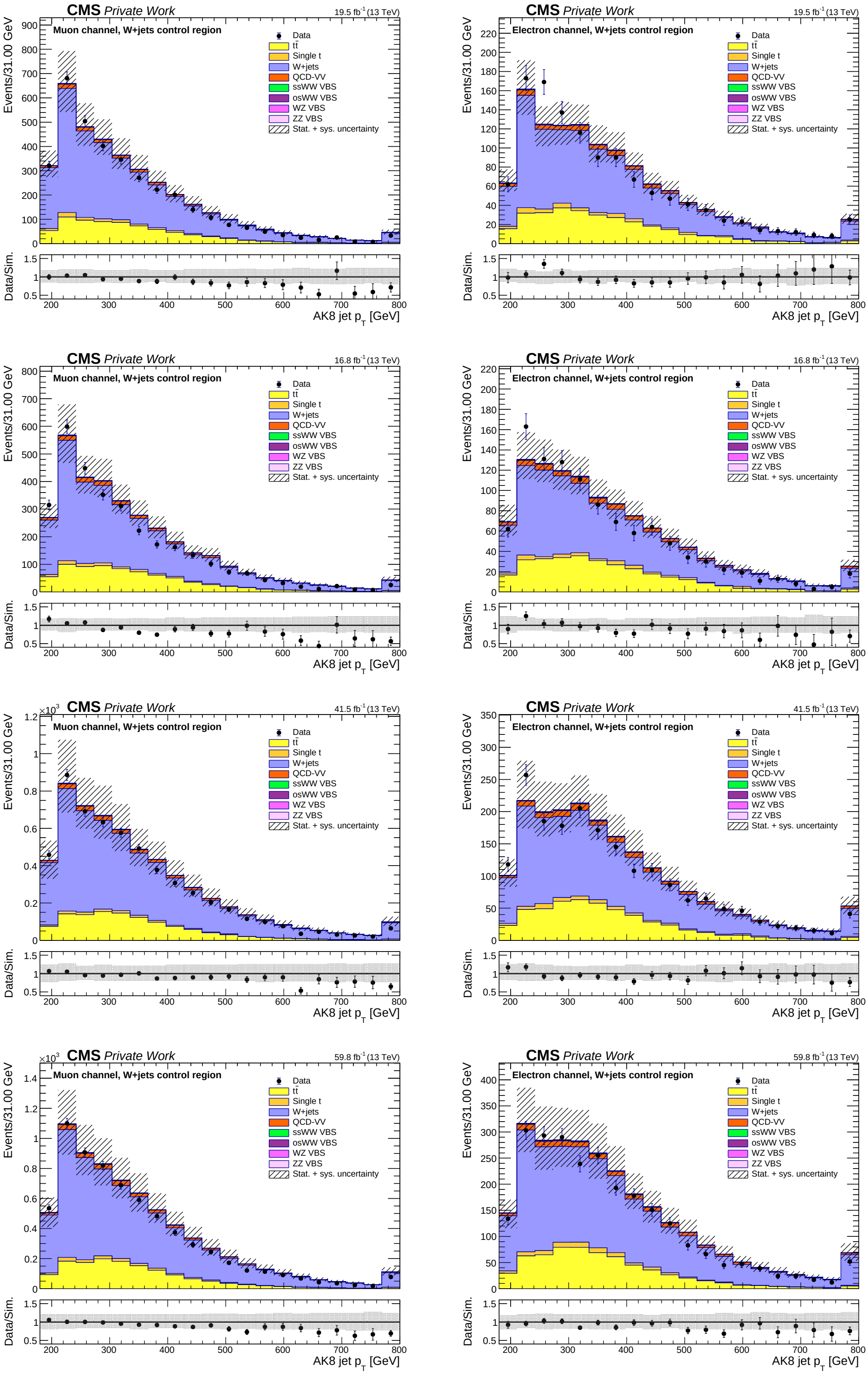


Figure F.19: Data and simulation comparison of the **transverse momentum of the AK8 jet** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

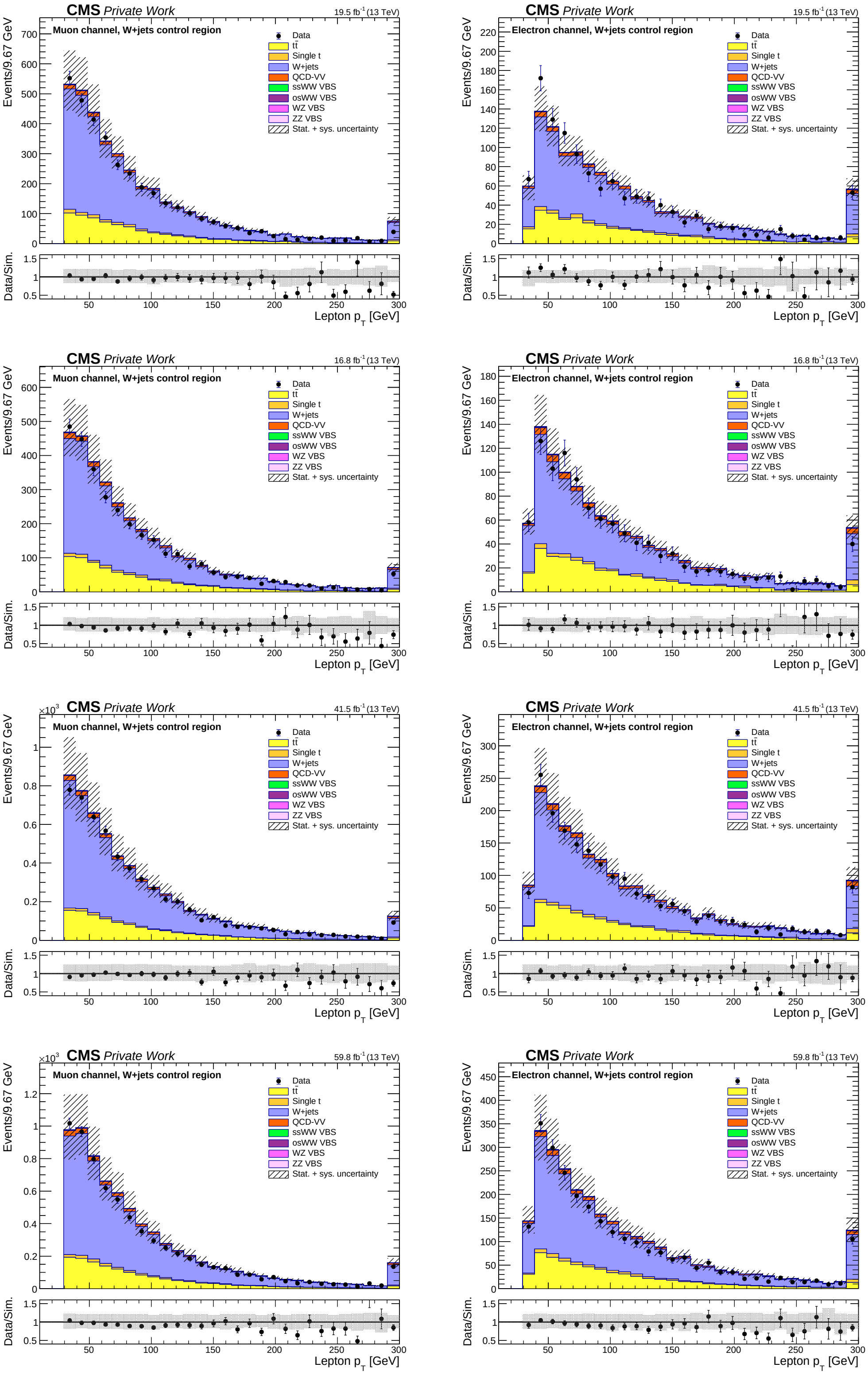


Figure F.20: Data and simulation comparison of the **transverse momentum of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

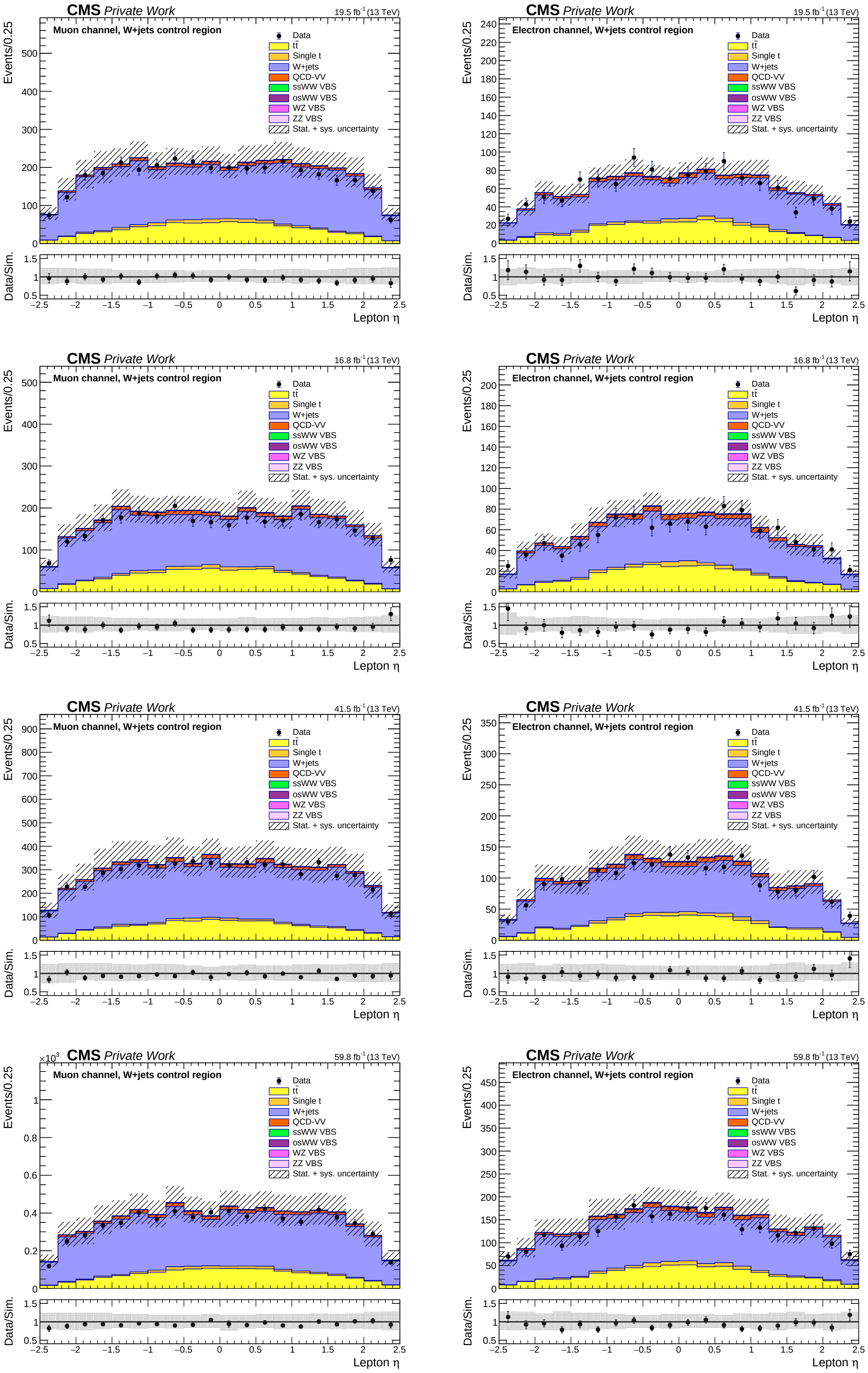


Figure F.21: Data and simulation comparison of the **pseudorapidity of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

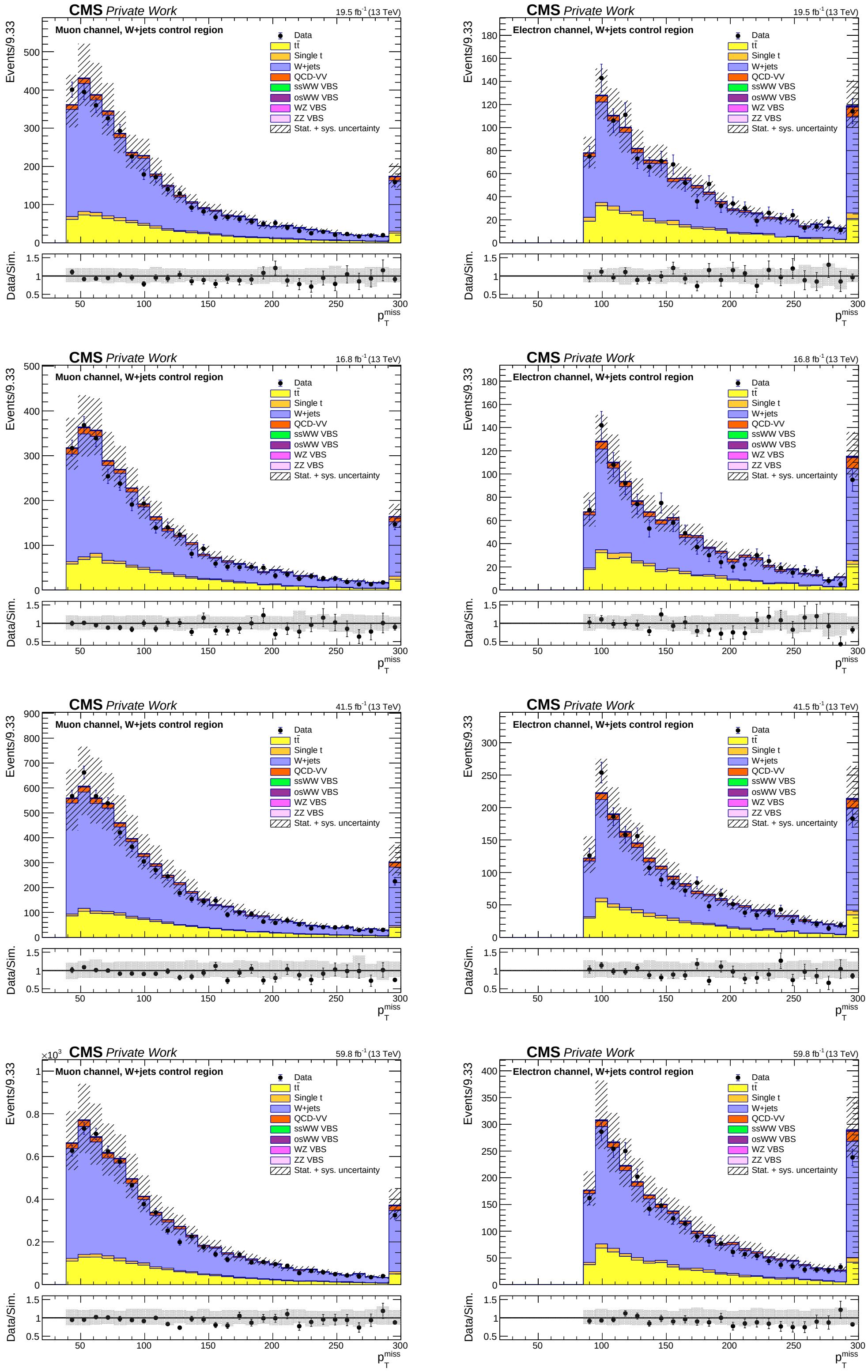


Figure F.22: Data and simulation comparison of the **missing transverse momentum** ($\mathbf{p_T}^{\text{miss}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

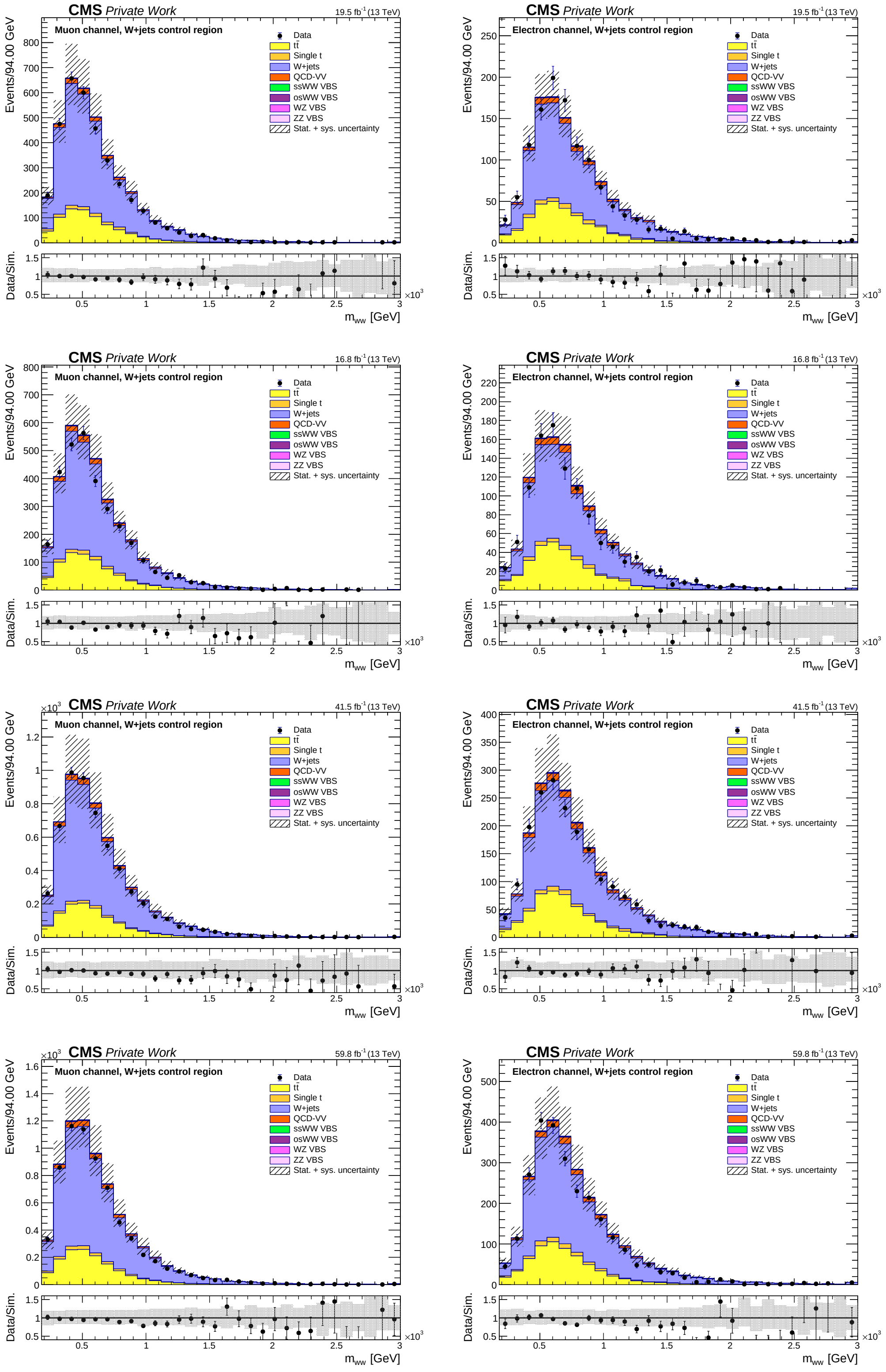


Figure F.23: Data and simulation comparison of the **invariant mass of the two W bosons** ($\mathbf{m}_{\mathrm{ww}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

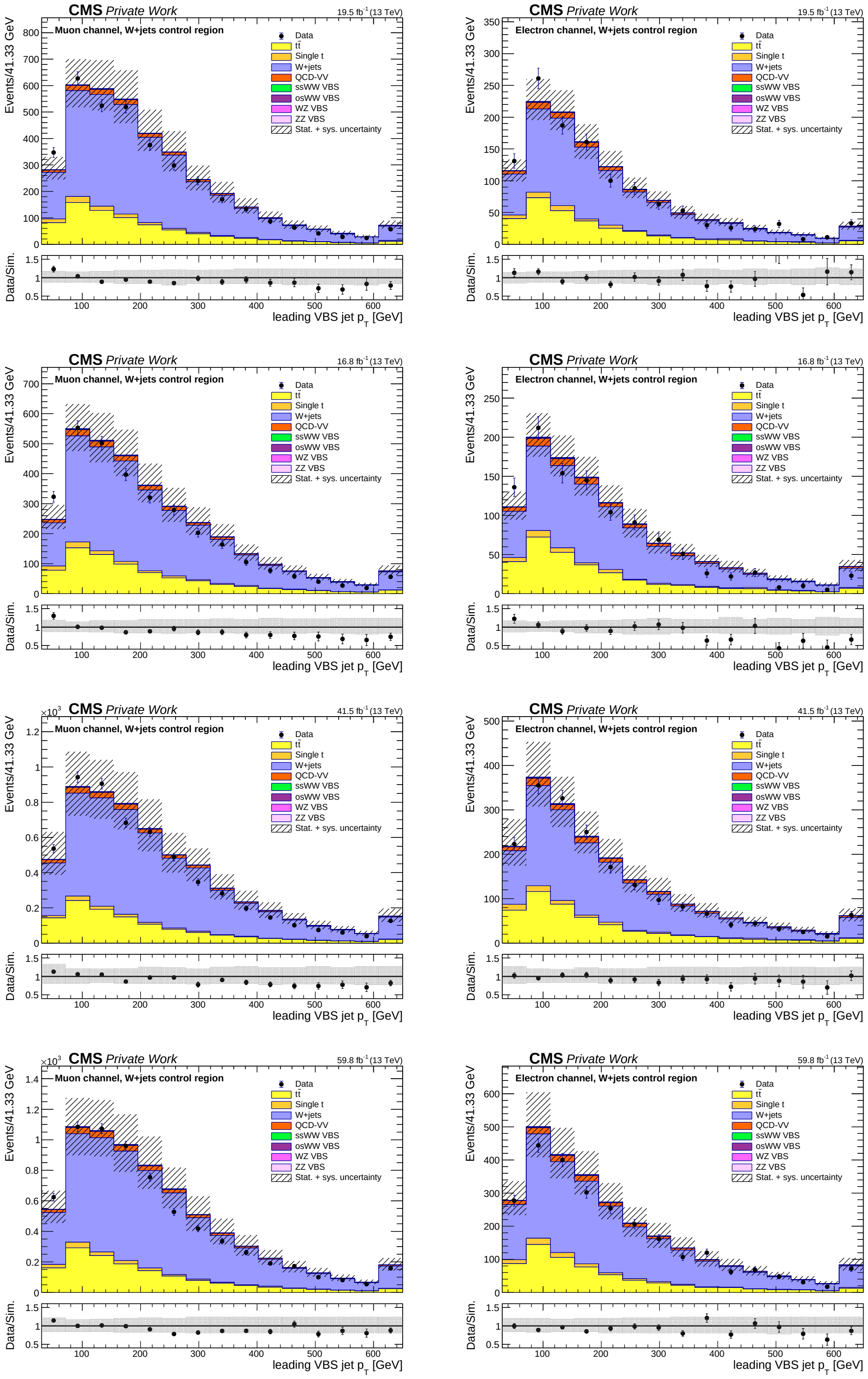


Figure F.24: Data and simulation comparison of the **leading VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

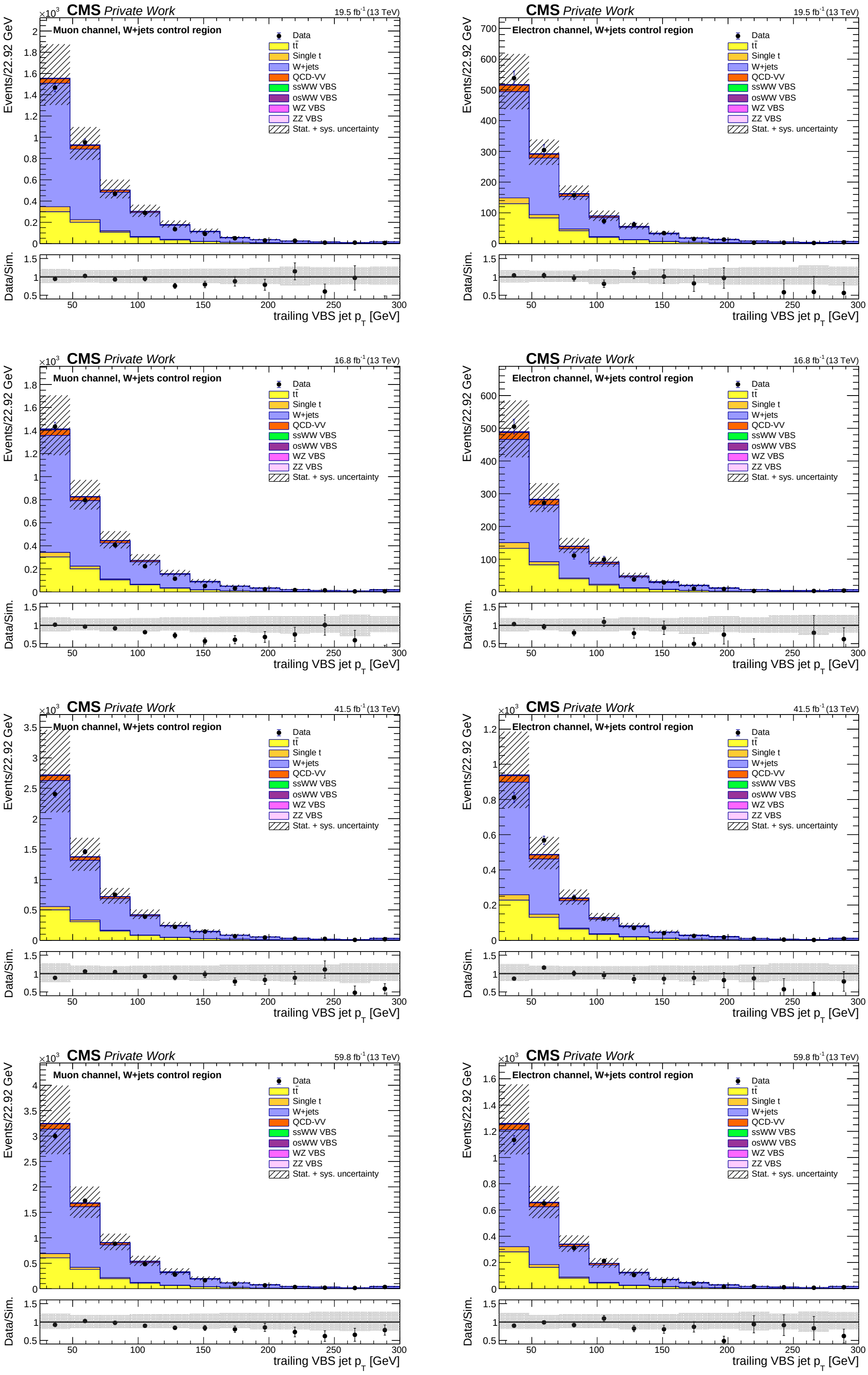


Figure F.25: Data and simulation comparison of the **trailing VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

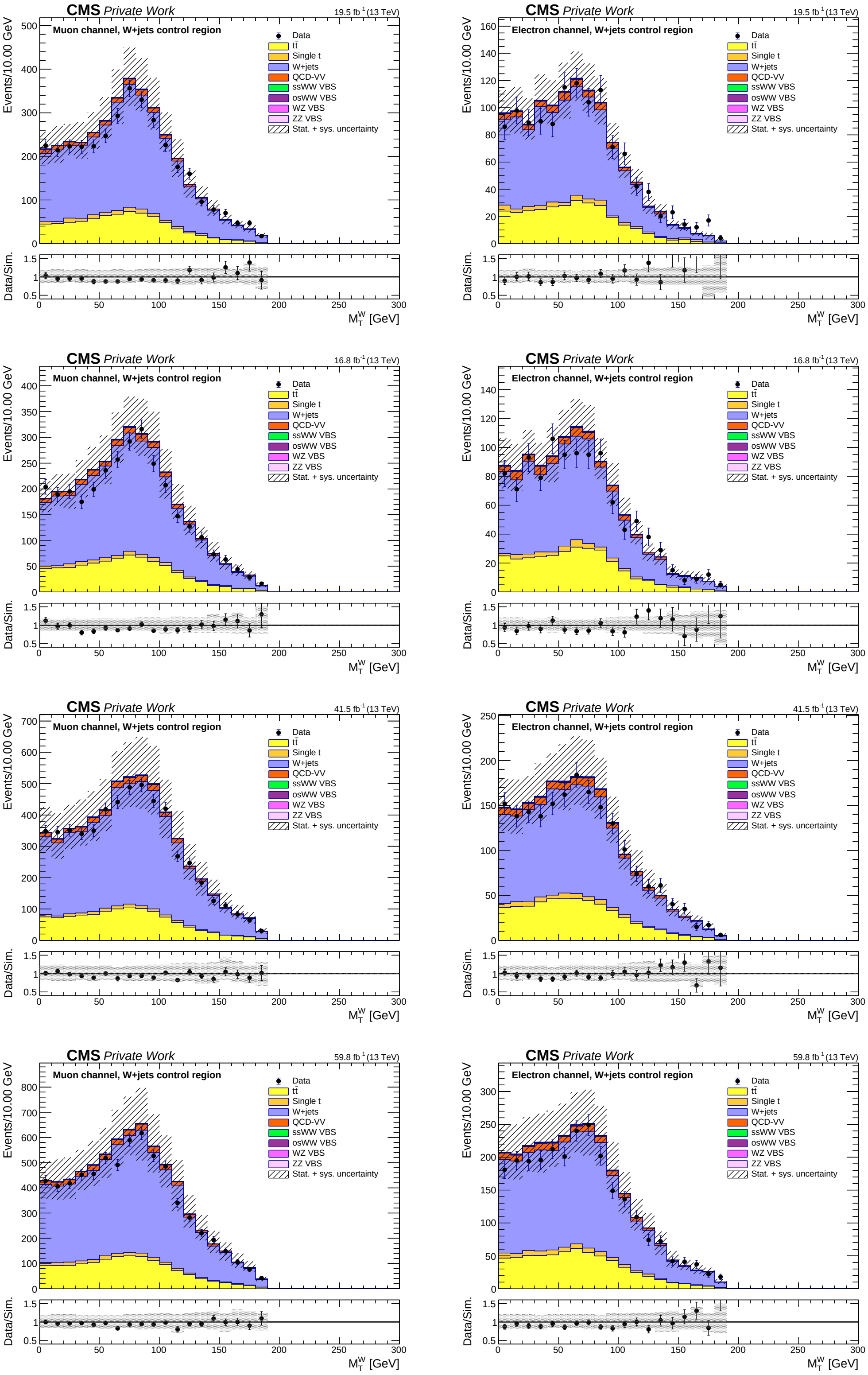


Figure F.26: Data and simulation comparison of the **transverse mass of leptonically decaying W boson** ($M_T{}^W$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

