Study of vector boson scattering in events with four leptons and two jets with CMS detector at the LHC

Giljanović, Duje

Doctoral thesis / Doktorski rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:742181

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-05-11**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

INSTITUT POLYTECHNIQUE DE PARIS

NNT : XXXXXXXX

Thèse de doctorat

Study of vector boson scattering in events with four leptons and two jets with CMS detector at the LHC

Thèse de doctorat de l'Institut Polytechnique de Paris et de l'Université de Zagreb
préparée à l'Ecole Polytechnique

École doctorale de l'Institut Polytechnique de Paris (ED IP Paris) n°XXX
Spécialité de doctorat: Physique des particules

Thèse présentée et soutenue à Palaiseau, le XX Decembre 2022, par

GILJANOVIĆ DUJE

Composition du Jury :

XXX YYY address	Président du Jury
XXX YYY address	Rapporteur
XXX YYY address	Rapporteur
XXX YYY address	Examinateur
XXX YYY address	Directeur de thèse
XXX YYY address	Co-directeur de thèse
XXX YYY address	Examinateur
XXX YYY address	Examinateur

Contents

1 The Standard Model and the vector boson scattering **9**

1.1 Preface to the chapter . . . 9

1.2 Introduction to the Standard Model . . . 9

1.2.1 The Lagrangian of the quantum electrodynamics . . . 11

1.2.2 The Lagrangian of the quantum chromodynamics . . . 12

1.2.3 Unification of the electromagnetic and the weak interaction . . . 14

1.3 Electroweak symmetry breaking . . . 17

1.3.1 Spontaneous symmetry breaking and the Goldstone theorem . . . 17

1.3.2 The Brout-Englert-Higgs mechanism . . . 18

1.4 Vector boson scattering . . . 19

1.4.1 Characteristics of VBS processes . . . 21

1.4.2 Effective field theory . . . 24

1.5 Overview of the experimental searches for vector boson scattering . . . 27

2 The large Hadron Collider and the CMS experiment **31**

2.1 Preface to the chapter . . . 31

2.2 The Large Hadron Collider (LHC) . . . 31

2.2.1 A brief history History of LHC . . . 31

2.2.2 The LHC machine and physics experiments . . . 33

2.3 The CMS experiment . . . 35

2.3.1 The Silicon Tracker system . . . 35

2.3.2 The Electromagnetic Calorimeter . . . 35

2.3.3 The Hadron Calorimeter . . . 35

2.3.4 The solenoid magnet . . . 35

2.3.5 The muon chambers . . . 35

2.3.6 The trigger system . . . 35

2.4 Physics objects reconstruction . . . 35

2.5 The future of CMS and LHC . . . 35

2.5.1 The high-granularity calorimeter . . . 35

2.5.2 High-luminosity LHC . . . 35

2.5.3 High-energy LHC . . . 35

3 Electron reconstruction and identification **37**

3.1 Preface to the chapter . . . 37

3.2 Electron reconstruction . . . 37

3.2.1 Clustering . . . 38
3.2.2 Track reconstruction . . . 38
3.2.3 Charge estimation . . . 41
3.2.4 Classification . . . 42
3.2.5 Energy corrections . . . 42
3.2.6 Combining energy and momentum measurements . . . 43
3.2.7 Integration with particle-flow framework . . . 44
3.3 Electron selection . . . 46
3.3.1 Kinematic and impact parameter selection . . . 46
3.3.2 Identification . . . 46
3.3.3 Isolation . . . 49
3.4 Electron efficiency measurements . . . 49
3.4.1 Tag and Probe method . . . 50
3.4.2 Electron selection efficiency in 2016, 2017 and 2018 . . . 52
3.5 Summary . . . 68

4 Search for the VBS in the 4l final state using the Run 2 data 69
4.1 Preface to the chapter . . . 69
4.2 Monte Carlo simulations and data sets . . . 70
4.2.1 Monte Carlo samples . . . 70
4.2.2 Data samples . . . 74
4.3 Event selection . . . 78
4.4 VBS observables . . . 82
4.5 Signal extraction and the cross-section measurement using the MELA discriminant . . . 90
4.5.1 The MELA discriminant . . . 90
4.5.2 Significance and cross-section measurement . . . 94
4.6 Signal extraction using Boosted Decision Trees . . . 96
4.6.1 A Tool for MultiVariate Analysis (TMVA) . . . 96
4.6.2 Introduction to Boosted Decision Trees . . . 97
4.6.3 Algorithm setup for the signal extraction . . . 101
4.6.4 Signal extraction using the BDT7 . . . 103
4.6.5 Signal extraction using the BDT28 . . . 107
4.7 Setting limits on anomalous quartic gauge couplings . . . 117
4.8 Systematic uncertainties . . . 121
4.9 Results . . . 123
4.10 Summary . . . 127

5 Prospective studies for the High-Lumi and the High-Energy LHC 129
5.1 Preface to the chapter . . . 129
5.2 Simulations of the signal and backgrounds . . . 130
5.2.1 Simulations of the EWK signal . . . 131
5.2.2 Simulations of the EWK backgrounds . . . 131
5.2.3 Simulations of the QCD backgrounds . . . 133
5.3 Event selection . . . 135
5.4 Cleaning of lepton-jets and effect of parton showering and pileup on the leading and subleading jets . 137

79 5.4.1 Lepton-jet cleaning . . . 137
80 5.4.2 Effect of parton showering on the leading and subleading jets . . . 140
81 5.4.3 Effect of pileup on the leading and subleading jets . . . 142
82 5.5 Kinematics at 14 and 27 TeV . . . 147
83 5.6 Signal extraction using a BDT and signal significance measurements . . . 154
84 5.6.1 The combined-background BDT and the 2D BDT methods for signal extraction . . . 154
85 5.6.2 Signal extraction and significance measurements at 14 TeV . . . 157
86 5.6.3 Signal extraction and significance measurements at 27 TeV . . . 166
87 5.7 Results . . . 176
88 5.8 Summary . . . 178

89 **6 Summary** **179**

90 **A Supporting plots for the analysis presented in chapter 4** **181**

91 **B Supporting plots for the analysis presented in chapter 5** **191**

Abstract

Studying Vector Boson Scattering is crucial for understanding the electroweak symmetry breaking mechanism and it provides a complementary tool for measuring Higgs boson couplings to vector bosons. In addition, using the effective field theory (EFT) framework, one can probe the Beyond Standard Model physics through modifications of certain quartic gauge couplings. This thesis reports the first evidence, with the CMS detector, of electroweak (EW) production of leptonically decaying Z boson pair accompanied by two hadronic jets with a vector boson scattering topology. The study analyses $137 fb^{-1}$ of proton-proton collisions produced at CERN Large Hadron Collider (LHC) at 13 TeV centre-of-mass energy. Additionally, a prospective study is presented on the longitudinal scattering in the same channel at High-Luminosity (HL) and High-Energy LHC (HE-LHC) conditions, corresponding to 14 and 27 TeV centre-of-mass energy, respectively, with full event kinematics simulated.

Although this channel is characterised by a fully reconstructable final state, the small cross section of EW signal compared to the QCD-induced background makes it challenging to measure. Efficient identification of final state leptons is essential since efficiencies on their measurement enter the analysis with a power of four. Measurement of electron selection efficiencies and derivation of scale factors for 2016, 2017 and 2018 data-taking periods was performed. Electron identification is done at CMS using the multivariate approach with a multivariate classifier retrained, for all three periods, using the ExtremeGradient Boost software and with electron isolation included in the training. Uncertainties on both electron selection efficiencies and scale factors were reduced across the p_T spectrum with special care towards reducing the uncertainties at low p_T.

The EW signal was extracted at 13 TeV using the Matrix Element Likelihood Approach (MELA) and the performance was cross-checked with the boosted decision tree (BDT) classifier. The EW production of two jets in association with two Z bosons was measured with an observed (expected) significance of 4.0 (3.5) standard deviations. The cross sections for the EW production were measured in three fiducial volumes and is $0.33^{(+0.11)}_{(-0.10)}(stat)^{(+0.04)}_{(-0.03)}(syst)\ fb$ in the most inclusive volume, in agreement with the Standard Model (SM) prediction of $0.275 \pm 0.021\ fb$. Limits on the anomalous quartic gauge couplings were derived in terms of EFT operators T0, T1, T2, T8, and T9.

The extraction of the longitudinal component of the Z bosons at the HL- and HE-LHC was performed using two multivariate approaches. A combined-background BDT was trained to separate the $Z_L Z_L$ signal from the mixture of $Z_L Z_T$, $Z_T Z_T$ and QCD-induced backgrounds. In addition, a more complex approach, referred to as the 2D BDT, was designed to increase signal sensitivity. Two BDTs were trained simultaneously to separate the $Z_L Z_L$ signal from the QCD-induced backgrounds and the $Z_L Z_L$ signal from the mixture of $Z_L Z_T$ and $Z_T Z_T$ backgrounds. The effect on signal significance when increasing electron acceptance from $|\eta| = 3$ to $|\eta| = 4$ was studied as well. With an increased electron acceptance, the longitudinal component is expected to be measured with a significance of 1.4 standard deviations at 14 TeV and with an integrated luminosity of $3000 fb^{-1}$. A measurement of the longitudinal scattering in the ZZ channel is expected at 27 TeV, corresponding to an integrated luminosity of $15000 fb^{-1}$, with

a signal significance of 4.6 standard deviations. With the extended electron acceptance, the first observation is expected with a significance of 5.4 standard deviations. Hence, this study demonstrates a significant benefit of further energy increase at the LHC for understanding the EW sector of the SM.

Chapter 1

The Standard Model and the vector boson scattering

1.1 Preface to the chapter

This chapter discusses theoretical foundations essential to follow the work presented in this thesis. Chapter starts with the short overview of the Standard Model. In sections 1.2.1 and 1.2.2 the Lagrangians of the quantum electrodynamics and quantum chromodynamics are derived from the local gauge invariance requirement. The next section discusses the unification of electromagnetic and weak interactions and derives the Lagrangian for the theory of electroweak interactions. In section 1.3 the origin of the weak vector boson masses is discussed through the mechanism of electroweak symmetry breaking and the Brout-Englert-Higgs mechanism. Section 1.4 introduces the theoretical concepts and phenomenology of the vector boson scattering. The emergence of the longitudinal polarization of vector bosons after the electroweak symmetry breaking is greatly important concept and is discussed in detail. Section 1.4.2 introduces the effective field theory framework withing which the beyond Standard Model physics is searched for via measurements of anomalous quartic gauge copulings. The chapter ends in chronological overview of the most important results published thus far by the CMS and ATLAS collaborations on the topic of vector boson scattering in various channels and at different energies. This section is envisioned as a compact summary of available results and will help reader see the contribution of this thesis work as a part of the important ongoing endeavor towards better understanding fundamental physics at the smallest scale.

1.2 Introduction to the Standard Model

The most complete theory, to date, of elementary particles and interactions between them is given by the Standard Model (SM) of particle physics. This is a relativistic quantum field theory with underlying $SU(3)_C \times SU(2)_L \times U(1)_Y$ structure where the first term denotes a group symmetry of the quantum chromodynamics (QCD), the second term defines a group symmetry of the week sector of the theory while $U(1)$ is a group symmetry of the quantum electrodynamics (QED). The subscripts C, L, and Y refer to *color*, *left*, and *hypercharge*. The building blocks of the theory are quantum fields whose excitations are identified as elementary particles.

The fundamental classification of elementary particles in the SM is based on the quantum mechanics observable *spin*. Therefore, elementary particles are divided into the half-integer spin particles called *fermions*, and the integer spin particles called *bosons*. All mater in the universe consists of fermions, while bosons govern the inter-

actions between them. Fermions are further divided into leptons and quarks while themselves are further grouped into three flavor generations. The first lepton generation consists of an electron (e) and an electron neutrino (ν_e). The other two generations, namely, muon (μ) with corresponding muon neutrino (ν_μ) and tau (τ) with corresponding tau neutrino (ν_τ), are then more massive replicas of the former. Only the first generation is stable, and can thus be observed in nature, while the other two are unstable and decay very fast into their first-generation counterparts.
Similarly, quarks are grouped into three generations. The first generation is comprised of the up (u) and down (d) quarks. The other two generations comprise of the charm (c) and strange (s) quarks and the top (t) and bottom (b) quark. Similarly to leptons, only the first generation of quarks is stable.
In addition to each fermion mentioned above, SM recognizes, for each fermion, its anti-particle. Each anti-particle differs from its matter counterpart only by an electromagnetic charge. The anti-particle is usually denoted with a *bar* above the particle designation. For example, an antimatter pair for the u quark is the anti-u quark denoted simply as $\bar{u}$.
Unlike leptons, which can be observed in nature as excitations of underlying fields, this is not the case for quarks. A single quark has never been observed in nature. Instead, only combinations of a (quark, anti-quark) pair, called *mezons*, or quark triplets, called *baryons*, have been observed. An example of a baryon are the (u, u, d) triplet or the (u, d, d). The former is known as the proton and the latter as the neutron. Baryons and mezons are usually referred to as *hadrons*. Together with an electron, they make up the atom which is the fundamental building block of life.

The electron is held in a bound state with a nucleus via electromagnetic interaction mediated through gauge bosons of the electromagnetic interaction. These are called *photons* and are a massless spin-1 bosons. On the other hand, the nucleus of the atom is held together by means of the strong force. The mediators of the strong force are massless, spin-1 gauge bosons called *gluons*. Heavy hadrons and leptons decay through the exchange of massive spin-1 gauge bosons of the weak force. These are $W^\pm$ and Z^0 bosons. Unlike the strong interaction, which affects only those particles which posses the color charge, or electromagnetic interaction which only affects fermions with non-vanishing electric charge, a weak interaction effects all aforementioned particles. In the high energy limit, the electromagnetic force and the weak force are unified into a single force - the electroweak force.

The final member of the elementary particle zoo is the massive, spin-0 particle predicted in theory in 1964 and discovered at CERN in 2012. As we will see in the following sections, this particle was introduced in order to resolve an important disagreement between the theory and the measured reality. Namely, according to the "original" SM, the gauge bosons should be massless. Although this is true for gauge bosons of the electromagnetic and the strong interactions, it contradicts the mass measurements of the gauge bosons in the weak sector of the theory. The particle is known as the Higgs boson and its origin will be discussed in section 1.3.2.

The full list of thus known elementary particles in the SM is summarized in Fig. 1.1.

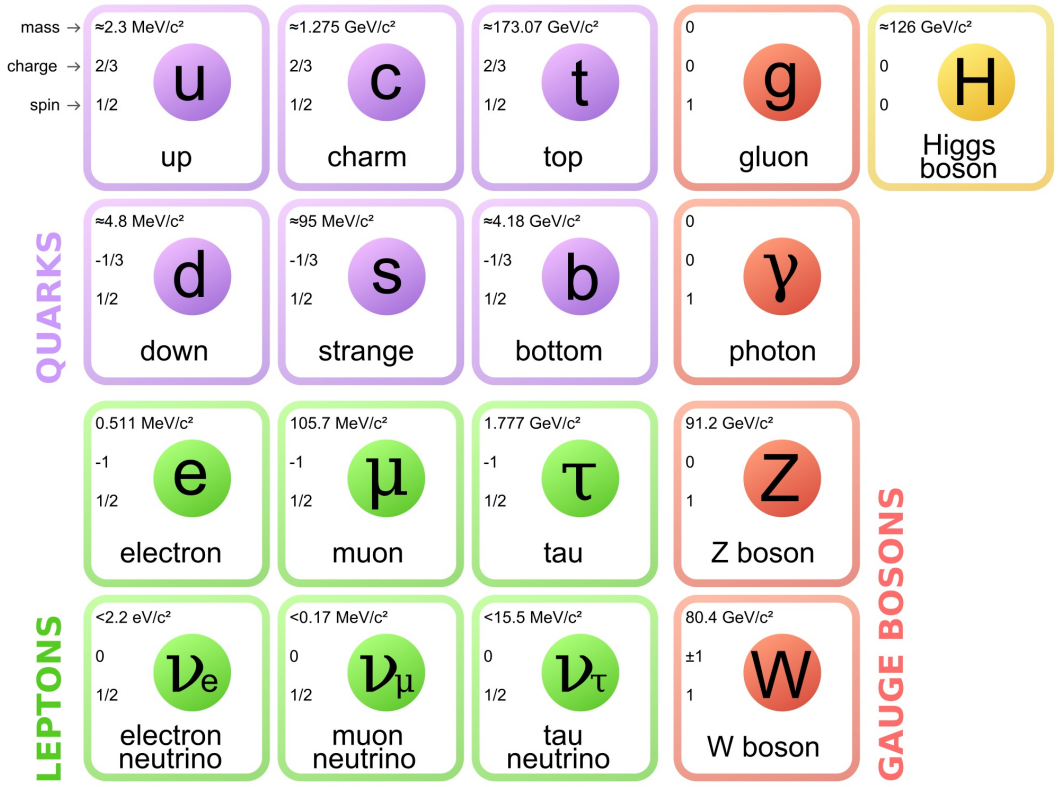


Figure 1.1: The list of known elementary particles in the Standard Model.

1.2.1 The Lagrangian of the quantum electrodynamics

The starting point for constructing the Lagrangian density (henceforth Lagrangian) of the QED is the free Dirac fermion:

$$\mathcal{L}_{free} = i\bar{\psi}(x)\gamma^{\mu}\partial_{\mu}\psi(x) - m\bar{\psi}(x)\psi(x) \tag{1.2.1}$$

It can be easily checked that $\mathcal{L}_{free}$ is invariant under *global* U(1) transformation

$$\psi(x) \to \psi'(x) = e^{iQ\theta}\psi(x) \tag{1.2.2}$$

where $Q\theta$ is an arbitrary real constant, by plugging the transformation 1.2.2 into $\mathcal{L}_{free}$:

$$\begin{aligned}
\mathcal{L}_{free} &= i\bar{\psi}'(x)\gamma^{\mu}\partial_{\mu}\psi'(x) - m\bar{\psi}'(x)\psi'(x) \\
&= ie^{-iQ\theta}\bar{\psi}(x)\gamma^{\mu}\partial_{\mu}e^{iQ\theta}\psi(x) - me^{-iQ\theta}\bar{\psi}(x)e^{iQ\theta}\psi(x) \\
&= ie^{-iQ\theta}\bar{\psi}(x)\gamma^{\mu}e^{iQ\theta}\partial\psi(x) - m\bar{\psi}(x)\psi(x) \\
&= i\bar{\psi}(x)\gamma^{\mu}\partial_{\mu}\psi(x) - m\bar{\psi}(x)\psi(x)
\end{aligned} \tag{1.2.3}$$

One would also like to have a similar behavior of the Lagrangian if the phase θ was the explicit function of the space-time coordinate $\theta = \theta(x)$. However, this is not the case because

$$\partial\psi_{\mu}(x) \to e^{iQ\theta}\left(\partial_{\mu} + iQ\partial_{\mu}\theta\right)\psi(x) \tag{1.2.4}$$

If one wants the $U(1)$ phase invariance to hold locally, a requirement know as the "gauge principle", one has to add another piece to the Lagrangian in a way that additional term in 1.2.4 ($\partial_\mu\theta$) will cancel out. This can be achieved by introducing a new spin-1 field $A_\mu(x)$ which transforms as

$$A_\mu(x) \to A'_\mu(x) \equiv A_\mu(x) + \frac{1}{e}\partial_\mu\theta \tag{1.2.5}$$

and replacing the usual derivative (∂_μ) with the *covariant derivative*

$$D_\mu\psi(x) \equiv \left[\partial_\mu - ieQA_\mu(x)\right]\psi(x) \tag{1.2.6}$$

that transforms in the same way as the field itself:

$$D_\mu\psi(x) \to (D_\mu\psi)'(x) \equiv e^{iQ\theta}D_\mu\psi(x) \tag{1.2.7}$$

The new Lagrangian

$$\mathcal{L} \equiv i\bar{\psi}(x)\gamma^\mu D\mu\psi(x) - m\bar{\psi}(x)\psi(x) = \mathcal{L}_{free} + eQA_\mu(x)\bar{\psi}(x)\gamma^\mu\psi(x) \tag{1.2.8}$$

is now invariant under local $U(1)$ transformations. The second term in Eq. 1.2.8 defines an interaction between the Dirac spinor $\psi(x)$ and the gauge field A_μ. Finally, in order for A_μ to be a true propagating field, one needs to add a gauge-invariant kinetic term in the Lagrangian:

$$\mathcal{L}_{kin} = -\frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x) \tag{1.2.9}$$

where $F_{\mu\nu}(x) \equiv \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic field strength tensor.
The complete Lagrangian describing an electron and a massless vector boson (photon) of spin 1 can be written as

$$\mathcal{L}_{QED} = \mathcal{L}_{free} + eQA_\mu(x)\bar{\psi}(x)\gamma^\mu - \frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x)\psi(x) \tag{1.2.10}$$

One could be tempted to introduce a mass term $\frac{1}{2}m^2A_\mu A_\nu$, but this is not possible since the gauge invariance of the Lagrangian would be violated. As a result, a photon field remains massless. The Lagrangian that was derived gives raise to a set of equations

$$\partial_\mu F^{\mu\nu} = J^\nu \tag{1.2.11}$$

where $J^\nu = -eQ\bar{\psi}\gamma^\nu\psi$ is the fermion electromagnetic current. These are know as Maxwell's equations for electromagnetism. Therefore, by only using gauge symmetry requirements, one can deduce the right QED Lagrangian from which the Maxwell equations follow. This points to the possibility that the QCD Lagrangian could be derived in the similar manner.

1.2.2 The Lagrangian of the quantum chromodynamics

Let q_f^α be a quark flavor field f and a color charge α. Using a vector notation in the color space, $q_f^T \equiv (q_f^1, q_f^2, q_f^3)$, the free QCD Lagrangian reads

$$\mathcal{L}_{free} = \sum_f \bar{q_f}(i\gamma^\mu\partial_\mu - m_f)q_f. \tag{1.2.12}$$

The Lagrangian is invariant under global $SU(3)_C$ transformations in color space

$$q_f^\alpha \to (q_f^\alpha)' = U_\beta^\alpha q_f^\beta \tag{1.2.13}$$

where U are unitary $SU(3)$ matrices that can be written as

$$U = e^{i\frac{\lambda^a}{2}\theta_a} \tag{1.2.14}$$

Here, θ_a are the arbitrary parameters, while $\frac{1}{2}\lambda^a$ $(a = 1, 2, ..., 8)$ are traceless Gell-Mann matrices that represent eight generators of the $SU(3)_C$ group. The matrices λ^a satisfy the commutation relations

$$\left[\frac{\lambda^a}{2}, \frac{\lambda^b}{2}\right] = if^{abc}\frac{\lambda^c}{2} \tag{1.2.15}$$

f^{abc} being the $SU(3)$ structure constant.

As was done in the derivation of the QED Lagrangian, one would like the QCD Lagrangian to be invariant under *local* $SU(3)_C$ transformations. Requiring again $\theta = \theta(x)$, one is forced to replace ordinary quark derivatives with covariant derivatives. Since the $SU(3)_C$ group has eight generators of symmetry, eight different gauge bosons, $G_a^\mu(x)$, are needed. These are identified with eight gluons that mediate the strong interaction. Hence,

$$D^\mu q_f \equiv \left[\partial^\mu - ig_s\frac{\lambda^a}{2}G_a^\mu(x)\right] q_f \equiv \left[\partial^\mu - ig_s G^\mu(x)\right] q_f \tag{1.2.16}$$

where $[G^\mu(x)_{\alpha\beta}] \equiv \left(\frac{\lambda^a}{2}\right)_{\alpha\beta} G_a^\mu(x)$ and g is a dimensionless coupling strength.

Similarly to what was done in the QED case, one requires that covariate derivatives of the color vectors, $D^\mu q_f$, transform as vectors themselves which fixes the transformation properties of the gauge fields:

$$\begin{aligned} D^\mu &\to (D^\mu)' = UD^\mu U^\dagger \\ G^\mu &\to (G^\mu)' = UG^\mu U^\dagger - \frac{i}{g_s}(\partial^\mu U)U^\dagger. \end{aligned} \tag{1.2.17}$$

One can show that, for the infinitesimal $SU(3)_C$ transformation, the gauge fields transform as

$$G_a^\mu \to (G_a^\mu)' = G_a^\mu + \frac{1}{g_s}\partial^\mu(\delta\theta_a) - f^{abc}(\delta\theta_b)G_c^\mu. \tag{1.2.18}$$

Because the $SU(3)_C$ group is not commutative, the gauge transformation of the gluon fields is more complicated than that obtained in QED for the photon field. In addition, non-commutativity of the $SU(3)_C$ gives rise to an additional term involving the gluon fields themselves. Finally, the coupling constant, g, which describes the strength of the interaction between the gluon fields and quarks, is constant.

In order to construct a kinetic term for the gluon fields, the corresponding field strengths are introduced:

$$G^{\mu\nu}(x) \equiv \frac{i}{g_s}[D^\mu, D^\nu] = \partial^\mu G^\nu - \partial^\nu G^\mu - ig_s[G^\mu, G^\nu] \equiv \frac{\lambda^a}{2}G_a^{\mu\nu}(x) \tag{1.2.19}$$

where $G_a^{\mu\nu}(x) = \partial^\mu G_a^\nu - \partial^\nu G_a^\mu + g_s f^{abc}G_b^\mu G_c^\nu$

Using a gauge transformation

$$G^{\mu\nu} \to (G^{\mu\nu})' = UG^{\mu\nu}U^\dagger \tag{1.2.20}$$

and taking a proper normalization for the gluon kinetic term, one can derive the $SU(3)_C$-invariant QCD Lagrangian:

$$\mathcal{L}_{QCD} = \frac{1}{4} G_a^{\mu\nu} G^a_{\mu\nu} + \sum_f \bar{q}_f (i\gamma^\mu D_\mu - m_f) q_f. \tag{1.2.21}$$

By decomposing the Lagrangian into separated components, one can get a better insight into the structure of the QCD:

$$\begin{aligned} \mathcal{L}_{QCD} = & -\frac{1}{4}(\partial^\mu G_a^\nu - \partial^\nu G_a^\mu)(\partial_\mu G^a_\nu - \partial_\nu G^a_\mu) + \sum_f \bar{q}_f^\alpha (i\gamma^\mu \partial_\mu - m_f) q_f^{alpha} \\ & + g_s G_a^\mu \sum_f \bar{q}_f^\alpha \gamma_\mu \left(\frac{\lambda^a}{2}\right)_{\alpha\beta} q_f^\beta \\ & - \frac{g_s}{2} f^{abc} (\partial^\mu G_a^\nu - \partial^\nu G_a^\mu) G^b_\mu G^c_\nu - \frac{g_s^2}{4} f^{abc} f_{ade} G_b^\mu G_c^\nu G_\mu^d G_\nu^e \end{aligned} \tag{1.2.22}$$

The first line contains the correct kinetic terms for the different fields which give rise to the quark propagators. The second line describes the interaction between quarks and gluons. The last line is a consequence of the non-Abelian structure of the $SU(3)_C$ group and includes the cubic and the quartic gluon self-interaction terms. The gluon self-interaction is a unique feature of the QCD theory whereby no such terms exist in the QED. This is the source of the emergent phenomena in the theory of QCD interactions such as the asymptotic freedom and the color confinement. The former ensures that the strong interaction becomes weaker at small distances, while the latter ensures that the strong interaction becomes stronger as quarks are being separated which results in only color-neutral states to be observed in nature.

1.2.3 Unification of the electromagnetic and the weak interaction

In the last two sections, the application of gauge invariance on the $SU(3)_C$ and $U(1)$ group led to the Lagrangian of the gauge fields of the QED and QCD sectors of the SM. It than makes sense to try the same approach to obtain the Lagrangian of the electroweak interaction. It is known that left- and right-handed fields exhibit different behavior. In addition, left-handed fermions appear in doublets, while the right-handed fermions appear as singlet states. In addition, the theory should produce three massive gauge bosons, corresponding to $W^\pm$ and Z^0 bosons, as well as the massless photon field. The simplest symmetry group that sattisfies these conditions is

$$SU(2)_L \otimes U(1)_Y \tag{1.2.23}$$

where the L and Y stand for *left* and *hypercharge* respectively.
For the single family of quarks we would have the following representation

$$\psi_1(x) = \begin{pmatrix} u \\ d \end{pmatrix}_L \qquad \psi_2(x) = u_R \qquad \psi_3(x) = d_R. \tag{1.2.24}$$

The same can be applied to leptons

$$\psi_1(x) = \begin{pmatrix} \mu_e \\ e^- \end{pmatrix}_L \qquad \psi_2(x) = \nu_{eR} \qquad \psi_3(x) = e^-_R. \tag{1.2.25}$$

We start from the free Lagrangian

$$\mathcal{L}_{free} = i\bar{u}(x)\gamma^\mu\partial_\mu u(x) + i\bar{d}(x)\gamma^\mu\partial_\mu d(x) = \sum_{j=1}^{j=3} i\bar{\psi}_j(x)\gamma^\mu\partial_\mu\psi_j(x) \tag{1.2.26}$$

which is invariant under global transformations in the flavor space:

$$\begin{aligned}
\psi_1(x) &\to \psi_1'(x) \equiv e^{iy_1\beta}U_L\psi_1(x),\\
\psi_2(x) &\to \psi_2'(x) \equiv e^{iy_2\beta}\psi_2(x),\\
\psi_3(x) &\to \psi_3'(x) \equiv e^{iy_3\beta}\psi_3(x).
\end{aligned} \tag{1.2.27}$$

Here, parameters y_i are hypercharges and $U_L \equiv e^{i\frac{\sigma_i}{2}\alpha^i}$ $(i = 1, 2, 3)$ is the non-Abelian matrix representing the $SU(2)_L$ transformation acting on the doublet field ψ_1.

Similarly to the QED case, we require the the Lagrangian to be invariant under the local gauge transformations by having $\alpha^i = \alpha^i(x)$ and $\beta^i = \beta^i(x)$. The first step is to introduce the covariant derivative. Since there are four generators of symmetry in the $SU(2)_L \otimes U(1)_Y$, there will laso be four gauge parameters:

$$\begin{aligned}
D_\mu\psi_1(x) &\equiv \left[\partial_\mu - ig\tilde{W}_\mu(x) - ig'y_1B_\mu(x)\right]\psi_1(x),\\
D_\mu\psi_2(x) &\equiv \left[\partial_\mu - ig'y_2B_\mu(x)\right]\psi_2(x),\\
D_\mu\psi_3(x) &\equiv \left[\partial_\mu - ig'y_3B_\mu(x)\right]\psi_3(x).
\end{aligned} \tag{1.2.28}$$

For easier readability, $SU(2)_L$ matrix field $\widetilde{W}_\mu(x) = \frac{\sigma_i}{2}W_\mu^i(x)$ was introduced. The three gauge fields W_μ and additional field B_μ give exactly four gauge fields as needed. However, these are not, at this point, identically identified with the $W^\pm$ and Z^0 bosons.

Similarly to what was done in the derivation of the QCD Lagrangian, one wants the $D_\mu\psi_j(x)$ to transform in the same manner as the $\psi_j(x)$ fields which fixes the transformation properties of the gauge fields:

$$\begin{aligned}
\widetilde{W}_\mu \to \widetilde{W}_\mu' &\equiv U_L(x)\widetilde{W}_\mu U_L^\dagger(x) - \frac{i}{g}\partial_\mu U_L(x)U_L^\dagger(x),\\
B_\mu(x) \to B_\mu'(x) &\equiv B_\mu(x) + \frac{1}{g'}\partial_\mu\beta(x),
\end{aligned} \tag{1.2.29}$$

where $U_L \equiv e^{i\frac{\sigma_i}{2}\alpha^i(x)}$. One can see that the B_μ field transform exactly the same as the photon field of the QED, while W_μ^i fields transform in a similar way to the gluon field of the QCD.

Finally, the free Lagrangian

$$\mathcal{L}_{free} = \sum_{j=1}^{j=3} i\bar{\psi}_j(x)\gamma^\mu D_\mu\psi_j(x) \tag{1.2.30}$$

is now invariant under local $SU(2)_L \otimes U(1)_Y$ gauge transformations. If one wants to built gauge-invariant kinetic term,

one can introduce corresponding field strengths:

$$\begin{aligned} B_{\mu\nu} &\equiv \partial_\mu B_\nu - \partial_\nu B_\mu \\ \widetilde{W}_{\mu\nu} &\equiv \frac{\sigma_i}{2} W^i_{\mu\nu}, \end{aligned} \tag{1.2.31}$$

where $W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g\epsilon^{ijk} W^j_\mu W^k_\nu$. The field $B_{\mu\nu}$ will remain invariant under the local gauge transformation, while $\widetilde{W}_{\mu\nu}$ will transform covariantly:

$$B_{\mu\nu} \to B_{\mu\nu}, \qquad \widetilde{W}_{\mu\nu} \to U_L \widetilde{W}_{\mu\nu} U_L^\dagger \tag{1.2.32}$$

The kinetic part of the Lagrangian is then

$$\mathcal{L}_{kin} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W_i^{\mu\nu} \tag{1.2.33}$$

Finally, the full Lagrangian for the electroweak interaction is then

$$\mathcal{L} = \sum_{j=1}^{j=3} i\bar{\psi}_j(x)\gamma^\mu D_\mu \psi_j(x) - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W_i^{\mu\nu} \tag{1.2.34}$$

Because the Lagrangian contains quadratic terms in $W^i_{\mu\nu}$, the cubic (ZWW, γWW) and quartic ($ZZWW$, γZWW, $\gamma\gamma WW$ and $WWWW$) self-interaction among gauge fields arise directlly. The strength of these interactions is determined by the $SU(2)_L$ coupling g. One can notice that there is always, at least, a pair of charged W bosons in the self-interaction terms since the non-Abelian structure of the $SU(2)_L$ doesn't generate neutral vertices containing only photons and Z bosons. The Lagrangian contains the interaction of the fermion fields with the gauge bosons and the $W^\pm$ and the Z^0 boson are obtained through linear combinations of the gauge bosons.

$$\begin{aligned} W^\pm_\mu &= \frac{1}{\sqrt{2}} \left(W^1_\mu \mp W^2_\mu \right) \\ Z^0_\mu &= cos\theta_w W^3_\mu - sin\theta_w B_\mu \end{aligned} \tag{1.2.35}$$

and the photon field A_μ

$$A_\mu = sin\theta_w W^3_\mu - cos\theta_w B_\mu \tag{1.2.36}$$

Additionally, one can see that the gauge invariance forbids the massive fermionic fields since this would give raise to a mixture of the left- and the right-handed fields through term $m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$ which would explicitly break the gauge symmetry of the Lagrangian. In the same manner, one cannot introduce the mass term for the gauge fields without explicitly breaking the gauge symmetry. Thus, the $SU(2)_L \otimes U(1)_Y$ only contains massless gauge fields. Before 1964, the origin of the gauge boson masses in the SM framework was one of the most urgent issues to be resolved since the experiments estimated the mass of the W and the Z bosons to be around $80.30\ GeV$ and $91.19\ GeV$ respectively. This was resolved through the mechanism of spontaneous symmetry breaking discussed in the next section.

1.3 Electroweak symmetry breaking

1.3.1 Spontaneous symmetry breaking and the Goldstone theorem

We start by introducing a complex scalard field ϕ, with Lagrangian

$$\mathcal{L} = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi), \tag{1.3.1}$$

where

$$V(\phi) = \mu^2 \phi^\dagger \phi + h\left(\phi^\dagger \phi\right)^2 \tag{1.3.2}$$

and the Lagrangian is invariant under global phase transformations of the scalar field

$$\phi(x) \to \phi'(x) \equiv e^{i\theta}\phi(x) \tag{1.3.3}$$

In order to find the ground state of the potential one can simply evaluate

$$\frac{\partial V(\psi)}{\partial \phi} \equiv \mu^2 \phi^\dagger + 2h\phi^\dagger \phi \phi^\dagger = 0 \Longrightarrow \phi^\dagger\left(\mu^2 + 2h\phi^\dagger \phi\right) = 0 \tag{1.3.4}$$

In order to have a ground state one must bound it from bellow and thus $h > 0$. The Eq. 1.3.4 can then only hold if

1. $\mu^2 > 0$: $\phi = 0$
2. $\mu^2 < 0$: $|\phi_0| = \sqrt{\frac{-\mu^2}{2h}} \equiv \frac{v}{\sqrt{2}} > 0$

The first solution is a trivial one which describes a scalar particle with mass μ and coupling h. This solution corresponds to the potential shape shown on the left-hand side of the Fig. 1.2.
The solution with $\mu < 0$ is more interesting and is shown on the right-hand side of the same figure. Due to the phase invariance of the Lagrangian, there is an infinite number of degenerate states of minimum energy, $\phi_0 = \frac{v}{\sqrt{2}} e^{i\theta}$, each corresponding to a single point in the "minima" valley depicted as the dashed circle in the figure. By choosing any specific solution, the symmetry of the ground state will be broken. This is referred to as the *spontaneous symmetry breaking*. One can chose $\theta = 0$ and parametrize the excitations above the ground state as

$$\phi(x) = \frac{1}{\sqrt{2}}\left[v + \phi_1(x) + i\phi_2(x)\right], \tag{1.3.5}$$

where ϕ_1 and ϕ_2 are real fields. The field ϕ_1 corresponds to the oscillations in the radial direction around the specified minimum. On the other side, the field ϕ_2 corresponds to the oscillation in the angular direction around the specified minimum, i.e. along the dashed circle in the figure. This then results in the potential of the form

$$V(\phi) = V(\phi_0) - \mu^2\phi_1^2 + hv\phi_1(\phi_1^2 + \phi_2^2) + \frac{h}{4}(\phi_1^2 + \phi_2^2)^2 \tag{1.3.6}$$

One can immediately notice that the mass term, $m_{\phi_1}^2 = -2\mu^2$, for the ϕ_1 field, describing the radial oscillations, pops out. At the same time, an additional massless field, ϕ_2, associated with the angular oscillations around the minimum, emerges as a consequence of the spontaneous symmetry breaking. In other terms, in addition to a massive particle, a massless particle emerged as a result of the spontaneously broken symmetry.

This finding is generalized through *Nambu-Goldstone theorem*: if the Lagrangian is invariant under a continuous group with M generators, but the vacuum is invariant only under a subgroup with N generators $(M > N)$, then

there must appear $M - N$ massles, spin-0 particles. In other words, one massless, spin-0 particle must appear for each generator of the symmetry that was lost. This particle is known as the *Nambu-Goldstone boson*, or, simply, the *Goldstone boson*. The idea of the spontaneous symmetry breaking, and, consequently, the emergence of Goldstone bosons, is central to the generation of masses of the $W^{\pm}$ and Z^0 bosons. This is discussed in the next section.

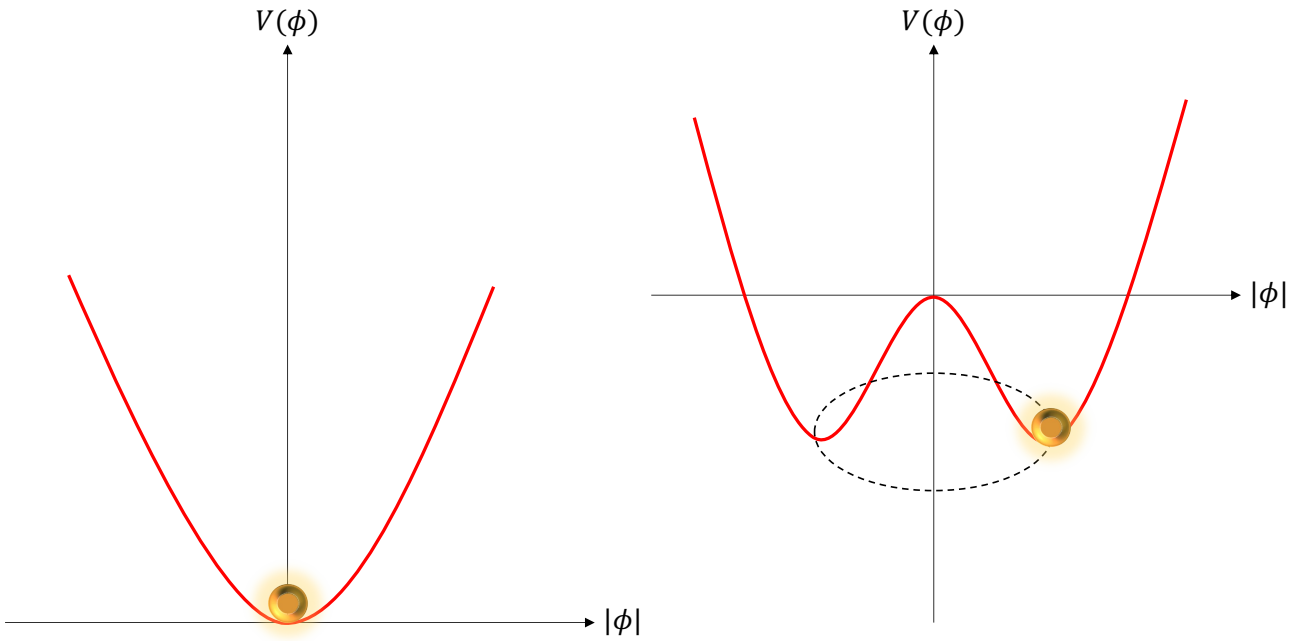


Figure 1.2: The shape of the scalar potential for $\mu^2 > 0$ (left) and $\mu^2 < 0$. The latter features an infinite set of degenerate vacua, corresponding to different phases θ, connected through a massless field excitation ϕ_2.

1.3.2 The Brout-Englert-Higgs mechanism

In order to explain the origin of masses of the weak gauge bosons, one must consider a doublet of complex scalar fields

$$\phi(x) = \begin{pmatrix} \phi^{(+)} = \phi_1 + i\phi_2 \\ \phi^{(0)} = \phi_3 + i\phi_4 \end{pmatrix} \tag{1.3.7}$$

The two components of the charged field, ϕ_1 and ϕ_2 will give raise to two Goldstone bosons that will be incorporated into two massive W bosons, while the ϕ_4 component of the neutral field will give raise to a third Goldstone boson that will be incorporated into a massive Z boson.

The corresponding Lagrangian for the the doublet of complex scalar fields than reads

$$\mathcal{L} = (D_\mu \phi)^\dagger D^\mu \phi - \mu^2 \phi^\dagger \phi - h(\phi^\dagger \phi)^2 \tag{1.3.8}$$

where

$$D^\mu \phi = \left[\partial^\mu - ig\widetilde{W}^\mu - \frac{ig'}{2} B^\mu \right] \phi, \tag{1.3.9}$$

is invariant under local $SU(2)_L \otimes U(1)_Y$ symmetry. As before, one requires that the potential be bound from bellow so that $h > 0$ and by choosing $\mu < 0$ obtains the potential similar to the one considered before. This gives rise to an

infinite set of degenerate ground states defined by

$$|\phi_0^{(0)}| = \sqrt{\frac{-\mu^2}{2h}} \equiv \frac{v}{\sqrt{2}} \tag{1.3.10}$$

where one must bear in mind that only a neutral field can acquire a vacuum expectation value due to electric charge conservation. By choosing any specific ground state one spontaneously breaks the $SU(2)_L \otimes U(1)_Y$ symmetry into electromagnetic subgroup $U(1)_Y$. Since the original symmetry with four generators has been broken into the symmetry with only one generator, the Goldstone theorem mandates that three Goldostone bosons must appear. One can now, similarly as before, parametrize the excitations above the ground state as

$$\phi(x) = e^{i\frac{\sigma_i}{2}\theta^i(x)} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.3.11}$$

where σ^i are generators of the $SU(2)_L$ algebra, $\theta^i(x)$ are the three massless Goldstone bosons, and $H(x)$ is the *Higgs field* whose excitation corresponds to the *Higgs boson*.

If one chooses the unitary gauge $\theta^i(x) = 0$, the kinetic part of the Lagrangian in Eq. 1.3.8 then becomes

$$(D_\mu \phi)^\dagger) D^\mu \phi \rightarrow \frac{1}{2}\partial_\mu H \partial^\mu H + (v+H)^2 \left[\frac{g^2}{4} W_\mu^\dagger W^\mu + \frac{g^2}{8cos^2\theta_w} Z_\mu Z^\mu \right] \tag{1.3.12}$$

As one can notice, the massless Goldstone boson fields have been incorporated into a new, massive, W and Z boson fields as a consequence of the non-vanishing vacuum value of the neutral scalar field and after a choice of an appropriate gauge requirement. After the spontaneous breaking of the electroweak symetry (EWSB), the mass of the electroweak bosons is

$$\begin{aligned} M_Z &= \frac{vg}{2cos\theta_w} \\ M_W &= \frac{1}{2} vg \end{aligned} \tag{1.3.13}$$

The photon remained massless after the EWSB because the $U(1)_{QED}$ is an unbroken symmetry. Before the EWSB the Lagrangian contained massless $W^\pm$ and Z^0 bosons which gives $3 \times 2 = 6$ degrees of freedom since massless, spin-1 fields can only have two values of polarization, 1 and -1, corresponding to the two transverse polarizations. However, after the EWSB, three Goldstone bosons have been "eaten" by the weak bosons giving them mass and, consequently, additional degree of freedom: longitudinal polarization.

1.4 Vector boson scattering

A new feature emerging from the EWSB mechanism is the longitudinal polarization of massive gauge bosons in the weak sector. By comparing the polarization vectors of the transversely polarized vector bosons

$$\begin{aligned} \epsilon_+^\mu &= \frac{1}{\sqrt{2}}(0\ 1\ i\ 0)^\mu \\ \epsilon_-^\mu &= \frac{1}{\sqrt{2}}(0\ 1\ -i\ 0)^\mu \end{aligned} \tag{1.4.1}$$

where ϵ_+^μ and ϵ_-^μ correspond to the right and left helicity states of the transverse polarization, respectively,

to the polarization vector of the longitudinally polarized vector boson of mass m and momentum $k^\mu = \frac{1}{m}(k_z\ 0\ 0\ E)^\mu$

$$\epsilon_L^\mu = \frac{1}{m}(k_z\ 0\ 0\ E)^\mu \tag{1.4.2}$$

one notices a striking difference between the two. While the transverse components remain constant as the scattering energy increases, the longitudinal component scales with scattering energy as E/m. The reason for the difference in the high-energy behavior of the two polarizations stems from the different origin of the two. The transverse polarization exist in the theory of the weak sector prior to the EWSB and corresponds to the massless gauge bosons. On the other hand, the longitudinal component of the polarization is the consequence of incorporating the Goldstone bosons of the EWSB into the gauge boson fields as a result of local symmetry and unitarity gauge requirement.

The difference in the behavior of the two polarizations in the high-energy limits suggests that, at high energies, a longitudinal component can be disentangled from the transverse component. While the longitudinal states are equivalent to the Goldstone Bosons of the EWSB, the transverse states correspond to the original electroweak gauge bosons. Using the Goldstone boson equivalence theorem [1–3], one can conclude that scattering of the longitudinal vector bosons is equivalent to the scattering of Goldstone bosons.
The importance of this statement becomes clear if one looks, for example, at the scattering amplitude of the $W_L^+W_L^- \to W_L^+W_L^-$ process [4]:

$$\mathcal{M}_{SM}(W_L^+W_L^- \to W_L^+W_L^-) \approx \mathcal{M}_{SM}(w^+w^- \to w^+w^-) \approx -i\frac{m_H^2}{v^2}\left[2 + \frac{m_H^2}{s - m_H^2} + \frac{m_H^2}{t - m_H^2}\right], \tag{1.4.3}$$

where $w^\pm$ are the Goldstone bosons, and s and t are the Mandelstam variables.
In the high-energy limit where $s,\ |t| >> m_H^2$, this expression becomes constant and the cross section falls proportionally with the scattering energy ($\sigma \sim \frac{1}{s}$).
On the other hand, without the Higgs boson in the SM ($m_H \to \infty$), the matrix element becomes

$$\mathcal{M}_{Higgsless}(W_L^+W_L^- \to W_L^+W_L^-) \approx i\frac{s+t}{v^2} \tag{1.4.4}$$

This shows that without the Higgs bosons, in the high-energy limit, the cross section will diverge and, therefore, the unitarity of the theory will be violated. This points to the significance of the cancellations between the contributions from pure gauge diagrams and Higgs interactions. More so, not only does the unitarity violation occur if there is no Higgs boson, but it occurs also around the energy of $\approx 1.2\ TeV$ if the couplings of the Higgs bosons to the vector bosons differs from the SM prediction [2, 5]. This result was the primary argument for the yet-unobserved physics at the TeV scale. It showed that either the Higgs boson will have to be found at the LHC, or some other phenomena would have to appear at the TeV scale in order to preserve the unitarity.

With the Higgs boson discovered, the problem of the unitarity violation was resolved given that the coupling of the Higgs boson to the vector bosons is as predicted by the SM. Thus, the scattering of longitudinal vector bosons proves to be an important tool for probing the scalar sector of the SM and studying EWSB mechanism. In addition, VBS enables one to study the non-Abelian structure of the electroweak (EW) sector by probing the quartic vertices. Finally, a beyond SM (BSM) phenomena could manifest themselves, for example, in modifications to the quartic gauge couplings that increase the production cross section. This will be discussed, in more detail, in section 1.4.2.

1.4.1 Characteristics of VBS processes

A typical example of a VBS process are two gauge bosons mediated by the two separate quark lines which then interact. VBS is defined at a tree-level as $\mathcal{O}(\alpha^6)$ process which includes the decay products of the heavy gauge bosons. Depending on the decay products of the two vector bosons, VBS processes are observed through the *fully leptonic* decay channel with four leptons and two hadronic jets in the final state, a *semi-leptonic* channel with two leptons and four jets in the final state and *fully hadronic* decay channel with six jets in the final state. Some example Feynman diagrams of the processes that lead to the $4l2j$ final state are shown in Figs. 1.3 - 1.6 where the vector bosons are denoted as V, fermions with f and quarks with q.

The Fig. 1.3 shows representative Feynman diagrams for the VBS production of the $4l2j$ final state. The top row shows the EW VBS production through the quartic (top-left) and two trilinear (top-right) coupling diagrams. The bottom diagram shows the production of the same final state through the t-channel exchange of the Higgs boson. The latter ensures unitarity through Higgs boson coupling to the vector boson fields.

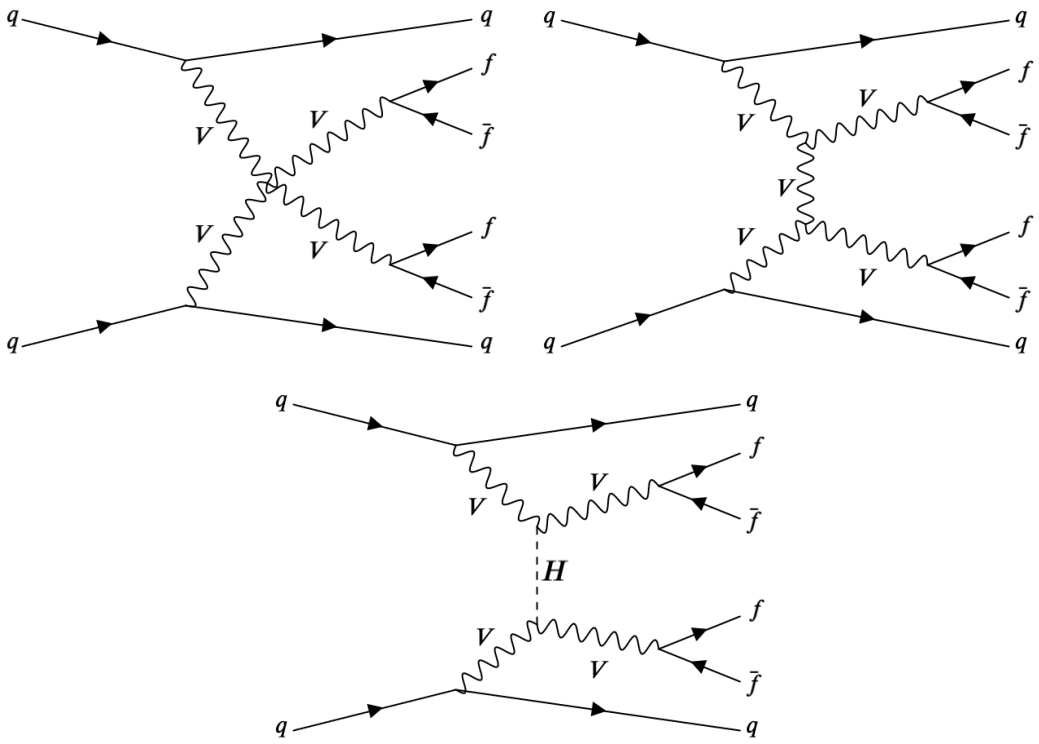


Figure 1.3: Representative Feynman diagrams for the VBS production of the *VVjj* final state with a scattering topology including the quartic (top-left) and two trilinear (top-right) vertices together with the t-channel exchange of the Higgs boson.