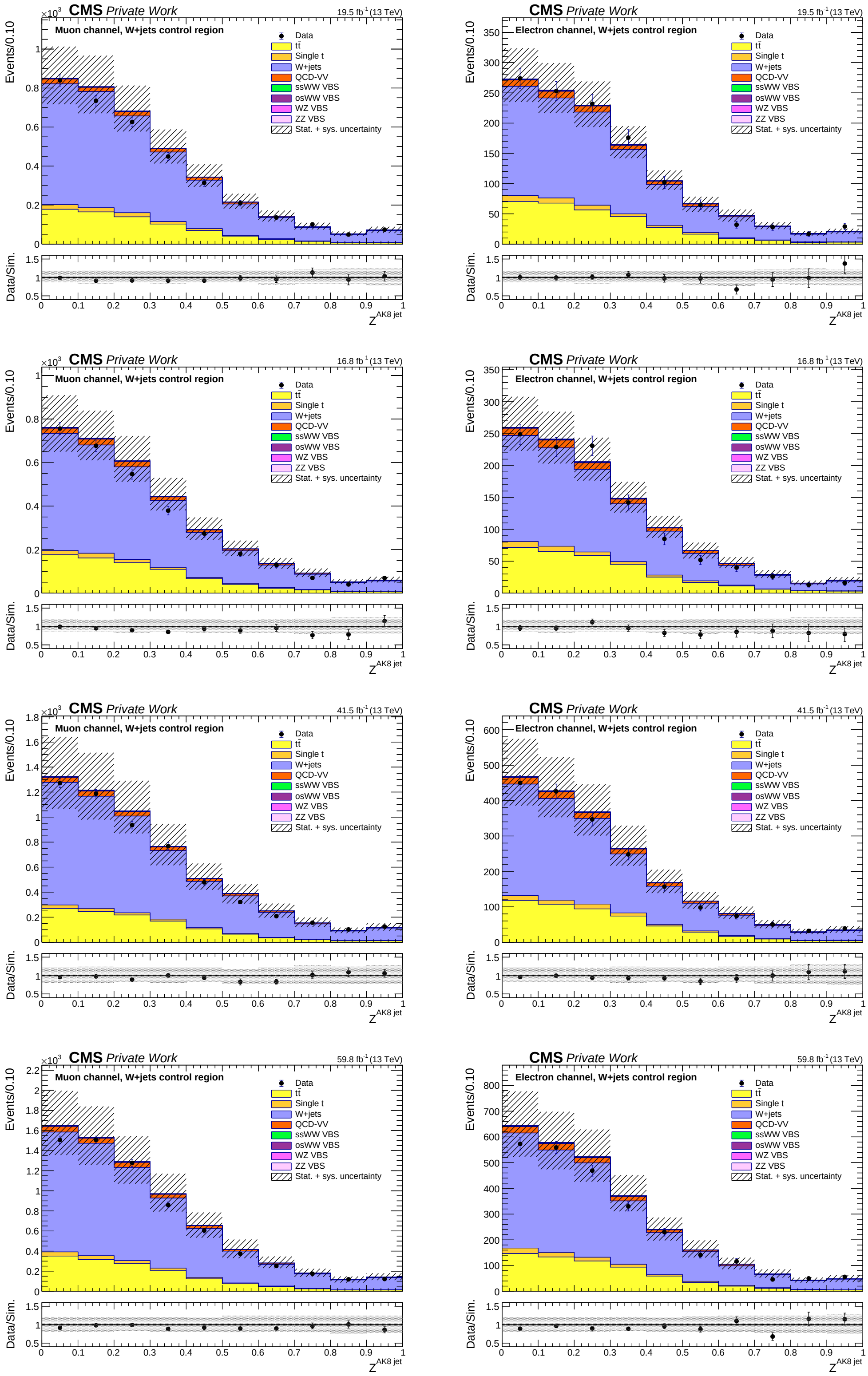


Figure F.27: Data and simulation comparison of **Zeppenfeld variable of the AK8 jet** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

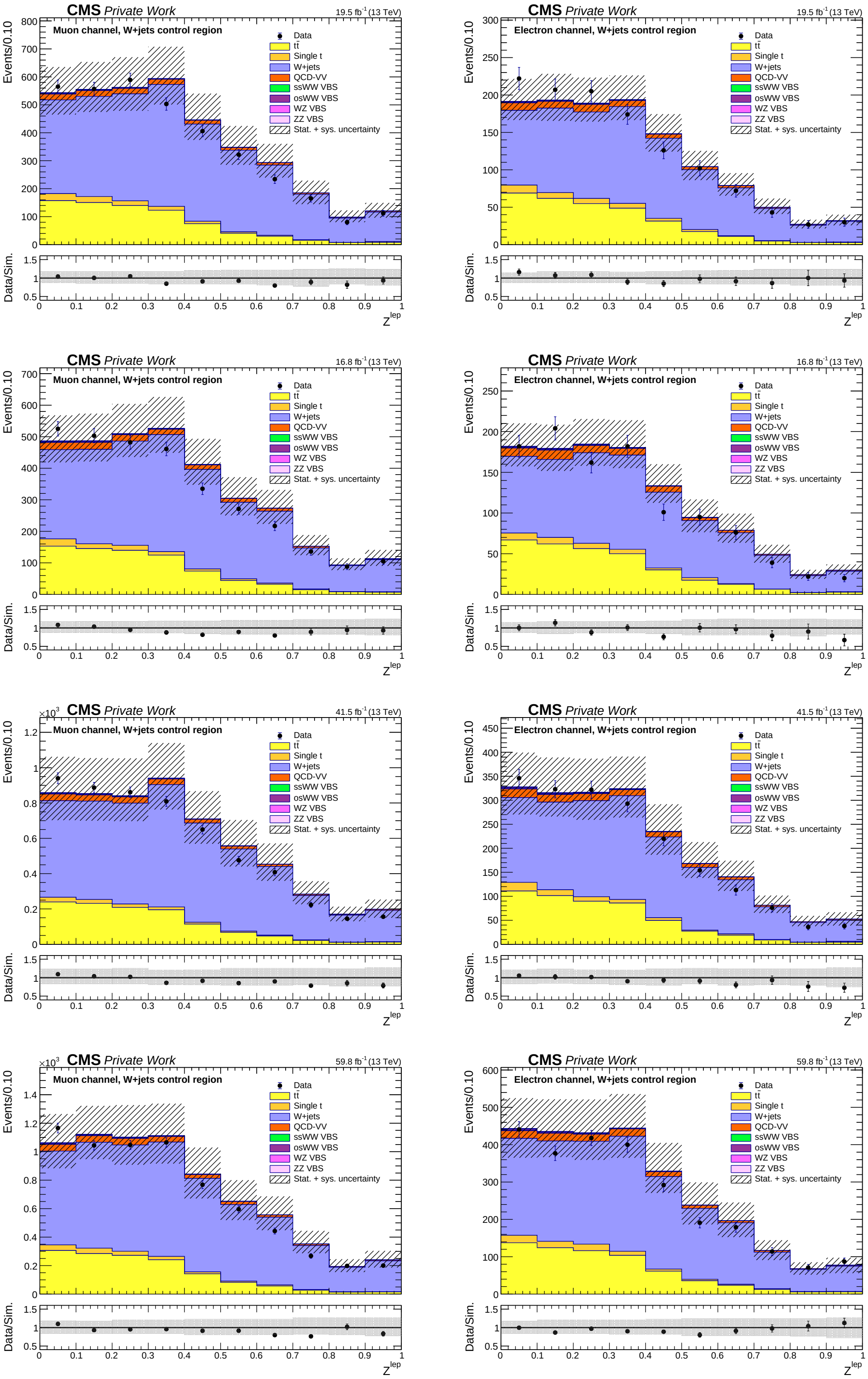


Figure F.28: Data and simulation comparison of **Zeppenfeld variable of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

G. Additional prefit plots

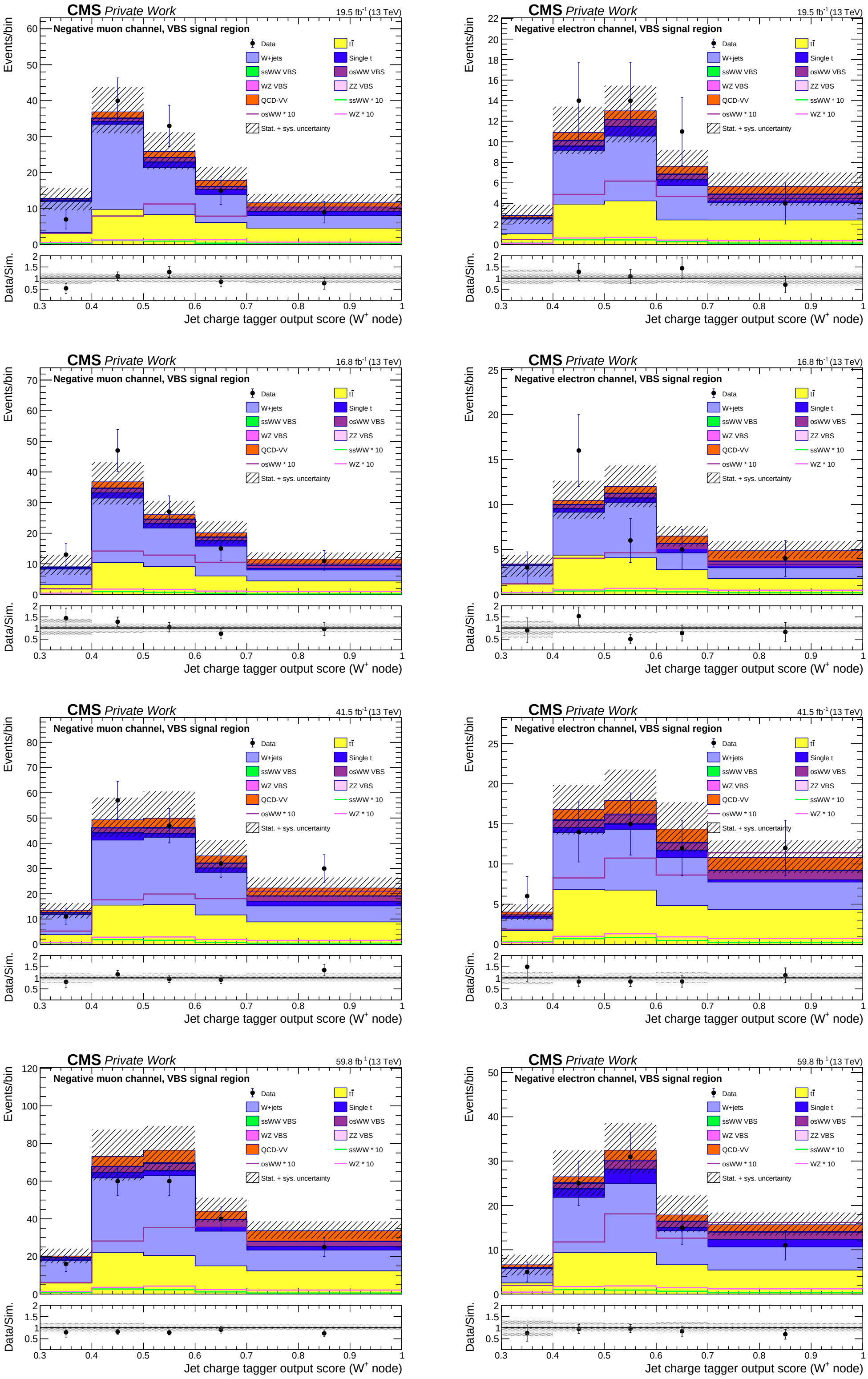


Figure G.1: Jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **opposite-sign WW** (W^+W^-) **VBS category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

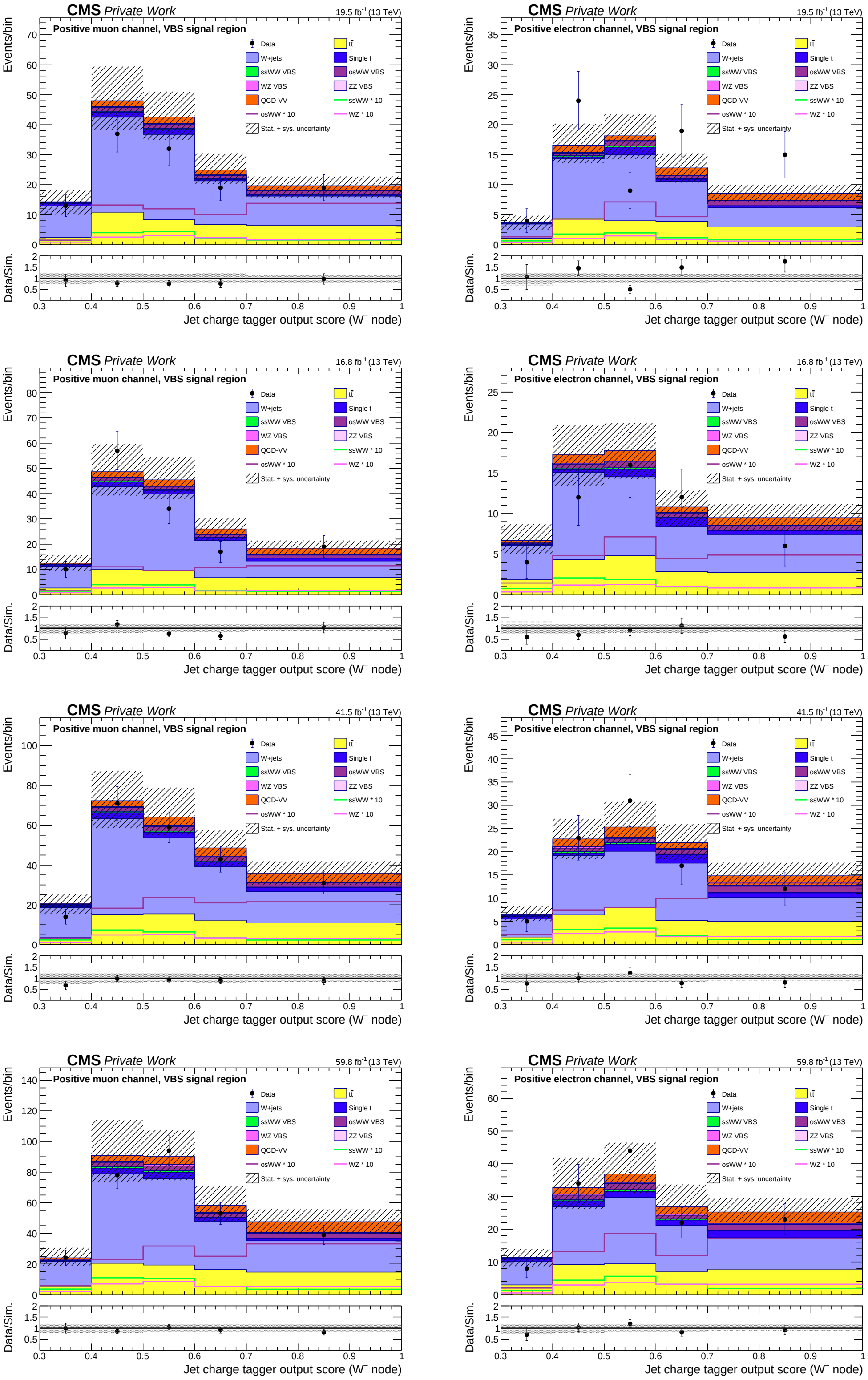


Figure G.2: Jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **opposite-sign WW (W^-W^+) VBS category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

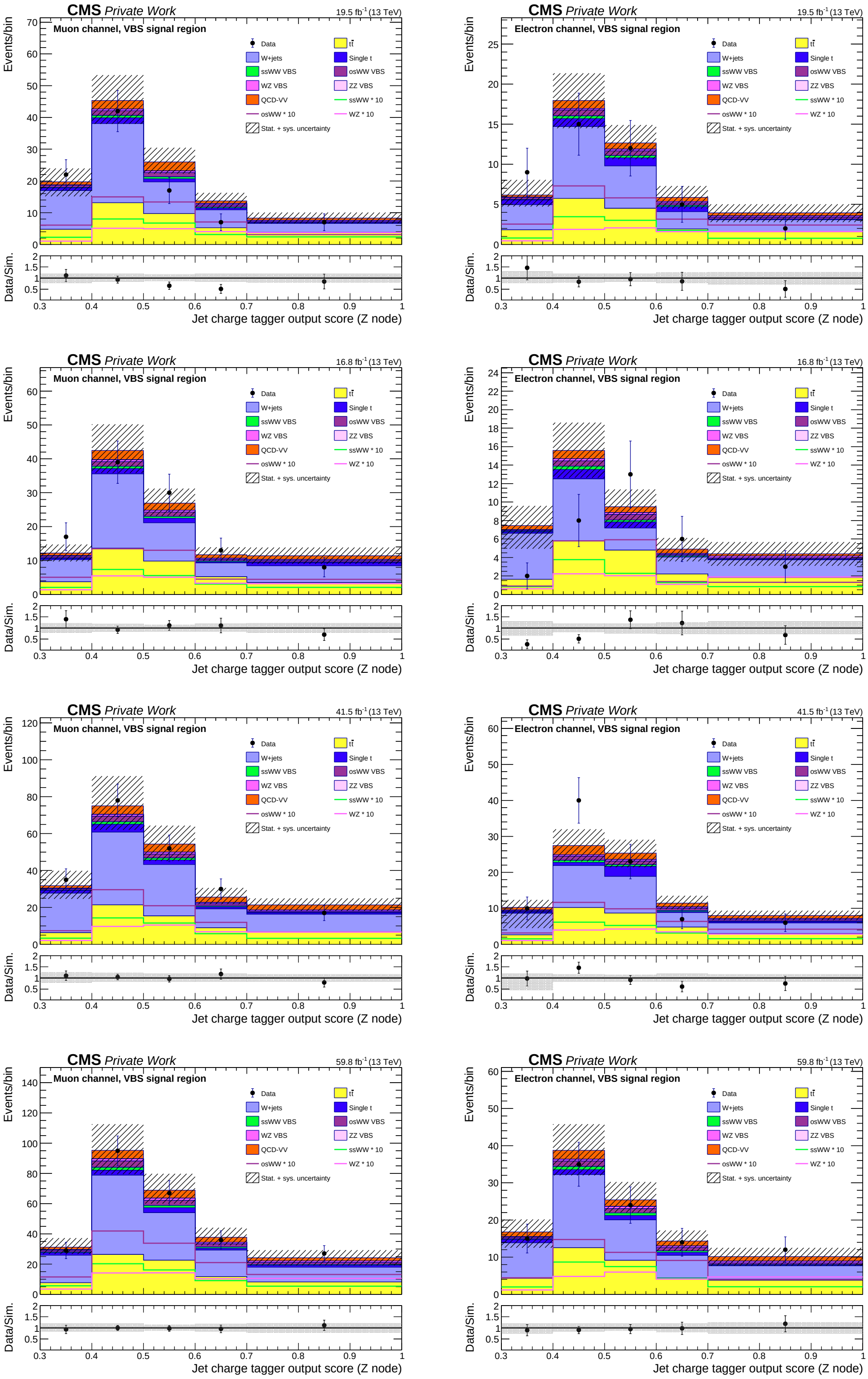


Figure G.3: Jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **WZ VBS category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

H. Additional postfit plots

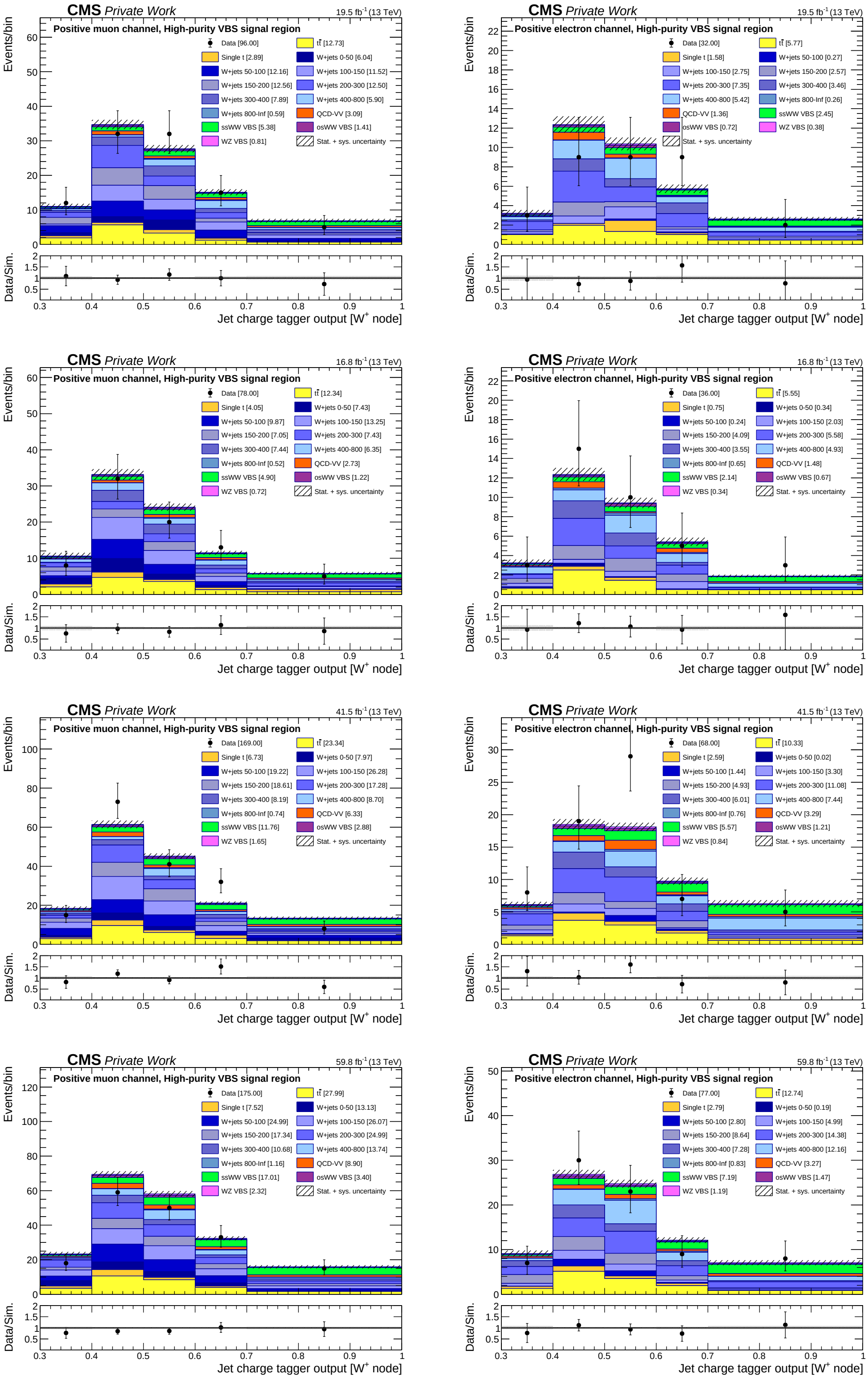


Figure H.1: Post-fit jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **same-sign WW** (W^+W^+) **VBS signal category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom). The VBS signal processes are also shown after rescaling using solid lines. For higher values of the tagger output score, the same-sign WW VBS process is dominant compared to the other VBS processes and QCD-VV.

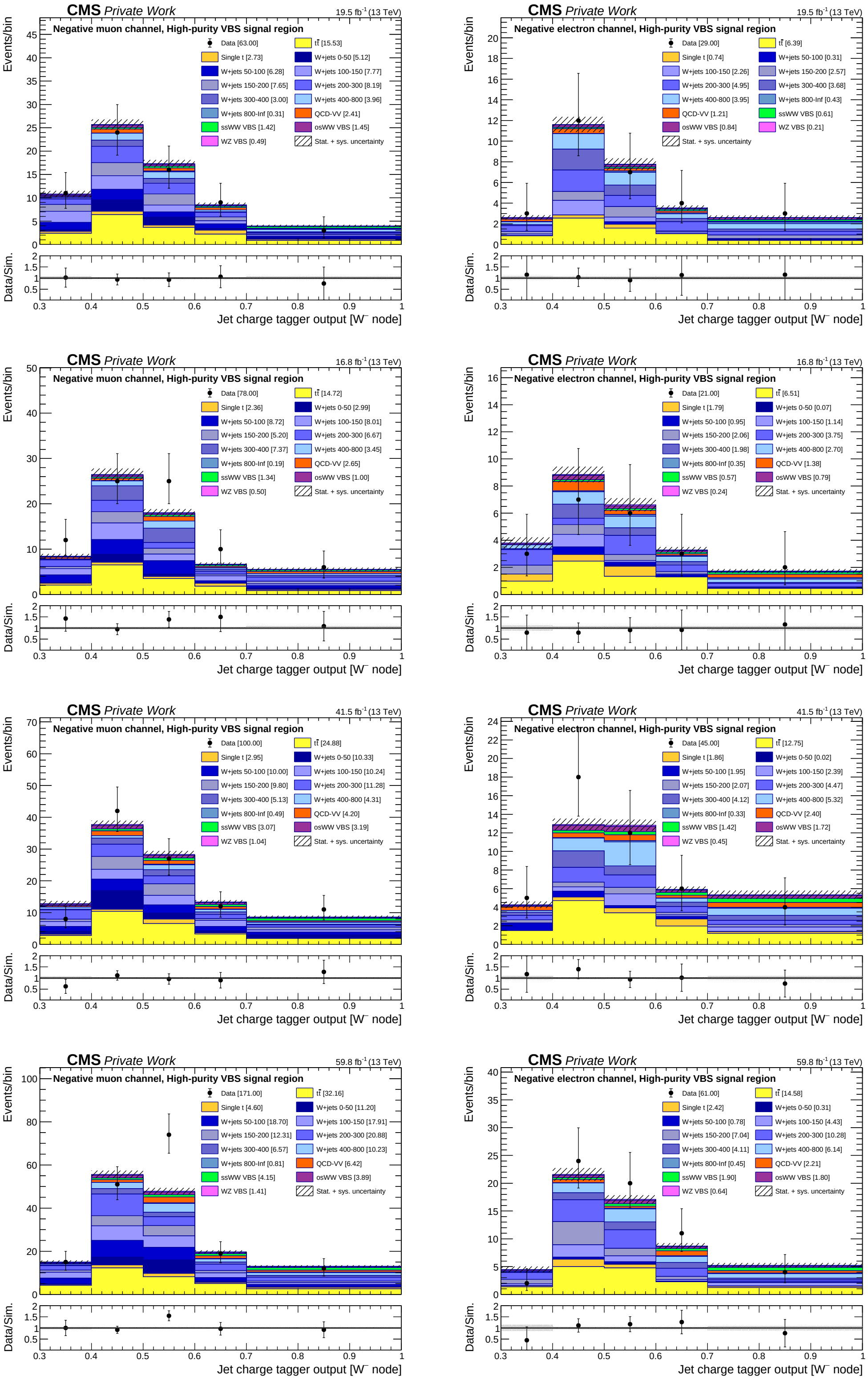


Figure H.2: Post-fit jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **same-sign WW ($\mathbf{W^{-}W^{-}}$) VBS signal category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom). The VBS signal processes are also shown after rescaling using solid lines. For higher values of the tagger output score, the same-sign WW VBS process is dominant compared to the other VBS processes and QCD-VV.