The Fig. 1.4 represents the non-VBS production with EW vertices only that cannot be separated from the VBS production and, thus, must be included. The left-hand side diagram is an example of one vector boson being radiated from the quark line, while on the right-hand side diagram an off-shell boson splits into the two final state bosons. These diagrams are needed for ensuring the gauge invariance.

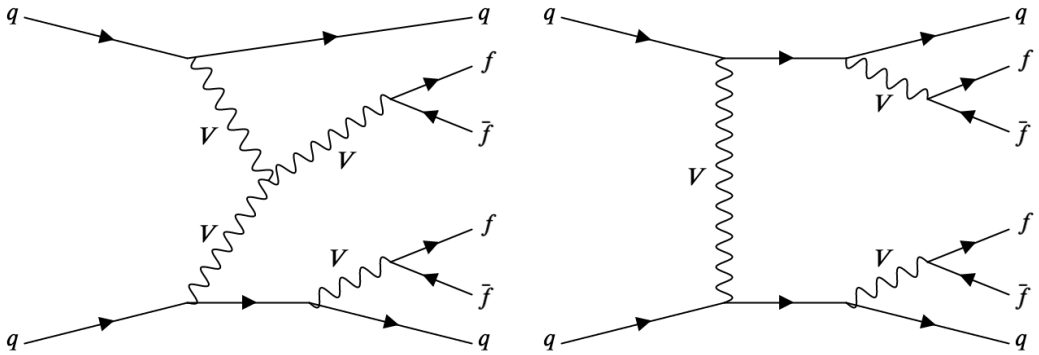


Figure 1.4: Representative Feynman diagrams for the non-VBS production of the *VVjj* final state with a scattering topology including the quartic (top-left) and two trilinear (top-right) vertices together with the t-channel exchange of the Higgs boson.

The Fig. 1.5 shows the pure EW diagrams that are not relevant for the study presented in this thesis and can be suppressed by appropriate phase space selection. The diagram shown on the left-hand side is an example of the non-resonant diagrams where the final state leptons originate from an off-shell Z boson and can be suppressed by requiring an on-shell Z boson. The right-hand side diagram shows the triboson production where one of the gauge bosons decay hadronically. The hadronic jets originating from such process will have dijet mass of around 100 GeV. For this reason, the $m_{jj} > 100\ GeV$ cut will be applied in the analysis.

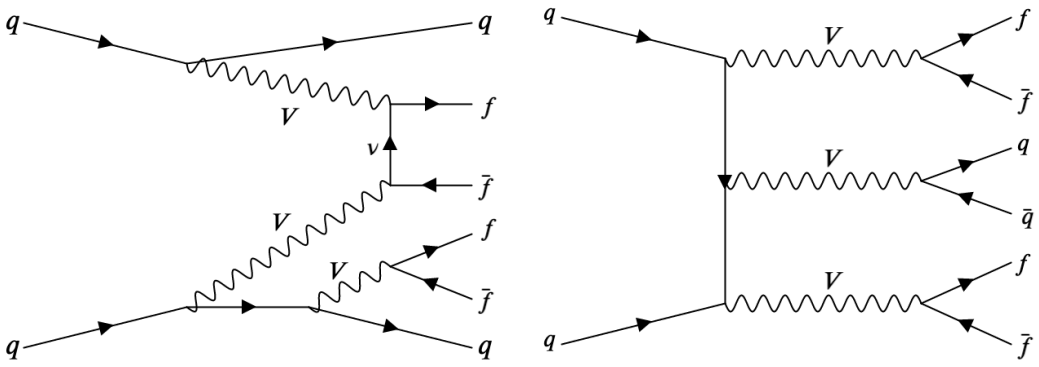


Figure 1.5: Feynman diagrams for the non-resonant production of the *VVjj* final state. These processes can be suppressed by appropriate pahe space selection.

Finally, the Fig. 1.6 shows two examples of the QCD-induced production of the $4l2j$ final state.

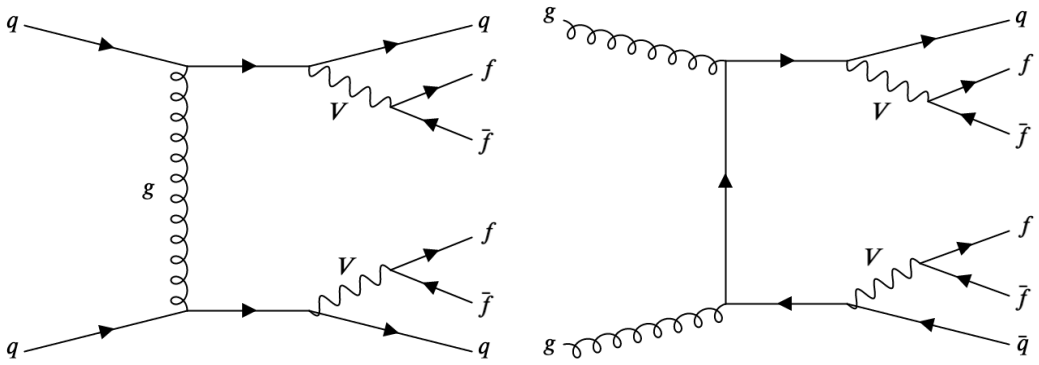


Figure 1.6: Representative Feynman diagrams for the QCD-induced production of the *VVjj* final state.

In addition to the purely EW contributions at order $\mathcal{O}(\alpha^6)$, VBS processes also include reducible contributions of order $\mathcal{O}(\alpha^5\alpha_s)$ and $\mathcal{O}(\alpha^4\alpha_s^2)$. These are reffered to as the *VBS signal*, *interference* and *QCD background* respectively. Because EW and QCD contributions behave differently, the maxima of dijet mass distributions will peak at different parts of the detector. This is shown in Fig. 1.7.

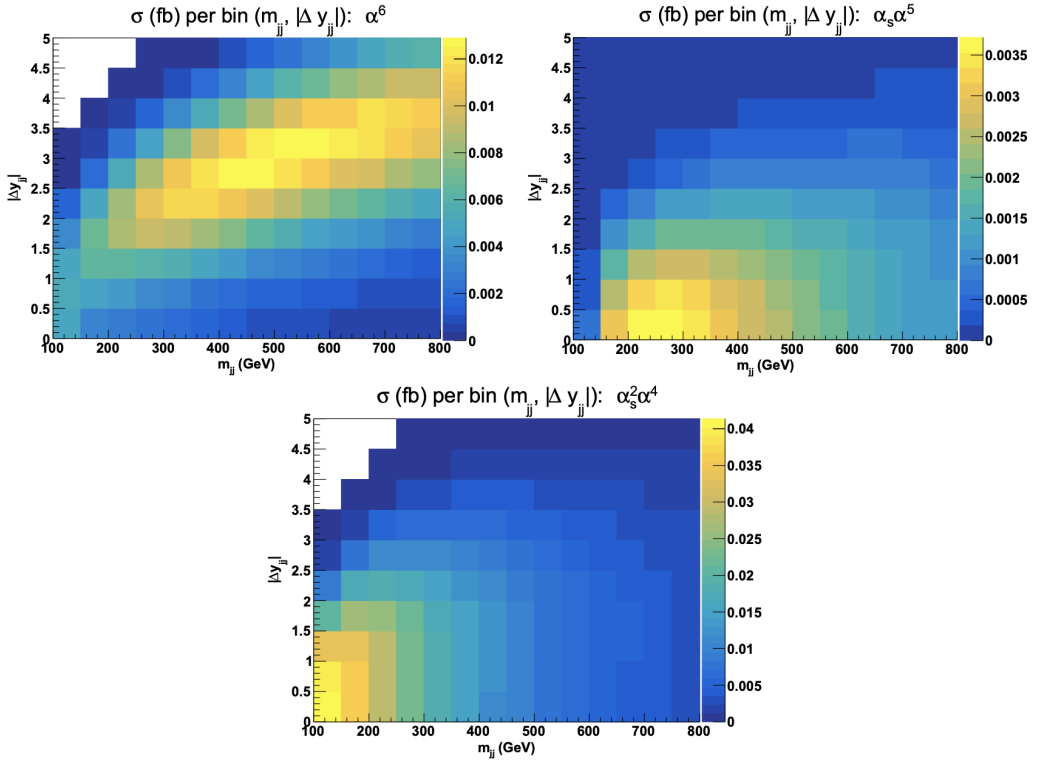


Figure 1.7: Double-differential distributions in the variables m_{jj} and $|\Delta y_{jj}|$ for the three LO contributions of orders $\mathcal{O}(\alpha^6)$ (top-left), $\mathcal{O}(\alpha^5\alpha_s)$ (top-right) and $\mathcal{O}(\alpha^4\alpha_s^2)$ (bottom) corresponding to the VBS signal, interference and QCD background contribution respectively. The figure is taken from [6] where one can also find the details on the selection criteria which was applied.

Since the EW contribution doesn't feature QCD exchanges between the two quark lines while the QCD component does, the differential cross section as a function of the dijet invariant mass or the rapidity difference between the two outgoing hadronic jets is different for the two components.

As one can see, the distinctive feature of the VBS processes are the two hadronic jets (referred to as the tagging jets) with large invariant masses and the large pseudorapidity gap between them. The latter can be also confirmed by investigating the expression for the square of the scattering amplitude:

$$|\mathcal{A}|^2 \sim \frac{p_1 \cdot p_2 \cdot p_3 \cdot p_4}{(q_1^2 - m_Z^2)^2 (q_2^2 - m_Z^2)^2}, \tag{1.4.5}$$

where p_1 and p_2 are the momenta of the incoming quarks, p_3 and p_4 are the momenta of the outgoing quarks, and $q_1 = p_1 - p_3$ and $q_2 = p_2 - p_3$ are the momenta of the intermediate gauge bosons.

The scattering amplitude is large if $m_{jj} \equiv p_3 \cdot p_4$ is large. One can show that

$$m_{jj} \approx 2 \cdot p_T(j_1) p_T(j_2) \left[cosh(\eta_{j_1} - \eta_{j_2}) - cos(\phi_{j_1} - \phi_{j_2}) \right] \tag{1.4.6}$$

Given the constant momenta of the outgoing jets, this expression is largest when the psudorapidity gap between the jets is large and when the jets are back-to-back ($\phi_{j_1} - \phi_{j_2} \to \pi$). This agrees with conclusions inferred from the distributions shown in Fig. 1.7

Additionally, the expression for the square of the scattering amplitude can be large if the denominator is small which occurs for small values of q_i. It can be shown that the square of q_1 can be written in terms of the scattering angle, θ_1, between $\vec{p}_1$ and $\vec{p}_3$, the energy of the incoming (E_1) and outgoing (E_3) quarks, and the transverse momentum of the outgoing quark ($p_{T,3}$):

$$q_1^2 = -\frac{2}{1 + cos\theta_1} \frac{E_1}{E_3} p_{T,3}^2 \tag{1.4.7}$$

This expression is smallest when the scattering angle is small ($\theta_1 \to 0$) or when the transverse momentum of the outgoing quark is small. Since the quarks will recoil against the vector bosons upon radiation, and since enough energy is needed to create the on-shell Z boson of the final state, the p_T of the outgoing jets will be of the order of the Z boson mass $p_T(j) \approx m_Z$.

An additional feature of the EW production of the $4l2j$ final state is the kinematics of the vector bosons with respect to the tagging jets. While the jets are found, preferably, at the low scattering angles, the gauge bosons tend to be found in the pseudorapidity gap between them.

Finally, due to the absence of the color flow between the interacting partons, hadron activity in the central region of the detector is suppressed [7].

1.4.2 Effective field theory

As was discussed in section 1.2.3, the $SU(2)_L \otimes U(1)_Y$ gauge symmetry gives rise to the trilinear and quartic gauge couplings of the vector bosons. Therefore, studying these interactions can further confirm the theoretical predictions, or point to some deviations from SM predictions that would give a hint to possible new physics at a higher scale. In particular, modifications of the vector boson couplings, either amongst themselves, or to the Higgs boson, could result in imperfect cancellation between the Feynman amplitudes including quartic gauge boson interactions, trilinear gauge boson couplings, and Higgs exchange. This would result in the increase of the cross section with energy that could be observed as an excess of events compared to the SM prediction.

One approach in the search for the beyond SM (BSM) physics exploits the effective field theory (EFT) approach. This approach comes in two flavours. In the model-dependent *top-down* approach one starts with an ultraviolet-complete (UV-complete) theory, finds its s low-energy behaviour, and, finally, tries to match it to the SM. On the other hand, in the *bottom-up* approach one starts with the SM and builds towards the UV regime. This approach does not provide the concrete predictions for the BSM scenarios, but gives tools to study new physics in the regime accessible from the SM. [8–11]

When building the bottom-up approach one starts from the SM Lagrangian with underlying $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ symmetry with all operators with dimension of up to four. Next, one seeks to extend the theory by adding operators of higher dimensions with coefficients of inverse power of mass therefore lifting the restriction of renormalizability. The EFT Lagrangian can be written as

$$\mathcal{L}_{EFT} = \mathcal{L}_{SM} + \sum_{d>4} \sum_{i} \frac{f_i^{(d)}}{\Lambda^{d-4}} \mathcal{O}_i^{(d)}, \tag{1.4.8}$$

where $\mathcal{O}_i^{(d)}$ are the d-dimensional BSM operators invariant under the symmetries of the SM, f_i are the corresponding Wilson coefficients or *coupling strengths* and Λ is the typical scale of new physics. Any evidence for a non-zero Wilson coefficient would represent a clear sign of new physics.

The dominant operator in the expansion will be the one with dimension five and is responsible for generating Majorana massed for neutrinos [12]. However, all odd-power operators lead to lepton or baryon number violation and are omitted [13]. We are, thus, left with dimension-6 operators followed by the dimension-8 operators.

The former are responsible for the emergence of anomalous triple gauge couplings (aTGCs), while later give rise to the anomalous quartig gauge couplings (aQGCs). The study presented in this thesis will focus on the operators that modify the quartic gauge couplings while simultaneously leaving trilinear gauge couplings intact. The reason behind this is the fact that VBS channel explored in this thesis is most sensitive to probing aQGCs. The list of all aQGC operators in the linear Higgs-doublet representation [14] can be found in Table 1.1. The modified field strength tensors are given by

$$\begin{aligned} \hat{W}^{\mu\nu} &= i g_w \frac{\sigma^j}{2} W^{j,\,\mu\nu} \\ \hat{B}^{\mu\nu} &= \frac{ig}{2} B^{\mu\nu} \end{aligned} \tag{1.4.9}$$

Table 1.2 shows which vertices are modified by individual operators.

The aQGCs that only involve the EW fields are given by the tensor operators $\mathcal{O}_T$, whereby the $ZZjj$ channel explored in this thesis work is most sensitive to the charged-current operators $\mathcal{O}_{T,0,1,2}$ as well as the neutral-current operators $\mathcal{O}_{T,8,9}$.

Class	Definition
Scalar involve only the scalar field	$\mathcal{O}_{S,0} = \left[(D_\mu\phi)^\dagger D_\nu\phi\right] \times \left[(D^\mu\phi)^\dagger D^\nu\phi\right]$
	$\mathcal{O}_{S,1} = \left[(D_\mu\phi)^\dagger D_\mu\phi\right] \times \left[(D_\nu\phi)^\dagger D^\nu\phi\right]$
	$\mathcal{O}_{S,2} = \left[(D_\mu\phi)^\dagger D_\nu\phi\right] \times \left[(D^\nu\phi)^\dagger D^\mu\phi\right]$
Tensor involve only the field strength tensor	$\mathcal{O}_{T,0} = Tr\left[\hat{W}_{\mu\nu},\, \hat{W}^{\mu\nu}\right] \times Tr\left[\hat{W}_{\alpha\beta},\, \hat{W}^{\alpha\beta}\right]$
	$\mathcal{O}_{T,1} = Tr\left[\hat{W}_{\alpha\nu},\, \hat{W}^{\mu\beta}\right] \times Tr\left[\hat{W}_{\mu\beta},\, \hat{W}^{\alpha\nu}\right]$
	$\mathcal{O}_{T,2} = Tr\left[\hat{W}_{\alpha\mu},\, \hat{W}^{\mu\beta}\right] \times Tr\left[\hat{W}_{\beta\nu},\, \hat{W}^{\nu\alpha}\right]$
	$\mathcal{O}_{T,5} = Tr\left[\hat{W}_{\mu\nu},\, \hat{W}^{\mu\nu}\right] \times \hat{B}_{\alpha\beta}\hat{B}^{\alpha\beta}$
	$\mathcal{O}_{T,6} = Tr\left[\hat{W}_{\alpha\nu},\, \hat{W}^{\mu\beta}\right] \times \hat{B}_{\mu\beta}\hat{B}^{\alpha\nu}$
	$\mathcal{O}_{T,7} = Tr\left[\hat{W}_{\alpha\mu},\, \hat{W}^{\mu\beta}\right] \times \hat{B}_{\beta\nu}\hat{B}^{\nu\alpha}$
	$\mathcal{O}_{T,8} = \hat{B}_{\mu\nu}\hat{B}^{\mu\nu} \times \hat{B}_{\alpha\beta}\hat{B}^{\alpha\beta}$
	$\mathcal{O}_{T,9} = \hat{B}_{\alpha\mu}\hat{B}^{\mu\beta} \times \hat{B}_{\beta\nu}\hat{B}^{\nu\alpha}$
Mixed involve the field strength tensor and the scalar field	$\mathcal{O}_{M,0} = Tr\left[\hat{W}_{\mu\nu},\, \hat{W}^{\mu\nu}\right] \times \left[(D_\beta\phi)^\dagger D^\beta\phi\right]$
	$\mathcal{O}_{M,1} = Tr\left[\hat{W}_{\mu\nu},\, \hat{W}^{\nu\beta}\right] \times \left[(D_\beta\phi)^\dagger D^\mu\phi\right]$
	$\mathcal{O}_{M,2} = \hat{B}_{\mu\nu}\hat{B}^{\mu\nu} \times \left[(D_\beta\phi)^\dagger D^\beta\phi\right]$
	$\mathcal{O}_{M,3} = \hat{B}_{\mu\nu}\hat{B}^{\nu\beta} \times \left[(D_\beta\phi)^\dagger D^\mu\phi\right]$
	$\mathcal{O}_{M,4} = (D_\mu\phi)^\dagger \hat{W}_{\beta\nu} D^\mu\phi \times \hat{B}^{\beta\nu}$
	$\mathcal{O}_{M,5} = (D_\mu\phi)^\dagger \hat{W}_{\beta\nu} D^\nu\phi \times \hat{B}^{\beta\mu}$
	$\mathcal{O}_{M,7} = (D_\mu\phi)^\dagger \hat{W}_{\beta\nu} \hat{W}^{\beta\mu} D^\nu\phi$

Table 1.1: Scalar, tensor and mixed dimension-eight operators in the EFT approach. Limits on the Wilson coefficients for the $\mathcal{O}_{T,0,1,2}$ as well as the $\mathcal{O}_{T,8,9}$ operators are derived in chapter 4. The table is taken from [15].

	$\mathcal{O}_{S,0}$ $\mathcal{O}_{S,1}$ $\mathcal{O}_{S,22}$	$\mathcal{O}_{M,0}$ $\mathcal{O}_{M,1}$ $\mathcal{O}_{M,7}$	$\mathcal{O}_{M,2}$ $\mathcal{O}_{M,3}$ $\mathcal{O}_{M,4}$ $\mathcal{O}_{M,5}$	$\mathcal{O}_{T,0}$ $\mathcal{O}_{T,1}$ $\mathcal{O}_{T,2}$	$\mathcal{O}_{T,5}$ $\mathcal{O}_{T,6}$ $\mathcal{O}_{T,7}$	$\mathcal{O}_{T,8}$ $\mathcal{O}_{T,9}$
$WWWW$	✓	✓		✓		
$WWZZ$	✓	✓	✓	✓	✓	
$ZZZZ$	✓	✓	✓	✓	✓	✓
$WWZ\gamma$		✓	✓	✓	✓	
$WW\gamma\gamma$		✓	✓	✓	✓	
$ZZZ\gamma$		✓	✓	✓	✓	✓
$ZZ\gamma\gamma$		✓	✓	✓	✓	✓
$Z\gamma\gamma\gamma$				✓	✓	✓
$\gamma\gamma\gamma\gamma$				✓	✓	✓

Table 1.2: List of vertices modified by a given aQGC operator

1.5 Overview of the experimental searches for vector boson scattering

The following section represents a chronological overview of the most important results, obtained by the CMS and ATLAS collaborations, on the vector boson scattering in different channels and center-of-mass energies. This section will help the reader understand the progress in the field and will put in perspective the work presented in this thesis. For brevity sake, many details are omitted. An interested reader can find an in-depth discussion on the selection criteria, fiducial region definitions, signal extraction methods and other details in the corresponding papers.

The first results on the scattering of two vector bosons in the VBS topology channels was reported by the CMS and ATLAS collaborations in 2014 at $8\ TeV$ centre-of-mass energy.
The CMS reported an observed (expected) significance of 2.0 (3.1) standard deviations for the **same sign W boson** production accompanied by the two hadronic jets with an integrated luminosity of $19.4\ fb^{-1}$. In addition, the cross section in fiducial region for $W^{\pm}W^{\pm}$ and WZ processes was also measured giving $\sigma_{fid}(W^{\pm}W^{\pm}jj) = 4.0^{+2.4}_{-2.0}(stat)^{+1.1}_{-1.0}(syst)\ fb$ with an expectation of $5.8 \pm 1.2\ fb$ for the former, and $\sigma_{fid}(WZjj) = 10.8 \pm 4.0(stat) \pm 1.3(syst)\ fb$ with an expectation of $14.4 \pm 4.0\ fb$ for the latter. Limits on aQGC operators S_0, S_1, M_0, M_1, M_6, M_7, T_0, T_1 and T_2 were reported as well [16].
The ATLAS collaboration reported the first evidence for the $\boldsymbol{W^{\pm}W^{\pm}jj}$ production and **electroweak-only** $\boldsymbol{W^{\pm}W^{\pm}jj}$ production with observed significance of 4.5 and 3.6 standard deviations respectively at an integrated luminosity of $20.3\ fb^{-1}$. In addition, the cross section measurements in the two fiducial regions is reported as well: inclusive region and the VBS region. The former is defined by requiring $p_T^l > 25\ GeV$, $|\eta| < 2.5$ and $\Delta R_{ll} > 0.3$. In addition, two jets with $p_T > 30\ GeV$ and $||\eta < 4.5$, separated from leptons by $\Delta R_{jl} > 0.3$ are also required. Jets are also required to have the invariant mass greater than $500\ GeV$. The VBS region is defined by requiring the two jets with the largest p_T to be separated in rapidity by $|\Delta y_{jj} > 2.4|$. The cross section of $\sigma_{fid} = 2.1 \pm 0.5(stat) \pm 0.3(syst)\ fb$ in the inclusive and $\sigma_{fid} = 1.3 \pm 0.4(stat) \pm 0.2(syst)\ fb$ in the VBS region is reported. The measured cross section in the VBS region was used to set limits on on aQGCs affecting vertices with four interacting W bosons [17].

In 2016 the ATLAS collaboration published their measurement of $\boldsymbol{W^{\pm}Z}$ production cross sections in pp collisions at $\sqrt{s} = 8\ TeV$ corresponding to the integrated luminosity of $20.3\ fb^{-1}$. In the VBS phase-space, the cross section was reported to be $\sigma_{VBS}(W^{\pm}Zjj) = 0.29^{+14}_{-12}(stat)^{+0.09}_{-0.1}(syst)\ fb$, where the SM prediction gives $0.13 \pm 0.01\ fb$. In addition, limits on anomalous gauge boson self-couplings were reported as well [18].
Another study was done by ATLAS in the same year aiming for the measurement of the $\boldsymbol{W^{\pm}W^{\pm}}$ production in events with two leptons (e or μ) with the same electric charge and at least two jets using the pp collision data at $\sqrt{s} = 8\ TeV$ and integrated luminosity of $20.3\ fb^{-1}$. In addition, the goal was to put more stringent limits on the aQGCs. The measured fiducial cross-section of $\sigma_{fid}^{incl.}(W^{\pm}W^{\pm}jj) = 2.3 \pm 0.6(stat) \pm 0.3(syst)\ fb^{-1}$ in the inclusive region was reported. The same was also measured in the VBS region giving $\sigma_{fid}^{VBS}(W^{\pm}W^{\pm}jj) = 1.5 \pm 0.5(stat) \pm 0.2(syst)\ fb^{-1}$. The expected sensitivity to α_4 and α_5 was improved significantly, compared to the previous ATLAS result, by selecting a phase-space region that is more sensitive to anomalous contributions to the $WWWW$ vertex. The paper reports the following expected (observed) limits: $-0.06 < \alpha_4 < 0.07$ ($-0.14 < \alpha_4 < 0.15$) and $-0.10 < \alpha_5 < 0.11$ ($-0.22 < \alpha_5 < 0.22$). The result constitutes a 35 % improvement in the expected aQGC sensitivity with respect to the previous results [19].
In the same year, the CMS collaboration reported a study on the electroweak-induced production of $\boldsymbol{W\gamma}$ with two jets in pp collisions at $\sqrt{s} = 8\ TeV$ and integrated luminosity of $19.7\ fb^{-1}$. The limits on the anomalous quartic gauge couplings were imposed as well. For the EW signal, the observed (expected) significance was found to be 2.7 (1.5) standard deviations, while for the EW+QCD signal significance of 7.7 (7.5) standard deviations was observed. The measured cross section in the fiducial region was found to be $10.8 \pm 4.1(stat) \pm 3.4(syst) \pm 0.3(lumi)\ fb$ for the

EW-induced $W\gamma$ + 2jets production and $23.2 \pm 4.3(stat) \pm 1.7(syst) \pm 0.6(lumi)\ fb$ for the total $W\gamma$ + 2jets production. Exclusion limits for aQGC parameters $f_{M,0-7}/\Lambda^4$, $f_{T,0-2}/\Lambda^4$ and $f_{T,5-7}/\Lambda^4$ were set at 95 % confidence level. This study reported the first limits on the $f_{M,4}/\Lambda^4$ and $f_{T,5-7}/\Lambda^4$ parameters [20].

In 2017 both the CMS and ATLAS collaborations reported the first measurements on the VBS in the $Z\gamma$ channel at $\sqrt{s} = 8\ TeV$ with the data corresponding to roughly $20\ fb^{-1}$.
The CMS reported an evidence for **EW $Z\gamma jj$** production with an observed (expected) significance of 3.0 (2.1) standard deviations. The fiducial cross section for EW $Z\gamma jj$ production was measured to be $\sigma_{fid}(Z\gamma) = 1.86^{+0.90}_{-0.75}(stat)^{+0.34}_{-0.26}(syst) \pm 0.05(lumi)\ fb^{-1}$. The fiducial cross section for combined EW and QCD $Z\gamma jj$ production of $\sigma_{fid}(Z\gamma) = 5.94^{+1.53}_{-1.35}(stat)^{+0.43}_{-0.37}(syst) \pm 0.13(lumi)\ fb^{-1}$ was reported as well. Both measurements are consistent with the theoretical predictions. In addition to previously imposed limits on the $f_{M0,1,2,3}$ and $f_{T0,1,2,9}$ parameters, the first observed (expected) limits on the neutral aQGC parameter f_{T8} were reported: $-1.8 < f_{T8}/\Lambda^4 < 1.8$ ($-2.7 < f_{T8}/\Lambda^4 < 2.7$). The limits on aQGC parameters are expressed in TeV^{-4} [21].
The ATLAS collaboration reported 2.0σ (1.8σ) observed (expected) significance for the production of the **EW $Z\gamma jj$** with the fiducial cross section of $\sigma_{fid}(Z\gamma) = 1.1 \pm 0.5(stat) \pm 0.4(syst)\ fb^{-1}$. The EWK+QCD cross section was also reported and quoted to be $\sigma_{fid}(Z\gamma) = 3.4 \pm 0.3(stat) \pm 0.4(syst)\ fb^{-1}$. Limits on the aQGCs are also discussed in the paper [22].
The first measurement of the **same-sign W** production at $\sqrt{s} = 13\ TeV$ was made by the CMS collaboration in 2017 using the data that corresponds to the integrated luminosity of $35.9\ fb^{-1}$. The observed significance of 5.5 standard deviations was reported, where a significance of 5.7 standard deviations was expected based on the standard model predictions. The ratio of the measured event yields to that expected from the standard model at leading-order was measured to be 0.90 ± 0.22. In addition, bounds were given on the structure of quartic vector boson interactions in the framework of dimension-eight effective field theory operators and on the production of doubly charged Higgs bosons [23].
In the same year the CMS collaboration did, for the first time, a search for the VBS in the **fully leptonic ZZjj** chanel at $\sqrt{13}\ TeV$. The process is measured with an observed (expected) significance of 2.7 (1.6) standard deviations. A fiducial cross section for the EW production is measured to be $\sigma_{EW}(ZZjj) = 0.40^{+0.21}_{-0.16}(stat)^{+0.13}_{-0.09}(syst)\ fb$ which is consistent with the SM prediction. Limits on the anomalous quartic gauge couplings were determined in terms of the EFT operators $f_{T0,1,2,8,9}$. These are shown in Table 1.3. More details on this study can bi found in [24]

Coupling	Exp. lower	Exp. upper	Obs. lower	Obs. upper	Unitarity bound
f_{T_0}/Λ^4	-0.53	0.51	-0.46	0.44	2.9
f_{T_1}/Λ^4	-0.72	0.71	-0.61	0.61	2.7
f_{T_2}/Λ^4	-1.4	1.4	-1.2	1.2	2.8
f_{T_8}/Λ^4	-0.99	0.99	-0.84	0.84	1.8
f_{T_9}/Λ^4	-2.1	2.1	-1.8	1.8	1.8

Table 1.3: Expected and observed lower and upper 95% CL limits on the couplings of the quartic operators T0, T1, and T2, as well as the neutral current operators T8 and T9. The unitarity bounds are also listed. All coupling parameter limits are in TeV^{-4}, while the unitarity bounds are in TeV. The table is taken from [24]

In 2018 the ATLAS collaboration reported their efforts in measuring the **EW WZ** boson pair production in association with two jets in pp collisions at $\sqrt{s} = 13\ TeV$, corresponding to an integrated luminosity of $36.1 fb^{-1}$. The observed (expected) significance of 5.3 (3.2) standard deviations was reported. The measured fiducial cross section for EW produc-

tion, including interference effects, was measured to be $\sigma_{fid}(W^{\pm}Z) = 0.57^{+0.14}_{-0.13}(stat)^{+0.05}_{-0.04}(syst)^{+0.01}_{-0.01}(lumi)\ fb$ [25].

In the following year the CMS collaboration published their results on the measurement of the **EW WZ** boson production and search for new physics in $WZ+\ 2jet$ events in pp collisions at $\sqrt{s} = 13\ TeV$. The measured (expected) significance of 2.2 (2.5) standard deviations was reported. The best-fit value for the $WZjj$ signal strength was used to obtain a cross section in the tight fiducial region and was measured to be $\sigma^{tight}_{fid}(WZjj) = 3.18^{+0.57}_{-0.52}(stat)^{+0.43}_{-0.36}(syst)\ fb$. This is compatible with the SM prediction of $\sigma_{pred} = 3.27^{+0.39}_{-0.32}(scale) \pm 0.15(PDF)\ fb$. In addition, results were also obtained in a looser fiducial region to simplify comparisons with theoretical calculations. The resulting WZjj loose fiducial cross section was measured to be $\sigma^{loose}_{fid}(WZjj) = 4.39^{+0.78}_{-0.72}(stat)^{+0.60}_{-0.50}(syst)\ fb$. This can be compared to the predicted value of $\sigma_{pred} = 4.51^{+0.78}_{-0.72}(scale) \pm 0.18(PDF)\ fb$. Finally, constraints on charged Higgs boson production and on aQGCs in terms of dimension-eight EFT operators were presented as well [26].
The first measurement of the **same-sign W** boson pair production at $\sqrt{s} = 13\ TeV$ by ATLAS collaboration was published in the same year. The background-only hypothesis was rejected with the significance of 6.5σ. The measurement of the fiducial cross section was reported as well giving $\sigma_{fid}(W^{\pm}W^{\pm}) = 2.89^{+0.51}_{-0.48}(stat)^{+29}_{-28}(syst)\ fb$ [27].
Another important paper in 2019 was published by the CMS collaboration where a report on a search for anomalous EW production of WW, WZ, **and** ZZ boson pairs in association with two jets in proton-proton collisions at $\sqrt{s} = 13\ TeV$ was presented. No excess of events, with respect to the SM background predictions, was observed. The events in the signal region were used to constrain aQGCs in the EFT framework. The study reported new constraints on operators $T_{S0,1}$, $T_{M0,1,6,7}$ and $T_{T0,1,2}$. This study was the first one to search for anomalous EW production of WW, WZ, and ZZ boson pairs in WV and ZV semi-leptonic channels at $13\ TeV$ and it improved the sensitivity of the previous CMS results at 13 TeV in fully leptonic decay channel by factors of up to seven, depending on the operator [28].

While the analysis presented in chapter 4 was being prepared for publishing, the ATLAS collaboration also published their results on VBS scattering in the channel with **two leptonically decaying Z bosons** accompanied by the two hadronic jets using the full Run 2 data at $\sqrt{s} = 13\ TeV$ corresponding to the integrated luminosity of $137\ fb^{-1}$. The background-only hypothesis was rejected with observed (expected) significance of 5.5σ (4.3σ). The fiducial cross section was reported to be $\sigma_{fid}(ZZjj) = 1.27 \pm 0.14\ fb$ where $1.14 \pm 0.04(stat) \pm 0.20(syst)$ was expected [29].
Another important study in 2020 was the first measurements of the **polarized same-sign W** boson pairs in association with two jets in pp collisions at $\sqrt{s} = 13\ TeV$ presented by the CMS collaboration. The signal was measured with an observed (expected) significance of 2.3 (3.1) standard deviations. An observed 95% confidence level upper limit on the production cross section for longitudinally polarized same-sign W boson production was reported to be $0.32^{+0.42}_{-40}\ fb$ in the $W^{\pm}W^{\pm}$ center-of-mass frame and $0.24^{+0.40}_{-0.37}\ fb$ in the parton-parton center-of-mass frame. Both measurements agree with theoretical predictions [30].

Finally, in 2021 the CMS collaboration measured the EW production of $Z\gamma$ associated with two jets at $\sqrt{s} = 13\ TeV$ with both expected and observed signal significance greater than five standard deviations. The fiducial cross section was reported to be $\sigma_{fid}(EWZ\gamma) = 5.21 \pm 0.52(stat) \pm 0.56(syst)\ fb$. Exclusion limits on the dimension-eight operators M_{0-7} and $T_{0-2,5-9}$ in the EFT framework at 95% confidence level was reported as well [31].

The work presented in this thesis work shows the first measurement, in the CMS collaboration, of the EW production of two leptonically decaying Z bosons accompanied with two jets. In addition, it shows the first prospective studies for the scattering of the longitudinal Z bosons at 14 as well as $27\ TeV$ conditions expected in future LHC runs.

Chapter 2

The large Hadron Collider and the CMS experiment

2.1 Preface to the chapter

In the first section of this chapter I will make a historic overview of events that led to construction of the World-largest particle detector in history; the Large Hadron Collider (LHC) at CERN. Next, I will briefly introduce the LHC machine and the largest experiments designed to collect and analyse the data produced in the proton-proton collisions at the LHC.

The analysis presented in this thesis uses data collected by the Compact Muon Solenoid (CMS) detector described in section 2.3. I will first describe the coordinate system that will be used throughout of this document. Next, I will briefly describe each of the subdetectors. I will finish the section with discussion of the trigger system used at CMS. In section 2.4 I will describe how muons and jets are reconstructed at CMS. This is done for electrons, in much more detail, in the next chapter. Finally, in section 2.5, I will introduce a reader with future plans for the LHC and the CMS detector. This section is a basis for following analysis discussed in Chapter 5.

2.2 The Large Hadron Collider (LHC)

2.2.1 A brief history History of LHC

By the 1970s, electron-positron colliders were very common in the high-energy physics community. However, hadron accelerators have been working in fixed-target mode. The exception for this was CERN's Intersecting Storage Rings (ISR), colliding protons with beam energies of up to 31.4 GeV [32].

In 1970 S. Glashow, J. Iliopoulos and L. Maiani postulated the existence of a fourth quark, the charm quark, which will only be discovered through the measurement of J/Ψ mesons in 1974 by electron-positron and fixed-target experiments.

A significant breakthrough in the QCD sector came in 1973 by D. Gross, D. Politzer and F. Wilczek who introduced a concept of asymptotic freedom, one of the emergent phenomena of the QCD responsible for the confinement of the quarks and gluons in the hadron. The third charged lepton, the τ, was discovered in 1975 by *SPEAR* collaboration and two years later, evidence for the beauty quark was obtained at *Fermilab*.

It was clear that there was a long way before the SM would be fully established and a new collider was

needed in order to investigate the origin of mass and to search for new physics beyond SM. The main agenda of the new machine would include the search for the Higgs boson, understanding the EWSB mechanism, the search for supersymmetry, the investigation of the phenomenology of the bottom and top quarks as well as the possible investigation of the quark-gluon plasma.

The option of a hadron collider in the tunnel of the Large Electron-Positron Collider (LEP) was first proposed by Sir John Adams in 1977. The former CERN director general suggested that the LEP tunnel be made wide enough to accommodate a superconducting proton collider of above 3 TeV beam energy. A serious discussion of a large proton-proton collider started with the first internal notes in 1983 and the CERN-ECFA workshop in March 1984. Already in May 1984, CERN director general Herwig Schopper presented plans at the meeting of the American Association for the Advancement of Science (AAAS) to build a superconducting proton-proton collider with the 5 TeV beam energy inside the existing LEP tunnel.

At the time, AAAS was discussing the possibility of building the Super-conducting Supercollider (SSC), so it was only natural for some rivalry between the two ideas to arise. On one hand, the cost of building the hadron collider at the CERN site was much lower than that needed for building the SSC. On the other hand, SSC supporters claimed that the SSC would have greater centre-of-mass energy, and thus is a better option physics-wise. While this was true, CERN argued that even with lower centre-of-mass energy, the collider at CERN would benefit from much higher beam luminosity. In the end, a new collider had to provide beam energy of at least 1 TeV.

In 1985, a Long-Range Planning Committee (LRPC) was established at CERN with Carlo Rubbia as chairman. Then, in January 1987, USA president Ronald Reagan approved the SSC which almost resulted in the CERN project being cancelled. In the same year, a second general workshop on the Physics at Future Accelerators was held at which the LRPC supported the project of building the LHC at CERN. The supposed LHC was planned to have luminosities between 10^{33} and $10^{34}\ cm^{-2}s^{-1}$ and maximum beam energy of 8 TeV. This plan was put to CERN Council in 1987 together with a plan to build 10 T dipole magnets. In a 1990 workshop Carlo Rubia, a CERN director general at the time, suggested beam energy of 7.7 TeV with a maximum luminosity of up to $5 \cdot 10^{34} cm^{-2}s^{-1}$ and the start of operation as early as 1998.

Three years later, CERN started the approval procedure for the LHC and in December it was presented in detail to CERN Council for the first time. In October 1993, the SSC was cancelled which put the LHC project in a doubt; a statement than can be best supported by quoting Chris Llewellyn-Smith, CERN director general at the time, who wrote: "I do not think that the LHC would have been approved if the SSC had not been cancelled. . . ".

In June 1994, approval was requested from the CERN Council for a machine of 14 TeV centre-of-mass energy with luminosity up to $10^{34}\ cm^{-2}s^{-1}$. Six months later, in December 1994, CERN Council approved a two-stage procedure for the LHC construction. A machine with a third of the magnets left out was to be constructed for commissioning in 2002 and operations at 9-10 TeV centre-of-mass energy were planned for 2004, with an upgrade to 14 TeV foreseen in 2008.

Within the next two years, several non-member states became interested in the project. In December 1997, the USA was ready to sign the agreement on contribution to the LHC. It seemed that even a single-stage machine was now possible. However, Germany unexpectedly decided to reduce its contribution to CERN by 9 % with the UK following its steps. In the end, a precedent amongst scientific projects happened; CERN was allowed to take out loans. This enabled the CERN Council to approve LHC construction in a single stage in December 1997. Finally, LHC was commissioned in 2008, with the first beams making a full circle around the ring on September the 10^{th}.

2.2.2 The LHC machine and physics experiments

Only a brief overview of the LHC operation procedure is discussed here. Much more detailed description of the LHC can be found elsewhere [33].

A full LHC accelerator complex is illustrated on Fig 2.1.

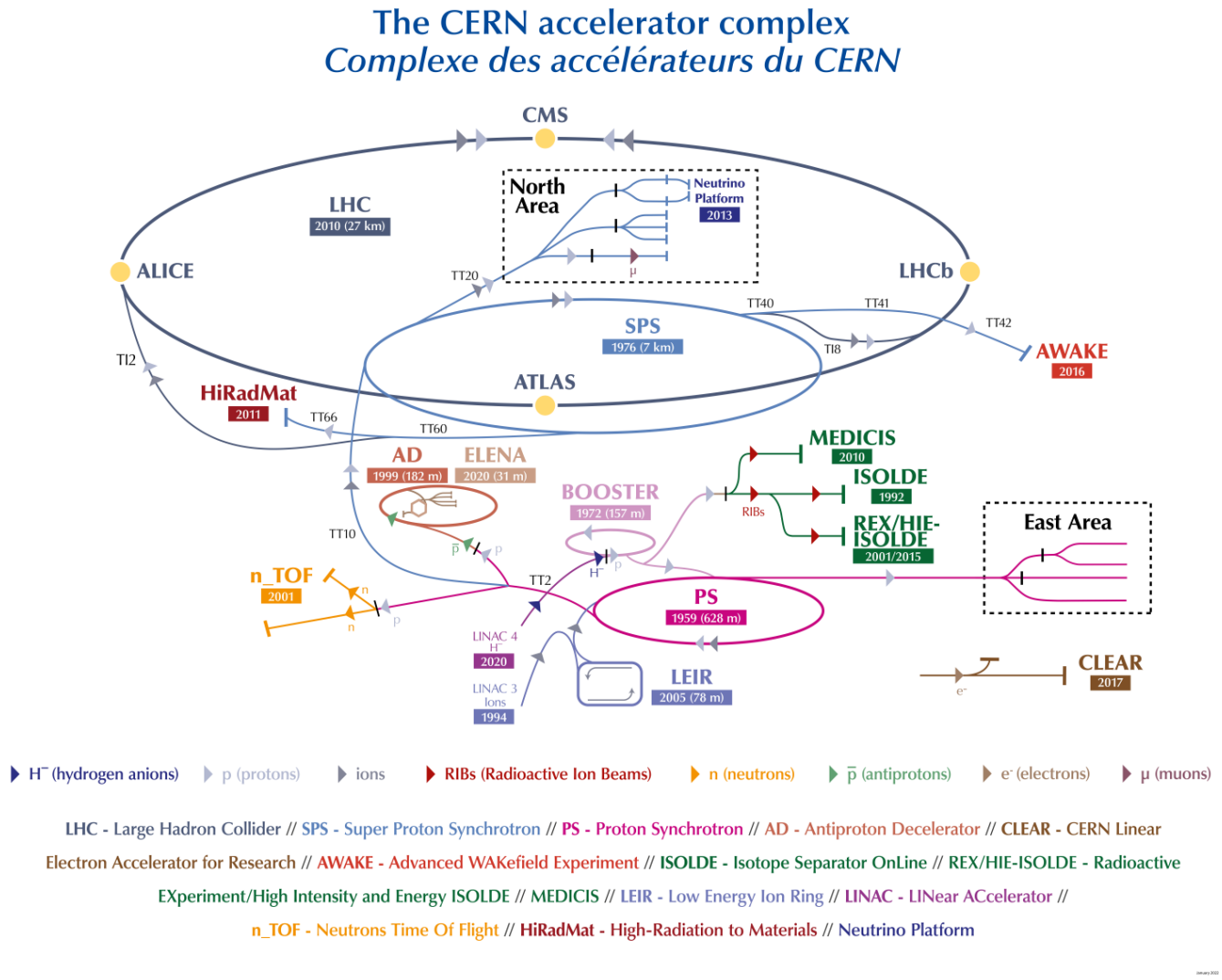


Figure 2.1: The LHC accelerator complex. Protons are first accelerated through the linear accelerator LINAC2. Before entering the largest ring, LHC, protons go through Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS). The illustration is taken from [34]

The process of particle acceleration starts by stripping electrons from the compressed hydrogen. This leaves only protons which are accelerated in the electric field. A DC voltage cannot be used as particles would be accelerated through the gap, but decelerated elsewhere. Thus, oscillating voltage is needed so that particles see accelerating voltage across the gap and, in the same time, the voltage has to cancel out as the particle goes around the rest of the accelerator. For this the radio frequency (RF) systems are used [35]. Initially, the protons are accelerated in the linear accelerator. In the following phase, protons enter Proton Synchrotron Booster where they are accelerated to the speed of 91.6 % of the speed of light. The next phase of acceleration is taking place in Proton Synchrotron (PS) where protons gain the speed of 99.9 % of the speed of light. Final phase of acceleration before the LHC is Super

Proton Synchrotron (SPS) which increases the energy of protons to 450 GeV.
Finally, protons are injected into the LHC which is placed roughly 170 m below the surface and has the circumference of 27 km. The LHC machine consists of two tubes in which protons circulate in opposite directions. In four locations the tubes cross and protons are collided.
To fill the LHC with protons, 12 cycles of SPS are needed. Each cycle of SPS required 3 to 4 cycles of PS. Since SPS and PS cycle times are 21.6 and 3.6 seconds respectively, LHC filling time is then around 4 minutes per beam. Since LHC requires additional 4 SPS cycles for the injection setup, and LHC operators need at least 2 minutes to adjust the machine settings, the injection time per beam for LHC then becomes approximately 16 minutes [36].

Protons are not spread uniformly along the beam, but are, instead, grouped together in, so called, *bunches*. Each bunch contains around $1.15 \cdot 10^{11}$ protons and is roughly 7.5 cm long and focused using quadrupole magnets into area of $16 \times 16\ \mu m^2$. At any given time, there are approximately 2000 proton bunches in a single beam.
Because of the small cross section of processes studied at the LHC, one would like to maximize the rate of events which depends on the cross section and the instantaneous luminosity, $\mathcal{L}$:

$$\mathcal{L} = \gamma \frac{f n_b N^2}{4\pi \epsilon_n \beta^*} R,$$

where $\gamma = \frac{E}{m}$ is the relativistic factor for protons, f is the revolution frequency, n_b is the number of bunches, N is the number of protons per bunch, ϵ_n is the normalized transverse beam emittance [37], β^* is the beam beta function at the collision point and R is a reduction factor due to the beam crossing angle at the interaction point [38]. Assuming nominal beam parameters, this yields instantaneous luminosity of order $10^{34}\ cm^2 s^{-1}$, two orders of magnitude larger than that of the Tevatron collider. The spacing between the two bunch crossings at the LHC is around 25 ns, which corresponds to the bunch crossing rate of 40 MHz.

Physics experiments at the LHC

One of the first meetings dedicated to physics experiments at the LHC was held in Barcelona in 1989 where the first predecessor of the Experiment for Accurate Gamma, Lepton and Energy measurements (EAGLE) experiment started forming. The next important workshop, Towards the LHC Experimental Programme, was held in Evian in 1992 where proto-collaborations described respective detector plans. In total, 12 proposals were made.
Four of proposals were made for general-purpose experiments: EAGLE, Apparatus with Superconducting Toroids (ASCOT), L3+1 and Compact Muon Solenoid (CMS). Next, three b-physics experiments were competing for approval: a Collider Beauty Experiment (COBEX), the LHB collaboration envisaged as a fixed-target experiment dedicated to the study of beauty hadrons and a CP-violation gas jet experiment (GAJET).
Three experiments were proposed for heavy-ion experiments: the one that will later be known as A Large Ion Collider (ALICE), the one that wanted to use the DELPHI detector from LEP, and the one that suggested a heavy-ion program for the CMS detector.
Amongst four multi-purpose detectors, only two would be accepted at the LHC. One of them was CMS. The other formed by merging ASCOT and EAGLE into A Toroidal LHC Apparatus (ATLAS) experiment. In January 1996, CMS and ATLAS were approved and the approval for construction was given on January 31^{st} 1997. Last large pieces of CMS and ATLAS were lowered into the experimental caverns on July 23^{rd} and February 29^{th} 2008, respectively.
In January 1994, the CERN LHC Experiments Committee (LHCC) recommended that COBEX, GAJET and LHB form a single collaboration. In September 1998, a technical proposal for the newly formed collaboration, LHCb, was accepted. Finally, ALICE was approved in February 1997.
After the four big experiments, **CMS**, **ATLAS**, **LHCb** and **ALICE**, were approved, three smaller experiments submitted

a Letter of Intent: the Total Cross Section, Elastic Scattering and Diffraction Dissociation Measurement at the LHC (**TOTEM**) experiment in 1997, Monopole and Exotics Detector at the LHC (**MoEDAL**) in 1998 and **LHCf** in 2003. In the following sections, design of the CMS detector is discussed.

2.3 The CMS experiment

The CMS detector is one of the two largest and general-purpose detectors at the CERN LHC. It is located roughly 100 meters bellow the surface near the French village of Cessy, between Lake Geneva and the Jura mountains. Many goals of the LHC include understanding the mechanism of EWSB and the Higgs mechanism as well as the search for a new physics that could manifest itself in terms of extra dimensions, forces, and symmetries. These, and many other phenomena, present strong arguments to investigate a TeV energy scale at the LHC.
Apart from the high energy conditions, a very high luminosity is expected at the LHC as well, with estimated 10^9 proton-proton collisions every second. This will result in around 1000 charged particles emerging from the interaction point every 25 ns. This results in very high levels of radiation requiring radiation-hard detectors and front-end electronics. Finally, the greatest challenge for the LHC now and in the future is the *pileup*, i.e. the average number of events per bunch crossing.

2.3.1 The Silicon Tracker system

2.3.2 The Electromagnetic Calorimeter

2.3.3 The Hadron Calorimeter

2.3.4 The solenoid magnet

2.3.5 The muon chambers

2.3.6 The trigger system

2.4 Physics objects reconstruction

2.5 The future of CMS and LHC

2.5.1 The high-granularity calorimeter

2.5.2 High-luminosity LHC

2.5.3 High-energy LHC

Chapter 3

Electron reconstruction and identification

3.1 Preface to the chapter

After discussing muons and jets at the end of the previous chapter, in this chapter, I will discuss electron reconstruction and selection. The focal point of this chapter is my work on electron efficiency measurements and derivation of electron scale factors for the full Run 2 period. These results, presented in section 3.4, are an important contribution to the HZZ working group and were used in the $H \to ZZ \to 4l$ analysis. The results presented here are also used in the VBS $ZZ \to 4l2j$ analysis presented in chapter 4 since the electron selection in the two analyses is identical.
In section 3.2.1 I will describe the formation of ECAL clusters and the importance of superclustering algorithm. In sections 3.2.2 through 3.2.7 I present an overview of the algorithms used to reconstruct electron trajectories, measure electron charge, momentum and energy. Finally, I describe energy corrections, combining momentum and energy measurements as well as the incorporation of discussed algorithms into the particle flow framework.
In section 3.3 I will describe vertex and impact parameter requirements on electrons as well as the identification and isolation algorithms. These are all used to define the electron selection criteria. Finally, in section 3.5 I will summarize the results of the electron efficiency measurements and scale factors.