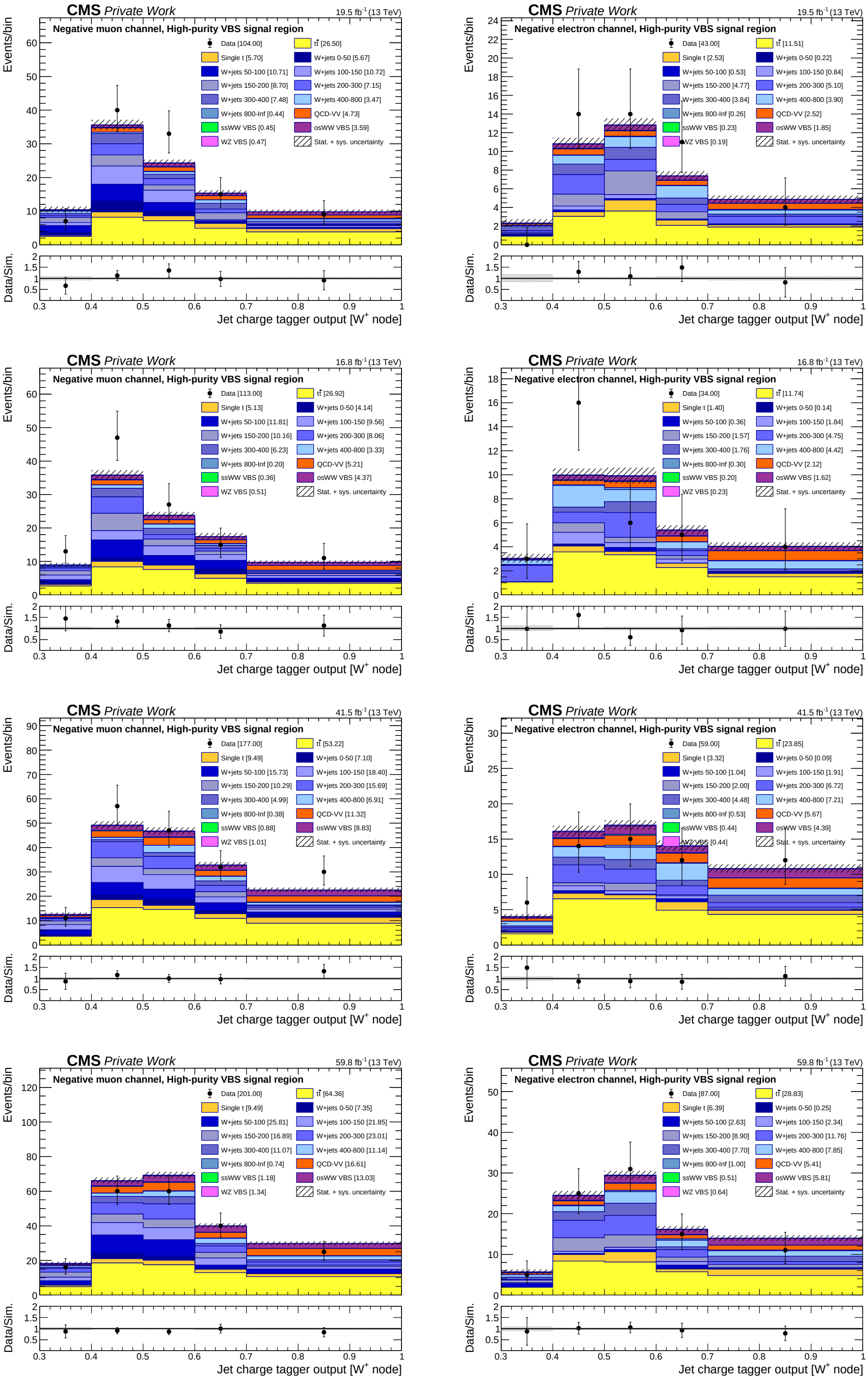


Figure H.3: Post-fit jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **opposite-sign WW** ($\mathrm{W^+W^-}$) **VBS category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

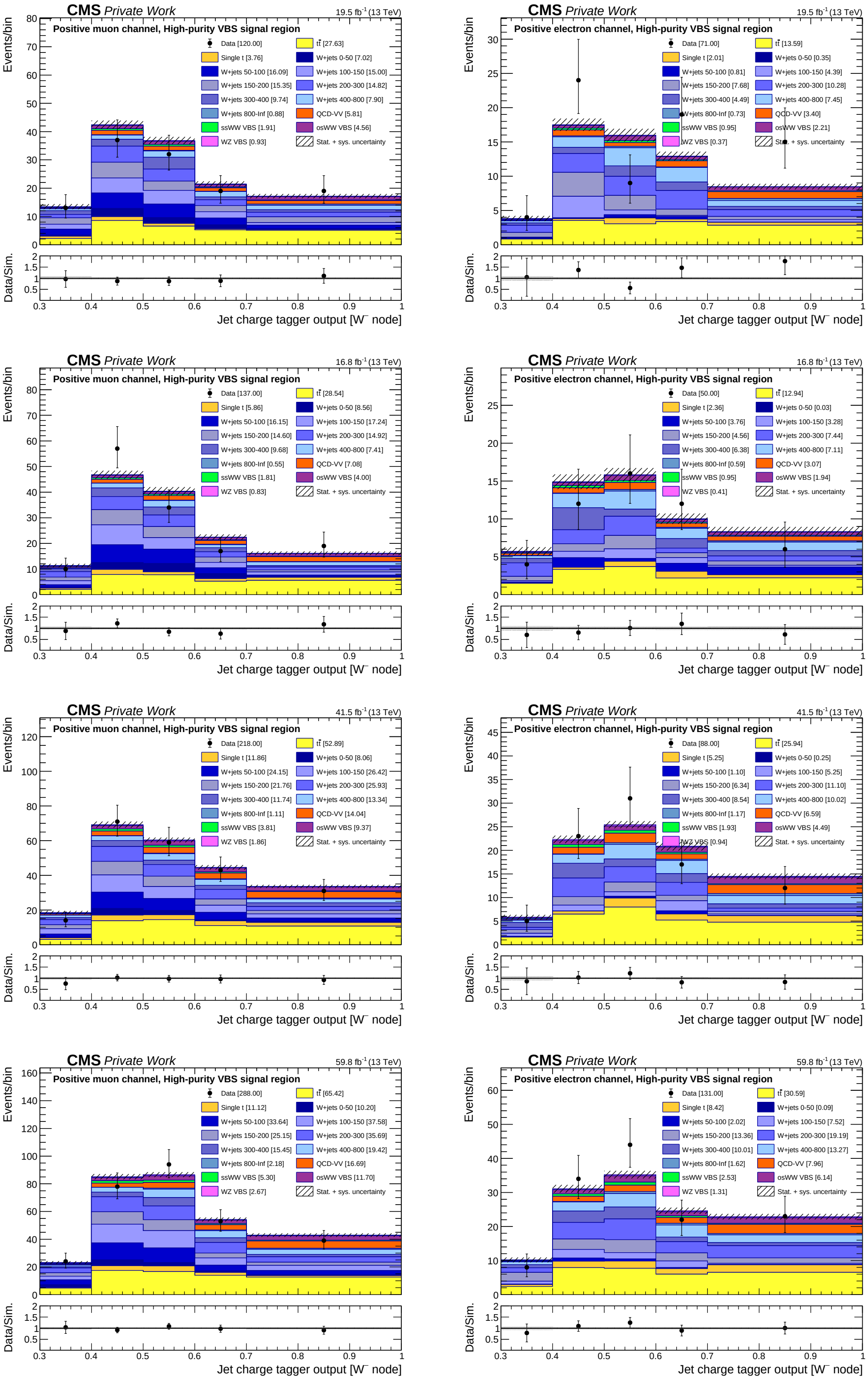


Figure H.4: Post-fit jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **opposite-sign WW (W^-W^+) VBS category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

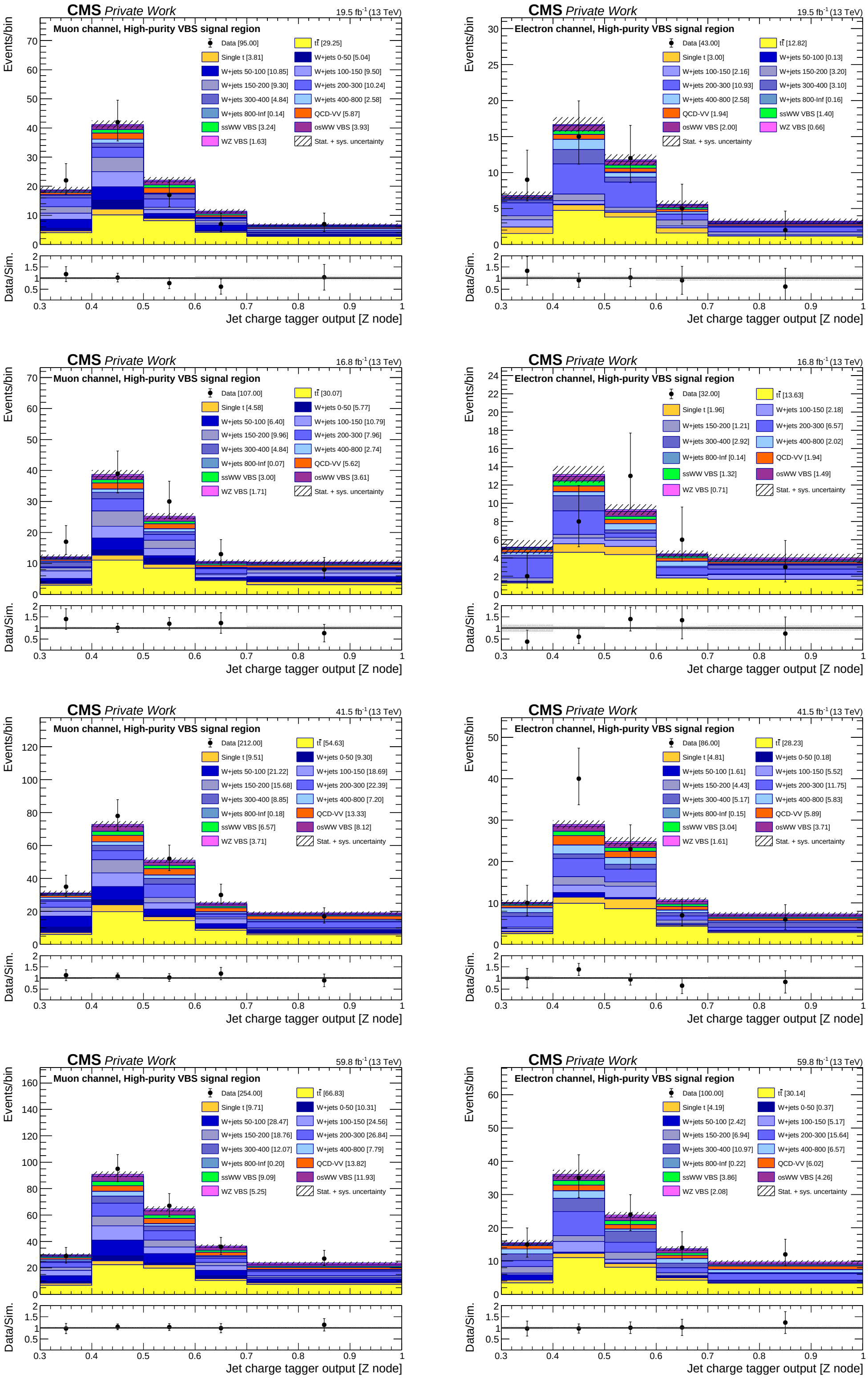


Figure H.5: Post-fit jet charge tagger output distributions for the muon channel (left) and electron channel (right) in the **WZ VBS category** for different reconstruction eras: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

I. Goodness of fit test

A goodness of fit test is performed to evaluate how well a statistical model describes the observed data. It helps determine whether the statistical model we have set up, including signal and background contributions, provides a satisfactory description of the data or not. The test is conducted using the saturated likelihood model, following the recommendations of the CMS Statistical Committee [146]. To properly initialize the nuisance parameters, toys are generated after performing a preliminary fit to data, with the signal strength fixed at 1. The resulting distribution of the test statistic t_{stat} obtained from toy simulations, is presented in Figure I.1. The observed data is found to be compatible with our statistical model.

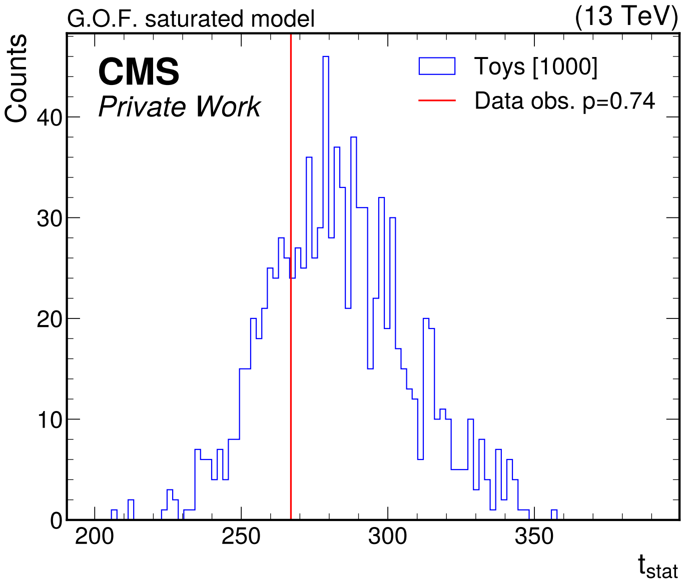
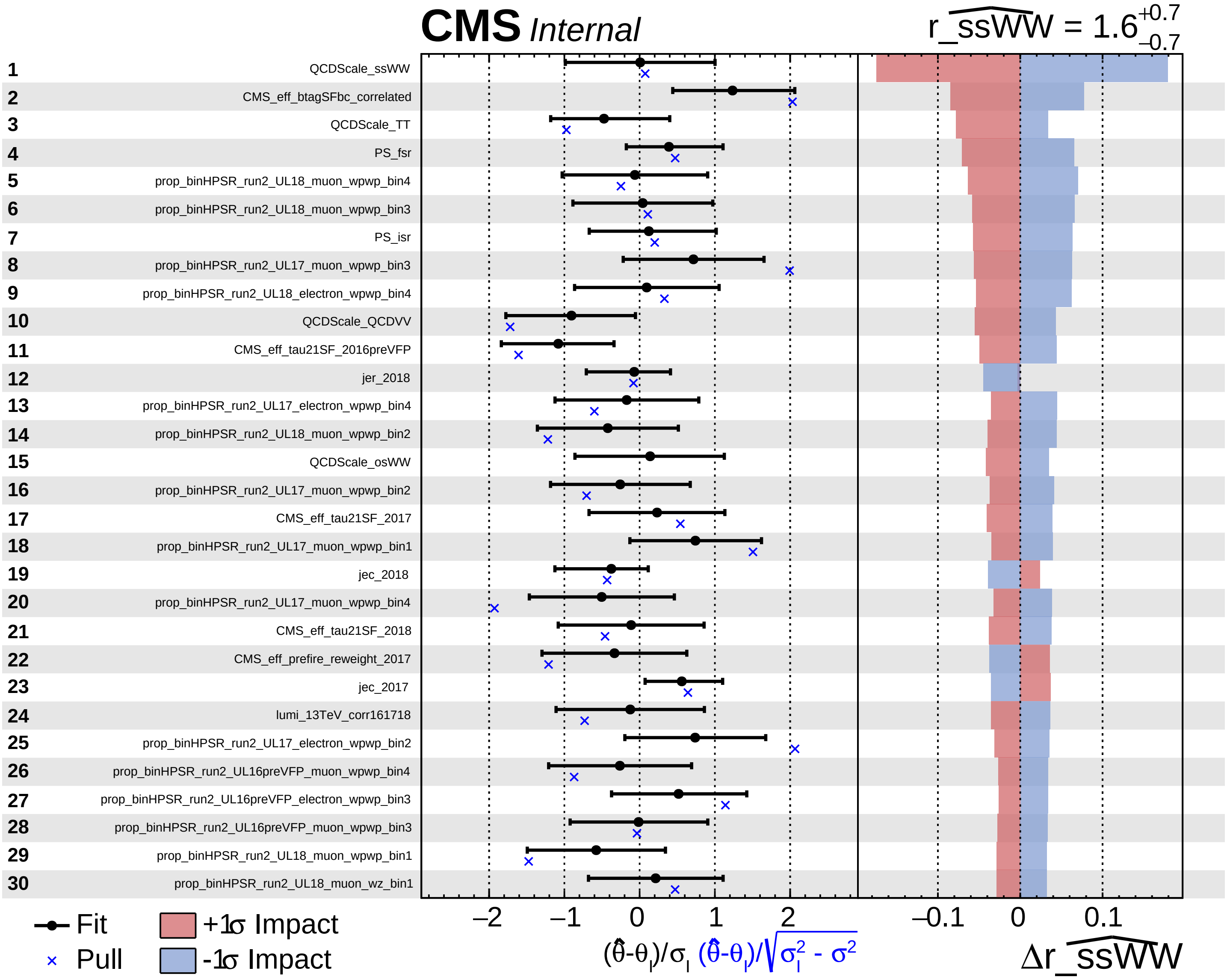
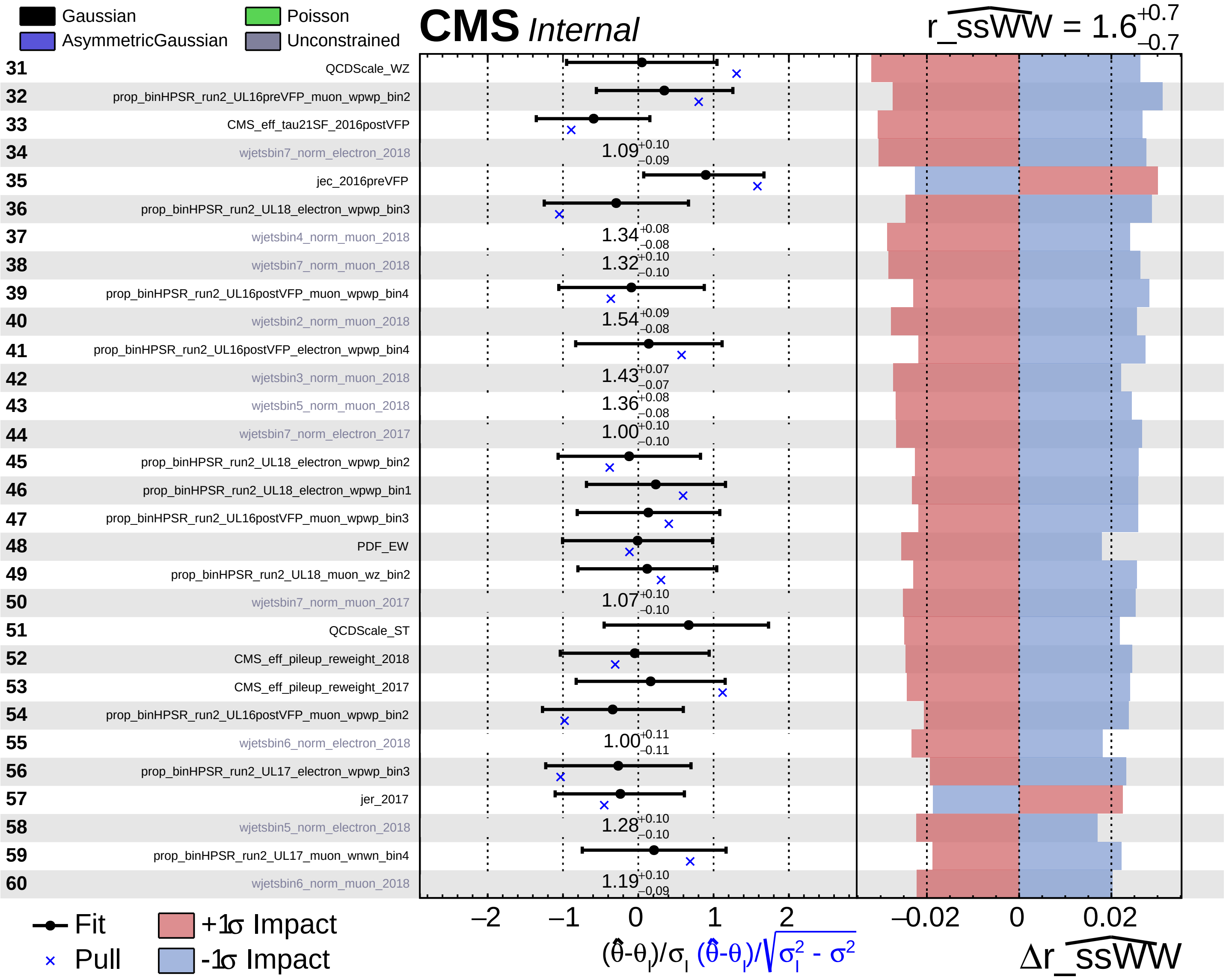
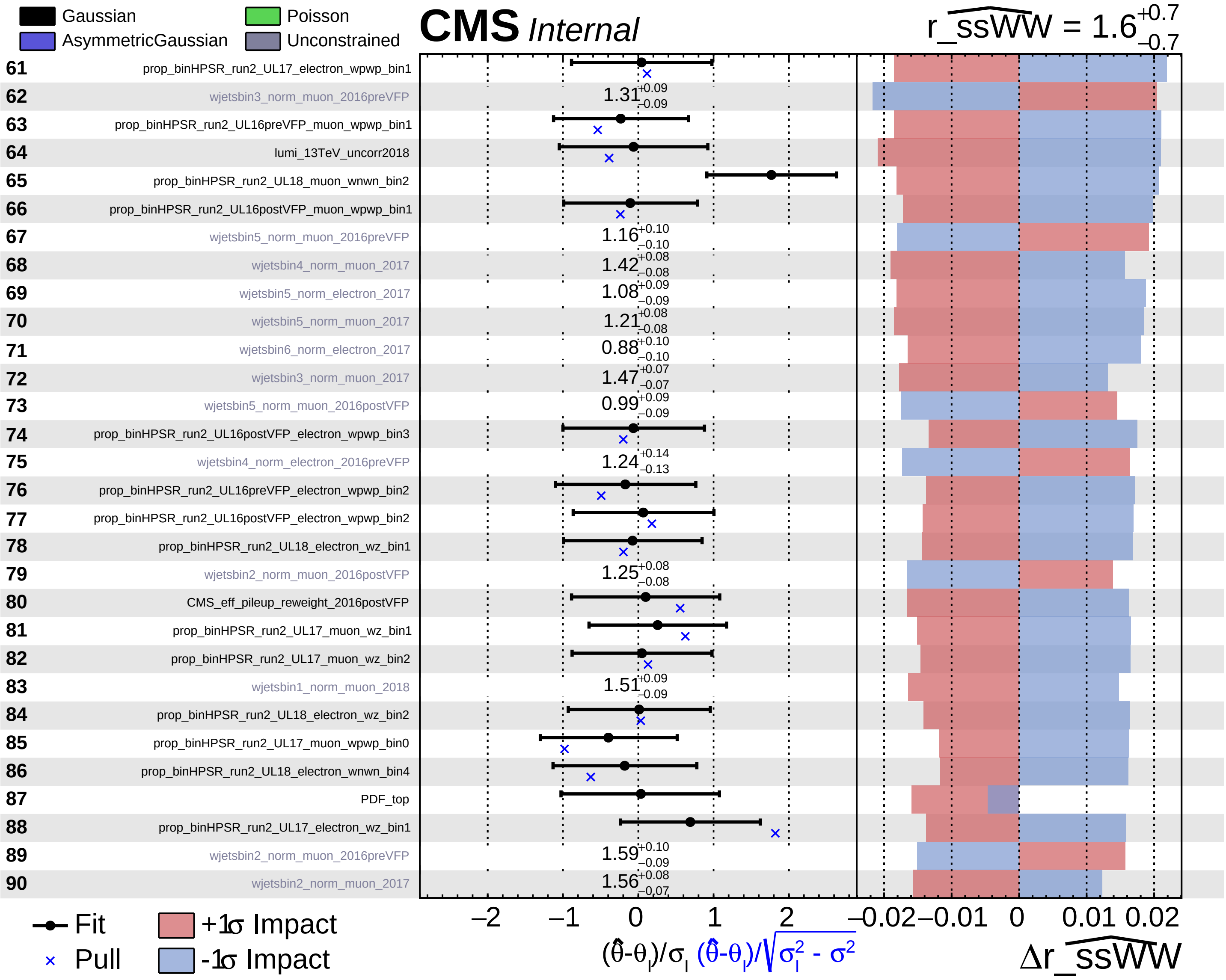


Figure I.1: The goodness-of-fit test statistic is evaluated using 1000 toy simulations. The observed test statistic from data is compared to the distribution obtained from these simulations. Since the data falls within the expected range, this indicates that the statistical model provides a satisfactory description of the observed data.

J. Nuisances impact plots


CMS Internal
r_ssWW = 1.6 +0.7 −0.7
1 QCDScale_ssWW
2 CMS_eff_btagSFbc_correlated
3 QCDScale_TT
4 PS_fsr
5 prop_binHPSR_run2_UL18_muon_wpwp_bin4
6 prop_binHPSR_run2_UL18_muon_wpwp_bin3
7 PS_isr
8 prop_binHPSR_run2_UL17_muon_wpwp_bin3
9 prop_binHPSR_run2_UL18_electron_wpwp_bin4
10 QCDScale_QCDVV
11 CMS_eff_tau21SF_2016preVFP
12 jer_2018
13 prop_binHPSR_run2_UL17_electron_wpwp_bin4
14 prop_binHPSR_run2_UL18_muon_wpwp_bin2
15 QCDScale_osWW
16 prop_binHPSR_run2_UL17_muon_wpwp_bin2
17 CMS_eff_tau21SF_2017
18 prop_binHPSR_run2_UL17_muon_wpwp_bin1
19 jec_2018
20 prop_binHPSR_run2_UL17_muon_wpwp_bin4
21 CMS_eff_tau21SF_2018
22 CMS_eff_prefire_reweight_2017
23 jec_2017
24 lumi_13TeV_corr161718
25 prop_binHPSR_run2_UL17_electron_wpwp_bin2
26 prop_binHPSR_run2_UL16preVFP_muon_wpwp_bin4
27 prop_binHPSR_run2_UL16preVFP_electron_wpwp_bin3
28 prop_binHPSR_run2_UL16preVFP_muon_wpwp_bin3
29 prop_binHPSR_run2_UL18_muon_wpwp_bin1
30 prop_binHPSR_run2_UL18_muon_wz_bin1
−2
−1
0
1
2
$(\hat{\theta}-\theta_I)/\sigma_I$ $(\hat{\theta}-\theta_I)/\sqrt{\sigma_I^2 - \sigma^2}$
−0.1
0
0.1
Δr_ssWW
Fit
Pull
+1σ Impact
-1σ Impact


Gaussian
Poisson
AsymmetricGaussian
Unconstrained
CMS Internal
$\widehat{r\_ssWW} = 1.6^{+0.7}_{-0.7}$
31 QCDScale_WZ
32 prop_binHPSR_run2_UL16preVFP_muon_wpwp_bin2
33 CMS_eff_tau21SF_2016postVFP
34 wjetsbin7_norm_electron_2018 $1.09^{+0.10}_{-0.09}$
35 jec_2016preVFP
36 prop_binHPSR_run2_UL18_electron_wpwp_bin3
37 wjetsbin4_norm_muon_2018 $1.34^{+0.08}_{-0.08}$
38 wjetsbin7_norm_muon_2018 $1.32^{+0.10}_{-0.10}$
39 prop_binHPSR_run2_UL16postVFP_muon_wpwp_bin4
40 wjetsbin2_norm_muon_2018 $1.54^{+0.09}_{-0.08}$
41 prop_binHPSR_run2_UL16postVFP_electron_wpwp_bin4
42 wjetsbin3_norm_muon_2018 $1.43^{+0.07}_{-0.07}$
43 wjetsbin5_norm_muon_2018 $1.36^{+0.08}_{-0.08}$
44 wjetsbin7_norm_electron_2017 $1.00^{+0.10}_{-0.10}$
45 prop_binHPSR_run2_UL18_electron_wpwp_bin2
46 prop_binHPSR_run2_UL18_electron_wpwp_bin1
47 prop_binHPSR_run2_UL16postVFP_muon_wpwp_bin3
48 PDF_EW
49 prop_binHPSR_run2_UL18_muon_wz_bin2
50 wjetsbin7_norm_muon_2017 $1.07^{+0.10}_{-0.10}$
51 QCDScale_ST
52 CMS_eff_pileup_reweight_2018
53 CMS_eff_pileup_reweight_2017
54 prop_binHPSR_run2_UL16postVFP_muon_wpwp_bin2
55 wjetsbin6_norm_electron_2018 $1.00^{+0.11}_{-0.11}$
56 prop_binHPSR_run2_UL17_electron_wpwp_bin3
57 jer_2017
58 wjetsbin5_norm_electron_2018 $1.28^{+0.10}_{-0.10}$
59 prop_binHPSR_run2_UL17_muon_wnwn_bin4
60 wjetsbin6_norm_muon_2018 $1.19^{+0.10}_{-0.09}$
−2
−1
0
1
2
$(\hat{\theta}-\theta_I)/\sigma_I$ $(\hat{\theta}-\theta_I)/\sqrt{\sigma_I^2 - \sigma^2}$
−0.02
0
0.02
$\Delta\widehat{r\_ssWW}$
Fit
+1σ Impact
Pull
-1σ Impact


Gaussian
AsymmetricGaussian
Poisson
Unconstrained
CMS Internal
$\widehat{r\_ssWW} = 1.6^{+0.7}_{-0.7}$
61 prop_binHPSR_run2_UL17_electron_wpwp_bin1
62 wjetsbin3_norm_muon_2016preVFP 1.31 +0.09 −0.09
63 prop_binHPSR_run2_UL16preVFP_muon_wpwp_bin1
64 lumi_13TeV_uncorr2018
65 prop_binHPSR_run2_UL18_muon_wnwn_bin2
66 prop_binHPSR_run2_UL16postVFP_muon_wpwp_bin1
67 wjetsbin5_norm_muon_2016preVFP 1.16 +0.10 −0.10
68 wjetsbin4_norm_muon_2017 1.42 +0.08 −0.08
69 wjetsbin5_norm_electron_2017 1.08 +0.09 −0.09
70 wjetsbin5_norm_muon_2017 1.21 +0.08 −0.08
71 wjetsbin6_norm_electron_2017 0.88 +0.10 −0.10
72 wjetsbin3_norm_muon_2017 1.47 +0.07 −0.07
73 wjetsbin5_norm_muon_2016postVFP 0.99 +0.09 −0.09
74 prop_binHPSR_run2_UL16postVFP_electron_wpwp_bin3
75 wjetsbin4_norm_electron_2016preVFP 1.24 +0.14 −0.13
76 prop_binHPSR_run2_UL16preVFP_electron_wpwp_bin2
77 prop_binHPSR_run2_UL16postVFP_electron_wpwp_bin2
78 prop_binHPSR_run2_UL18_electron_wz_bin1
79 wjetsbin2_norm_muon_2016postVFP 1.25 +0.08 −0.08
80 CMS_eff_pileup_reweight_2016postVFP
81 prop_binHPSR_run2_UL17_muon_wz_bin1
82 prop_binHPSR_run2_UL17_muon_wz_bin2
83 wjetsbin1_norm_muon_2018 1.51 +0.09 −0.09
84 prop_binHPSR_run2_UL18_electron_wz_bin2
85 prop_binHPSR_run2_UL17_muon_wpwp_bin0
86 prop_binHPSR_run2_UL18_electron_wnwn_bin4
87 PDF_top
88 prop_binHPSR_run2_UL17_electron_wz_bin1
89 wjetsbin2_norm_muon_2016preVFP 1.59 +0.10 −0.09
90 wjetsbin2_norm_muon_2017 1.56 +0.08 −0.07
−2 −1 0 1 2
$(\hat{\theta}-\theta_I)/\sigma_I$ $(\hat{\theta}-\theta_I)/\sqrt{\sigma_I^2 - \sigma^2}$
−0.02 −0.01 0 0.01 0.02
$\Delta\widehat{r\_ssWW}$
Fit
Pull
+1σ Impact
-1σ Impact

K. Additional W+jets control plots without corrections

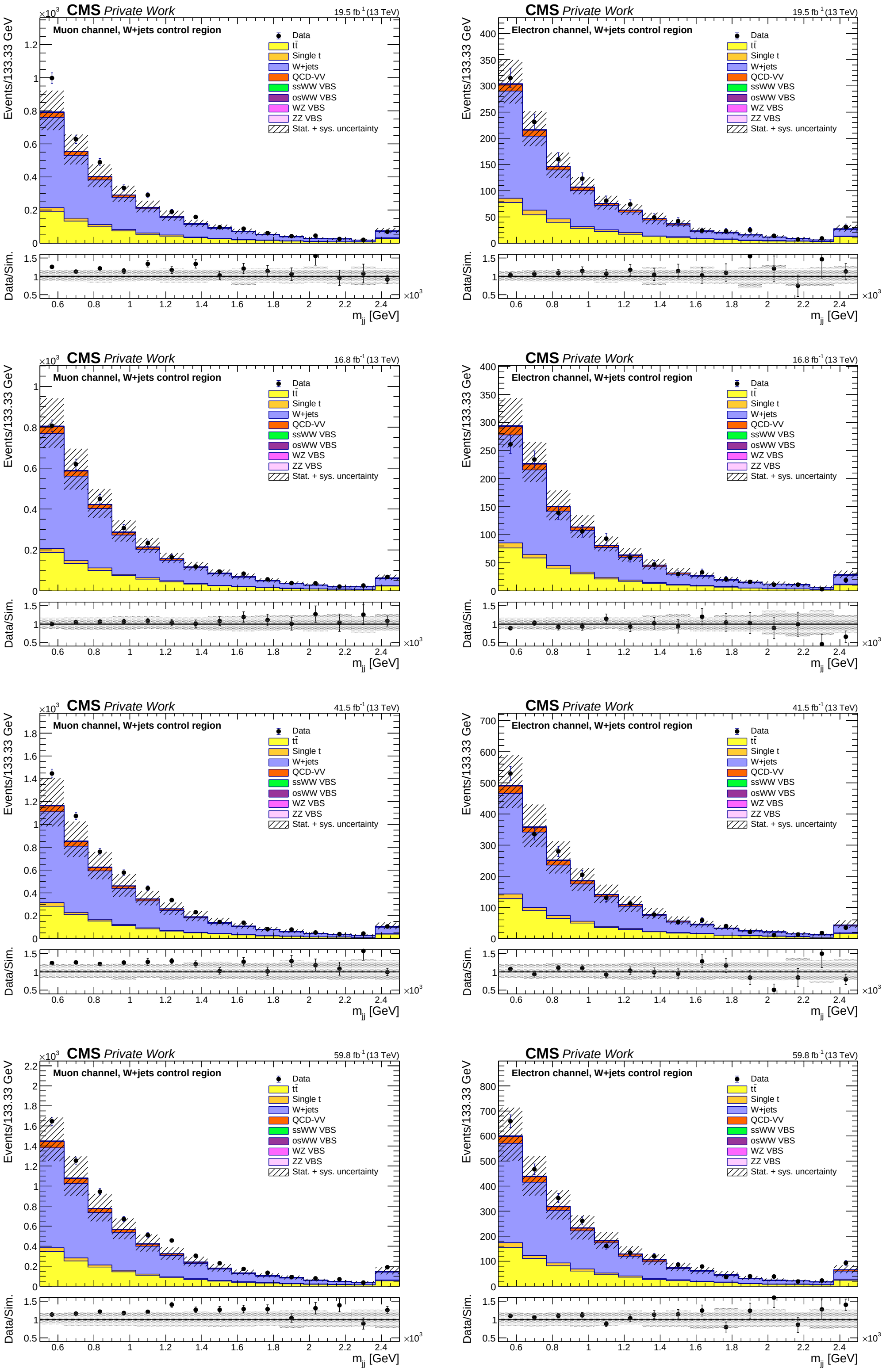


Figure K.1: Data and simulation comparison of **dijet invariant mass** ($\mathbf{m_{jj}}$) of the two VBS jets in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

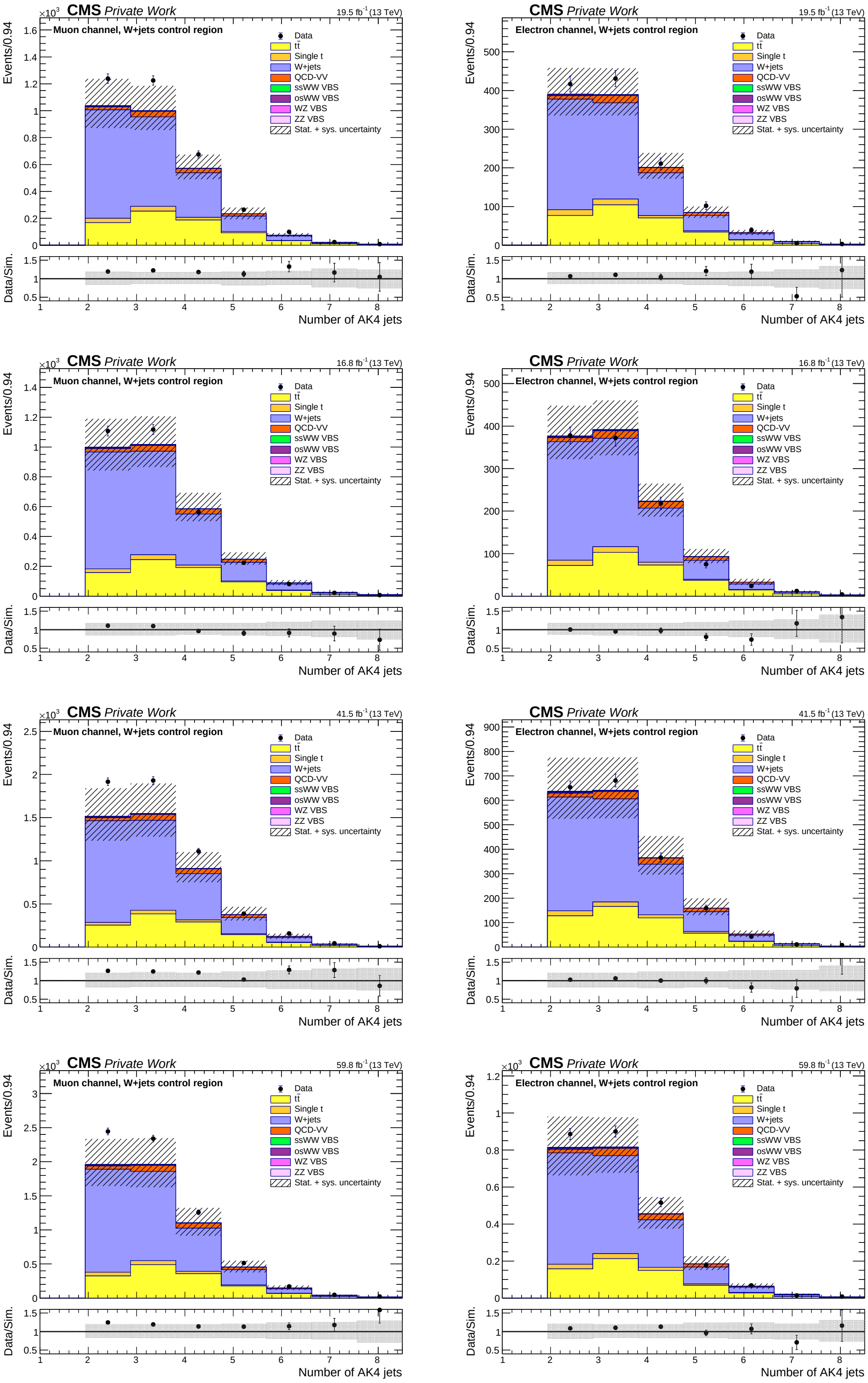


Figure K.2: Data and simulation comparison of **number of AK4 jets** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

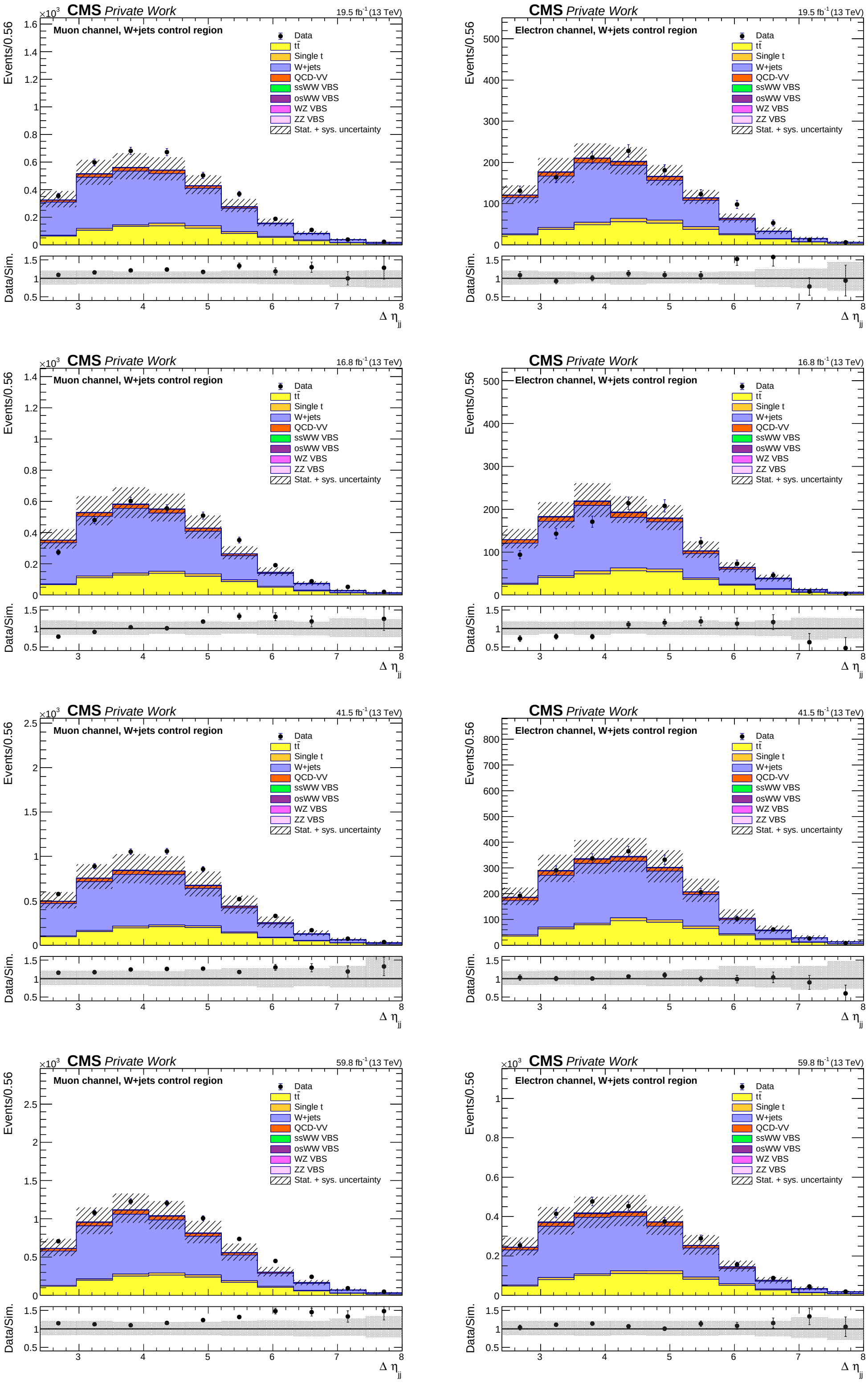


Figure K.3: Data and simulation comparison of **psuedorapidity separation** ($\boldsymbol{\Delta}\eta_{\text{jj}}$) of the two VBS jets in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

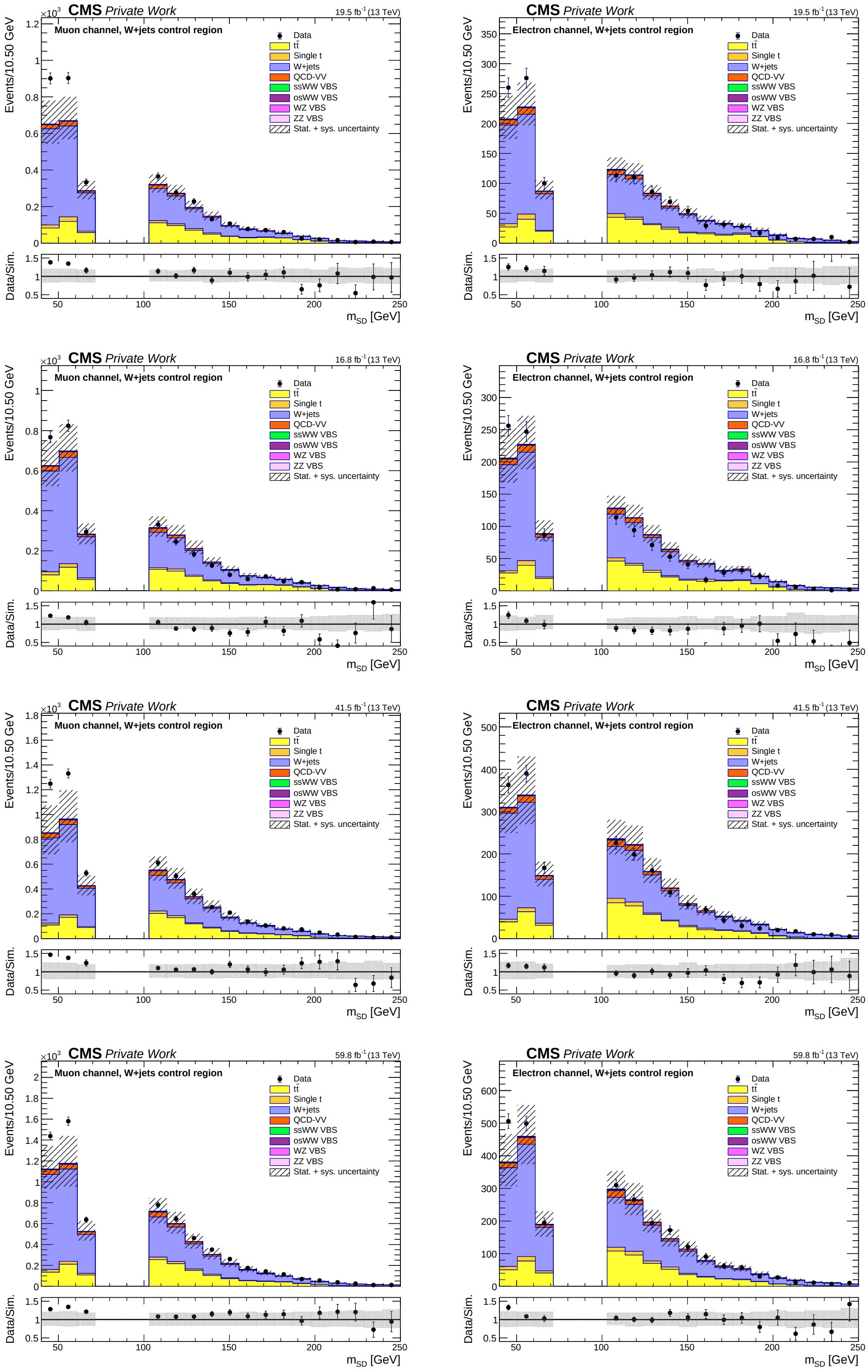


Figure K.4: Data and simulation comparison of the **soft drop mass of the AK8 jet** (**m_{SD}**) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

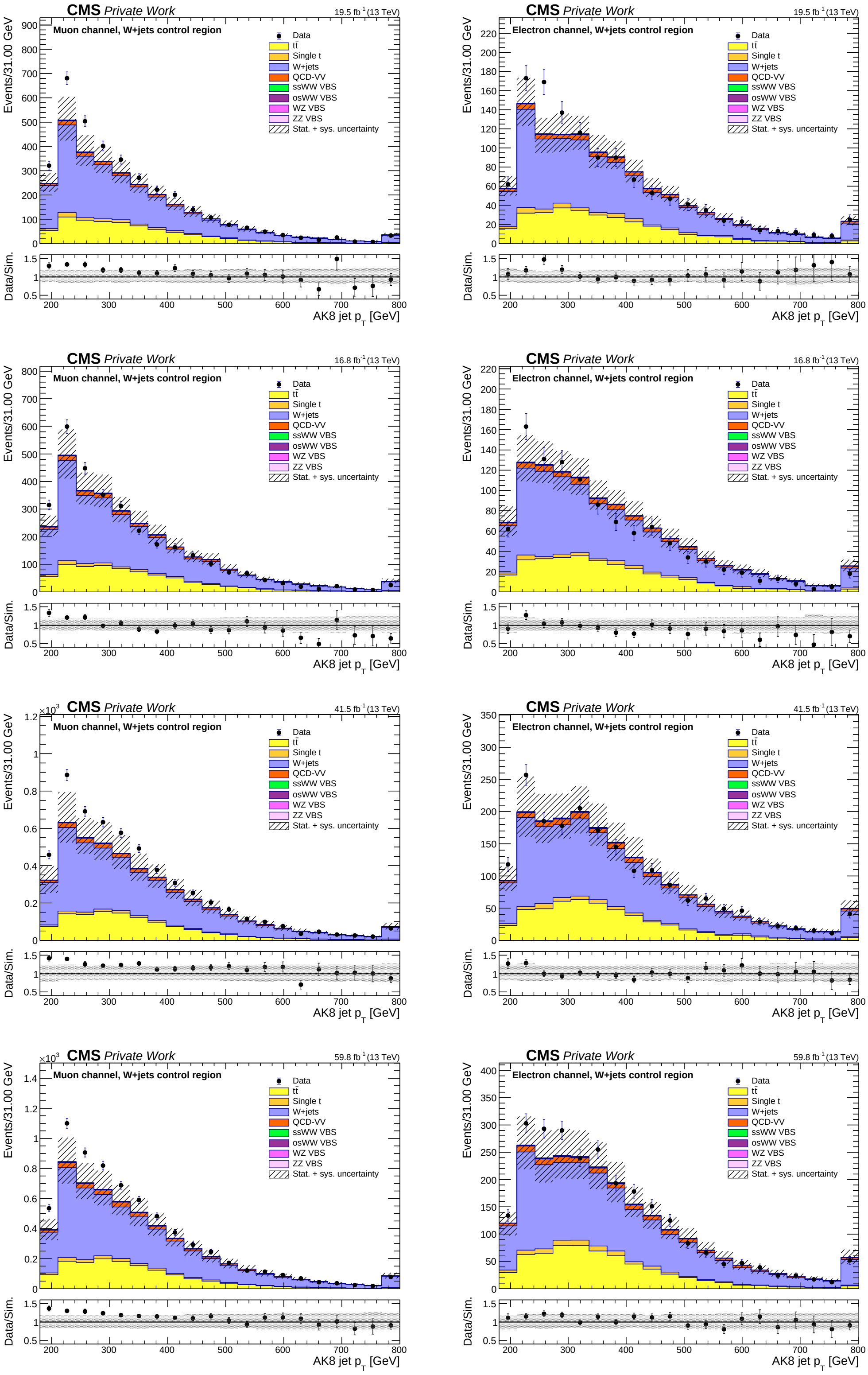


Figure K.5: Data and simulation comparison of the **transverse momentum of the AK8 jet** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

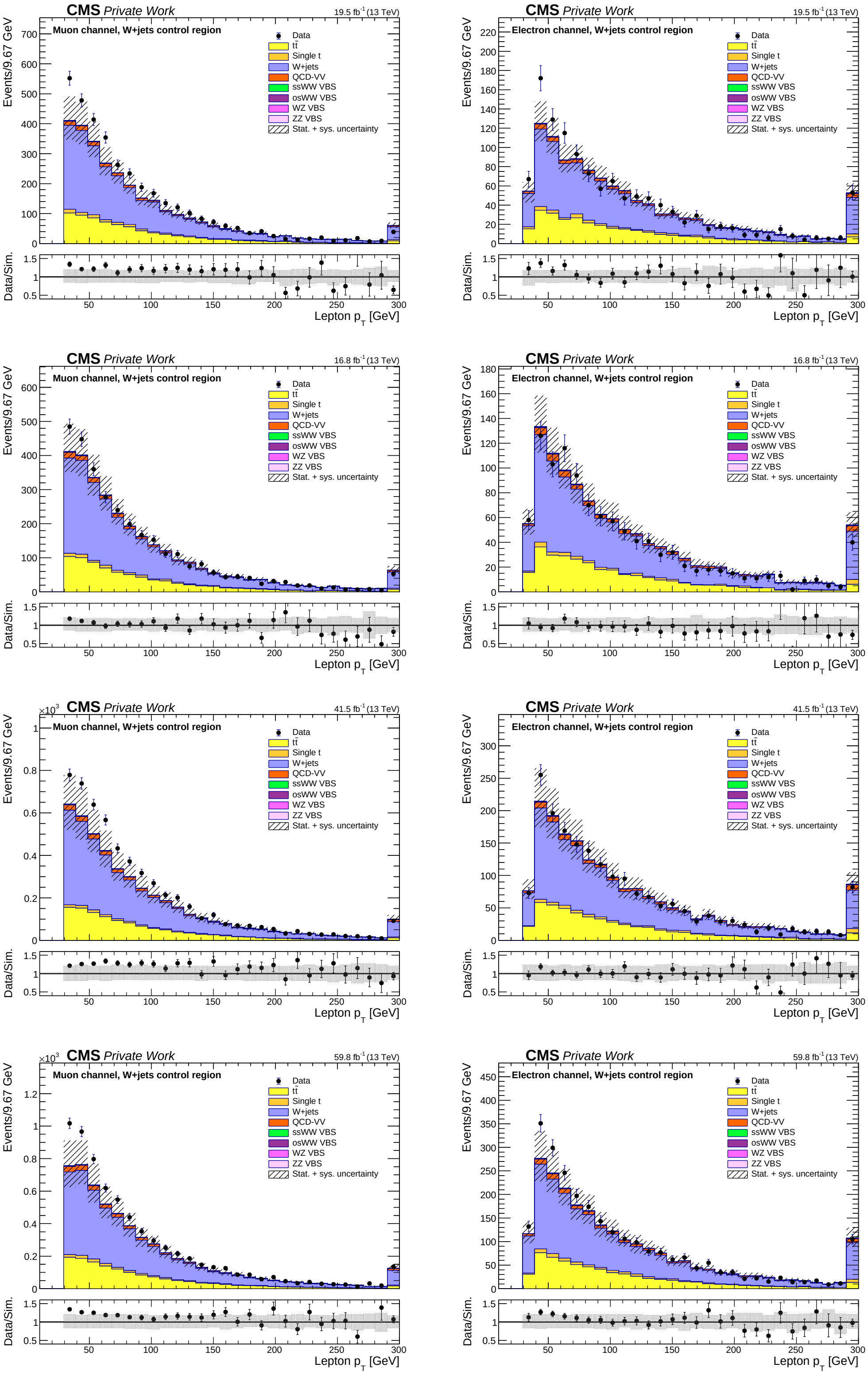


Figure K.6: Data and simulation comparison of the **transverse momentum of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

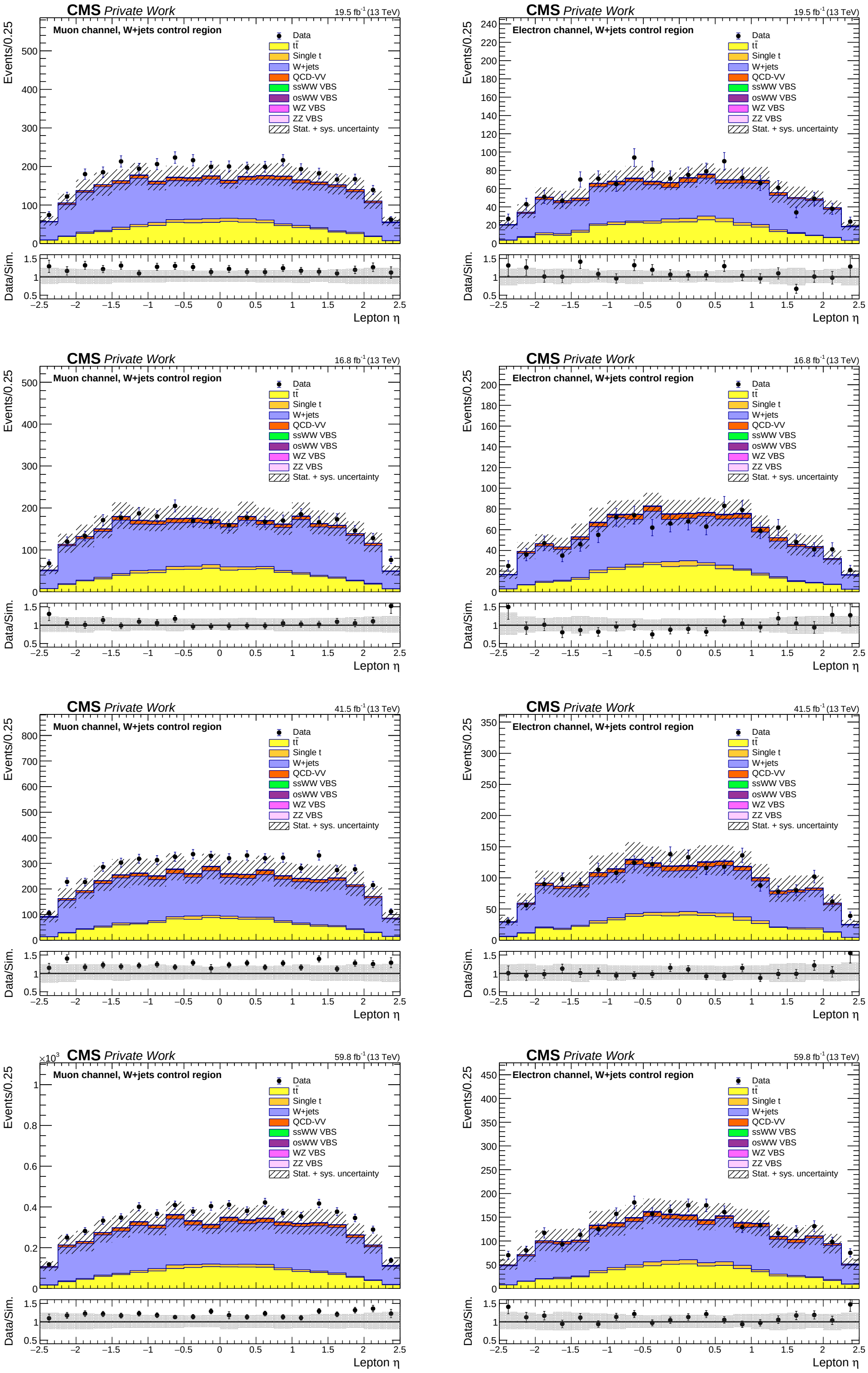


Figure K.7: Data and simulation comparison of the **pseudorapidity of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