3.2 Electron reconstruction

Data obtained using a 120 GeV electron test beam showed that electron impinging directly on the center of the ECAL crystal will leave 97% of its energy in a 5x5 crystal array centered around the hit crystal [39]. However, due to a large material budget in front of the ECAL, a single electron will often produce a shower of particles through bremsstrahlung and photon conversions before reaching it. Energy loss due to bremsstrahlung is directly dependent on the thickness of the material electron traverses. An electron will lose, on average, 33% of its energy before reaching ECAL if it propagates through the region with the least material budget that corresponds to $\eta \approx 0$. On the other hand, this goes up to 86%, on average, for electrons traversing through the region with the highest material budget, around $|\eta| \approx 1.4$. The first effect of bremsstrahlung is the spread of electron energy depositions in ECAL along the ϕ direction. In order to cope with this, several algorithms were studied in CMS. Additionally, radiation results in a sizable change of curvature of the electron trajectory along the preshower and tracker detectors. All this makes an energy measurement associated with the original electron a challenging task.

3.2.1 Clustering

In order to measure the energy of the primary electron, it is imperative to collect the energy of all particles from the shower produced by it's interaction with detector material. Due to the solenoidal magnetic field, the energy reaching ECAL will be spread along the ϕ direction. The spread in η will usually be negligible except for very low-p_T electrons ($p_T < 3\ GeV$). Two algorithms have been developed to recover the energy spread in ϕ: the "hybrid" algorithm for the ECAL barrel and the "multi-5x5" algorithm for the ECAL endcaps.

The hybrid algorithm exploits the geometry of the ECAL barrel and the shape of the shower to collect the energy deposits in a small window in η and extended window in ϕ [40]. The algorithm starts from the most energetic crystal in a region that has a transverse energy deposit larger than a predefined threshold ($E_T^{seed} > E_{T,\ min}^{seed}$). The crystal is referred to as the *seed crystal*. From here, 5x1 crystal "dominos" are added around the seed crystal in $\phi > 0$ and $\phi < 0$ direction as long as the transverse energy contained in the domino is larger than another threshold ($E_T^{5x1\ domino} > E_{T,\ min}^{5x1\ domino}$). Contiguous dominos around the seed crystal that contain an energy greater than a threshold $E_{min}^{domino-array}$ are grouped within, so called, *clusters*.

The multi-5x5 algorithm starts by finding crystal seeds defined as the ones having the highest energy amongst the four direct neighbours. Around each seed, starting with the one containing the highest energy, the energy is collected in clusters of 5x5 crystals. Since crystals in different clusters can overlap, a Gaussian shower profile is used to determine the fraction of the energy deposit to be assigned to each of the clusters [41].

In order to collect all the energy contained in the shower, corresponding to the energy of original electron and spread in ϕ direction, one final step must be done. In this step, clusters spread in ϕ are joined into *superclusters* (SCs). Two algorithms are combined in CMS for this task.

The first of the two, the "mustache" algorithm, especially useful for measuring very low energy deposits, relies solely on the information from the ECAL and the preshower detector. The algorithm starts by identifying the *seed cluster* around which other clusters are added if they fall in certain $\Delta\eta$ - $\Delta\phi$ window. Because of the solenoid magnet, the $\Delta\eta$ - $\Delta\phi$ region has a slight bend since the energy spread is more pronounced along ϕ than η, hence the algorithm name. The region defined by the mustache SC is optimized to contain 98% of the shower energy in several bins of cluster seed energy and position along the detector [42].

The second superclustering algorithm, the "refined" algorithm, exploits the tracker information to extrapolate the trajectories of bremsstrahlung photons and the tracks of converted electron pairs in order to decide whether a given cluster should belong to the SC. Although it uses a mustache algorithm as a starting point, it is capable of increasing or decreasing the number of cluster in the SC. The refined algorithm ultimately determines all ECAL-based quantities of electron and photon objects. An illustration of superclustering algorithm is shown in the top row of Fig. 3.1. The bottom row shows the reconstructed to generated energy ratio with and without the superclustering algorithm in the barrel and endcap regions.

3.2.2 Track reconstruction

When the energy loss due to bremsstrahlung radiation is significant, a classic KF approach [I will discuss Kalman filter in the global track reconstruction approach at the end of the 2nd chapter] will not be able to follow the changes in the curvature of the track and, thus, the tracks reconstruction efficiency will suffer. In order to better cope with non-Gaussian bremsstrahlung radiation losses, a dedicated algorithm, based on the *Gaussian Sum Filtering* (GSF), has been developed [44]. In essence, unlike KF which uses a single Gaussian to model the radiation loss, the GSF approach relies on mixing multiple Gaussians which provide to approximate the energy loss distribution. In essence, the electron trajectory is reconstructed by collecting the hits that belong to a track and fitting the track parameters using the GSF algorithm. In the end, the backward fit is applied in order to optimize the trajectory parameters.

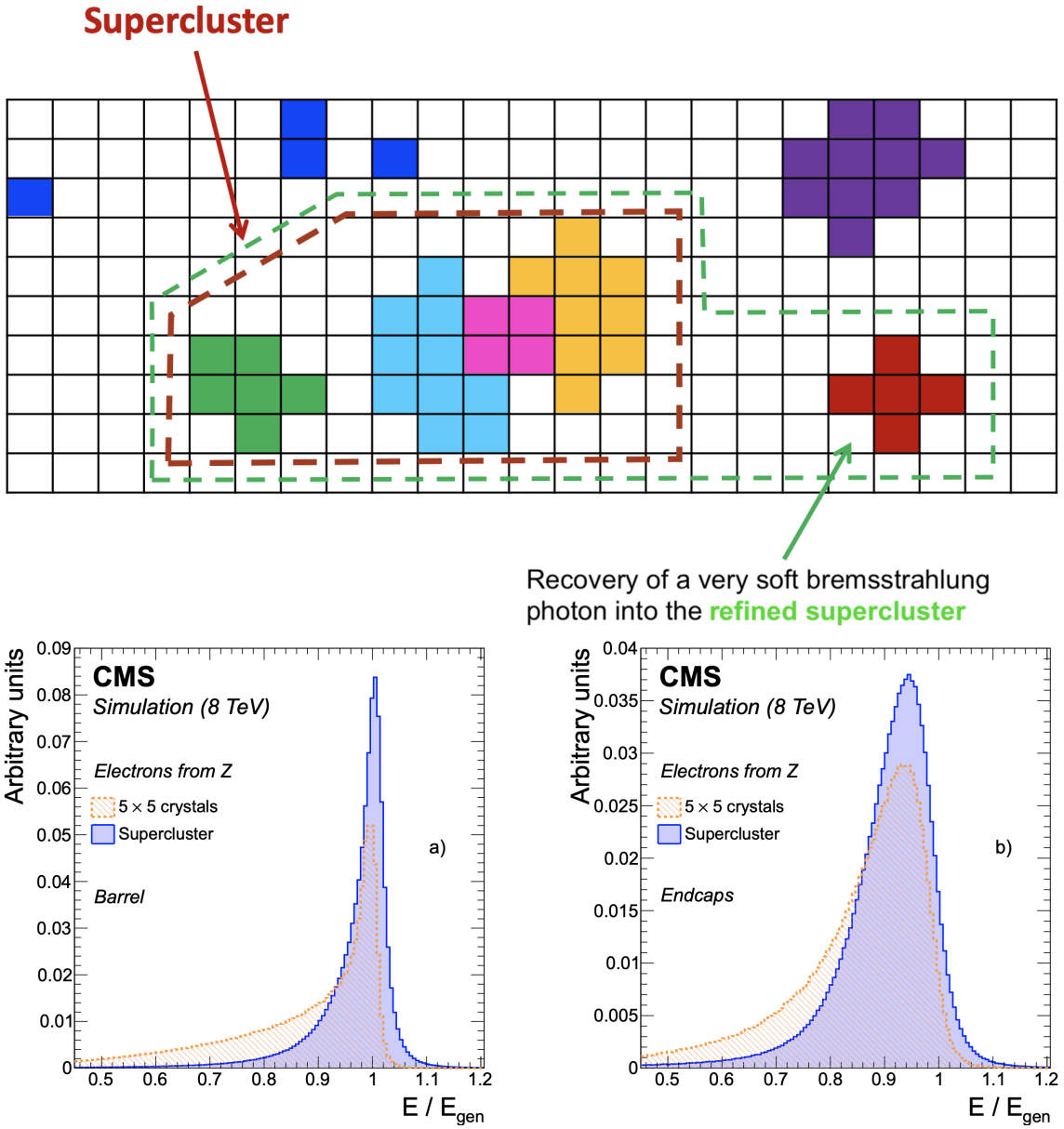


Figure 3.1: The top row shows an illustration of the superclustering algorithm. The bottom shows the comparison of the distributions of the ratio of reconstructed over generated energy for simulated electrons from the Z boson decays in the barrel (left), and the endcaps (right), for energies reconstructed using superclustering (solid histogram) and a matrix of 5x5 crystals (dashed histogram). No energy correction is applied to any of the distributions. The bottom plot is taken from Ref. [43]

Seeding

Due to more complex and, thus, CPU-intensive nature of the GSF algorithm, the track parameter estimation cannot be performed on all tracks reconstructed in the tracker. The first step in the track reconstruction is finding two or three hits in the tracker from which the track can be initiated. This is referred to as the *track seeding* and is of high importance since it can affect the reconstruction efficiency. The trajectory seeding can be either "ECAL-driven" or "tracker-driven".

The ECAL-driven approach first selects mustache SCs with transverse energy $E_{SC,\,T} > 4\ GeV$ and with $H/E_{SC} < 0.15$ where the E_{SC} is the SC energy and H is the sum of the HCAL tower energies within a cone of $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.15$ centered at the SC position. Hits in the pixel layers are predicted using the energy-weighted position of SCs, assuming the helical trajectory of electrons in the magnetic field (and therefore no radiation losses) [43]. Here, both positive and negative charge hypothesis is tested. The first hit is searched for starting from the innermost pixel layer outward until it is found. When two hits of a tracker seed are matched within a certain $\Delta z \times \Delta\phi$ ($\Delta r \times \Delta\phi$) window for the barrel pixel detectors (forward pixel disks and endcap tracker) to the SC-predicted trajectory, they are selected for seeding a GSF track. The $\Delta z \times \Delta\phi$ ($\Delta r \times \Delta\phi$) windows are defined to take into account the fact that the trajectories of electrons deviate from perfect helices due to radiation losses.

The tracker-driven trajectory seeding starts by going through all generic tracks (not limited to electrons) with $p_T > 2\ GeV$ obtained using the KF approach. A multivariate algorithm is then used to check whether any of these tracks are compatible with either SC position. If so, their seeds are used to initiate a GSF track.

The ECAL-driven approach is more suited for the high-p_T isolated electrons while the tracker-driven approach is designed to recover efficiency for low-p_T or nonisolated electrons. In the end, the two approaches are combined to give an overall $> 95\%$ seeding efficiency for simulated electrons originating from the Z boson decay. The performance of the seeding algorithms is checked with the data showing a good agreement [43].

Trajectory building

The collection of trajectory seeds obtained by combining the ECAL-driven and tracker-driven approach is used to initiate the reconstruction of electron tracks. Starting from each track seed, compatible hits in the next layers are searched for using the KF algorithm to iteratively build the electron trajectory, with the electron energy loss modeled using a Bethe-Heitler distribution [45]. This is done until the last tracker layer, unless no hit is found in the two consecutive layers. A minimum of five hits is required to create a track. For each layer, the compatibility between the predicted and measured hit is calculated using the χ^2 test. No cut on the χ^2 is imposed for electrons. Instead, many trajectories are grown in parallel and only the two best candidates, with the smallest values of χ^2, are kept in the end. It can happen that a tracker hit is assigned to multiple electron trajectories. In this case, the trajectory with less hits is dropped. Alternatively, if the number of the hits is the same, the track with higher χ^2 is dropped.

Track parameter estimation

When all the hits are collected, the GSF fit is performed to estimate the track parameters. For each hit, the GSF algorithm uses the parameters of all gaussians that enter the mixture to model the energy loss in that layer. One possible approach for the electron momentum estimate is to take the weighted mean of all the components. An alternative is to take only the most probable value (i.e. the mode) of the probability density function. The "weighted mean" approach provides the best sensitivity to the momentum change along the track due to radiation emission, while the "mode" approach is better suited for obtaining an estimation, least affected by bremsstrahlung emission, of the most probable track parameters [46]. The two approaches are compared in Figure 3.2 using the p_T/p_T^{gen} ratio for simulated electrons from the Z boson decay [41]. As can be seen from the figure, the peak of the GSF mean distribution is slightly biased towards the higher values of the p_T/p_T^{gen} spectrum. This shows that the bulk of the non-radiating electrons will have the wrongly assigned value of the transverse momentum in this approach. On the other hand, the GSF mode approach gives a better resolution around the peak. In addition, even though the p_T/p_T^{gen} distribution shows a pronounced tail towards the lower values of the spectrum, which is expected since

photon emission results in a more curved track than predicted from the most probable value, it is peaking exactly at unity meaning that, for electrons that don't radiate a lot, it assigns the correct value of the transverse momentum. For these reasons, the mode approach is used to characterize all the parameters of electron tracks.

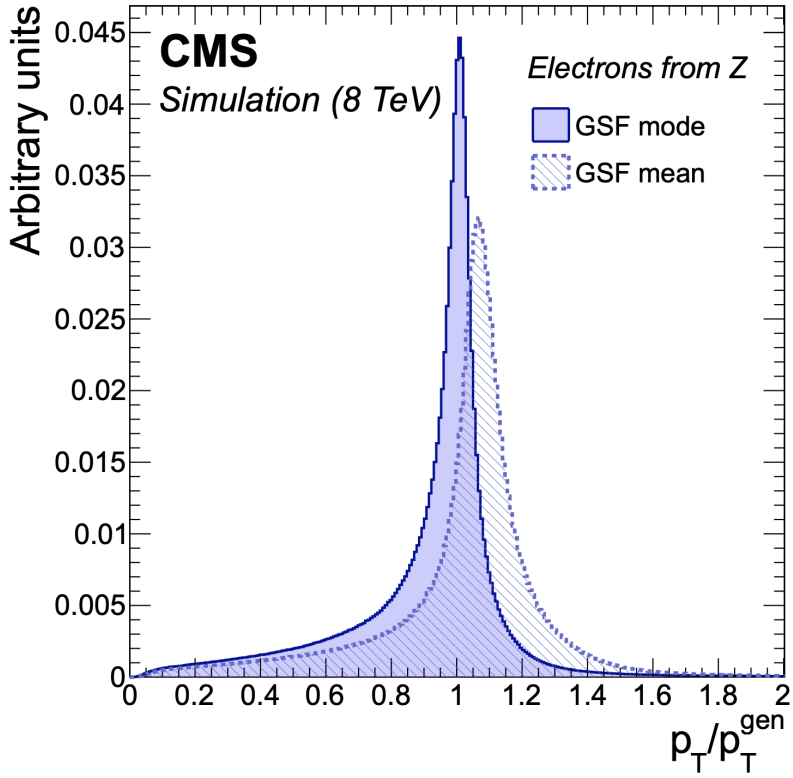


Figure 3.2: Distribution of the ratio of reconstructed over generated electron p_T in simulated $Z \to e^+e^-$ events, reconstructed through the most probable value of the GSF track components (solid histogram) and its weighted mean (dashed histogram). The figure is taken from Ref. [43].

Since the described trajectory building approach enables to collect hits up to the outermost layers of the tracker, it is possible to extract track parameters close to the surface of the ECAL. This is used to assess the fraction of the energy lost due to bremsstrahlung radiation using the momentum at the innermost layer position (p_{in}) and the momentum at the outermost layer position (p_{out}). This variable, defined as $f_{brem} = 1 - \frac{p_{out}}{p_{in}}$, is used to define electron classes (see section 3.2.4) and, also, in the MVA-based electron identification (see section 3.3.2). Finally, it is used to assess whether the material budget is simulated properly as a function of η (since it measures the amount of bremsstrahlung).

3.2.3 Charge estimation

The electron charge measurement can become more complex in case of early bremsstrahlung followed by photon conversion. The resulting electromagnetic showers can lead to very complex hit patterns, and the contributions from conversions legs can be wrongly included in the fitting of the electron track. Thus, three methods are combined in CMS to minimize probability of mismeasuring electron charge:

1. sign of the GSF track curvature
2. curvature of the associated KF track matched to a GSF track when at least one hit is shared in the innermost region

3. sign of the difference in ϕ between vector joining the beam spot to the SC position and the vector joining the beam spot and the first hit of the electron GSF track

The electron charge is the majority vote of the three charge measurements. The misidentification probability is predicted by the simulation to be 1.5% for reconstructed electrons from Z boson decays and is an improvement by a factor two with respect to GSF track curvature measurement only. In addition, misidentification probability at very large values of $|\eta|$ are predicted to be below 7%. Even higher purity can be achieved, at the price of a p_T- and η-dependent efficiency loss, by requiring all three charge measurements to agree. In that case, a misidentification probability of less than 0.2% in the central part of the barrel, less than 0.5% in the outer part of the barrel, and less than 1% in the endcaps are achieved. This comes at the price of $\approx 7\%$ efficiency loss for electrons coming from Z boson decays. All predictions discussed above closely match the observations in data [43].

3.2.4 Classification

The previously defined variable, f_{brem}, together with a bremsstrahlung fraction in the ECAL defined as $f_{brem}^{ECAL} = 1 - \frac{E_{ele}^{PF}}{E_{SC}^{PF}}$ are used to define five classes of electrons. Here, E_{ele}^{PF} and E_{SC}^{PF} are the electron-cluster energy and SC energy measured with PF algorithm respectively. [PF algo will be discussed in ch. 2]

1. The "golden" electrons are those with little bremsstrahlung and thus will provide the most accurate estimation of momentum. They are defined by a SC built from a single ECAL cluster and $f_{brem} < 0.5$.

2. "Big-brem" electrons have a large amount of bremsstrahlung radiated in a single step, either very early or very late along the electron trajectory. They are defined by a SC built from a single ECAL cluster and $f_{brem} > 0.5$.

3. "Showering" electrons have a large amount of bremsstrahlung radiated all along their trajectory. They are defined by a SC built from several ECAL clusters.

4. "Crack" electrons are defined by a SC seed crystal adjacent to an η boundary between the modules of the ECAL barrel, between the ECAL barrel and endcaps, or at the high $|\eta|$ edge of the endcaps.

5. "Bad track" electrons are defined by a significantly larger calorimetric bremsstrahlung fraction compared to the track bremsstrahlung fraction ($f_{brem}^{ECAL} - f_{brem} > 0.15$). These are electrons with a poorly fitted track in the innermost part of the trajectory.

3.2.5 Energy corrections

The idea behind clustering energy deposits in SCs is to reduce energy losses due to bremsstrahlung and photon conversions and thus improve upon the energy estimation of the primary electron. However, several effects can impact the estimation of SC energy. These are the energy leakage in ϕ or η outside SC, the energy leakage into the gaps between the crystals, modules, supermodules, as well as the transition region between the barrel and the endcaps, the energy leakage into the HCAL, the energy loss due to interactions in the material before the ECAL and the additional energy coming from pileup interactions. All these effects result in systematic variations of the energy measured in the ECAL and degrade the electron energy measurement. In order to improve the resolution, different multivariate techniques have been developed in CMS. The regression technique uses simulated events only, while the energy scale and resolution corrections are based on the comparison between data and simulation. Since the details of this procedure are not essential for understanding the work presented in this thesis, only the key elements are discussed here. An interested reader can find more details in Ref. [41].

Energy corrections with multivariate regressions

The multivariate regression for the SC energy correction defines a target as the ratio between the true energy of an electron and its reconstructed energy. Therefore, the regression prediction is used as the correction factor applied to the measured energy to obtain the best estimate of the true energy. The regression is implemented via a gradient-boosted decision tree (BDTG) (for details on BDTG see section 4.6.2) with a double-sided Crystal ball (DSCB) function [47] used in the regression algorithm. Through the training phase, the regression algorithm performs an estimate of the parameters of the DCB probability density as a function of the input vector of the object and event characteristics. The electron energy correction is obtained by applying the regression algorithm in three steps. A first regression gives the correction of the SC energy, a second regression gives an estimate of the SC energy resolution and the last regression yields the final energy value correcting the combined energy estimate from the SC and the electron track information.

Energy scale and smearing corrections

Even after introducing energy corrections with the multivariate approach discussed above, small differences remain between the data and the simulation an example being a resolution which is better in the simulation than in the data. Hence, an additional smearing has to be applied to the electron energy in simulations so that the peak position of the Z boson mass in the simulation matches that in the data. The electron energy scale is corrected by varying the scale in the data to match that observed in simulated events. The magnitude of the final correction is below 1.5% with an uncertainty as small as 0.1% for the barrel and 0.3% for the endcap.
These corrections are obtained using the "fit method" and the "smearing method", both developed in Run 1. In the former, an analytic fit is performed to the invariant mass distribution of the Z boson by convoluting the Breit-Wigner (BW) and the one-sided Crystal ball (OSCB) function. The latter utilizes the simulated Z boson invariant mass distribution as a PDF in a maximum likelihood fit to the data. The difference in width between the data and simulation is described by an energy smearing function applied to the simulation.

The final electron energy resolution, after all corrections are applied, ranges from 2 - 5% depending on the electron η and the amount of energy lost due to the bremsstrahlung. The performance of energy corrections in data is shown in Figure 3.3 with the $Z \to ee$ mass distribution before and after corrections. The result is a peak in data that is better matched to the one in the simulation. The improvement is more pronounced in the endcap region. Additionally, one can see on the same figure an improvement in the energy resolution after applying energy corrections.

3.2.6 Combining energy and momentum measurements

The electron momentum estimate can be improved by combining the corrected energy measurements with the track momentum measurement. At low electron energies ($\lesssim 15\ GeV$), and for electrons near gaps in detectors, the track momentum is, in general, more precisely measured than the ECAL SC energy. The two approaches are combined using a regression technique that defines a weight w that multiplies the track momentum in a linear combination with the estimated SC energy as $p = wp + E_{SC} \cdot (1 - w)$. The variables used to train the regression BDT are the corrected ECAL energy, the track momentum estimate, the uncertainties of the two, the ratio of the corrected ECAL energy over the track momentum as obtained from the track fit, the uncertainty in this ratio, and the electron category, based on the amount of bremsstrahlung [48].
After combining the two estimates, the bias in the electron momentum is reduced in all regions and all electron classes. An exception are the showering electrons in the endcaps, where the bias becomes slightly worse. The effective

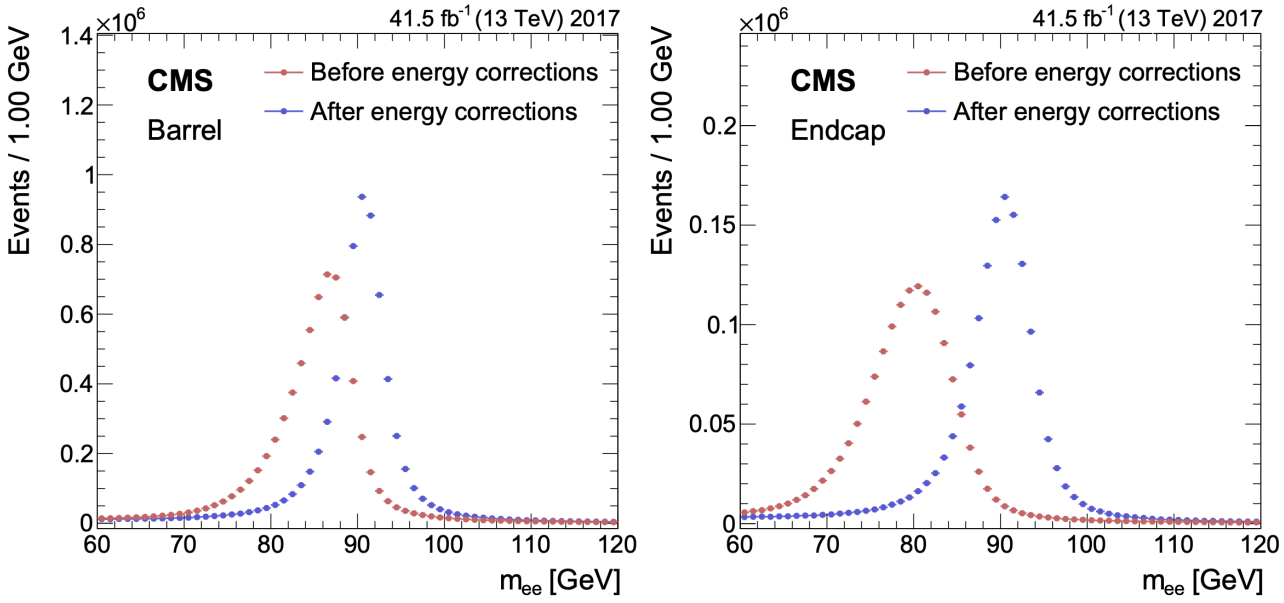


Figure 3.3: Dielectron invariant mass distribution in data before and after energy corrections (regression and scale corrections) for barrel (left) and endcap (right) regions for $Z \to ee$ events. The figure is taken from Ref. [41]

resolution, defined as the smallest interval around the peak position containing $\approx 68\%$ of the distribution, in the combined electron momentum can be seen in Figure 3.4 as a function of its p_T compared to the effective resolution of the corrected SC energy for golden electrons in the barrel and for showering electrons in the endcaps. The improvement is around 25% for electrons with $p_T \approx 15\ GeV$ in the barrel. For the golden electrons with $p_T < 10\ GeV$, this can reach 50%. More details on this topic can be found in Ref. [43].