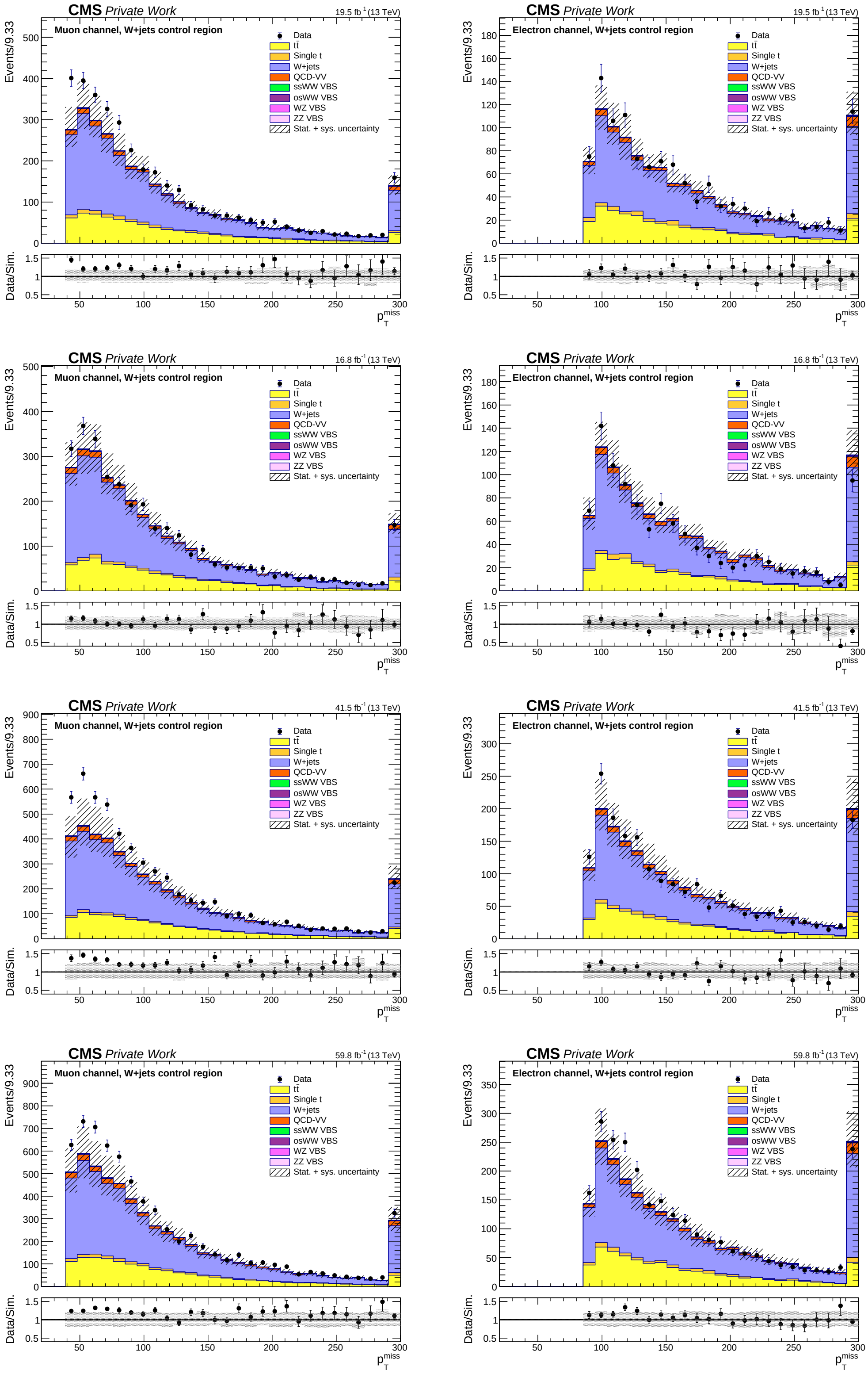


Figure K.8: Data and simulation comparison of the **missing transverse momentum** (**p_T^{miss}**) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

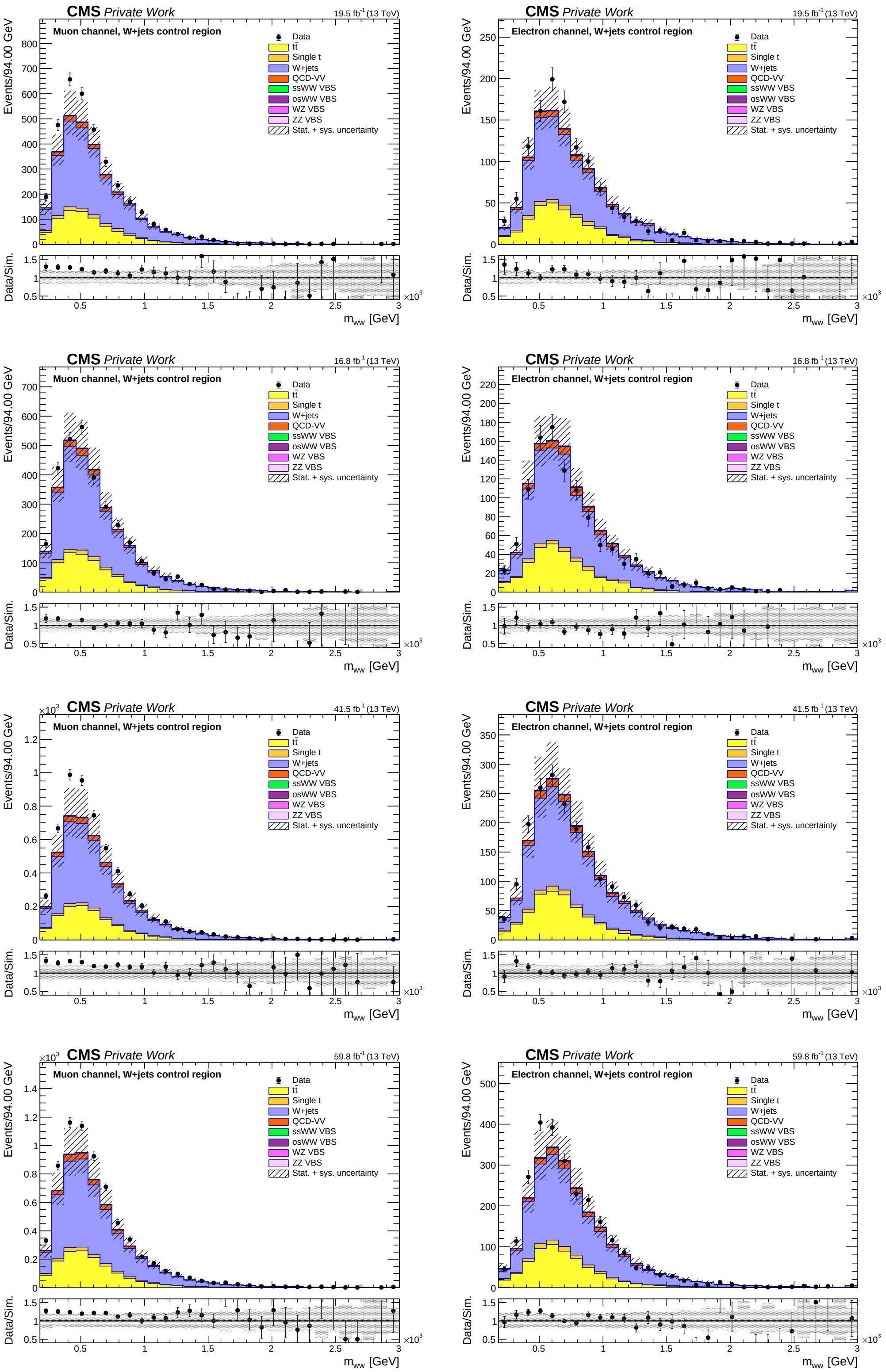


Figure K.9: Data and simulation comparison of the **invariant mass of the two W bosons ($\mathbf{m}_{\mathrm{ww}}$)** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

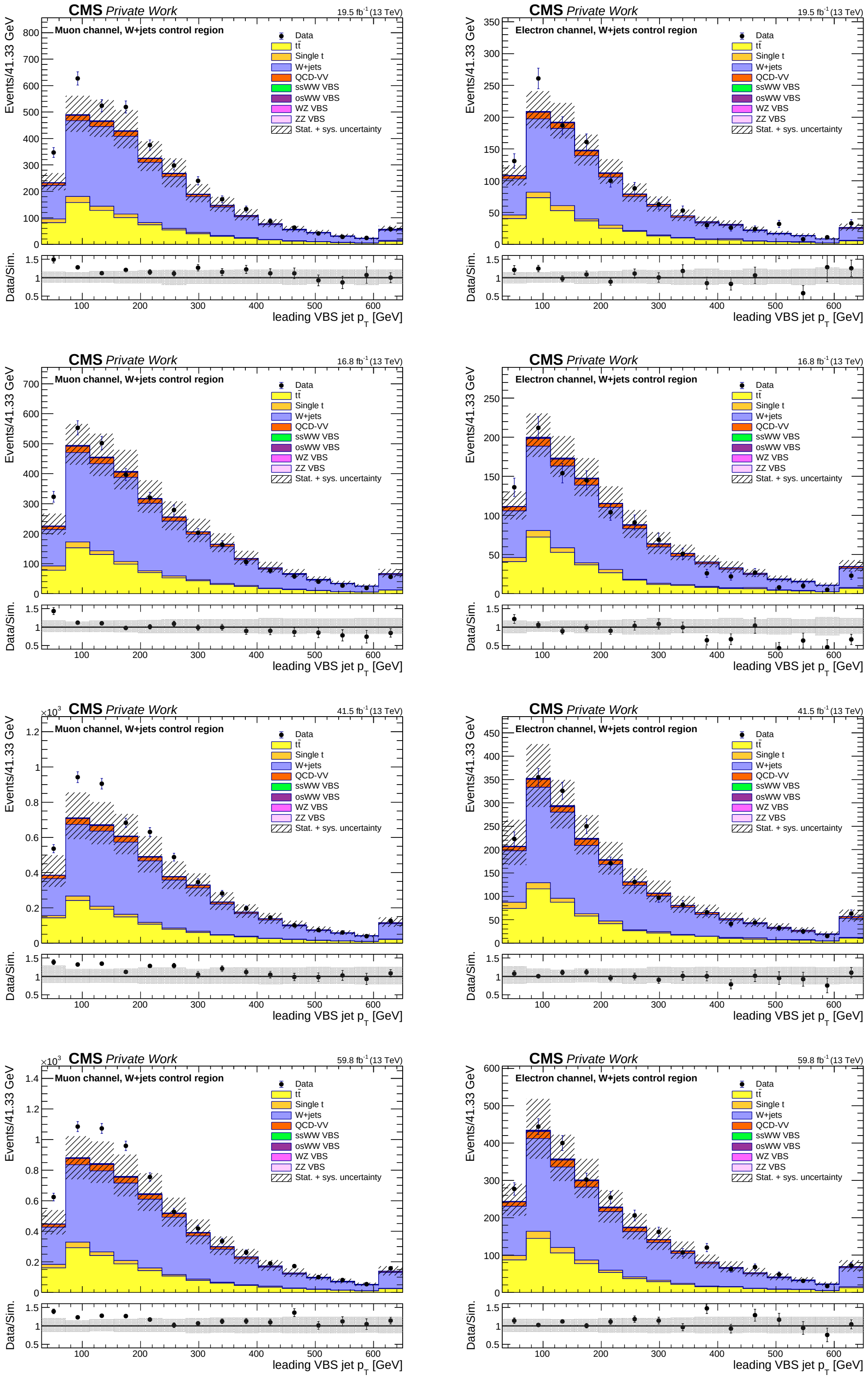


Figure K.10: Data and simulation comparison of the **leading VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

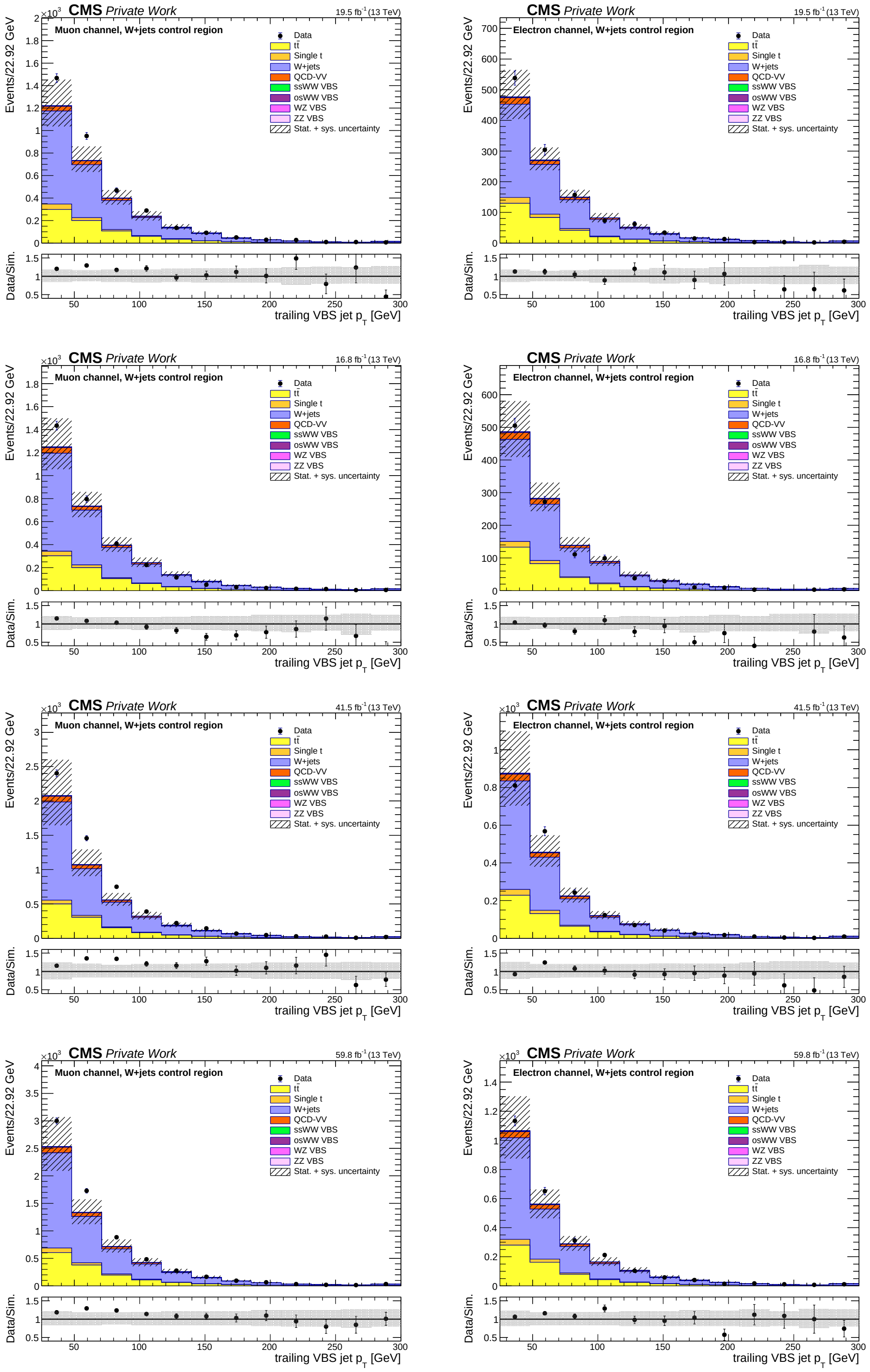


Figure K.11: Data and simulation comparison of the **trailing VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

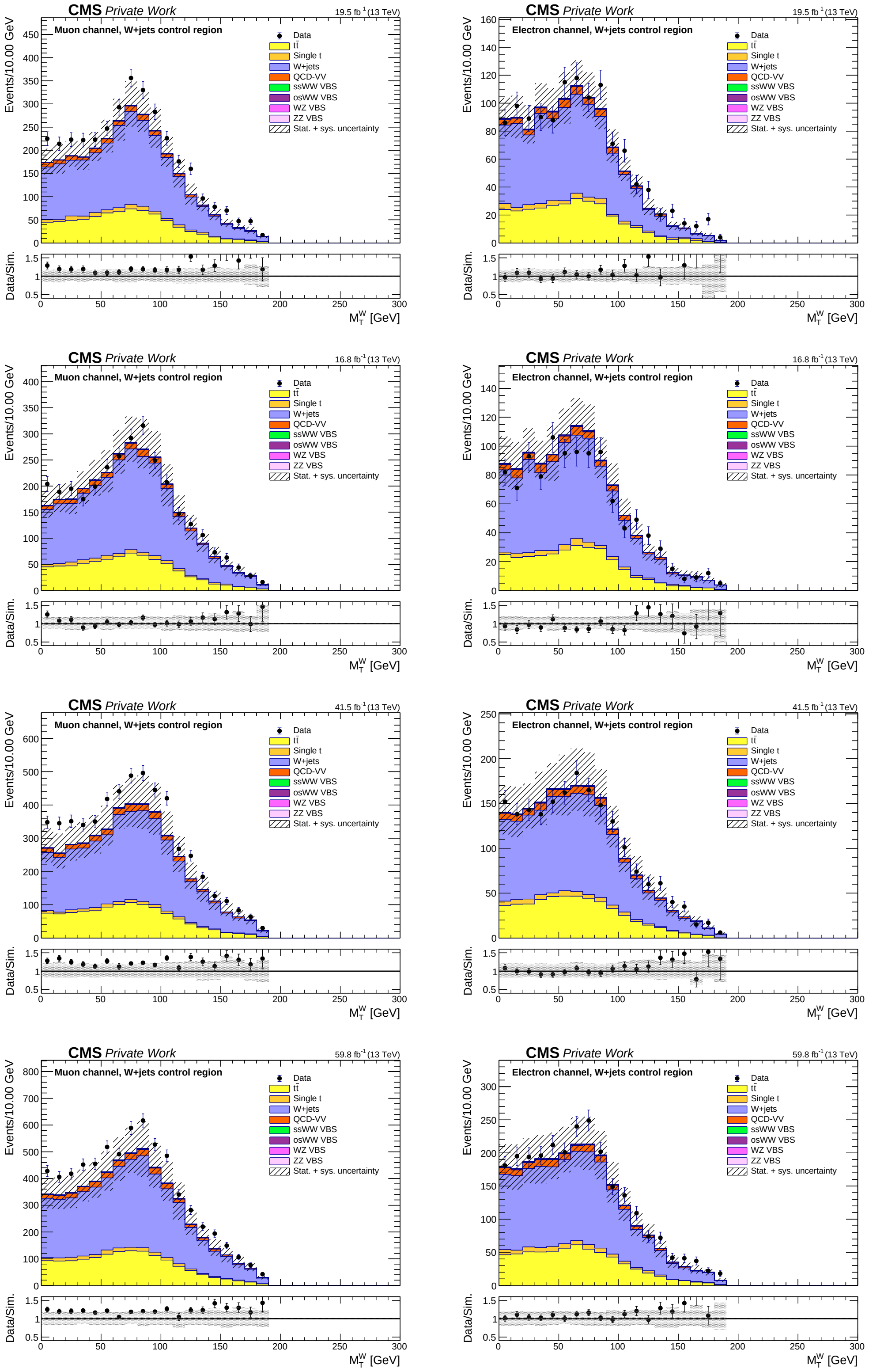


Figure K.12: Data and simulation comparison of the **transverse mass of leptonically decaying W boson** ($M_T{}^W$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

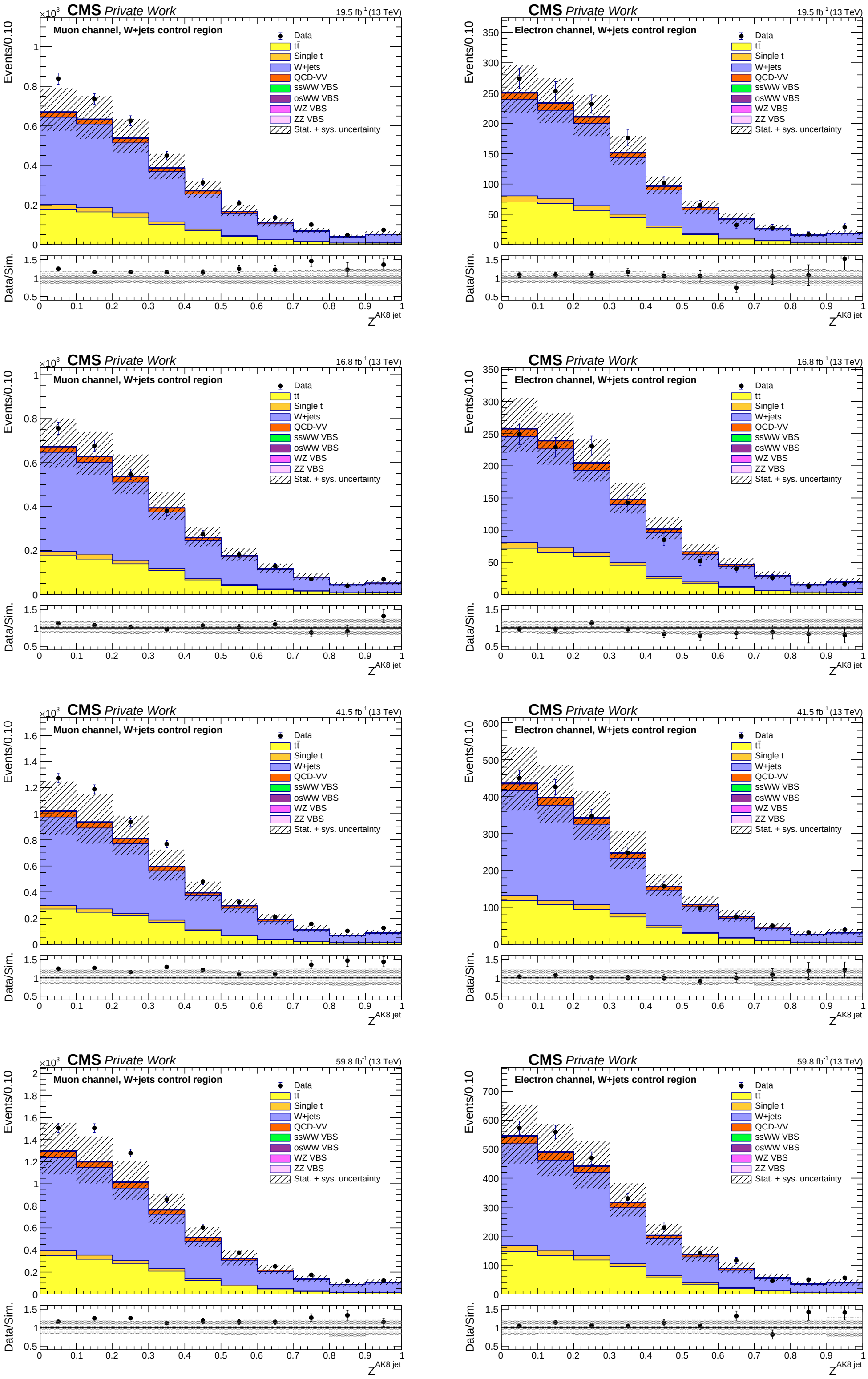


Figure K.13: Data and simulation comparison of **Zeppenfeld variable of the AK8 jet** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

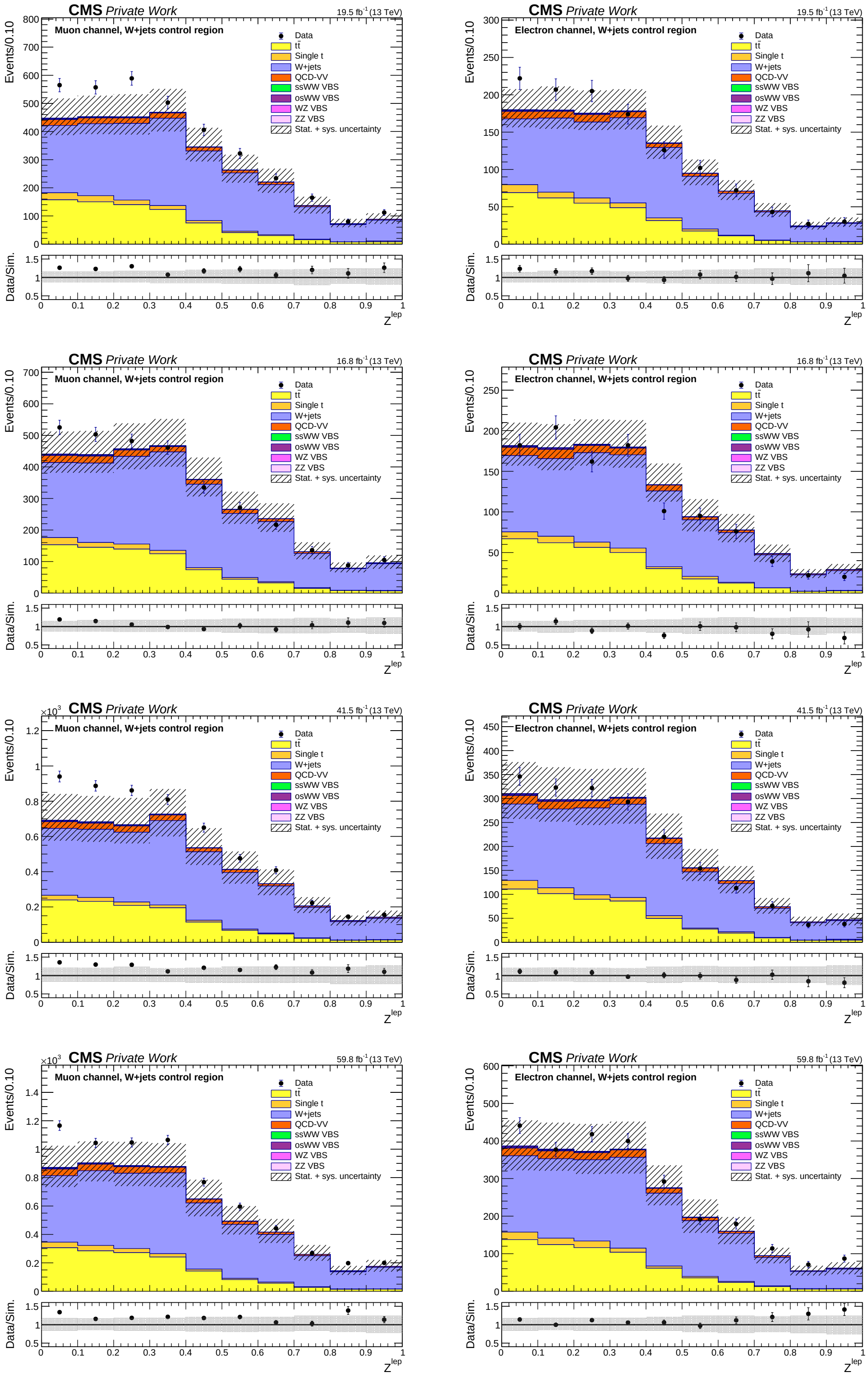


Figure K.14: Data and simulation comparison of **Zeppenfeld variable of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

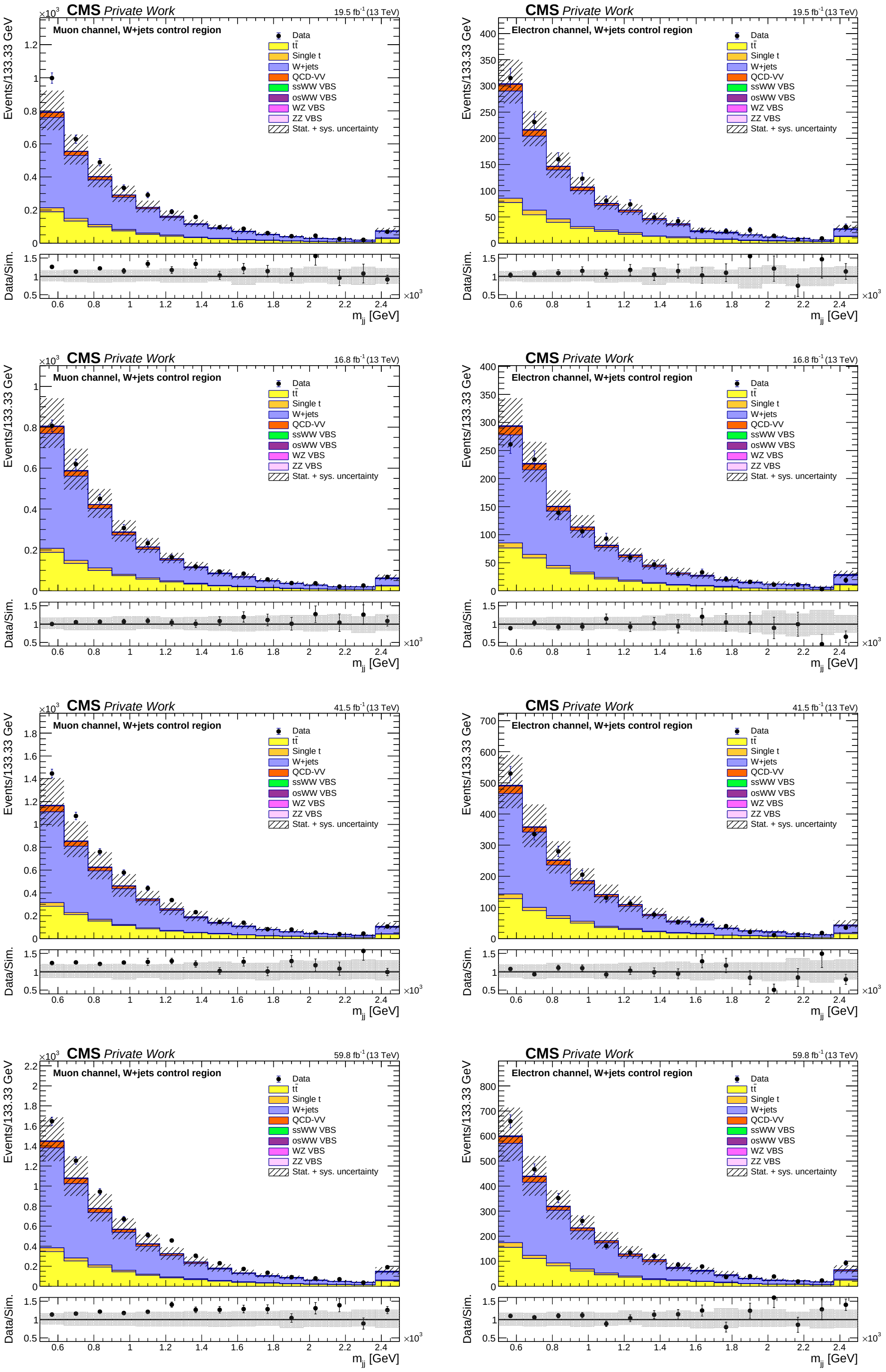


Figure K.15: Data and simulation comparison of **dijet invariant mass** ($\mathbf{m}_{\mathrm{jj}}$) of the two VBS jets in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