3.2.7 Integration with particle-flow framework

Contrary to the Run 1, where different reconstruction algorithms were used for electrons, electron reconstruction in CMS is now fully integrated into the PF framework. ECAL clusters, SCs, GSF tracks and generic tracks associated with electrons, as well as the conversion tracks and associated clusters, are all imported into the PF algorithm that links the elements together into blocks of particles. These blocks are resolved into electron and photon objects, starting from either a GSF track or a SC, respectively. No difference between electrons and photons exist at this stage. Electron and photon objects are built from the refined SCs based on loose selection criteria (for clarification on selection criteria see section 3.3). All objects that pass the selection criteria, and have an associated GSF track, are labeled as electrons. Objects that pass the selection criteria but don't have a GSF track associated with them are identified as photons. This collection is referred to as the e/γ collection.

To separate electrons and photons from hadrons in the PF framework, a tighter selection is applied to decide if they are accepted as an electron or an isolated photon. If the object passes both the electron and the photon selection criteria, its object type is determined by whether it has a GSF track with a hit in the first layer of the pixel detector. If it fails the electron and photon selection criteria, its ECAL clusters and generic tracks are considered to form neutral hadrons, charged hadrons or nonisolated photons in the PF framework.

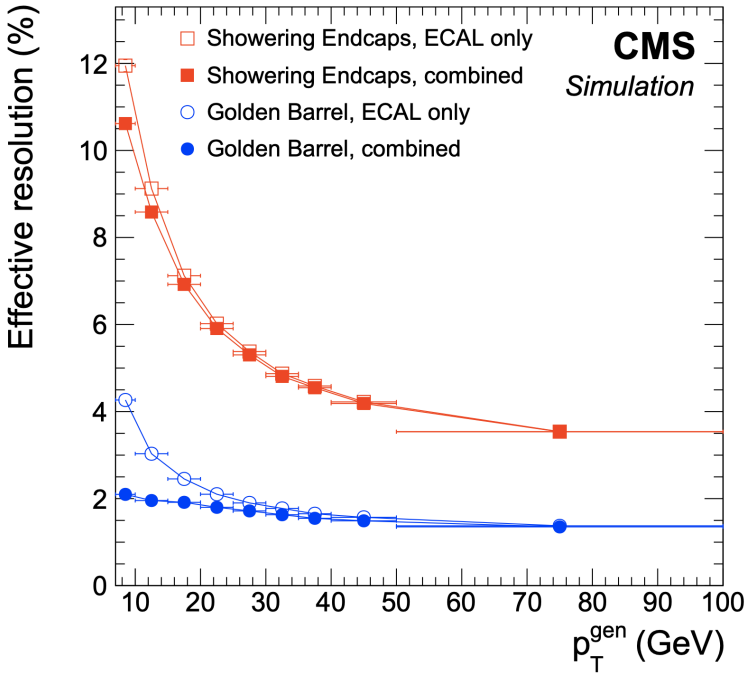


Figure 3.4: Effective resolution, as a function of the generated electron p_T, in electron momentum after combining the corrected SC energy and momentum estimates (solid symbols) compared to that of the corrected SC energy (open symbols). Golden electrons in the barrel (circles) and showering electrons in the endcaps (squares) are shown as examples. Electrons are generated with uniform distributions in η and ϕ and the resolution is shown after applying the spreading corrections. The figure is taken from Ref. [43]

3.3 Electron selection

The main goal of the electron selection is to reduce the rate of fake electrons coming from various sources and thus contaminating the analysis. The selection criteria described in this section is used for the $H \to ZZ^* \to 4l$ analysis, where the lepton efficiency enters the selection with the power of four. Full details on selection criteria can be found on [49, 50]. Only the main points needed to understand the electron efficiency measurements discussed in section 3.4 are outlined here. In general, electron selection can be split into three blocks: kinematic and impact parameter selection, electron identification and electron isolation.

3.3.1 Kinematic and impact parameter selection

Because of the tracker acceptance, only electrons with $|\eta| < 2.5$ are considered in the analysis. Additionally, in order to mitigate the effect of the background, especially in the very low p_T region, as well as to account for the difficulties in reconstructing tracks and measuring momentum in this region, only electrons with $p_T > 7\ GeV$ are kept.

Loose vertex requirements defined as

$$|d_{xy}| < 0.5\ cm$$

$$|d_z| < 1\ cm$$

where $|d_{xy}|$ refers to the absolute value of the impact parameter, with respect to the primary vertex, in the transverse plane, and $|d_z|$ is the absolute value of the impact parameter along the z axis are imposed on electron candidates.

Next, impact parameter selection is introduced in order to reduce the background that doesn't originate from the primary vertex but, rather, from bremsstrahlung photons, photon conversions and heavy flavor decays. In general, tracks of these secondary electron candidates (background in this analysis) will not point to the primary vertex and this can be used to separate them from primary electrons. The *impact parameter*, IP_{3D}, is defined as the algebraic distance, in the 3-dimensional space, between an electron candidate and the primary vertex. However, instead of the impact parameter, the significance of the impact parameter, SIP_{3D}, is used by dividing the impact parameter by its uncertainty. The selection than requires

$$|SIP_{3D}| = \frac{|IP_{3D}|}{\sigma_{IP_{3D}}} < 4$$

3.3.2 Identification

By imposing the selection on the significance of the impact parameter, backgrounds originating from secondary vertices are suppressed. However, hadronic jets (and remaining photon conversions) can mimic genuine electron energy depositions in the calorimeter. In order to distinguish signal electrons from the backgrounds such as reconstructed tracks from $\pi^{\pm}$ in vicinity of an electromagnetic cluster from $\pi^0 \to \gamma\gamma$, a complex *electron identification* algorithm was designed. In the CMS, two approaches are used for the electron identification: the *cut-based* approach and the *MVA-based* approach.

Cut-based electron identification

In the cut-based approach one applies cuts on a set of tracker and ECAL related variables. Four working

points, corresponding to different signal efficiencies, are used in CMS. The "veto" working point corresponds to an average signal efficiency of about 95%. The "loose" working point corresponds to a signal efficiency of around 90% and is used in analyses with low backgrounds to electrons. The "medium" working point corresponds to an average signal efficiency of around 80%. Finally, the "tight" working point corresponds to roughly 70% signal efficiency and is used in analyses where large background contamination is expected.

MVA-based electron identification

Since the $H \rightarrow ZZ^* \rightarrow 4l$ channel requires a high signal efficiency, a loose ID, capable of reducing fake electrons, in particular in the low-p_T region, was developed. It uses a set of variables, summarized in Table 3.1, to produce a single MVA classifier using boosted decision tree (BDT) techniques. Three main categories of variables enter the training of the BDT:

- observables based on the shape of the ECAL clusters, example being the width of the cluster, specifically in the η direction
- observables based on the tracking information such as f_{brem} describing the energy lost through bremsstrahlung
- observables that describe the quality of the matching between the supercluster and the track, example being the ratio of the supercluster energy over the track momentum

The output of the BDT training is the score for each electron candidate, which is peaking close to unity for signal electrons and to zero for background electrons.

	Observable	Definition
Cluster shape	$\sigma_{i\eta i\eta}$	Energy-weighted standard deviation along η within a 5×5 block of crystals centered on the highest energy crystal of the seed cluster
	$\sigma_{i\phi i\phi}$	Similar to $\sigma_{i\eta i\eta}$ but in the ϕ direction
	η width	SC width along η
	ϕ width	SC width along ϕ
	$1 - E_{5\times1}/E_{5\times5}$	$E_{5\times5}$ is the energy computed in the 5×5 block of crystals centered on the highest energy crystal of the seed cluster, and $E_{5\times1}$ is the energy computed in the strip of crystals containing it
	R_9	Energy sum in the 3×3 block of crystals centered on the highest energy crystal, divided by the SC energy
	H/E	Energy collected by the HCAL towers within a cone of $\Delta R = 0.15$ centered on the SC position, divided by the SC energy
	E_{PS}/E_{raw}	Energy fraction deposited in the preshower detector divided by the raw SC energy
Tracking	$f_{brem} = 1 - p_{out}/p_{in}$	Fractional momentum loss as measured by the GSF fit. The momenta p_{in} and p_{out} are the innermost and outermost estimates respectively.
	N_{KF}	Number of hits of the KF track (when reconstructed)
	N_{GSF}	Number of hits of the GSF track
	χ^2_{KF}	Goodness of fit of the KF track (when reconstructed)
	χ^2_{GSF}	Goodness of fit of the GSF track
	$N_{miss.\ hits}$	Number of expected but missing inner hits in the first tracker layers
	$P_{conv.}$	Fit probability for a conversion vertex associated with the electron track
Track-cluster matching	E_{SC}/p_{in}	Ratio of the supercluster energy to the track momentum at the innermost track position
	E_{ele}/p_{out}	Ratio of the energy of the cluster closest to the electron track and the track momentum at the outermost track position
	$\frac{1}{E_{SC}} - \frac{1}{p}$	Deviation of the SC energy from the electron momentum obtained by combining ECAL and tracker information
	$\Delta\eta_{in} = \lvert\eta_{SC} - \eta_{in}\rvert$	Distance between the energy-weighted center of the SC and the expected shower position as extrapolated from the GSF trajectory state at the vertex
	$\Delta\phi_{in} = \lvert\phi_{SC} - \phi_{in}\rvert$	Same as $\Delta\eta_{in}$, but in the ϕ direction
	$\Delta\eta_{seed} = \lvert\eta_{seed} - \eta_{out}\rvert$	Distance between the η of the seed cluster and the expected shower position as extrapolated from the GSF trajectory state of the outermost hit

Table 3.1: List of input variables, divided into three categories, that enter the BDT training for the MVA-based electron identification used in the $H \to ZZ \to 4l$ analysis.

3.3.3 Isolation

Fake electrons from hadronic jets can be mitigated by means of *isolation*. Prompt electrons are characterized by the absence of activity around them. The isolation can be defined using the PF candidates reconstructed with a momentum direction within predefined isolation cone.
The isolation variables are obtained by summing the transverse momenta of charged hadrons, neutral hadrons and photons within an isolation cone defined by $\Delta R = 0.3$ and subtracting the contribution of the pileup. The combined per-electron isolation is constructed by combining different isolation related observables:

$$I = \sum_{\substack{charged \\ hadrons}} p_T + max\left[0, \sum_{\substack{neutral \\ hadrons}} p_T + \sum_{photons} p_T - p_T^{PU}\right]$$

where $p_T^{PU} = \rho \times A_{eff}$ is the pileup correction for electrons calculated following the *FASTJET* technique [51–53].

The problem with using isolation variable as defined above comes from the consideration of fake electrons in the background. For example, the p_T of the fake lepton inside a jet increases with the energy of the jet. If the energy of the jet is small, the activity surrounding the fake electron will be small and cutting simply on the p_T could lead to fake electron being wrongly classified as an isolated electron. Therefore, the thresholds applied on the isolation quantities should depend on the particle energy. For this reason, the *relative isolation* is introduced and defined by

$$I_{rel} = \frac{I}{p_T^e}$$

Electrons with $I_{rel} < 0.35$ are considered isolated. Those electrons that also satisfy the impact parameter and identification requirements are used to select Z boson candidates in the $H \to ZZ \to 4l$ analysis.

3.4 Electron efficiency measurements

In the previous section, electron selection requirements were defined. Depending on the analysis, one may need different selection criteria, which lead to different electron efficiency. Therefore, it is crucial to quantify the efficiency of the chosen selection criteria since these effects have to be included in the analysis. The same has to be done for the reconstruction procedure discussed in the first part of the chapter. One approach can be to estimate efficiencies using the simulations. However, because the detector effects aren't described perfectly by the simulation, this can lead to undesired bias in the estimation of the reconstruction or selection efficiency. In order to circumvent this issue, efficiencies are extracted directly from the data using the *Tag and Probe* (TnP) approach. For the electron efficiency measurements, the $Z \to ee$ channel is used to estimate the electron selection efficiencies.
In addition, the agreement between efficiencies in the data and simulation varies between the different regions of the detector and for different values of the electron p_T. This results in some disagreement, in most variables used in the analysis, between the simulation and the data. The differences in efficiency between the data and simulation are measured in various η and p_T bins using the TnP approach and *scale factors* are obtained by dividing the efficiency in the data by that in the simulation. These are applied to the simulation in order to correct for the efficiency difference.

3.4.1 Tag and Probe method

In order to measure the efficiency of a desired selection, one needs a pure sample of electrons. This can be achieved by using the decay products of a familiar resonance such as the Z boson which ensures a high purity. The Tag and Probe (TnP) approach is used in this analysis to measure the electron selection efficiency.

The TnP method starts with selecting a set of Z bosons that decay into pairs of oppositely charged electrons. These pairs of electrons are required to have a mass within a window $60\ GeV < m_{ee} < 120\ GeV$ which ensures that genuine $Z \to ee$ decays are selected. However, some background events, coming mainly from the W+jets or QCD multijet processes, may pass this requirement as well. In order to make sure that the efficiency is measured for signal electrons, one electron, referred to as the *tag*, is required to pass a very tight selection. The corresponding opposite sign electron, referred to as the *probe*, is used to probe the efficiency of the selection under consideration. The efficiency of the selection criteria is defined as the number of probes that pass the selection with respect to the total number of probes:

$$\epsilon_{sel.} = \frac{N_P}{N_P + N_F}$$

where N_P is the number of the passing probes and N_F is the number of the failing probes. The probes are first split into several p_T and η bins defined in a way that ensures enough statistics inside every bin. The efficiency is then calculated for each bin separately.

One way to implement the efficiency measurement is to use the cut-and-count approach in which one simply counts the number of probes passing the selection and the number of probes that fail the selection. The efficiency is than easily calculated from the expression above. This can be a good approach when one is certain that there is no background contamination. Since this is the case in the simulation, the cut-and-count approach is used as the nominal method to estimate the efficiency in the simulation.
However, this is, in general, not the case in the data since very loose requirements are imposed on the probe. Therefore, another technique is used as the nominal signal efficiency measurement approach in the data. In this approach, both passing and failing probes are fitted, for each bin separately, using either the analytical function or the template extracted from the simulation. The nominal signal model is based on the Drell-Yan simulation used to obtain the template which is convoluted with a Gaussian distribution to account for the differences in resolution between the simulation and the data.

If no kinematic restrictions would be imposed on the tag and probe pairs, the dilepton mass distribution away from the resonance would be described nicely by a falling exponential function. However, cuts imposed on kinematic variables distort the invariant mass, m_{ee}, distribution in every bin in a way that is accounted for by using an error function. Thus, the background is described by a falling exponential function multiplied with an error function:

$$f(m_e e) = \mathrm{erf}(a - m_{ee}) \cdot e^{-d \cdot (m_{ee} - c)}$$

where a and c (b and d) are expressed in units of GeV (GeV^{-1}) and are free parameters in the fit.

The uncertainty on each efficiency measurement is obtained from the quadratic sum of the statistical uncertainty obtained from the fit and a systematic uncertainty. The leading source of the systematic uncertainty is the modeling of the signal and background contributions. The uncertainty in the signal model is obtained by replacing the template fit with a Breit-Wigner function convoluted with a one-sided Crystal ball (OSCB) function, while the

uncertainty in the background model is obtained by using a falling exponential function instead of the product of a falling exponential function and an error function. For some low-p_T bins, a Chebyshev polynomial multiplied by a Gaussian (nominal signal), a Gaussian convoluted by a CB function (alternative signal), or a Gaussian multiplied by an exponential function (alternative background) was used in order to obtain a better fit.

The number of passing and failing probes in each bin is defined by the area between the signal and background functions. Examples of nominal signal fits in the data are shown on the top of the Fig. 3.5 for the (passing probe, failing probe) distributions for two different $(p_T,\ \eta)$ bins. The alternative signal fits in the simulation are shown on the bottom of the figure for the (passing probe, failing probe) distributions in the same $(p_T,\ \eta)$ bins. The fitted signal contributions are shown in red, while the fitted background contributions are shown is blue.

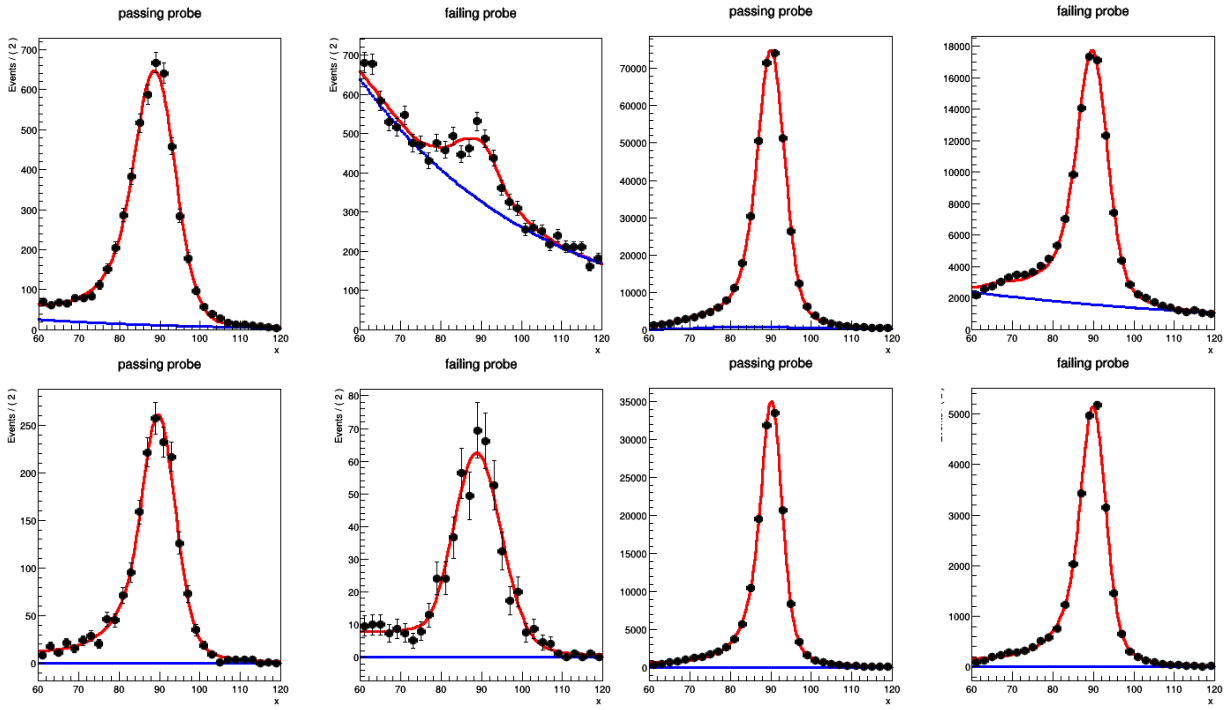


Figure 3.5: Example of the nominal signal fits in the data are shown on the top of the figure for the (passing probe, failing probe) distributions for two different $(p_T,\ \eta)$ bins. The alternative signal fits in the simulation are shown on the bottom of the figure for the (passing probe, failing probe) distributions in the same $(p_T,\ \eta)$ bins. The left-hand side plots show the (passing probe, failing probe) distributions in the $(2.00 < |\eta| < 2.5,\ 7\ GeV < p_T < 11\ GeV)$ bin and the right-hand side plots show the same in the $(2.00 < |\eta| < 2.5,\ 20\ GeV < p_T < 35\ GeV)$ bin. The fitted signal contributions are shown in red, while the fitted background contributions are shown in blue.

The efficiency measurements in each bin for the data and the simulation are used to derive *scale factors* (SFs) which are defined as the per-bin ratio of the efficiency under study obtained in the data divided by the efficiency in the simulation:

$$SF(p_T,\ \eta) = \frac{\epsilon_{data}(p_T,\ \eta)}{\epsilon_{MC}(p_T,\ \eta)}$$

These are used to scale the simulations to account for the different efficiency between the data and the simulation and therefore mitigate any discrepancies between the two left from the imperfect modeling.

Finally, the overall electron efficiency can be expressed as the product of the trigger efficiency, reconstruc-

tion efficiency and selection efficiency. The discussion about the trigger is rather involved and is not needed to follow the study presented here. An interested reader can found the details on the trigger performance in [54, 55]. The reconstruction efficiency is also measured using the TnP technique where the tag is an electron coming from the decay of the Z boson, and the second leg of the TnP are the SCs used to measure the efficiency (probes). One then counts the number of SCs that are promoted to electron (passing probe) with respect to the total number of probes. The largest source of uncertainty in the reconstruction efficiency measurements comes from the association of the SCs to the track. Since every analysis in CMS uses reconstruction efficiencies, these are produced centrally by the CMS collaboration and provided to all analyses containing electrons in the final state.

3.4.2 Electron selection efficiency in 2016, 2017 and 2018

The selection efficiency was derived for each data-taking period separately. The working points (WPs) for the electron ID were optimized for the 2016 data-taking period in a way that corresponds to around 98% signal efficiency. The WPs for the 2017 and 2018 IDs were adjusted to reproduce the same signal efficiency. For all three data-taking periods, the electron ID included the isolation variables in the training of the multivariate classifier.

A first contribution to the electron ID was the measurement of the electron selection efficiency for the 2018 data-taking period using the recently improved MVA-based electron ID. Prior to this, the MVA training for the electron ID was based on the Toolkit for MultiVariate Analysis (TMVA) tool [56] and did not include isolation variables.
The retrained ID was obtained using the eXtreme Gradient Boosting (XGBoost) package [57] with the isolation variables included in the training. While the performance of the retrained ID was already demonstrated for the 2017 data-taking period [58], this was not yet done for the 2016 and 2018 periods. The efficiency of the retrained ID for the 2018 period was prepared and presented, for the first time, for the 2019 Moriond conference. An improvement for the 2017 data-taking period was presented on the conference as well. The efficiency of the retrained electron ID for the 2017 and 2018 periods are first discussed in this section.

Table 3.2 shows the list of data and MC samples used for both 2017 and 2018 periods. The nominal MC efficiencies for both periods are evaluated from the leading order (LO) MadGraph [59] Drell-Yan sample, corresponding to a generic $q\bar{q} \to Z/\gamma^* \to e^+e^-$ production, while the next-to leading order (NLO) MadGraph_AMCatNLO sample is used to assess the systematic uncertainty related to the generator being used.
For both the 2017 and 2018 periods the same requirements on the tag are imposed:

- trigger matched to HLT_Ele32_WPTight_Gsf_L1DoubleEG_v*
- $p_T^{tag} > 30\ GeV$, $|\eta_{SC}^{tag}| < 2.17$ and $q^{tag} \cdot q^{probe} < 0$

The first bullet ensures the geometrical matching of the tag to the leg of a single electron HLT object, ensuring that probes do not have any trigger selection cuts. Otherwise, the measurement of the ID efficiency would be biased. The second bullet defines the p_T and η cut on the tag and requires an opposite-sign electron pair. Since the single electron trigger is restricted to $|\eta_{SC}| < 2.17$ because of the high background rates in the forward region of the detector, the same cut is imposed on the tag selection.

2017	
data	
/SingleElectron/Run2017B-17Nov2017-v1/MINIAOD	
/SingleElectron/Run2017C-17Nov2017-v1/MINIAOD	
/SingleElectron/Run2017D-17Nov2017-v1/MINIAOD	
/SingleElectron/Run2017E-17Nov2017-v1/MINIAOD	
/SingleElectron/Run2017F-17Nov2017-v1/MINIAOD	
MC	
sample	**usage**
/DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/RunIIFall17MiniAOD-RECOSIMstep_94X_mc2017_realistic_v10-v1/MINIAODSIM	nominal
/DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/RunIIFall17MiniAOD-RECOSIMstep_94X_mc2017_realistic_v10_ext1-v1/MINIAODSIM)	nominal
/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/RunIIFall17-MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM	systematics
2018	
data	
/EGamma/Run2018A-17Sep2018-v2/MINIAOD	
/EGamma/Run2018B-17Sep2018-v2/MINIAOD	
/EGamma/Run2018C-17Sep2018-v2/MINIAOD	
/EGamma/Run2018D-17Sep2018-v2/MINIAOD	
MC	
sample	**usage**
/DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/RunII-Autumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/MINIAODSIM	nominal
DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/RunII-Spring18MiniAOD-100X_upgrade2018_realistic_v10-v1/MINIAODSIM	systematics

Table 3.2: Data and MC samples used for the measurement of the electron selection efficiency for the 2017 and 2018 data-taking periods.

For the low p_T bins of the probe ($< 20\ GeV$), additional requirements were imposed in order to reject electrons coming from the W boson decays:

$$trigMVA_{tag} > 0.92,\ \sqrt{2 \cdot PFMET \cdot p_T^{tag} \cdot [1 - cos(\phi_{PFMET} - \phi_{tag})]}$$

For both periods, the selection under study is defined by the $H \to ZZ \to 4l$ MVA-based ID (mvaEleID-Fall17-iso-V2-wpHZZ) and the requirements on the vertex parameters and SIP as defined in section 3.3.1. Since electrons that end up in the region between the barrel and the endcap (henceforth referred to as the *gap electrons*) are expected to be reconstructed with a lower efficiency, they are treated separately in the efficiency measurements. Therefore, the selection efficiency and SFs are first derived for the non-gap electrons followed by the same analysis for the gap electrons only. The same selection on the tag and probe pairs is imposed in both cases.

Figure 3.6 shows the measured selection efficiencies (top pad in the figure) and SFs (bottom pad in the figure) in the different p_T bins for the two periods. The binning in η was chosen to be the same as the one used in the 2017 results already approved by CMS prior to this analysis. Gap electrons are excluded.