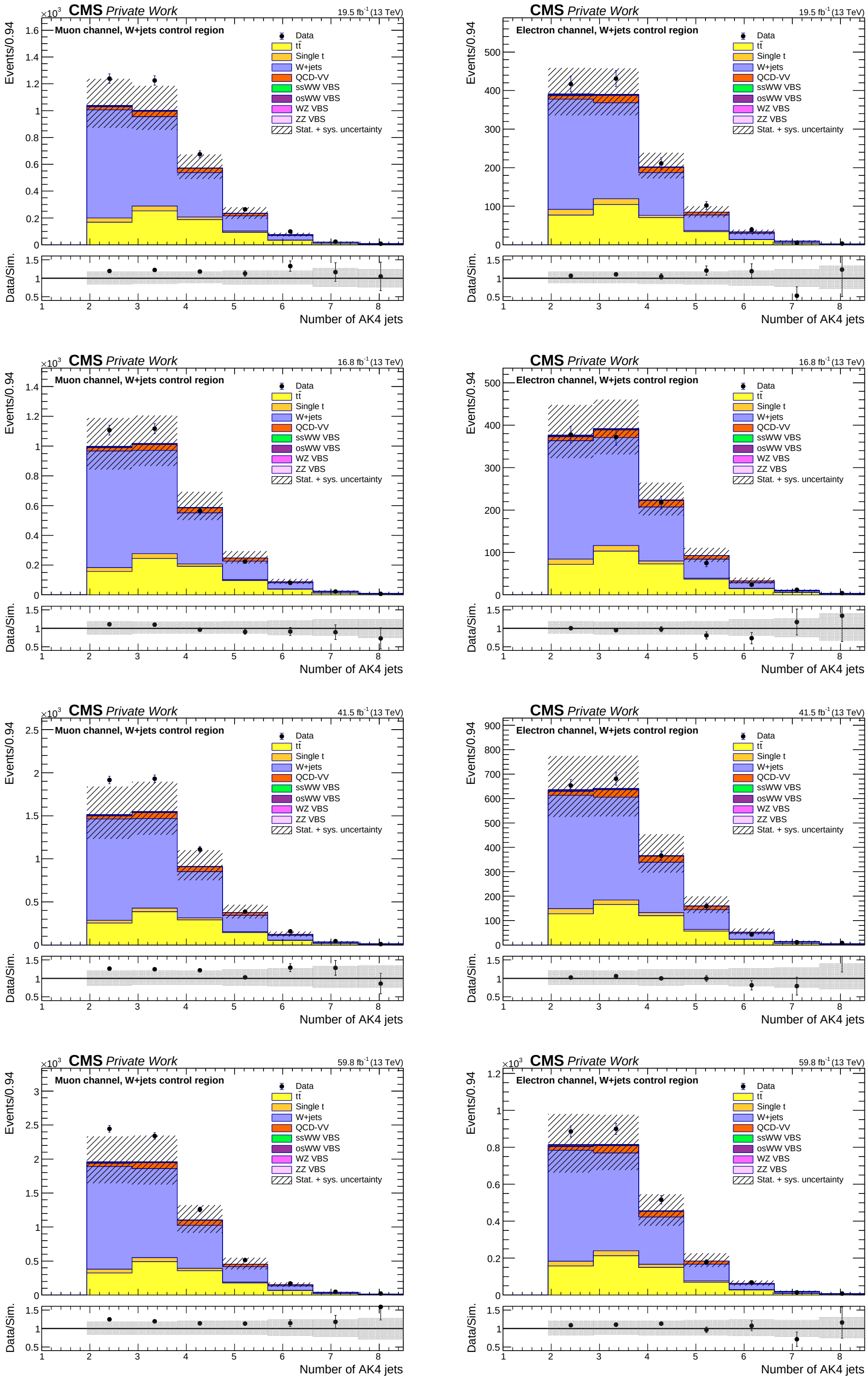


Figure K.16: Data and simulation comparison of **number of AK4 jets** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

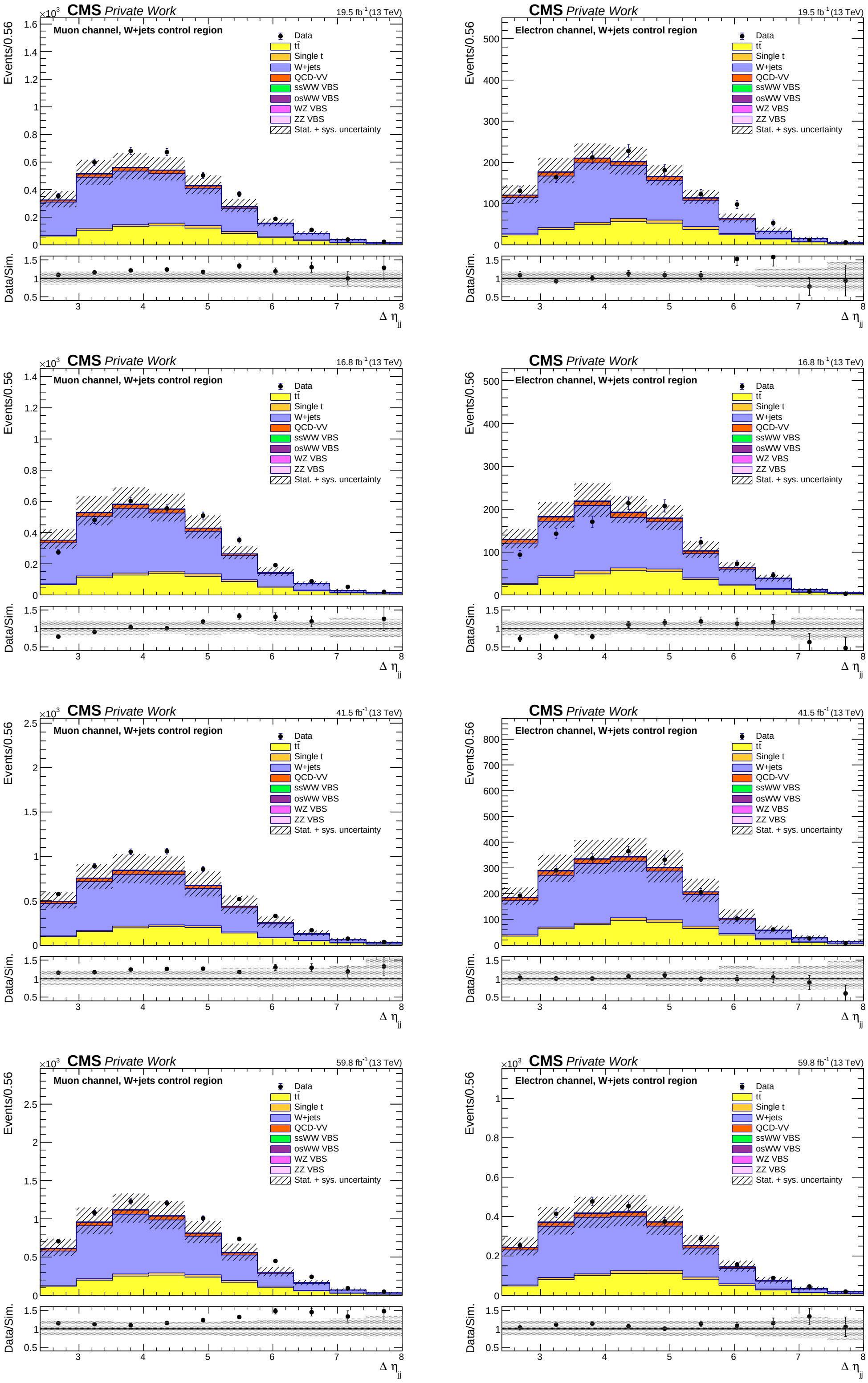


Figure K.17: Data and simulation comparison of **psuedorapidity separation** ($\boldsymbol{\Delta}\eta_{\rm jj}$) of the two VBS jets in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

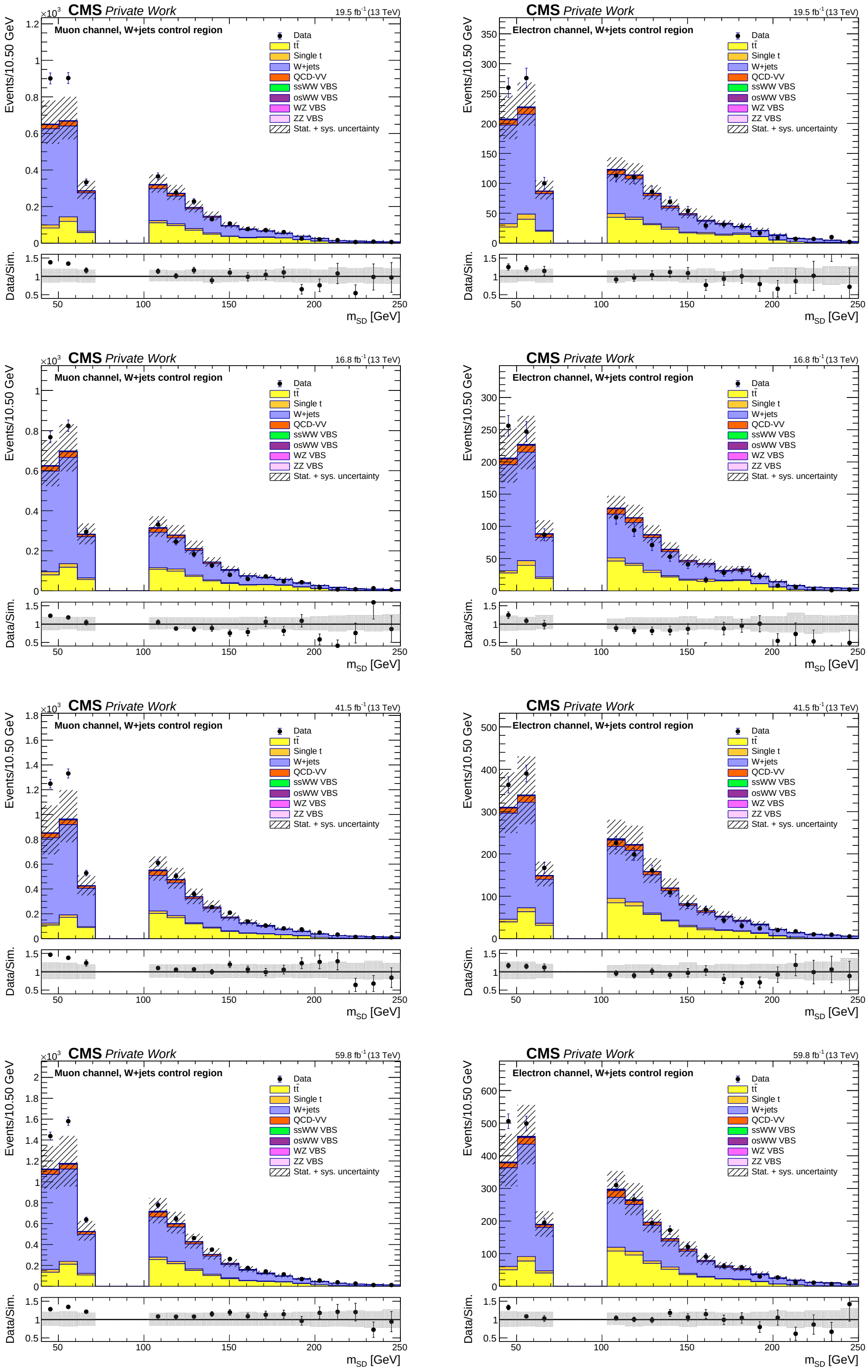


Figure K.18: Data and simulation comparison of the **soft drop mass of the AK8 jet** (**m**$_{\text{SD}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

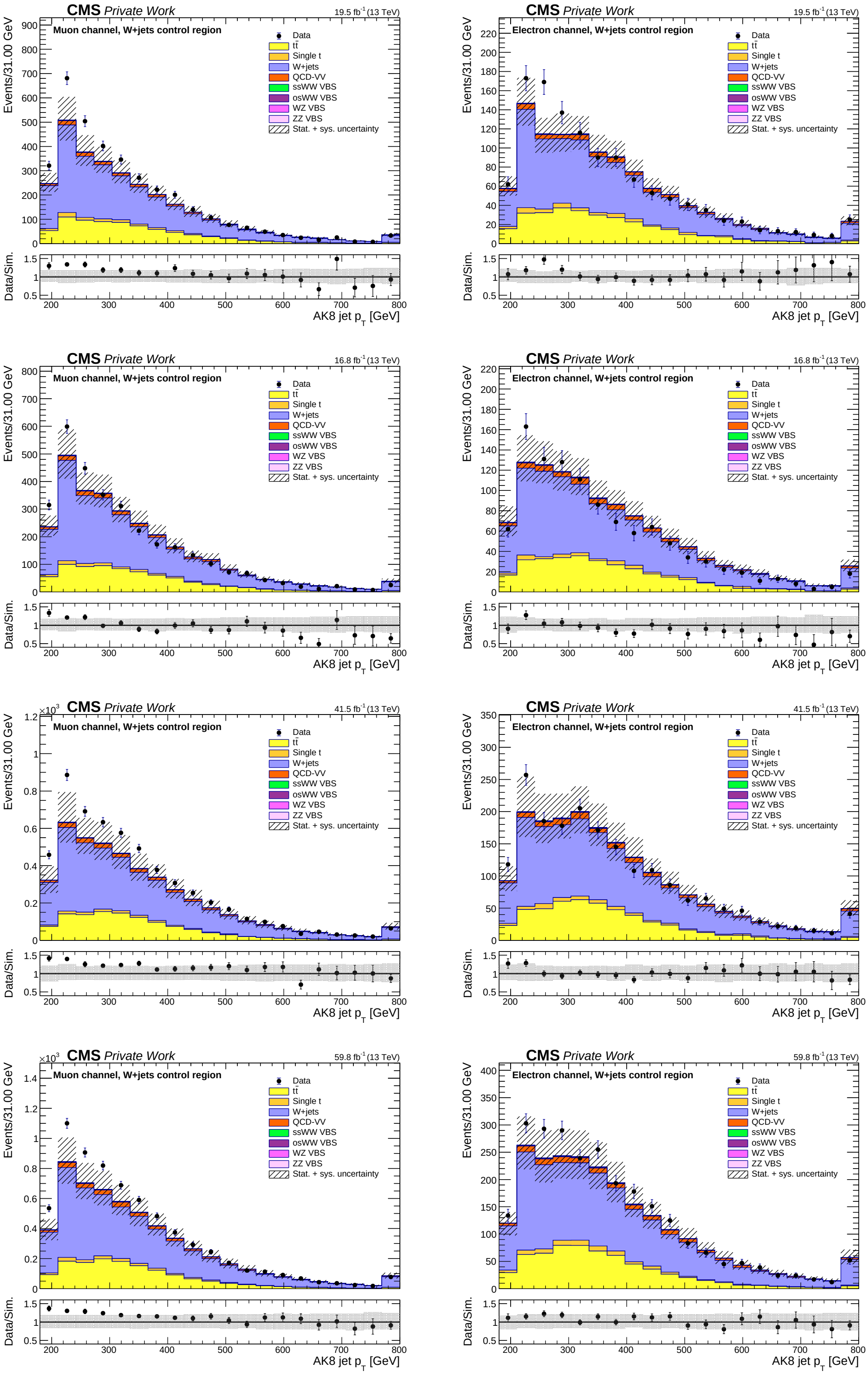


Figure K.19: Data and simulation comparison of the **transverse momentum of the AK8 jet** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

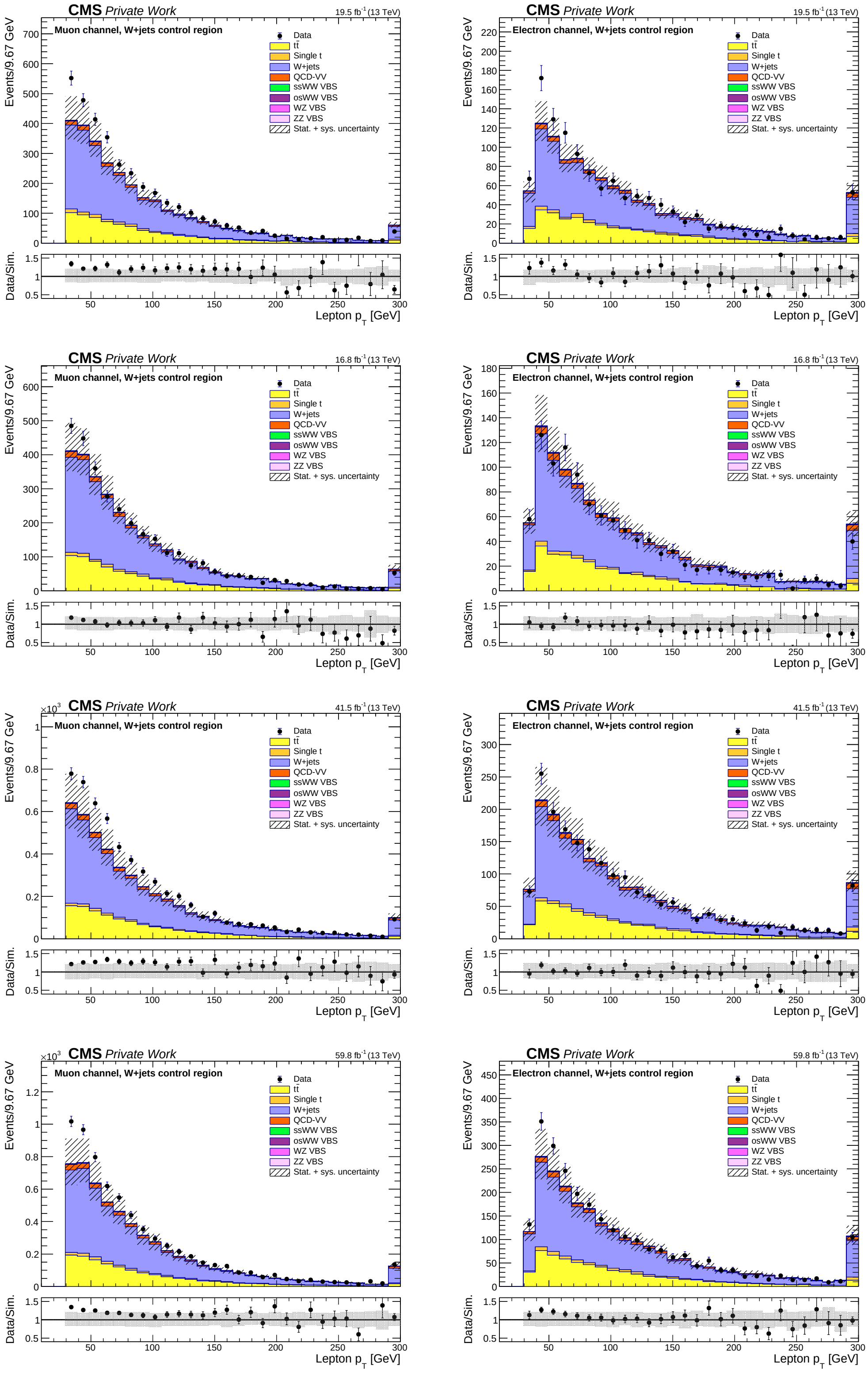


Figure K.20: Data and simulation comparison of the **transverse momentum of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

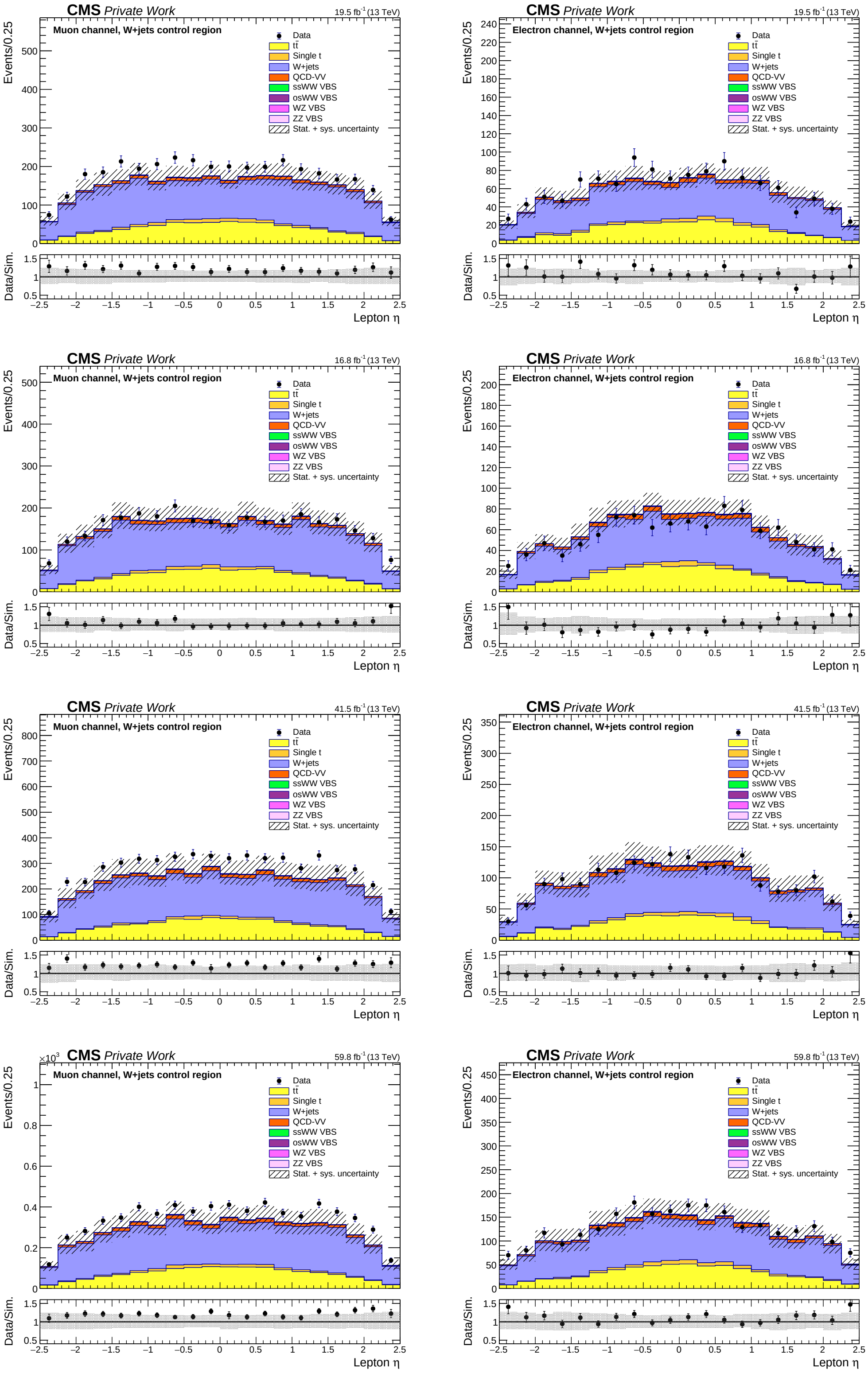


Figure K.21: Data and simulation comparison of the **pseudorapidity of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

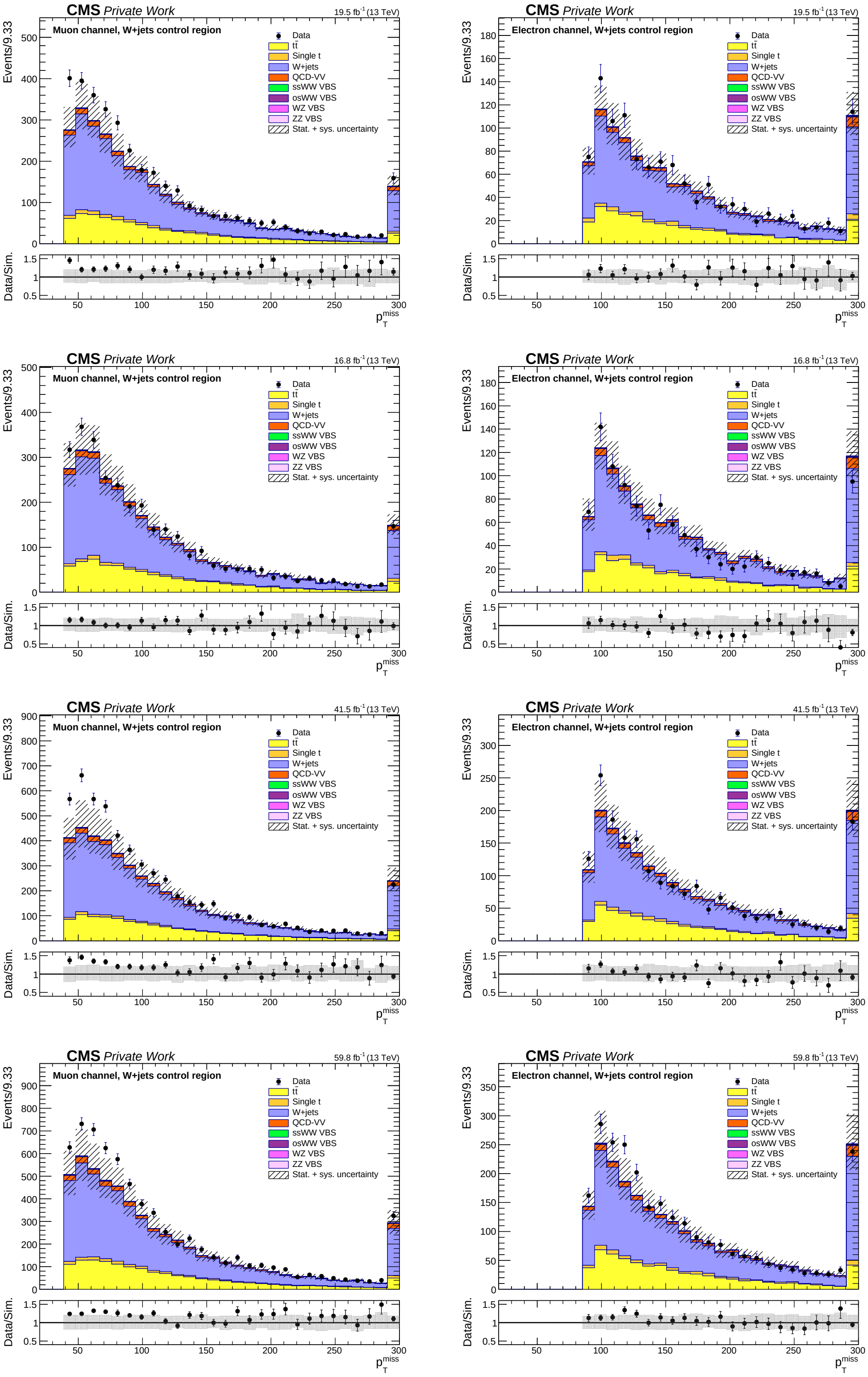


Figure K.22: Data and simulation comparison of the **missing transverse momentum** ($\mathbf{p_T}^{\text{miss}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

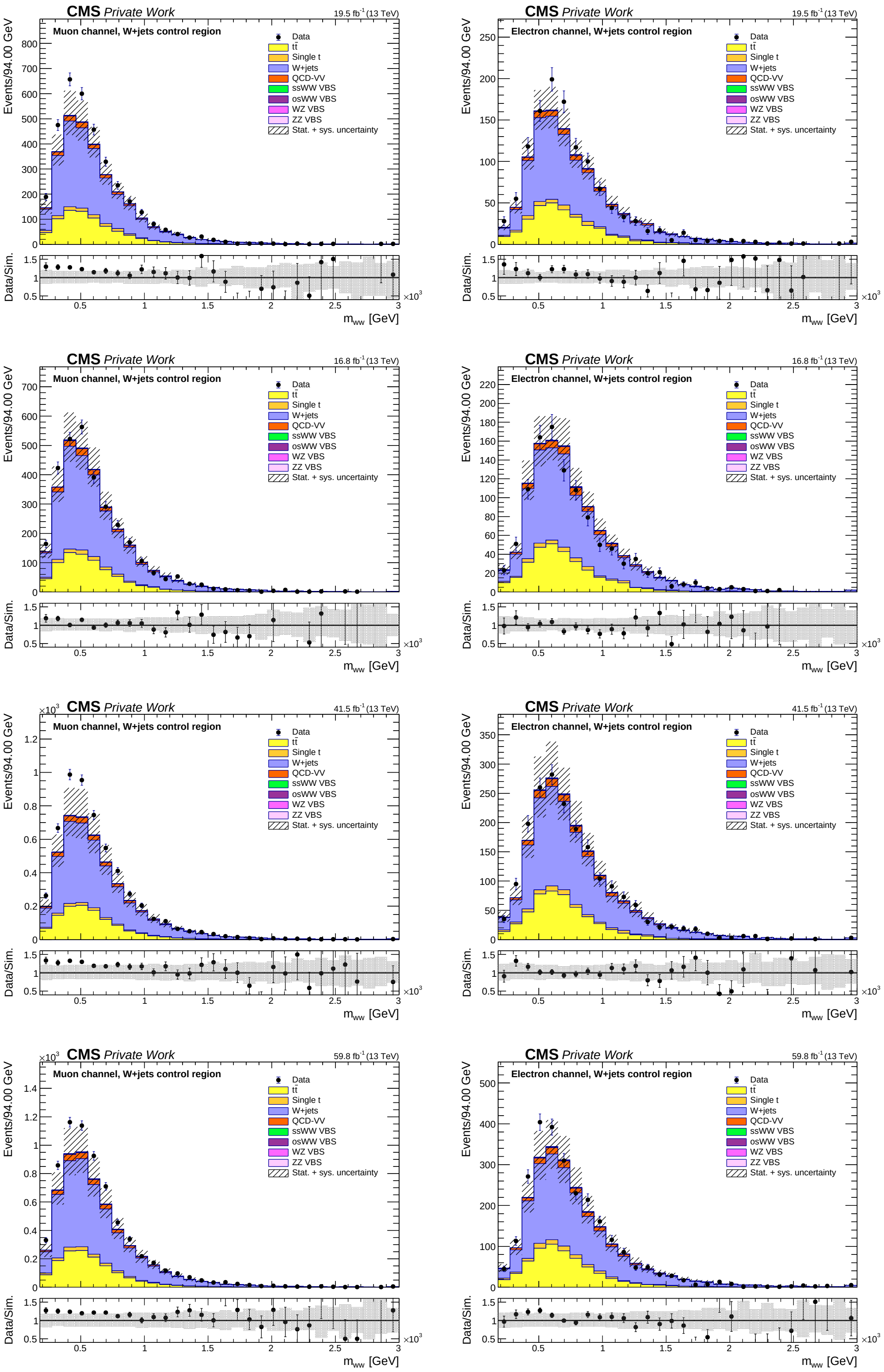


Figure K.23: Data and simulation comparison of the **invariant mass of the two W bosons** (**m**$_{\text{ww}}$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

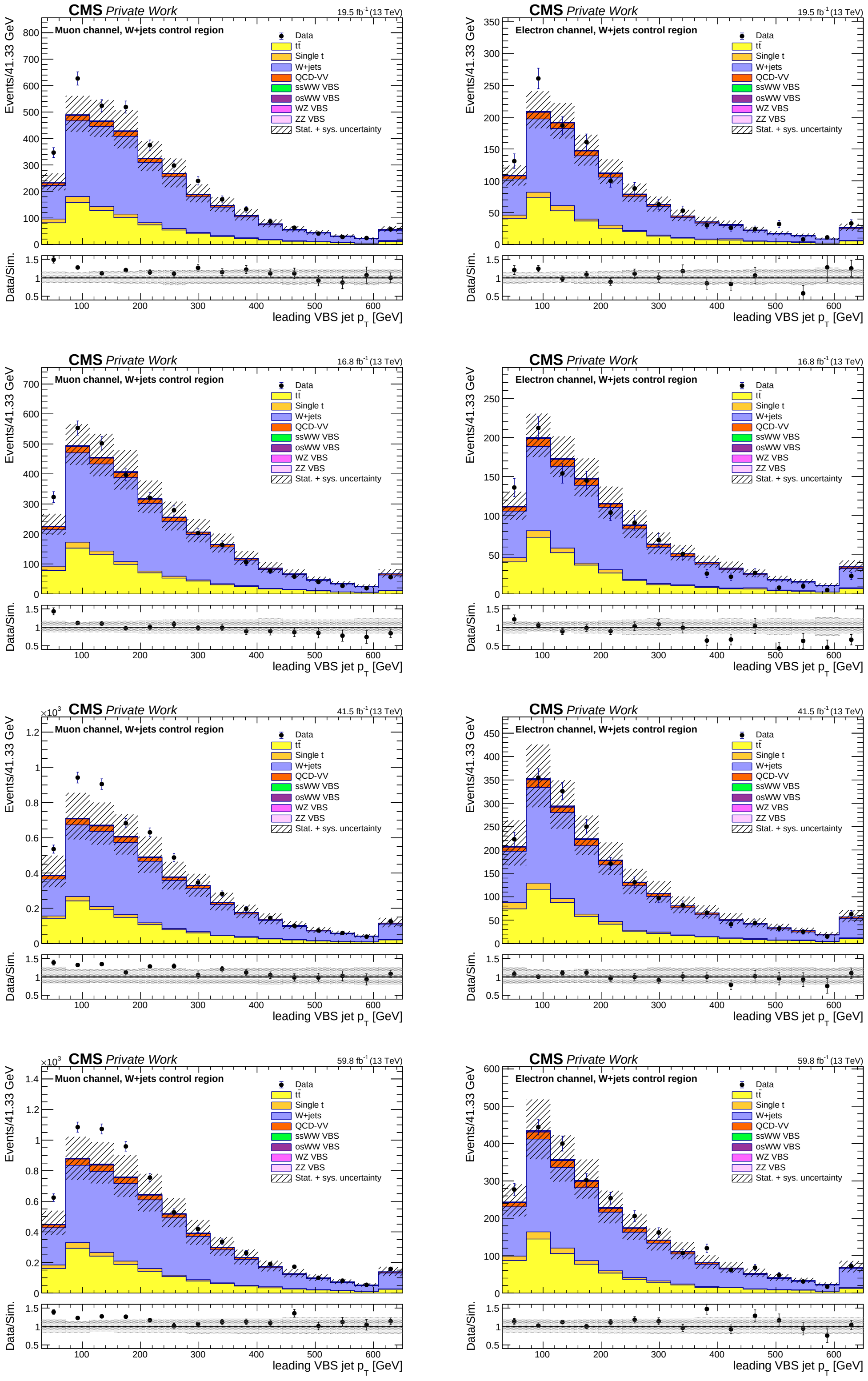


Figure K.24: Data and simulation comparison of the **leading VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

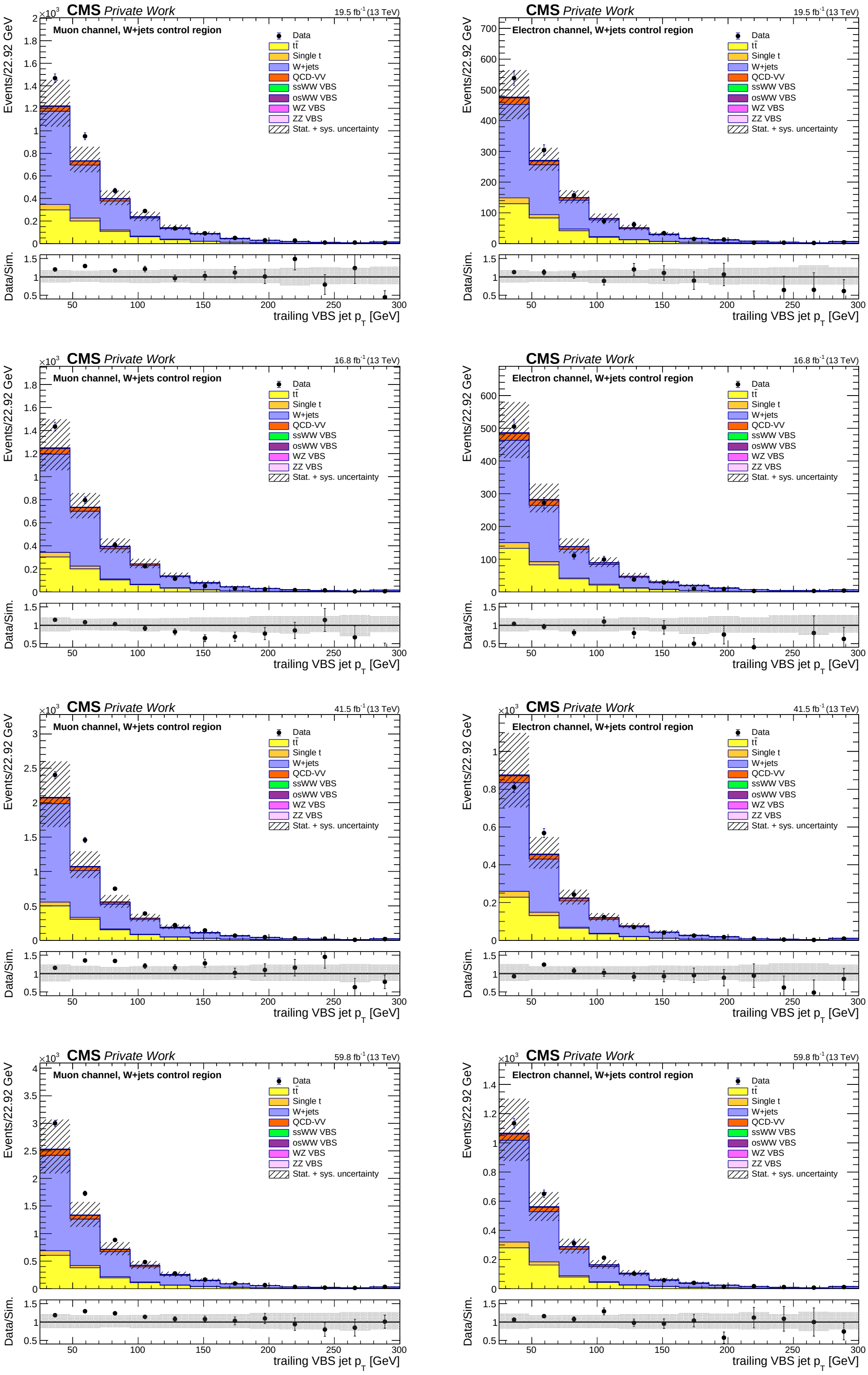


Figure K.25: Data and simulation comparison of the **trailing VBS jet transverse momentum** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

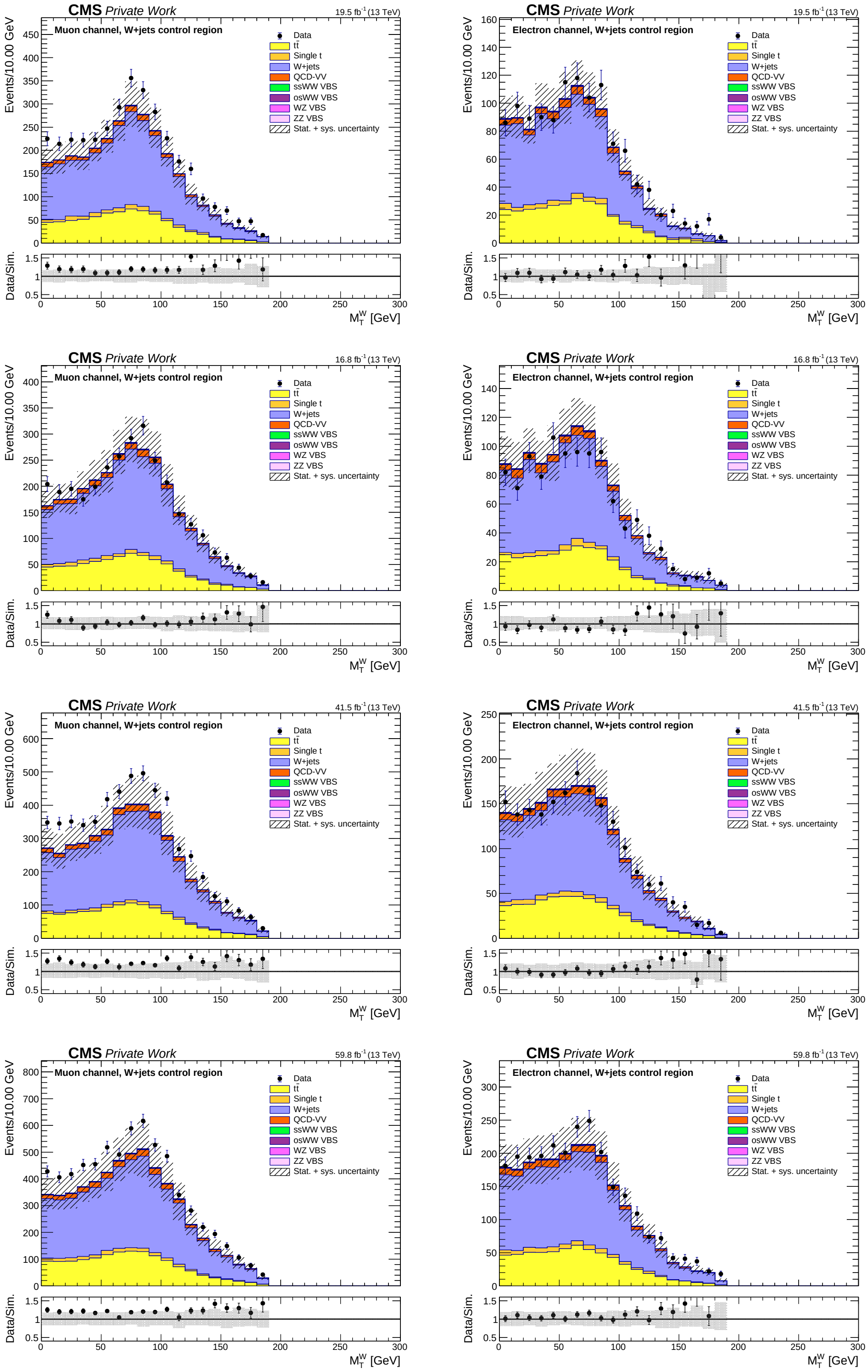


Figure K.26: Data and simulation comparison of the **transverse mass of leptonically decaying W boson** ($M_T{}^W$) in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

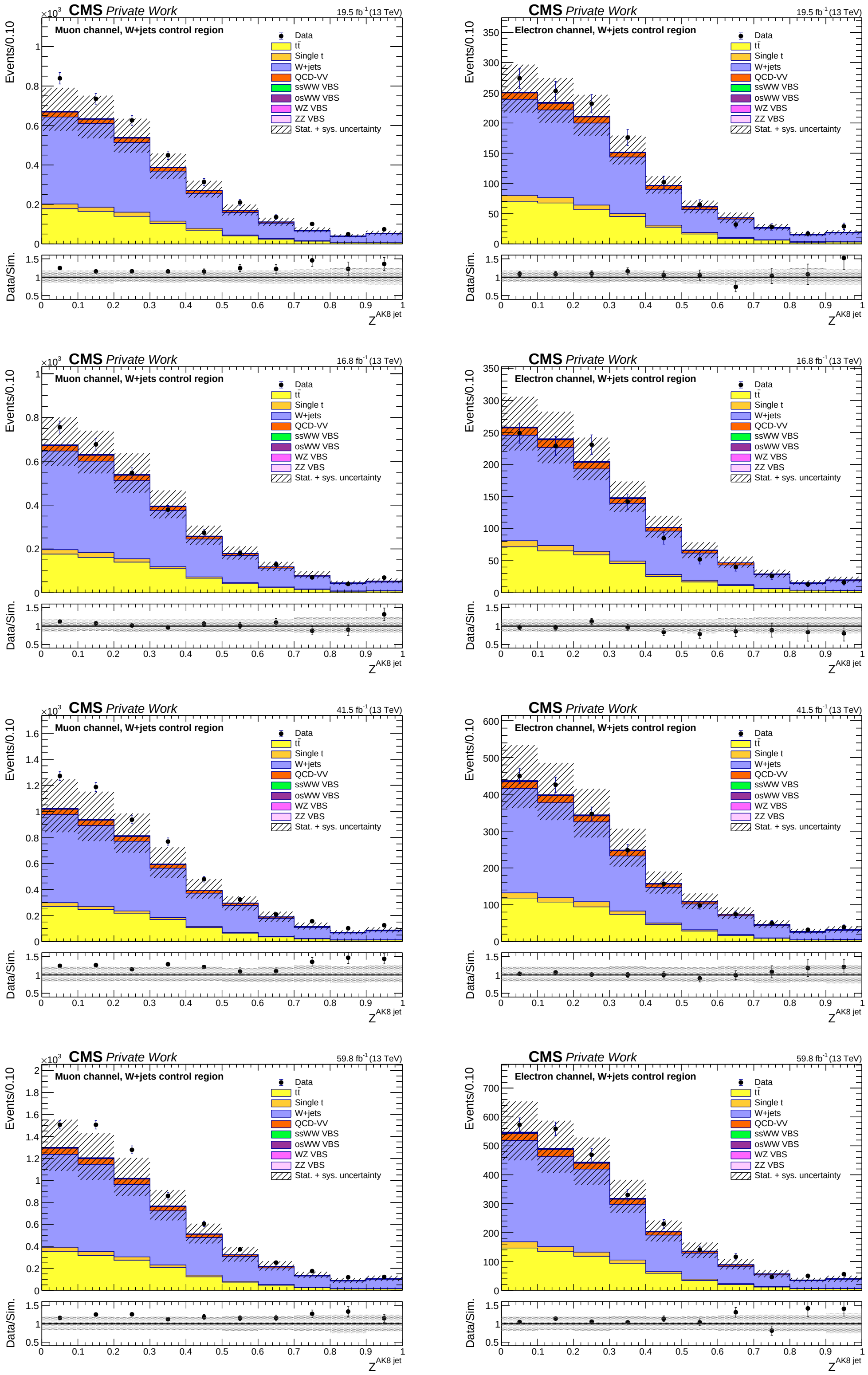


Figure K.27: Data and simulation comparison of **Zeppenfeld variable of the AK8 jet** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

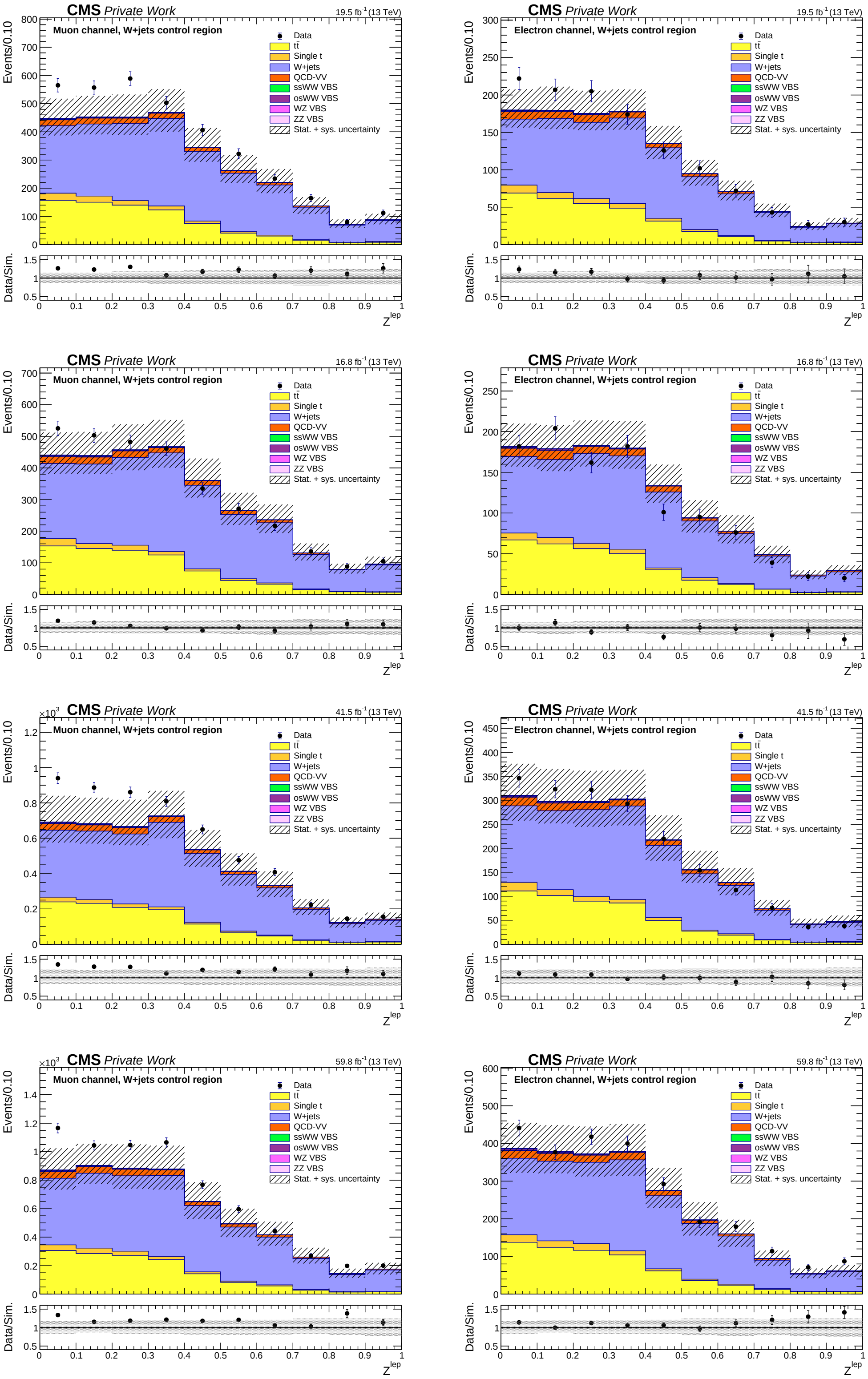


Figure K.28: Data and simulation comparison of **Zeppenfeld variable of the lepton** in the muon channel (left) and the electron channel (right) in the W+jets control region for different years: 2016 preVFP, 2016 postVFP, 2017, and 2018 (from top to bottom).

Bibliography

[1] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[2] 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 (Sept. 2012), pp. 30–61. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.021. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.021.

[3] A. Zannoni. "On the Quantization of the Monoatomic Ideal Gas". 1999. arXiv: cond-mat/9912229 [cond-mat.stat-mech]. URL: https://arxiv.org/abs/cond-mat/9912229.

[4] P. A. M. Dirac. "Quantum theory of emission and absorption of radiation". In: *Proc. Roy. Soc. Lond. A* 114 (1927), p. 243. DOI: 10.1098/rspa.1927.0039.

[5] S. Bose, O. Theimer, and B. Ram. "The beginning of quantum statistics: A translation of "Planck's law and the light quantum hypothesis"". In: *American Journal of Physics* 44.11 (Nov. 1976), pp. 1056–1057. ISSN: 0002-9505. DOI: 10.1119/1.10584. URL: https://doi.org/10.1119/1.10584.

[6] A. Einstein. "Quantentheorie des einatomigen idealen Gases". In: *Albert Einstein: Akademie-Vorträge*. John Wiley & Sons, Ltd, 2005, pp. 237–244. ISBN: 9783527608959. DOI: https://doi.org/10.1002/3527608958.ch27. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/3527608958.ch27. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/3527608958.ch27.

[7] I. G. Kaplan. "Pauli Exclusion Principle and its theoretical foundation". 2019. arXiv: 1902.00499 [quant-ph]. URL: https://arxiv.org/abs/1902.00499.

[8] B. C. Hall. "An Elementary Introduction to Groups and Representations". 2000. arXiv: math-ph/0005032 [math-ph]. URL: https://arxiv.org/abs/math-ph/0005032.

[9] Y. F. e. a. Super-Kamiokande Collaboration. "Evidence for Oscillation of Atmospheric Neutrinos". In: *Physical Review Letters* 81.8 (Aug. 1998), pp. 1562–1567. ISSN: 1079-7114. DOI: 10.1103/physrevlett.81.1562. URL: http://dx.doi.org/10.1103/PhysRevLett.81.1562.

[10] Q. R. Ahmad et al. "Measurement of the Rate of $\nu_e + d \rightarrow p + p + e^-$ Interactions Produced by ^8B Solar Neutrinos at the Sudbury Neutrino Observatory". In: *Phys. Rev. Lett.* 87 (7 2001), p. 071301. DOI: 10.1103/PhysRevLett.87.071301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.87.071301.

[11] P. D. G. Collaboration. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2022.8 (Aug. 2022), p. 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptac097. eprint: https://academic.oup.com/ptep/article-pdf/2022/8/083C01/49175539/ptac097.pdf. URL: https://doi.org/10.1093/ptep/ptac097.

[12] H. Suganuma. "Quantum Chromodynamics, Quark Confinement, and Chiral Symmetry Breaking: A Bridge Between Elementary Particle Physics and Nuclear Physics". In: *Handbook of Nuclear Physics*. Ed. by I. Tanihata, H. Toki, and T. Kajino. Singapore: Springer Nature Singapore, 2020, pp. 1–48. ISBN: 978-981-15-8818-1. DOI: 10.1007/978-981-15-8818-1_22-1. URL: https://doi.org/10.1007/978-981-15-8818-1_22-1.

[13] D. J. Gross. "Asymptotic Freedom and QCD–a Historical Perspective". In: *Nuclear Physics B - Proceedings Supplements* 135 (2004). Loops and Legs in Quantum Field Theory. Proceedings of the 7th DESY Workshop on Elementary Particle Theory, pp. 193–211. ISSN: 0920-5632. DOI: https://doi.org/10.1016/j.nuclphysbps.2004.09.049. URL: https://www.sciencedirect.com/science/article/pii/S0920563204004104.

[14] R. H. Dalitz. "The $\tau - \theta$ puzzle". In: *AIP Conf. Proc.* 300 (1994). Ed. by A. K. Mann and D. B. Cline, pp. 141–158. DOI: 10.1063/1.45424.

[15] C. S. Wu et al. "Experimental Test of Parity Conservation in Beta Decay". In: *Phys. Rev.* 105 (4 Feb. 1957), pp. 1413–1415. DOI: 10.1103/PhysRev.105.1413. URL: https://link.aps.org/doi/10.1103/PhysRev.105.1413.

[16] F. J. Hasert et al. "Observation of Neutrino Like Interactions Without Muon Or Electron in the Gargamelle Neutrino Experiment". In: *Phys. Lett. B* 46 (1973), pp. 138–140. DOI: 10.1016/0370-2693(73)90499-1.

[17] P. W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[18] 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. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[19] J. Ellis. "Higgs Physics". In: *2013 European School of High-Energy Physics*. 2015, pp. 117–168. DOI: 10.5170/CERN-2015-004.117. arXiv: 1312.5672 [hep-ph].

[20] J. Goldstone, A. Salam, and S. Weinberg. "Broken Symmetries". In: *Phys. Rev.* 127 (3 Aug. 1962), pp. 965–970. DOI: 10.1103/PhysRev.127.965. URL: https://link.aps.org/doi/10.1103/PhysRev.127.965.

[21] M. Gonzalez-Garcia and M. Maltoni. “Phenomenology with massive neutrinos”. In: *Physics Reports* 460.1–3 (Apr. 2008), pp. 1–129. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2007.12.004. URL: http://dx.doi.org/10.1016/j.physrep.2007.12.004.

[22] A. de Gouvêa, D. Hernández, and T. M. P. Tait. “Criteria for natural hierarchies”. In: *Physical Review D* 89.11 (June 2014). ISSN: 1550-2368. DOI: 10.1103/physrevd.89.115005. URL: http://dx.doi.org/10.1103/PhysRevD.89.115005.

[23] G. Branco et al. “Theory and phenomenology of two-Higgs-doublet models”. In: *Physics Reports* 516.1–2 (July 2012), pp. 1–102. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2012.02.002. URL: http://dx.doi.org/10.1016/j.physrep.2012.02.002.

[24] H. Georgi and M. Machacek. “Doubly charged Higgs bosons”. In: *Nuclear Physics B* 262.3 (1985), pp. 463–477. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(85)90325-6. URL: https://www.sciencedirect.com/science/article/pii/0550321385903256.

[25] A. Ballestrero et al. “Precise predictions for same-sign W-boson scattering at the LHC”. In: *The European Physical Journal C* 78.8 (Aug. 2018). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-018-6136-y. URL: http://dx.doi.org/10.1140/epjc/s10052-018-6136-y.

[26] M. Szleper. “The Higgs boson and the physics of WW scattering before and after Higgs discovery”. 2015. arXiv: 1412.8367 [hep-ph]. URL: https://arxiv.org/abs/1412.8367.

[27] B. W. Lee, C. Quigg, and H. B. Thacker. “Weak interactions at very high energies: The role of the Higgs-boson mass”. In: *Phys. Rev. D* 16 (5 Sept. 1977), pp. 1519–1531. DOI: 10.1103/PhysRevD.16.1519. URL: https://link.aps.org/doi/10.1103/PhysRevD.16.1519.