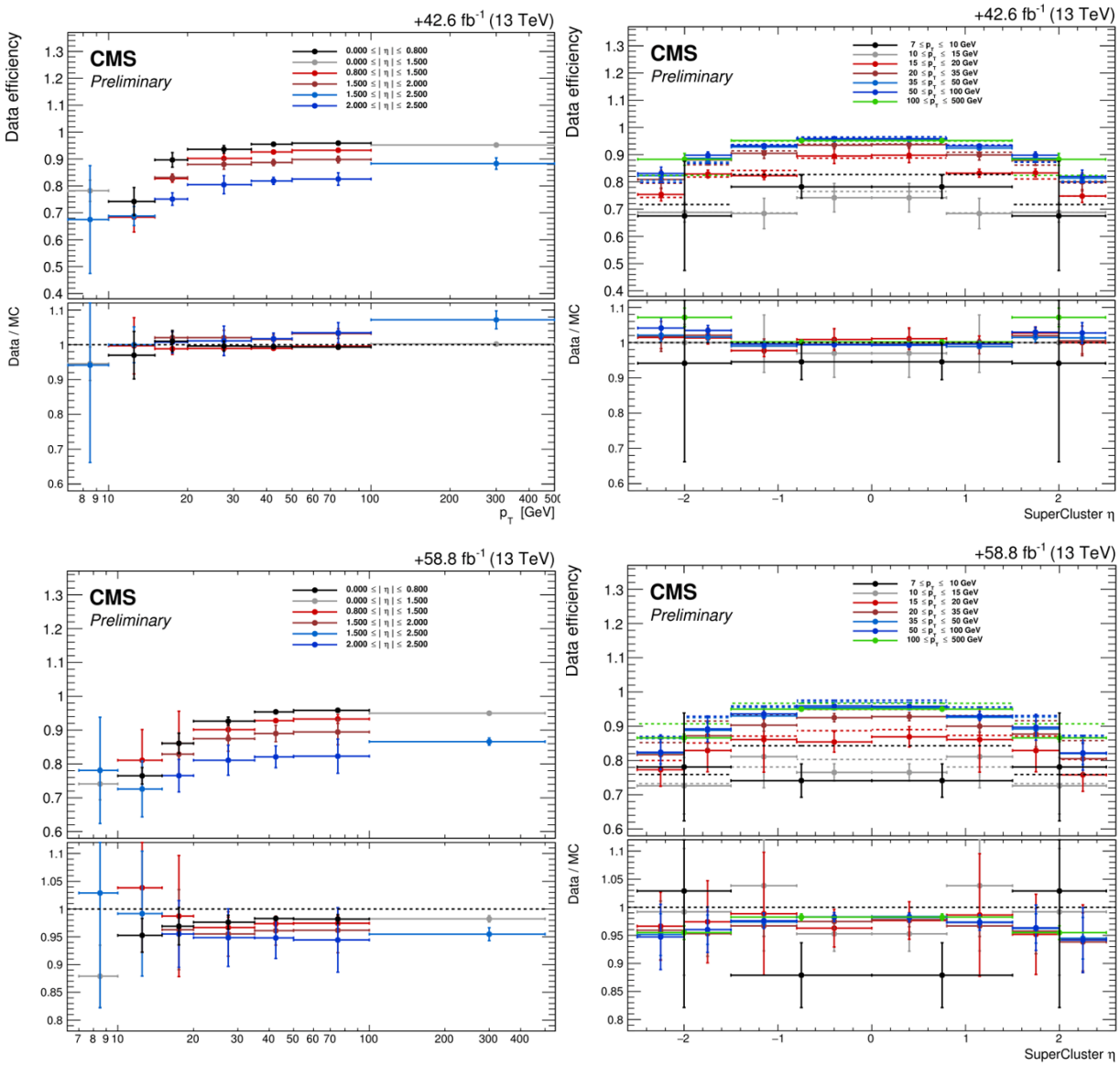


Figure 3.6: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2017 (top row) and 2018 (bottom row) data-taking periods. The left-hand side plots show the results for different p_T bins, while the right-hand side plots shows the same for different η bins.

Due to a lower statistics in the very low-p_T bin ($< 10\ GeV$) and high-p_T bin ($> 100\ GeV$), the efficiencies and SFs were calculated only for the combined barrel (light grey histogram) and the endcap (light blue histogram) region. The middle-p_T range is split into several η bins in order to gain insight into possible η-dependent structure of SFs.

One feature that can be seen on the efficiency plots versus the electron p_T is the increase of the efficiency in the low-p_T region until the plateau is reached. This is the consequence of the bremsstrahlung which causes the loss of

efficiency at low values of p_T.

An additional feature, especially pronounced in the 2018 period, is a consistent offset of SFs from unity over the entire p_T range. This was studied and traced back to the $|SIP| < 4$ cut. If the SIP cut requirement is removed from the selection, keeping other things unchanged, this feature disappears. This can be seen on Fig. 3.7 where the SFs are now consistent with unity. This behavior was afterward cured by the Ultra legacy (UL) reprocessing of the data and the MC samples.

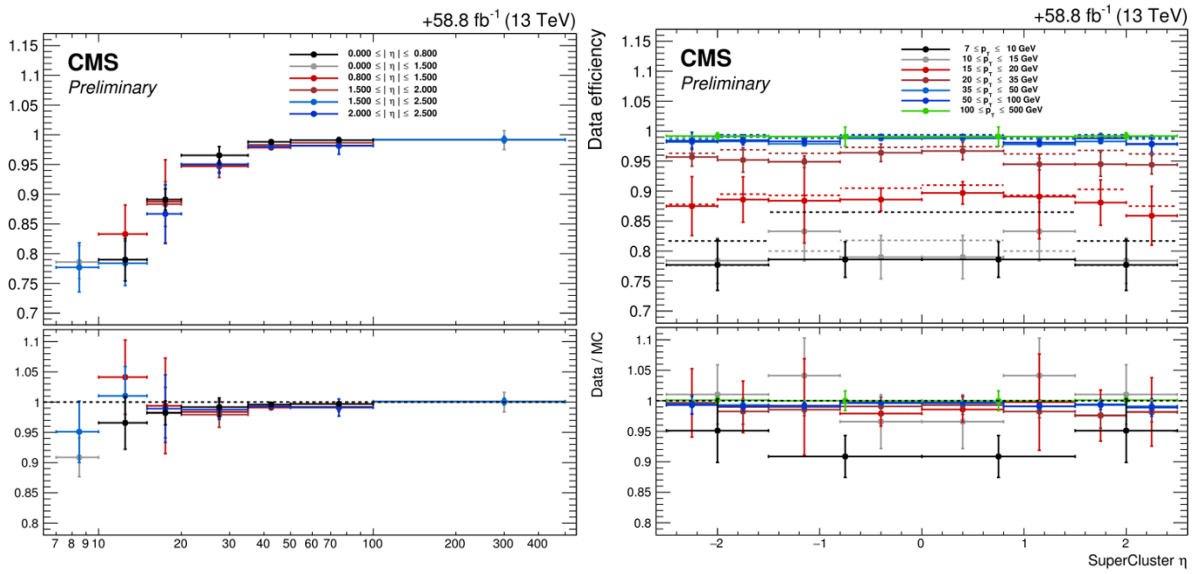


Figure 3.7: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2018 data-taking periods. Left-hand side plot show the results for different p_T bins, while the right-hand side plot shows the same for different η bins. The only change with respect to the bottom row plots in Fig. 3.6 is the removal of the $|SIP| < 4$ cut.

Comparing the uncertainties obtained for the 2017 and 2018 periods, one can see that these are larger for the latter. This can be more easily seen on Fig. 3.8 showing the SFs and corresponding uncertainties in all p_T and η bins.

Fig. 3.9 and Fig. 3.10 show the election efficiency, scale factors and corresponding uncertainty for the gap electrons in the 2018 data-taking period. The same plots for the 2017 period were obtained in CMS prior to this analysis and are thus omitted. Only three p_T bins were used in order to keep sufficient statistics in each bin. In addition, on the right-hand side plot, the bins are split in $|\eta|$, rather than in η.

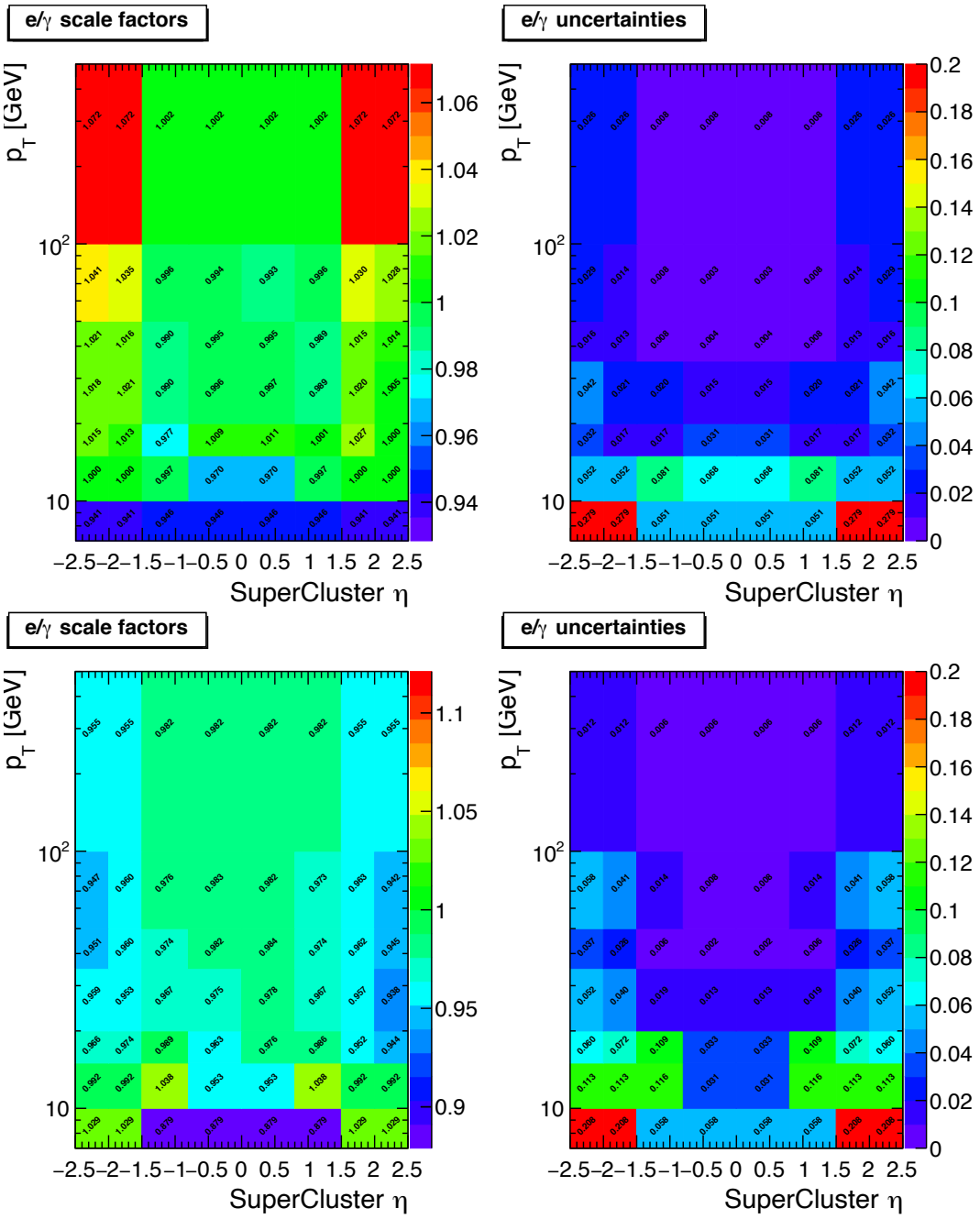


Figure 3.8: Electron SFs (left row) and corresponding overall uncertainty (right row) for all p_T and η bins shown in Fig. 3.6. Results for the 2017 (2018) period is shown in the top (bottom) row.

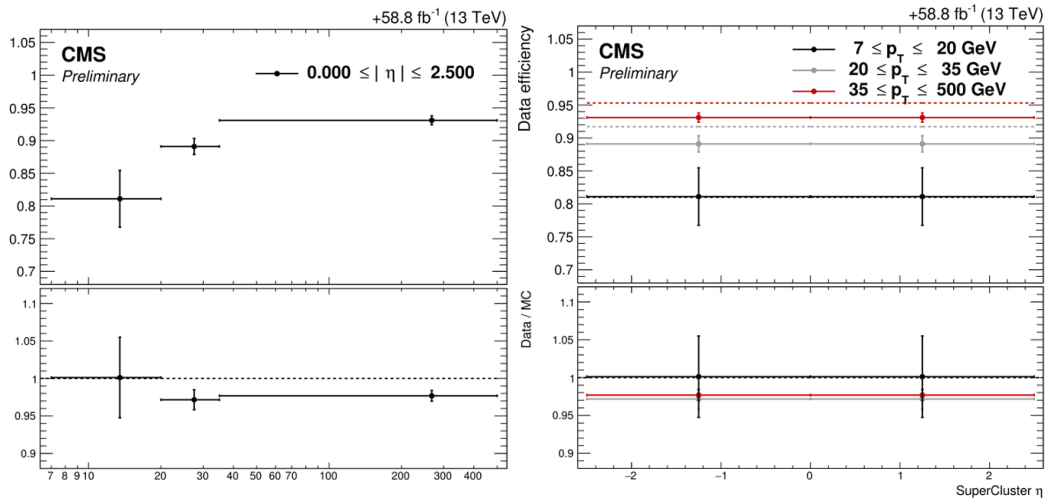


Figure 3.9: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2018 data-taking periods. The left-hand side plot shows the results for different p_T bins, while the right-hand side plot shows the same for different η bins. Only gap electrons are considered.

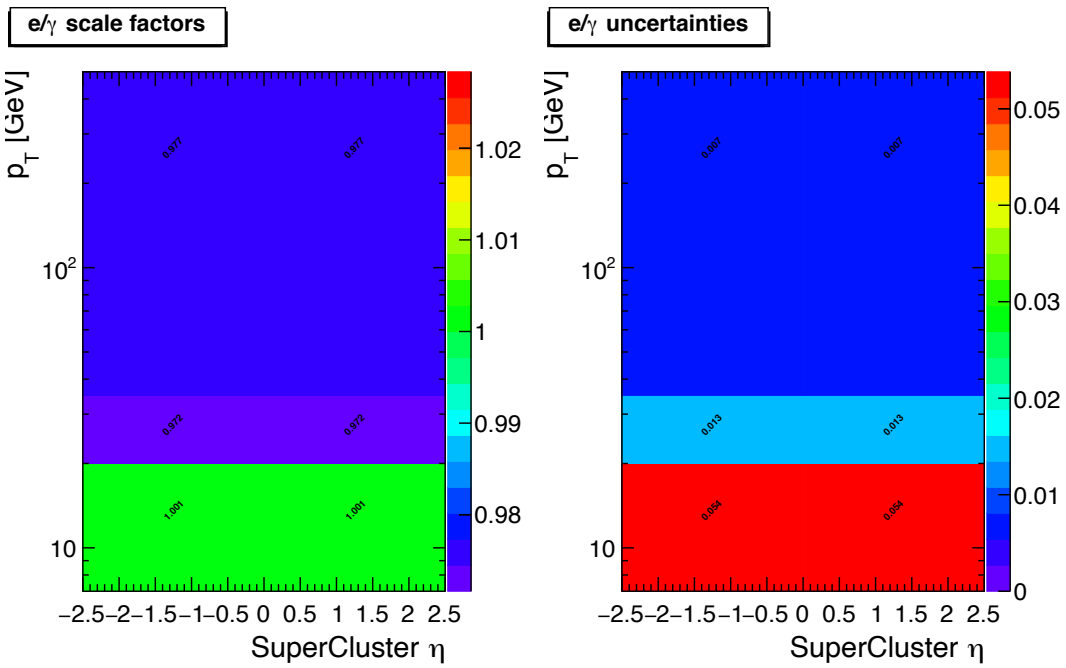


Figure 3.10: Electron SFs (left row) and corresponding overall uncertainty (right row) for all p_T and η bins shown in Fig. 3.9. Results for the 2018 period gap electrons are shown.

In order to prepare for the $H \to ZZ \to 4l$ Run 2 legacy paper [60], it was decided to retrain the electron ID for the 2016 period. This meant replacing the older ID which didn't incorporate isolation variable in the training and which was trained using the TMVA package with the new ID (mvaEleID-Summer16-ID-ISO-HZZ) that included isolation in the training and was trained using the XGBoost package.
In the analysis discussed thus far in this section, the retrained ID used for the 2017 data-taking period was also used for the 2018 period. A dedicated ID retrained for the 2018 period was not essential at the time because it was shown that the performance of the 2017 training on the 2018 data was satisfactory. However, in the meantime, a dedicated ID was retrained also for the 2018 period by the CMS collaboration for consistency sake. Since the training of the IDs is not a direct contribution of this thesis work, the WPs and corresponding signal and background efficiencies for all three Run 2 periods are merely summarized in Table 3.3.

The electron efficiency measurements and the SFs discussed in the following part of the section were rederived using the retrained electron IDs for the 2016, 2017 and 2018 periods. The goal of this analysis was to further reduce the uncertainty in the low-p_T region and study the η-dependent structure of SFs. The former was especially needed since the leading source of uncertainty in the $H \to ZZ \to 4l$ analysis is the uncertainty on electron efficiency measurements that mostly originates from the measurement uncertainty of the low-p_T electrons that are present in the analysis due to the off-shell Z boson.
Data and simulations used in the analysis are listed in Table 3.4 for all three periods. For the 2016 period, the nominal MC efficiencies are evaluated from the leading order (LO) MadGraph Drell-Yan sample, while the next-to leading order (NLO) MadGraph_AMCatNLO sample is used to access the systematic uncertainty. The only change in the 2018 period, with respect to the previously discussed analysis, is the use of the *POWHEG* [61–63] sample for accessing the systematic uncertainties instead of the (NLO) MadGraph_AMCatNLO sample. The reason for this change is the higher statistics in the POWHEG sample. As before, efficiency measurements for the non-gap electrons are shown first, followed by the measurements for the gap electrons.

2016 (mvaEleID-Summer16-ID-ISO-HZZ)			
$\|\eta\| < 0.8$			
	WP	ϵ_{sig} [%]	ϵ_{bkg} [%]
$5\ GeV < p_T < 10\ GeV$	1.8409	81.64	3.93
$p_T > 10\ GeV$	0.3902	97.44	2.17
$0.8 < \|\eta\| < 1.479$			
$5\ GeV < p_T < 10\ GeV$	1.7830	80.31	3.63
$p_T > 10\ GeV$	0.3484	96.68	2.75
$\|\eta\| > 1.479$			
$5\ GeV < p_T < 10\ GeV$	1.7559	74.37	3.06
$p_T > 10\ GeV$	-0.6518	96.62	7.66
2017 (mvaEleID-Fall17-iso-V2-wpHZZ)			
$\|\eta\| < 0.8$			
	WP	ϵ_{sig} [%]	ϵ_{bkg} [%]
$5\ GeV < p_T < 10\ GeV$	1.4499	81.64	5.66
$p_T > 10\ GeV$	0.0081	97.44	3.26
$0.8 < \|\eta\| < 1.479$			
$5\ GeV < p_T < 10\ GeV$	1.4856	80.31	4.74
$p_T > 10\ GeV$	-0.0374	96.68	4.05
$\|\eta\| > 1.479$			
$5\ GeV < p_T < 10\ GeV$	1.6901	74.37	3.59
$p_T > 10\ GeV$	-0.7497	96.62	8.10
2018 ((mvaElectronID_Autumn18_ID_ISO)			
$\|\eta\| < 0.8$			
	WP	ϵ_{sig} [%]	ϵ_{bkg} [%]
$5\ GeV < p_T < 10\ GeV$	0.8962	81.64	5.66
$p_T > 10\ GeV$	0.0279	97.45	3.28
$0.8 < \|\eta\| < 1.479$			
$5\ GeV < p_T < 10\ GeV$	0.9070	80.31	4.69
$p_T > 10\ GeV$	-0.0024	96.68	4.12
$\|\eta\| > 1.479$			
$5\ GeV < p_T < 10\ GeV$	0.9396	74.37	3.26
$p_T > 10\ GeV$	-0.5983	96.62	8.06

Table 3.3: Working points together with corresponding signal and background efficiencies for the BDT training of the electron ID for the three data-taking periods.

2016	
data	
/SingleElectron/Run2016B-17Jul2018_ver2-v1/MINIAOD	
/SingleElectron/Run2016C-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016D-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016E-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016F-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016G-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016H-17Jul2018-v1/MINIAOD	
MC	
sample	**usage**
/DYJetsToLL_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/RunII-Summer16-MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v2/MINIAODSIM	nominal sample
/DYJetsToLL_M-50_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/RunIISummer16-MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext2-v1/MINIAODSIM	systematics
2017	
data	
same as in Table 3.2	
MC	
same as in Table 3.2	
2018	
data	
same as in Table 3.2	
MC	
sample	**usage**
same as in Table 3.2	nominal sample
/DYToEE_M-50_NNPDF31_TuneCP5_13TeV-powheg-pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/MINIAODSIM	systematics

Table 3.4: Data and MC samples used for the measurement of the electron selection efficiency and SFs for the $H \to ZZ \to 4l$ Run 2 legacy paper.

The same selection on the tag is applied on for three period and is, in most part, the same as defined before. In order to try to reduce the uncertainties, the p_T requirement on the tag was increased to 50 GeV for the lower p_T bins of the probe ($< 20\ GeV$). In addition, the requirement that all three charge measurements, defined in section 3.2.3, agree was required for the same bins. Finally, the coarser binning of the m_{ee} distribution, using 30 bins instead of 60, was used in order to further stabilize the fits.

The new requirements on the tag resulted in a slightly more clear peak around the nominal Z boson mass which

resulted in better precision and lower uncertainty for these bins. This can be seen in the right column in Fig. 3.11 which shows the nominal signal fit in data for one low-p_T bin ($11\ GeV < p_T < 15\ GeV$ and $0 < \eta < 0.5$). It was checked that no bias is introduced in the efficiency measurement by doing so.

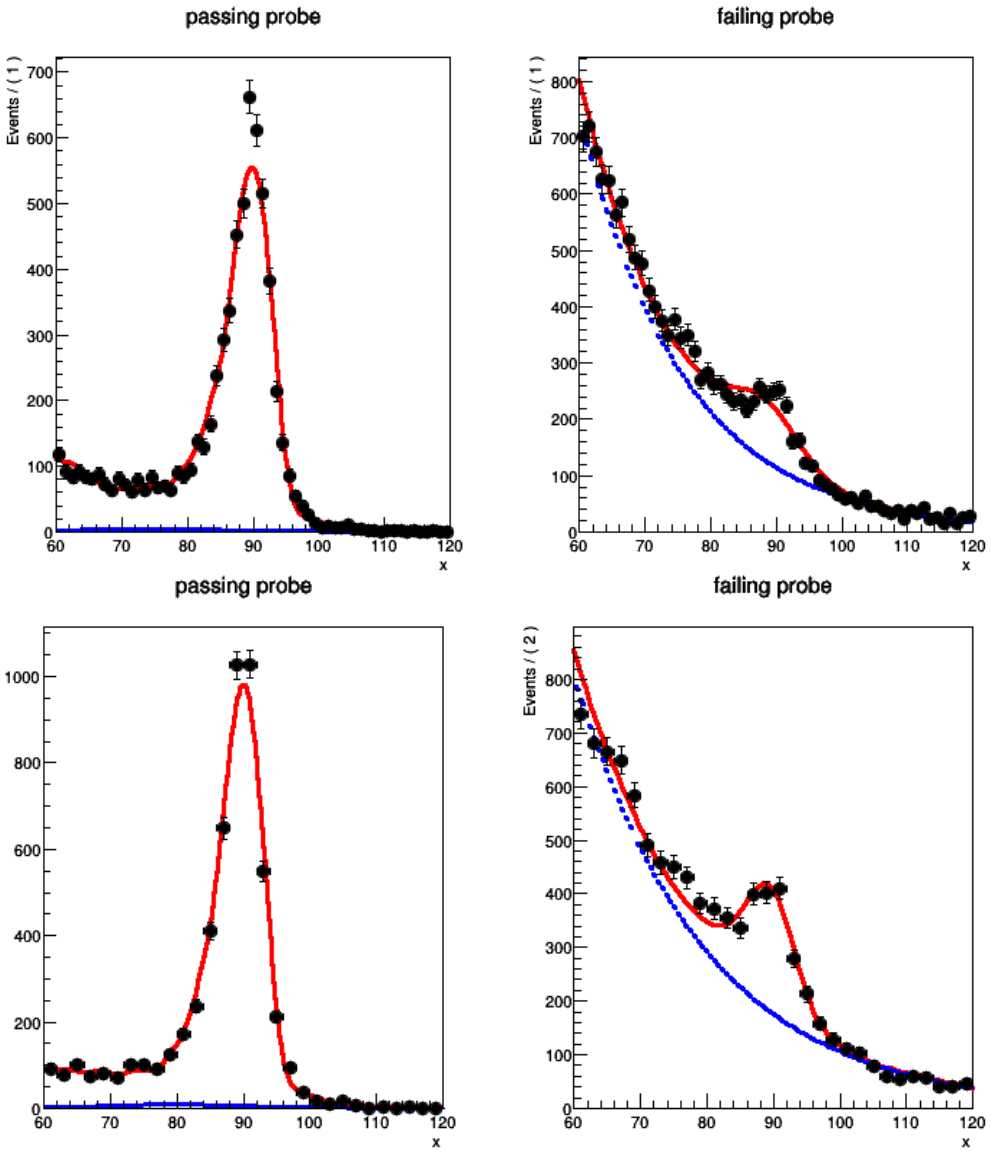


Figure 3.11: The m_{ee} distribution for one low-p_T bin ($11\ GeV < p_T < 15\ GeV$ and $0 < \eta < 0.5$) before (top row) and after (middle row) tightening the tag selection for the low-p_T bins of probe. The nominal fit in the data is shown in both figures.

Another consequence of the tighter tag selection was the appearance of the excess of events (the "bump") in the low mass tail of the m_{ee} distribution of the failing probes for $15 < p_T < 20$ GeV bins. This bump comes from the signal electrons that migrated from the passing probe group before tightening the cut to the failing probe group after tightening the cut. In order to successfully fit the bump, the function for the signal model in the failing probes had to be modified. It was found that a good fit for the signal can be achieved with a help of additional Gaussian. To achieve the convergence of the fit for the background, a default model was modified by introducing a Chebyshev polynomial. This is shown in Fig. 3.12 for the nominal fit in data for one bin ($15\ GeV < p_T < 20\ GeV$ and $1 < \eta < -0.5$). The left-hand side plot shows the bad fit in the failing probes before the modification of the fitting function, while the right-hand side

plot shows the improved fit.