[28] M. A. Iqbal. “Looking for new physics: Search for anomalous gauge couplings in WW and WZ production in lepton + jet events in proton-proton collisions at 13 TeV with the CMS experiment”. PhD thesis. Karlsruher Institut für Technologie (KIT), 2020. 193 pp. DOI: 10.5445/IR/1000121030.

[29] ATLAS Collaboration. “Observation of electroweak production of a same-sign W boson pair in association with two jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”. In: *Phys. Rev. Lett.* 123.16 (2019), p. 161801. DOI: 10.1103/PhysRevLett.123.161801. arXiv: 1906.03203 [hep-ex].

[30] A. M. Sirunyan et al. “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), p. 081801. DOI: 10.1103/PhysRevLett.120.081801. arXiv: 1709.05822 [hep-ex].

[31] CMS Collaboration. “Measurement of same sign WW VBS processes at CMS with one hadronic tau in the final state”. In: *PoS* EPS-HEP2023 (2024), p. 322. DOI: 10.22323/1.449.0322.

[32] CMS Collaboration. "Evidence for WW/WZ vector boson scattering in the decay channel $\ell\nu$qq produced in association with two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Physics Letters B* 834 (Nov. 2022), p. 137438. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2022.137438. URL: http://dx.doi.org/10.1016/j.physletb.2022.137438.

[33] CERN. "Linear accelerator 4". 2020. URL: https://home.cern/science/accelerators/linear-accelerator-4.

[34] CERN. "Linear accelerator 2". 1978 - 2018. URL: https://home.cern/science/accelerators/linear-accelerator-2.

[35] CERN. "Radiofrequency cavities". URL: https://cds.cern.ch/record/1997424.

[36] CERN. "The Proton Synchrotron Booster". URL: https://home.cern/science/accelerators/proton-synchrotron-booster.

[37] CERN. "The Proton Synchrotron". URL: https://home.cern/science/accelerators/proton-synchrotron.

[38] CERN. "The Super Proton Synchrotron". URL: https://home.cern/science/accelerators/super-proton-synchrotron.

[39] G. Arnison et al. "Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$ GeV". In: *Phys. Lett. B* 122 (1983), pp. 103–116. DOI: 10.1016/0370-2693(83)91177-2.

[40] G. Arnison et al. "Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c**2 at the CERN SPS Collider". In: *Phys. Lett. B* 126 (1983), pp. 398–410. DOI: 10.1016/0370-2693(83)90188-0.

[41] P. Bagnaia et al. "Evidence for $Z^0 \to e^+e^-$ at the CERN $\bar{p}p$ Collider". In: *Phys. Lett. B* 129 (1983), pp. 130–140. DOI: 10.1016/0370-2693(83)90744-X.

[42] M. Banner et al. "Observation of Single Isolated Electrons of High Transverse Momentum in Events with Missing Transverse Energy at the CERN anti-p p Collider". In: *Phys. Lett. B* 122 (1983), pp. 476–485. DOI: 10.1016/0370-2693(83)91605-2.

[43] CERN. "The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022". 2022. URL: https://cds.cern.ch/record/2800984.

[44] CERN. "Cross section of an LHC dipole in the tunnel." 2011. URL: https://cds.cern.ch/record/1365795.

[45] CMS Collaboration. "Public CMS Luminosity Information". URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[46] CERN. "ATLAS experiment at the LHC". URL: https://home.cern/science/experiments/atlas.

[47] CERN. “CMS experiment at the LHC”. URL: https://home.cern/science/experiments/cms.

[48] CERN. “ALICE experiment at the LHC”. URL: https://home.cern/science/experiments/alice.

[49] CERN. “LHCb experiment at the LHC”. URL: https://home.cern/science/experiments/lhcb.

[50] CERN. “TOTEM experiment at the LHC”. URL: https://home.cern/science/experiments/totem.

[51] CERN. “LHCf experiment at the LHC”. URL: https://home.cern/fr/science/experiments/lhcf.

[52] CERN. “MoEDAL-MAPP experiment at the LHC”. URL: https://home.cern/fr/science/experiments/moedal-mapp.

[53] CERN. “FASER experiment at the LHC”. URL: https://home.cern/science/experiments/faser.

[54] CERN. “SND@LHC experiment at the LHC”. URL: https://home.cern/science/experiments/sndlhc.

[55] CMS Collaboration. “The CMS Experiment at the CERN LHC”. In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[56] T. Sakuma and T. McCauley. “Detector and Event Visualization with SketchUp at the CMS Experiment”. In: *J. Phys. Conf. Ser.* 513 (2014). Ed. by D. L. Groep and D. Bonacorsi, p. 022032. DOI: 10.1088/1742-6596/513/2/022032. arXiv: 1311.4942 [physics.ins-det].

[57] CMS Collaboration. “The CMS Phase-1 Pixel Detector Upgrade”. Tech. rep. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2745805.

[58] CMS Collaboration. “The CMS silicon strip detector - mechanical structure and alignment system”. In: *Nucl. Instrum. Meth. A* 511 (2003). Ed. by S. L. Olsen and D. Bortoletto, pp. 52–57. DOI: 10.1016/S0168-9002(03)01750-9.

[59] CMS Collaboration. “CMS tracker detector performance results”. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/DPGResultsTRK.

[60] CMS Collaboration. “Simulation of the Silicon Strip Tracker pre-amplifier in early 2016 data”. In: (2020). URL: https://cds.cern.ch/record/2740688.

[61] CMS Collaboration. “Performance of the CMS electromagnetic calorimeter and its role in the hunt for the Higgs boson in the two-photon channel”. In: *Journal of Physics: Conference Series* 455.1 (Aug. 2013), p. 012028. DOI: 10.1088/1742-6596/455/1/012028. URL: https://dx.doi.org/10.1088/1742-6596/455/1/012028.

[62] CMS Collaboration. “CMS Physics: Technical Design Report Volume 1: Detector Performance and Software”. Technical design report. CMS. There is an error on cover due to a technical problem for some items. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/922757.

[63] CMS Collaboration. "Energy resolution of the barrel of the CMS electromagnetic calorimeter". In: *JINST* 2 (2007), P04004. DOI: 10.1088/1748-0221/2/04/P04004.

[64] CMS Collaboration. "The CMS ECAL performance with examples". Tech. rep. Geneva: CERN, 2014. DOI: 10.1088/1748-0221/9/02/C02008. URL: https://cds.cern.ch/record/1632384.

[65] CMS Collaboration. "The CMS hadron calorimeter project: Technical Design Report". Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/357153.

[66] CMS Collaboration. "Measurement of the Pion Energy Response and Resolution in the CMS HCAL Test Beam 2002 Experiment". Tech. rep. Geneva: CERN, 2004. URL: https://cds.cern.ch/record/800406.

[67] CMS Collaboration. "Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays". In: *Journal of Instrumentation* 5.03 (Mar. 2010), T03021. DOI: 10.1088/1748-0221/5/03/T03021. URL: https://dx.doi.org/10.1088/1748-0221/5/03/T03021.

[68] CMS Collaboration. "The CMS muon project: Technical Design Report". Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/343814.

[69] CMS Collaboration. "The performance of the CMS muon detector in proton-proton collisions at s = 7 TeV at the LHC". In: *Journal of Instrumentation* 8.11 (Nov. 2013), P11002. DOI: 10.1088/1748-0221/8/11/P11002. URL: https://dx.doi.org/10.1088/1748-0221/8/11/P11002.

[70] CMS Collaboration. "CMS data processing workflows during an extended cosmic ray run". In: *Journal of Instrumentation* 5.03 (Mar. 2010), T03006. DOI: 10.1088/1748-0221/5/03/T03006. URL: https://dx.doi.org/10.1088/1748-0221/5/03/T03006.

[71] CMS Collaboration. "Data Scouting and Data Parking with the CMS High level Trigger". In: *PoS* EPS-HEP2019 (2020), p. 139. DOI: 10.22323/1.364.0139.

[72] K. Bos et al. "LHC computing Grid: Technical Design Report. Version 1.06 (20 Jun 2005)". Technical design report. LCG. Geneva: CERN, 2005. URL: https://cds.cern.ch/record/840543.

[73] CMS Collaboration. "CMS computing model". July 2018. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookComputingModel.

[74] E. Martelli and S. Stancu. "LHCOPN and LHCONE: Status and Future Evolution". In: *Journal of Physics: Conference Series* 664.5 (Dec. 2015), p. 052025. DOI: 10.1088/1742-6596/664/5/052025. URL: https://dx.doi.org/10.1088/1742-6596/664/5/052025.

[75] CMS Collaboration. “Description and performance of track and primary-vertex reconstruction with the CMS tracker”. In: *Journal of Instrumentation* 9.10 (Oct. 2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. URL: https://dx.doi.org/10.1088/1748-0221/9/10/P10009.

[76] P. Billoir. “Progressive track recognition with a Kalman-like fitting procedure”. In: *Computer Physics Communications* 57.1 (1989), pp. 390–394. ISSN: 0010-4655. DOI: https://doi.org/10.1016/0010-4655(89)90249-X. URL: https://www.sciencedirect.com/science/article/pii/001046558990249X.

[77] P. Billoir and S. Qian. “Simultaneous pattern recognition and track fitting by the Kalman filtering method”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 294.1 (1990), pp. 219–228. ISSN: 0168-9002. DOI: https://doi.org/10.1016/0168-9002(90)91835-Y. URL: https://www.sciencedirect.com/science/article/pii/016890029091835Y.

[78] K. Rose. “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2210–2239. DOI: 10.1109/5.726788.

[79] W. Waltenberger, R. Frühwirth, and P. Vanlaer. “Adaptive vertex fitting”. In: *Journal of Physics G: Nuclear and Particle Physics* 34.12 (Nov. 2007), N343. DOI: 10.1088/0954-3899/34/12/N01. URL: https://dx.doi.org/10.1088/0954-3899/34/12/N01.

[80] CMS Collaboration. “Particle-flow reconstruction and global event description with the CMS detector”. In: *Journal of Instrumentation* 12.10 (Oct. 2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. URL: https://dx.doi.org/10.1088/1748-0221/12/10/P10003.

[81] CMS Collaboration. “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at s=13 TeV”. In: *Journal of Instrumentation* 13.06 (June 2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. URL: https://dx.doi.org/10.1088/1748-0221/13/06/P06015.

[82] CMS Collaboration. “Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC”. Tech. rep. Geneva: CERN, 2005. URL: https://cds.cern.ch/record/815410.

[83] CMS Collaboration. “Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC”. In: *Journal of Instrumentation* 16.05 (May 2021), P05014. DOI: 10.1088/1748-0221/16/05/P05014. URL: https://dx.doi.org/10.1088/1748-0221/16/05/P05014.

[84] M. Cacciari, G. P. Salam, and G. Soyez. “The anti-kt jet clustering algorithm”. In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. URL: https://dx.doi.org/10.1088/1126-6708/2008/04/063.

[85] S. D. Ellis and D. E. Soper. “Successive combination jet algorithm for hadron collisions”. In: *Physical Review D* 48.7 (Oct. 1993), pp. 3160–3166. ISSN: 0556-2821. DOI: 10.1103/physrevd.48.3160. URL: http://dx.doi.org/10.1103/PhysRevD.48.3160.

[86] CMS Collaboration. “A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging”. Tech. rep. Geneva: CERN, 2009. URL: https://cds.cern.ch/record/1194489.

[87] M. Cacciari, G. P. Salam, and G. Soyez. “FastJet user manual: (for version 3.0.2)”. In: *The European Physical Journal C* 72.3 (Mar. 2012). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-012-1896-2. URL: http://dx.doi.org/10.1140/epjc/s10052-012-1896-2.

[88] D. Bertolini et al. “Pileup per particle identification”. In: *Journal of High Energy Physics* 2014.10 (Oct. 2014). ISSN: 1029-8479. DOI: 10.1007/jhep10(2014)059. URL: http://dx.doi.org/10.1007/JHEP10(2014)059.

[89] 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. URL: https://dx.doi.org/10.1088/1748-0221/15/09/P09018.

[90] CMS Collaboration. “Pileup-per-particle identification: optimisation for Run 2 Legacy and beyond”. Tech. rep. CMS-DP-2021-001. CMS Detector Performance Summary. CERN, 2021. URL: https://cds.cern.ch/record/2751563.

[91] CMS Collaboration. “Jet ID at 13 TeV Ultra Legacy (UL)”. Accessed: 2024-10-28. 2024. URL: https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetID13TeVUL.

[92] A. J. Larkoski et al. “Soft drop”. In: *Journal of High Energy Physics* 2014.5 (May 2014). ISSN: 1029-8479. DOI: 10.1007/jhep05(2014)146. URL: http://dx.doi.org/10.1007/JHEP05(2014)146.

[93] H. Qu and L. Gouskos. “Jet tagging via particle clouds”. In: *Physical Review D* 101.5 (Mar. 2020). ISSN: 2470-0029. DOI: 10.1103/physrevd.101.056019. URL: http://dx.doi.org/10.1103/PhysRevD.101.056019.

[94] H. Qu, C. Li, and S. Qian. “Particle Transformer for Jet Tagging”. 2024. arXiv: 2202.03772 [hep-ph]. URL: https://arxiv.org/abs/2202.03772.

[95] J. Thaler and K. Van Tilburg. “Identifying boosted objects with N-subjettiness”. In: *Journal of High Energy Physics* 2011.3 (Mar. 2011). ISSN: 1029-8479. DOI: 10.1007/jhep03(2011)015. URL: http://dx.doi.org/10.1007/JHEP03(2011)015.

[96] E. Bols et al. “Jet flavour classification using DeepJet”. In: *Journal of Instrumentation* 15.12 (Dec. 2020), P12012–P12012. ISSN: 1748-0221. DOI: 10.1088/1748-0221/15/12/p12012. URL: http://dx.doi.org/10.1088/1748-0221/15/12/P12012.

[97] CMS Collaboration. "CMS b-tagging & Vertexing POG". https://btv-wiki.docs.cern.ch. Accessed: 2024-11-14. 2018.

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

[99] V. Gribov and L. Lipatov. "Deep inelastic electron scattering in perturbation theory". In: *Physics Letters B* 37.1 (1971), pp. 78–80. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(71)90576-4. URL: https://www.sciencedirect.com/science/article/pii/0370269371905764.

[100] G. Altarelli and G. Parisi. "Asymptotic freedom in parton language". In: *Nuclear Physics B* 126.2 (1977), pp. 298–318. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(77)90384-4. URL: https://www.sciencedirect.com/science/article/pii/0550321377903844.

[101] H. D. Politzer. "Asymptotic Freedom: An Approach to Strong Interactions". In: *Phys. Rept.* 14 (1974), pp. 129–180. DOI: 10.1016/0370-1573(74)90014-3.

[102] R. D. Ball et al. "Parton distributions for the LHC run II". In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep04(2015)040. URL: http://dx.doi.org/10.1007/JHEP04(2015)040.

[103] R. D. Ball et al. "Parton distributions from high-precision collider data: NNPDF Collaboration". In: *The European Physical Journal C* 77.10 (Oct. 2017). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-017-5199-5. URL: http://dx.doi.org/10.1140/epjc/s10052-017-5199-5.

[104] V. V. Sudakov. "Vertex parts at very high-energies in quantum electrodynamics". In: *Sov. Phys. JETP* 3 (1956), pp. 65–71.

[105] M. L. Mangano et al. "Matching matrix elements and shower evolution for top-pair production in hadronic collisions". In: *Journal of High Energy Physics* 2007.01 (Jan. 2007), pp. 013–013. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2007/01/013. URL: http://dx.doi.org/10.1088/1126-6708/2007/01/013.

[106] R. Frederix and S. Frixione. "Merging meets matching in MC@NLO". In: *Journal of High Energy Physics* 2012.12 (Dec. 2012). ISSN: 1029-8479. DOI: 10.1007/jhep12(2012)061. URL: http://dx.doi.org/10.1007/JHEP12(2012)061.

[107] B. Andersson et al. "Parton Fragmentation and String Dynamics". In: *Phys. Rept.* 97 (1983), pp. 31–145. DOI: 10.1016/0370-1573(83)90080-7.

[108] B. Andersson. "The Lund Model". Vol. 7. Cambridge University Press, 1998. ISBN: 978-1-009-40129-6, 978-1-009-40125-8, 978-1-009-40128-9, 978-0-521-01734-3, 978-0-521-42094-5, 978-0-511-88149-7. DOI: 10.1017/9781009401296.

[109] D. Amati and G. Veneziano. "Preconfinement as a Property of Perturbative QCD". In: *Phys. Lett. B* 83 (1979), pp. 87–92. DOI: 10.1016/0370-2693(79)90896-7.

[110] CMS Collaboration. “Extraction and validation of a new set of CMS Pythia8 tunes from underlying-event measurements”. In: *Eur. Phys. J. C* 80.1 (2020), p. 4. DOI: 10.1140/epjc/s10052-019-7499-4.

[111] S. Agostinelli et al. “GEANT4 - A Simulation Toolkit”. In: *Nucl. Instrum. Meth. A* 506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[112] J. Allison et al. “Geant4 Developments and Applications”. In: *IEEE Transactions on Nuclear Science* 53 (Feb. 2006), pp. 270–278. DOI: 10.1109/TNS.2006.869826.

[113] J. de Favereau et al. “DELPHES 3: a modular framework for fast simulation of a generic collider experiment”. In: *Journal of High Energy Physics* 2014.2 (Feb. 2014). ISSN: 1029-8479. DOI: 10.1007/jhep02(2014)057. URL: http://dx.doi.org/10.1007/JHEP02(2014)057.

[114] J. Alwall et al. “A standard format for Les Houches Event Files”. In: *Computer Physics Communications* 176.4 (Feb. 2007), pp. 300–304. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2006.11.010. URL: http://dx.doi.org/10.1016/j.cpc.2006.11.010.

[115] J. Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. In: *Journal of High Energy Physics* 2014.7 (July 2014). ISSN: 1029-8479. DOI: 10.1007/jhep07(2014)079. URL: http://dx.doi.org/10.1007/JHEP07(2014)079.

[116] J. Alwall et al. “MadGraph 5: going beyond”. In: *Journal of High Energy Physics* 2011.6 (June 2011). ISSN: 1029-8479. DOI: 10.1007/jhep06(2011)128. URL: http://dx.doi.org/10.1007/JHEP06(2011)128.

[117] S. Frixione and B. R. Webber. “Matching NLO QCD computations and parton shower simulations”. In: *Journal of High Energy Physics* 2002.06 (July 2002), p. 029. DOI: 10.1088/1126-6708/2002/06/029. URL: https://dx.doi.org/10.1088/1126-6708/2002/06/029.

[118] P. Artoisenet et al. “Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations”. In: *JHEP* 03 (2013), p. 015. DOI: 10.1007/JHEP03(2013)015. URL: https://doi.org/10.1007/JHEP03(2013)015.

[119] S. Alioli et al. “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”. In: *Journal of High Energy Physics* 2010.6 (June 2010). ISSN: 1029-8479. DOI: 10.1007/jhep06(2010)043. URL: http://dx.doi.org/10.1007/JHEP06(2010)043.

[120] S. Frixione, P. Nason, and C. Oleari. “Matching NLO QCD computations with parton shower simulations: the POWHEG method”. In: *Journal of High Energy Physics* 2007.11 (Nov. 2007), pp. 070–070. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2007/11/070. URL: http://dx.doi.org/10.1088/1126-6708/2007/11/070.

[121] T. Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (June 2015), pp. 159–177. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2015.01.024. URL: http://dx.doi.org/10.1016/j.cpc.2015.01.024.

[122] CMS Collaboration. "Utilities for Accessing Pileup Information for Data". https://twiki.cern.ch/twiki/bin/view/CMS/PileupJSONFileforData?rev=28#Pileup_JSON_Files_For_Run_II. Accessed: 2025-01-07. 2024.

[123] CMS Collaboration. "Reweighting recipe to emulate Level 1 ECAL and Muon prefiring". https://twiki.cern.ch/twiki/bin/viewauth/CMS/L1PrefiringWeightRecipe. Accessed: 2025-01-31. 2024.

[124] CMS Collaboration. "Determination of jet energy calibration and transverse momentum resolution in CMS". In: *Journal of Instrumentation* 6.11 (Nov. 2011), P11002. DOI: 10.1088/1748-0221/6/11/P11002. URL: https://dx.doi.org/10.1088/1748-0221/6/11/P11002.

[125] CMS Collaboration. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *Journal of Instrumentation* 12.02 (Feb. 2017), P02014–P02014. ISSN: 1748-0221. DOI: 10.1088/1748-0221/12/02/p02014. URL: http://dx.doi.org/10.1088/1748-0221/12/02/P02014.

[126] CMS Collaboration. "Jet Energy Resolution". https://twiki.cern.ch/twiki/bin/view/CMS/JetResolution. Accessed: 2025-01-31. 2024.

[127] E. Bols et al. "Jet Flavour Classification Using DeepJet". In: *JINST* 15.12 (2020), P12012. DOI: 10.1088/1748-0221/15/12/P12012. arXiv: 2008.10519 [hep-ex].

[128] CMS Collaboration. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *Journal of Instrumentation* 13.05 (May 2018), P05011–P05011. ISSN: 1748-0221. DOI: 10.1088/1748-0221/13/05/p05011. URL: http://dx.doi.org/10.1088/1748-0221/13/05/P05011.

[129] CMS Collaboration. "Methods to apply b-tagging efficiency scale factors". https://twiki.cern.ch/twiki/bin/view/CMS/BTagSFMethods#1a_Event_reweighting_using_scale. Accessed: 2025-01-31. 2024.

[130] CMS Collaboration. "W/Z-tagging of Jets". https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetWtagging#2018_scale_factors_and_correctio. Accessed: 2025-01-31. 2024.

[131] CMS Collaboration. "A novel approach for discriminating hadronically decaying $\mathbf{W}^+$, $\mathbf{W}^-$, and $\mathbf{Z}$ bosons in the CMS experiment". In: (2024). URL: https://cds.cern.ch/record/2904357.

[132] CMS Collaboration. "Summary table of samples produced for the 1 Billion campaign, with 25ns bunch-crossing". 2022. URL: https://twiki.cern.ch/twiki/bin/view/CMS/SummaryTable1G25ns.

[133] M. Czakon and A. Mitov. “NNLO+NNLL top-quark-pair cross sections. ATLAS-CMS recommended predictions for top-quark-pair cross sections using the Top++v2.0 program”. 2015. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO.

[134] M. Aliev et al. “HATHOR - HAdronic Top and Heavy quarks crOss section calculatoR”. In: *Computer Physics Communications* 182 (2011). DOI: 10.1016/j.cpc.2010.12.040.

[135] S. Marzani, G. Soyez, and M. Spannowsky. “Looking Inside Jets: An Introduction to Jet Substructure and Boosted-object Phenomenology”. Springer International Publishing, 2019. ISBN: 9783030157098. DOI: 10.1007/978-3-030-15709-8. URL: http://dx.doi.org/10.1007/978-3-030-15709-8.

[136] CMS Collaboration. “Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques”. In: (2020). URL: http://cds.cern.ch/record/2707946.

[137] Y. Wang et al. “Dynamic Graph CNN for Learning on Point Clouds”. In: *ACM Trans. Graph.* 38.5 (Oct. 2019). ISSN: 0730-0301. DOI: 10.1145/3326362. URL: https://doi.org/10.1145/3326362.

[138] Serkin, Leonid. “Treatment of top-quark backgrounds in extreme phase spaces: the ”top p_T reweighting” and novel data-driven estimations in ATLAS and CMS”. Tech. rep. Proceeding for 13th International Workshop on Top Quark Physics. Geneva: CERN, 2021. arXiv: 2105.03977. URL: https://cds.cern.ch/record/2765412.

[139] CMS Collaboration. “The CMS Statistical Analysis and Combination Tool: Combine”. In: *Computing and Software for Big Science* 8.1 (Nov. 2024). ISSN: 2510-2044. DOI: 10.1007/s41781-024-00121-4. URL: http://dx.doi.org/10.1007/s41781-024-00121-4.

[140] S. S. Wilks. “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses”. In: *Annals Math. Statist.* 9.1 (1938), pp. 60–62. DOI: 10.1214/aoms/1177732360.

[141] J. Butterworth et al. “PDF4LHC recommendations for LHC Run II”. In: *Journal of Physics G: Nuclear and Particle Physics* 43.2 (Jan. 2016), p. 023001. ISSN: 1361-6471. DOI: 10.1088/0954-3899/43/2/023001. URL: http://dx.doi.org/10.1088/0954-3899/43/2/023001.

[142] CMS Collaboration. “Btagging Scale Factor Uncertainties and Correlations Across Years”. https://btv-wiki.docs.cern.ch/PerformanceCalibration/SFUncertaintiesAndCorrelations. Accessed: 2025-01-07. 2024.

[143] CMS Collaboration. “The modeling of the top quark pT”. https://twiki.cern.ch/twiki/bin/view/CMS/TopPtReweighting. Accessed: 2025-01-07. 2024.

[144] R. J. Barlow and C. Beeston. “Fitting using finite Monte Carlo samples”. In: *Comput. Phys. Commun.* 77 (1993), pp. 219–228. DOI: 10.1016/0010-4655(93)90005-W.

[145] G. Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (Feb. 2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0. URL: http://dx.doi.org/10.1140/epjc/s10052-011-1554-0.

[146] R. D. Cousins. "Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms". 2013. URL: https://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf.

Acknowledgements

The journey of earning a doctorate has been transformative in many ways. During my first year, I welcomed my beautiful daughter, **Zirwa Ansar**, into the world. She changed my life completely, filling it with immeasurable love and joy. However, raising her during the challenges of the pandemic while managing my research was far from easy. There were moments when I felt it was impossible to succeed. Yet, looking back now, I see what once seemed unattainable become a reality. This accomplishment is not mine alone; it is the result of the unwavering support of many wonderful people to whom I am deeply grateful.

First and foremost, I sincerely thank my supervisor, **Prof. Thomas Müller**, for being an incredible mentor. Your consistent support and belief in me made this journey possible. Thank you for giving me the opportunity to work with you at the Institute of Experimental Particle Physics (ETP) at KIT, for your encouragement, and for standing by me through every challenge, especially during the time of my daughter's birth and the months that followed. Your guidance has profoundly shaped my skills and career, and I am honoured to have been your student. I also extend my deepest gratitude to my co-supervisor, **Prof. Abideh Jafari**, for her invaluable scientific guidance and mentorship. Despite your demanding schedule, you were always available to answer my questions — both technical and physics-related. Your dedication to nuturing my growth has been truly inspiring.

A special thanks goes to **Nils Faltermann** and **Thorsten Chwalek**, who have been my day-to-day supervisors at ETP. Your constant support, patience, and willingness to help me navigate this work have been invaluable. I owe much of what I know today to you both. Furthermore, thank you for taking the time out to proofread this dissertation. I also wish to thank **Andreas Hinzmann**, my silent mentor throughout this journey. Though our interactions were not frequent, your insights and efforts to promote my work have been deeply appreciated.

I am grateful to my colleagues at the University of Hamburg whose framework I used for my analysis. Thank you, **Matthias Schröder**, **Christopher Matthies**, and **Finn Jonathan Labe**, for your technical support in setting up the UHH2 framework.

I would also like to acknowledge the German Academic Exchange Service (DAAD) for financially supporting this work. In particular, I am grateful to **Dagmar Beerwerth** for ensuring a smooth visa and funding process. I sincerely appreciate the

Karlsruhe School for Particle and Astroparticle Physics: Science and Technology (KSETA) for accepting me as a doctoral student under the GSSP program funded by DAAD. I extend my gratitude to **Prof. Ulrich Nierste**, who served as my official mentor during my Ph.D. Additionally, I would like to express my appreciation to the administrative staff at ETP and KSETA — **Bärbel Bräunling**, **Barbara Lepold**, and **Raquel Lujan** — for their invaluable assistance in handling administrative matters seamlessly.

Life has presented its challenges, particularly back in Pakistan when my father fell seriously ill, and I had to take on responsibilities at a young age. Education became my path forward. During my time studying physics at Quaid-i-Azam University, Islamabad, a person stepped into my life and gave me the strength to pursue my dreams — my now husband, **Ansar Iqbal**. You not only believed in me but also provided support, making it possible for me to chase my aspirations. Your constant support and encouragement have meant everything to me.

I dedicate this thesis to my father, **Tauqeer Ahmed Tariq** whose unconditional love and support shaped the person I am today. His last words to me were, "I pray to live long enough to see you earn your Ph.D.". Though he is no longer with me, I hope this achievement would have made him proud. This dissertation is also dedicated to my loving mother **Zulaikha Tauqeer**, who held my hand and raised me with love and care after I lost my birth mother at the age of six. I would not be the person I am today without you. My every success — in education, career, and family — belongs to you. To my younger brother and sister, whom I cared for as my own children, I love you deeply. I look forward to seeing you achieve your dreams, just as I have, and I hope to witness your success stories one day.

To my friends, your support has meant the world to me. **Max Neukum** and **Marco Link**, thank you for helping me integrate into the group at KIT. **Ayesha Shahid** and **Myra Khalid**, your unwavering belief in me and your kindness during my most difficult moments helped me persevere.

This journey has been filled with challenges, but thanks to the incredible people in my life, I have overcome them. To each of you, I extend my heartfelt gratitude.