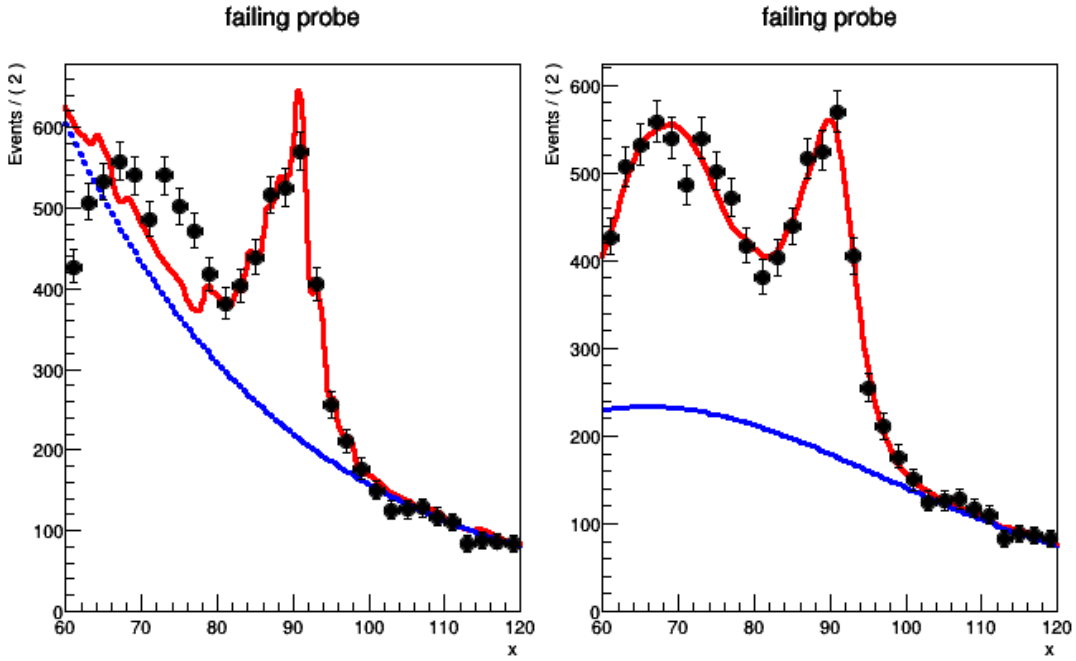


Figure 3.12: The m_{ee} distribution for one bin ($15\ GeV < p_T < 20\ GeV$ and $1 < \eta < -0.5$) after tightening the tag selection for the low-p_T bins of probe. The bad fit for the failing probes (left) was resolved by adding an additional Gaussian in the signal model and introducing a Chebyshev polynomial in the background model (right).

Fig. 3.13 shows that this treatment reduced uncertainties in the selection efficiency measurement, especially for the low-p_T and high-η region. Here, the gap electrons are not included.

In addition to the tighter requirements on the tag, one can see that the binning has been changed in order to try improving on the η dependency of the SFs. This feature is visible on the bottom-right plot in Fig. 3.13. While studying different binning scenarios for the 2017 data-taking period, it was found that better results can be achieved by using a finer η binning. This is shown in the top row in Fig. 3.14 where a more pronounced η structure in SFs is observed. The "umbrella" shape in efficiency (top pad on the figure) is the result of inefficiencies in electron reconstruction and identification in the more forward regions of the detector. The top row shows the results for the 2017 period, while the bottom row shows the results for the 2016 period. Fig. 3.15 shows the SFs and the corresponding uncertainty for the three data-taking periods.

Finally, Figs. 3.16 and 3.17 show the efficiency, SFs and the overall uncertainty for the gap electrons for the 2016, 2017 and 2018 data-taking period.

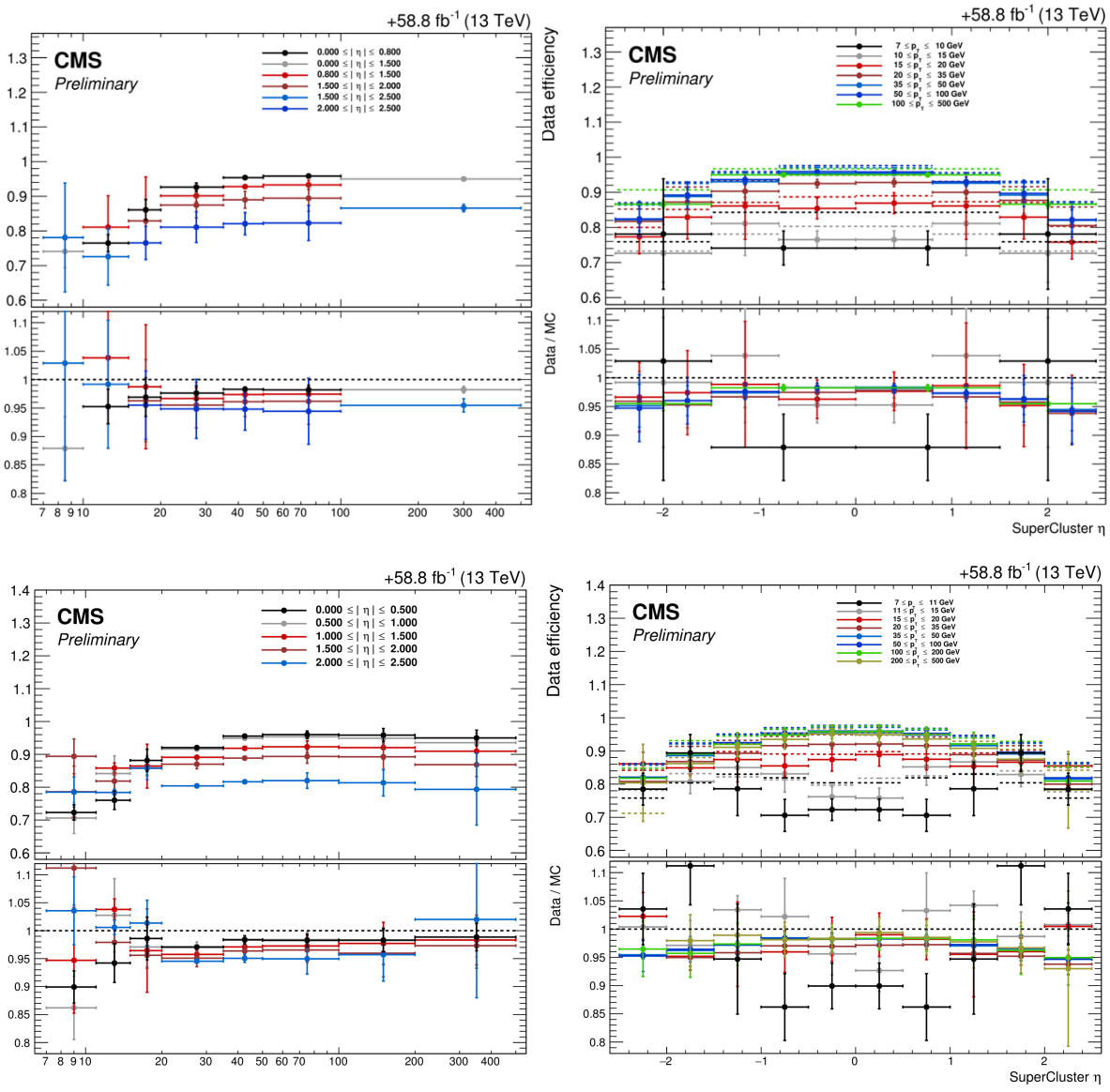


Figure 3.13: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2018 period with the original tag selection (top row) and the same period with tighter tag selection introduced for the low-p_T bins of the probe (bottom row). The left-hand side plots show the results for different p_T bins, while the right-hand side plots shows the same for different η bins.

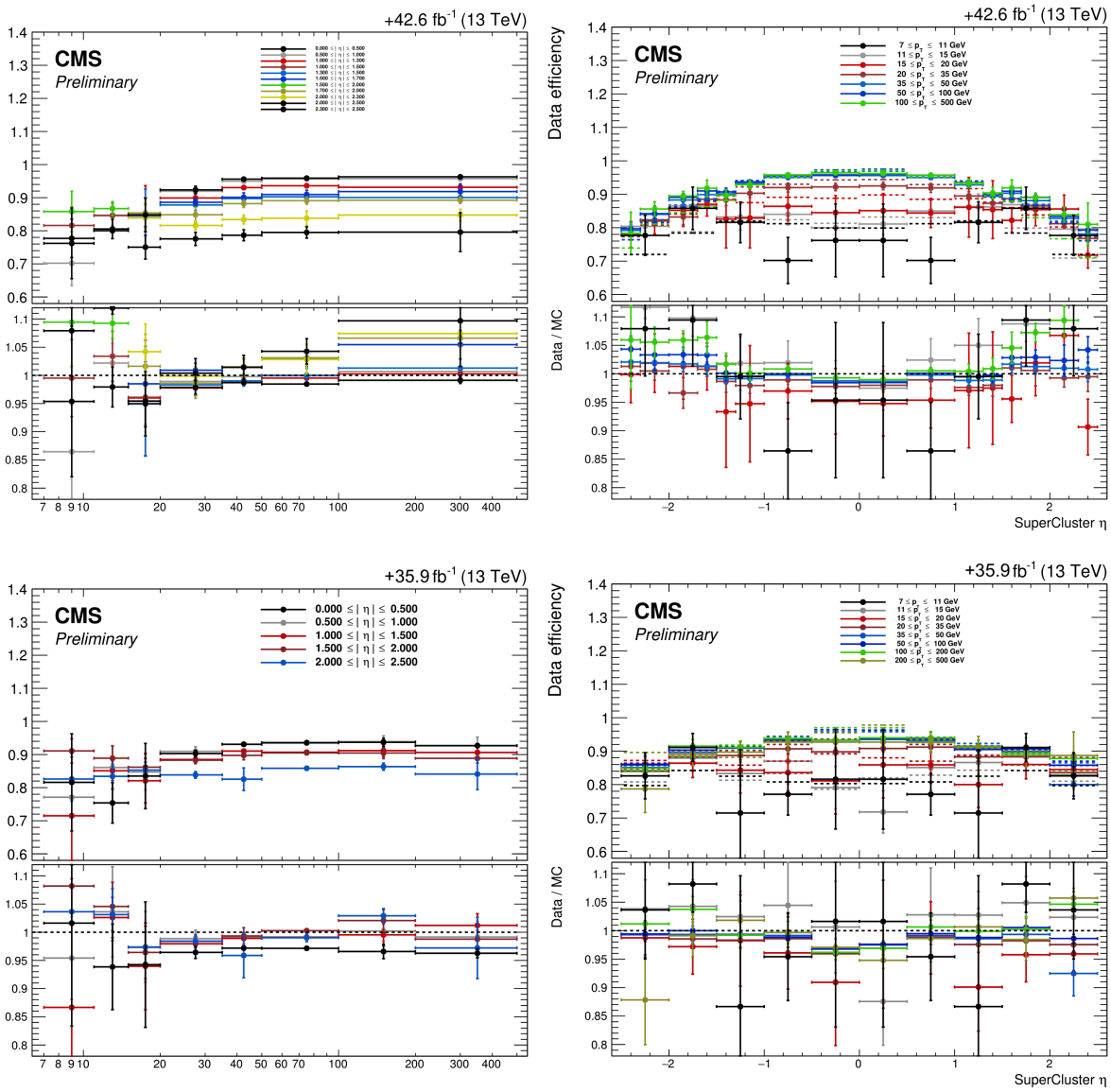


Figure 3.14: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2017 (top row) and 2016 (bottom row) periods using the retrained electron ID and the tighter tag selection for the low p_T bins of the probe. The left-hand side plots show the results for different p_T bins, while the right-hand side plots shows the same for different η bins.

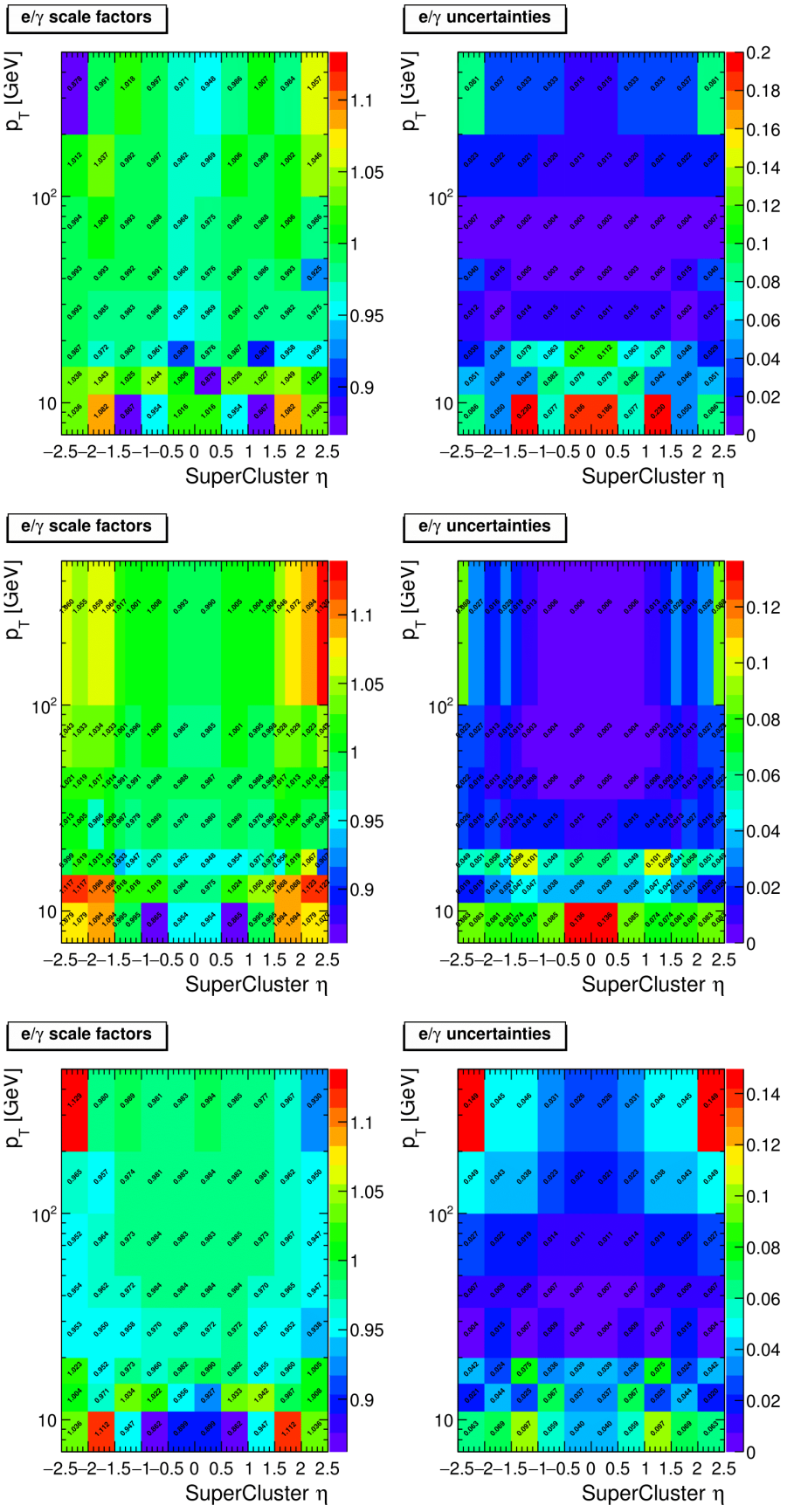


Figure 3.15: Electron SFs (left column) and corresponding overall uncertainties (right column) for all p_T and η bins shown in the bottom row in Fig. 3.13 and in Fig. 3.14. Results for the 2016, 2017 and 2018 periods are shown in the top, middle and bottom row respectively.

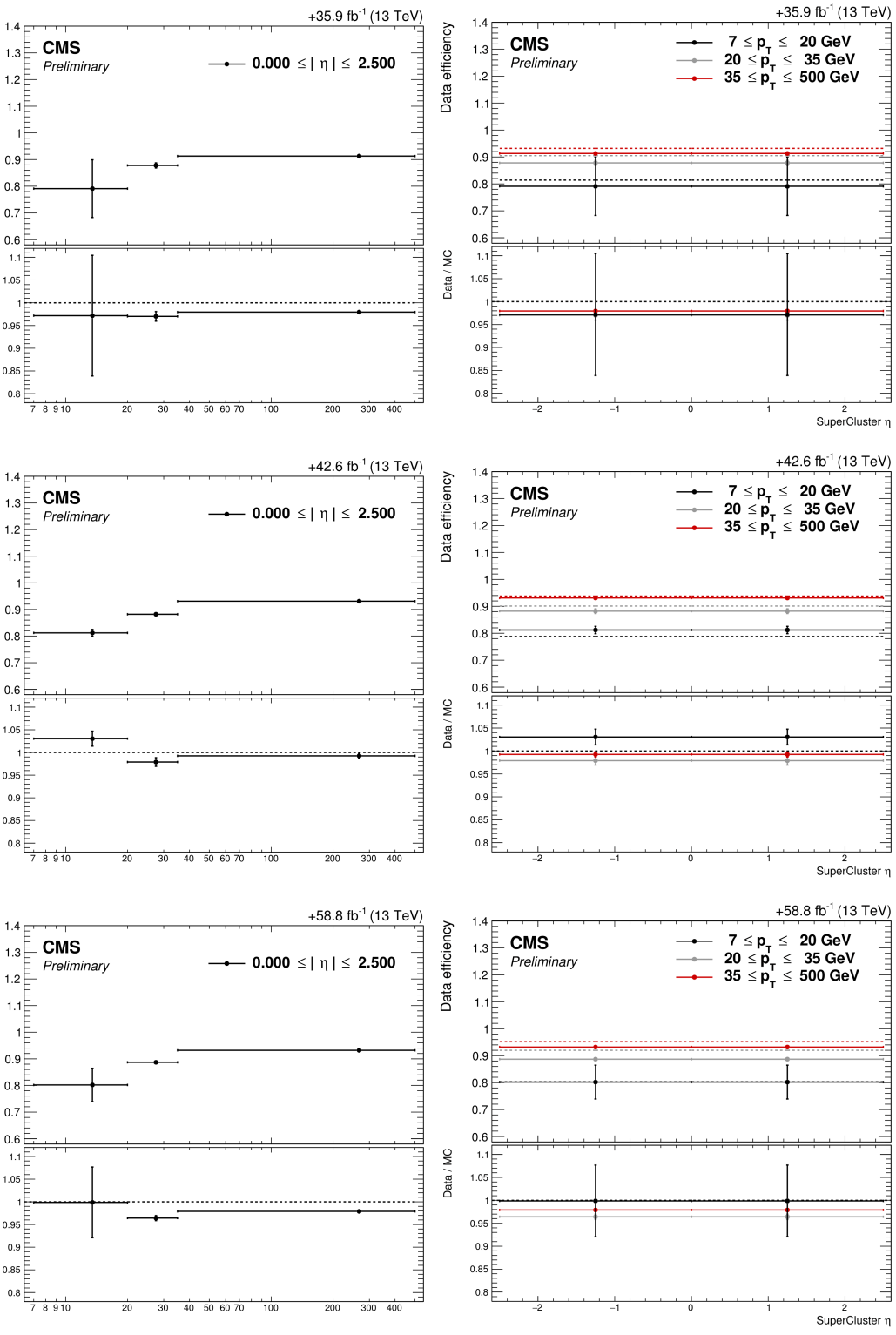


Figure 3.16: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2016 (top row), 2017 (middle row row) and 2018 (bottom row) periods using the retrained electron IDs and the tighter tag selection for the low p_T bins of the probe.. The left-hand side plots show the results for different p_T bins, while the right-hand side plots shows the same for different η bins. Results for gap electrons are shown.

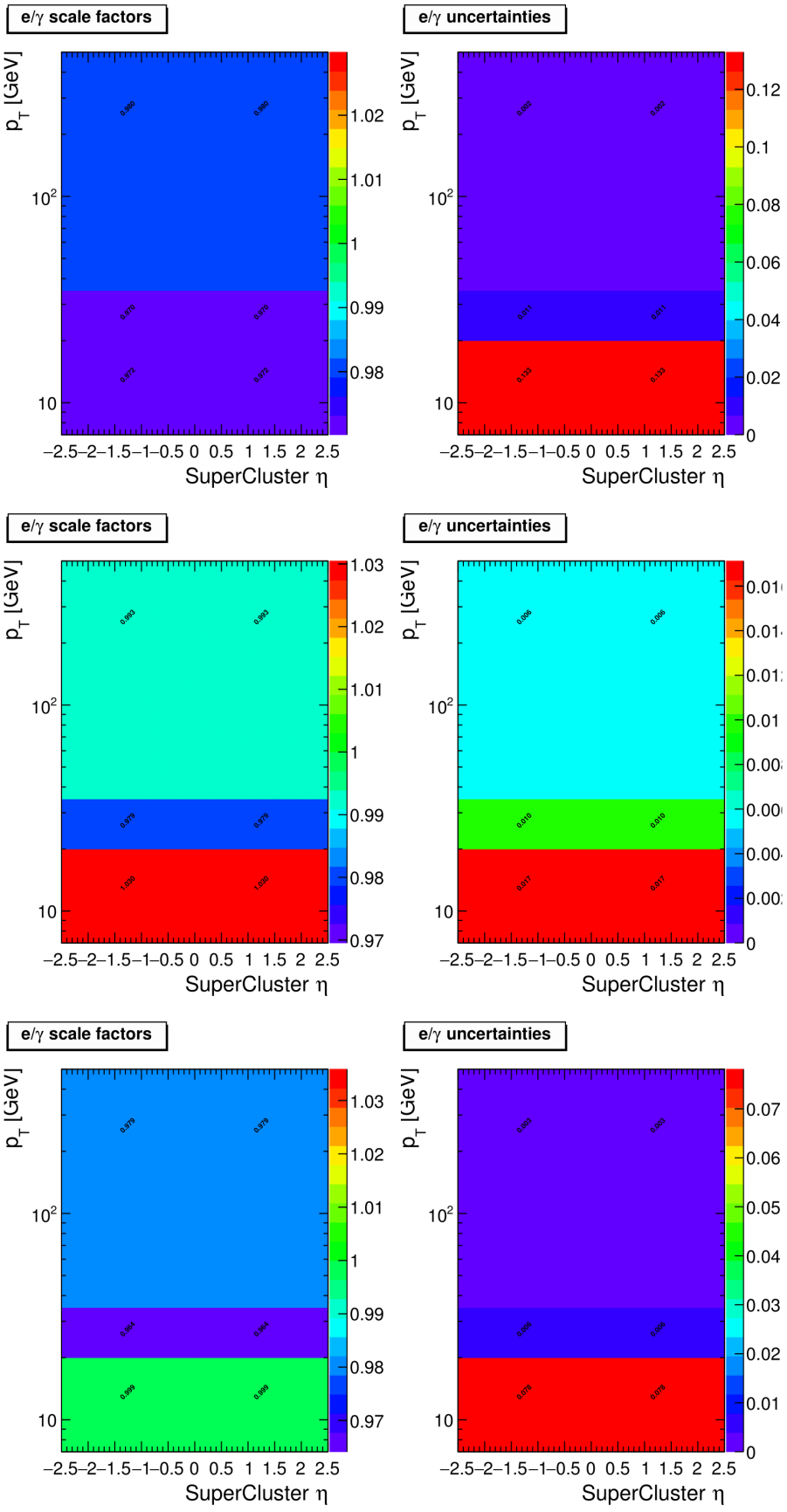


Figure 3.17: Electron SFs (left column) and corresponding overall uncertainties (right column) for all p_T and η bins shown in Fig. 3.16 for the 2016 (top row), 2017 (middle row) and 2018 (bottom row) period. Results for gap electrons are shown.

3.5 Summary

An overview of electron reconstruction and identification (ID) in CMS was discussed followed by the measurements of the electron selection efficiency for the 2016, 2017 and 2018 data-taking periods using the Tag and Probe (TnP) method.

The efficiency measurements and the scale factors (SFs) were first derived for the 2018 period using the electron ID trained on the 2017 data. The training of the ID was done centrally in CMS, the novelty being the incorporation of the isolation variables in the training of the multivariate classifier. The new ID training was improved by switching to the XGBoost package instead of the TMVA.

In order to prepare for the $H \rightarrow ZZ \rightarrow 4l$ Run 2 legacy paper, it was decided to retrain the electron ID for the 2016 data-taking period using the XGBoost package and including the isolation variables in the training. The same was done for 2018 period for consistency sake. The retraining of the IDs for both data-taking periods was done centrally in CMS. A new electron efficiency measurements, together with the SFs, were rederived for all three periods with a goal of reducing the uncertainties in the low-p_T region and studying the η structure in SFs.

In order to reduce the uncertainties in the low-p_T region, a tighter selection on the tag was applied for the low-p_T bins of probe. The requirement that all three charge measurement must agree was also added. The new requirements on the tag gave rise to a slightly more clear peak around the nominal Z boson mass. This resulted in in a better precision and lower uncertainty for these bins and was first shown for the 2018 period. The same conclusion was found to be true for the 2016 and 2017 periods as well. An additional consequence of the tighter tag selection was the appearance of the excess of events (the "bump") in the low mass tail of the m_{ee} distribution of the failing probes for $15 < p_T < 20\ GeV$ bins. It was found that the bump was populated by signal electrons that migrated to the failing probe group after tightening the tag selection. In order to fit the bump in the m_{ee} distribution for the failing probes, the fitting function for the signal and background contributions had to be modified. A better fit further reduced the uncertainty in these bins.

In addition to tightening the tag selection, the binning for the 2018 period was changed in order to try to improve on the η dependency of the SFs. This was more studied for the 2017 period where it was shown that a further improvement can be achieved by choosing even finer η binning. The "umbrella" shape in the efficiencies was observed due to a more challenging reconstruction and identification of electrons in the forward regions of the detector.

The results presented in this chapter were used in the publication of the $H \rightarrow ZZ \rightarrow 4l$ analysis. In addition, these are the integral part of the VBS $ZZ \rightarrow 4l2j$ analysis discussed in the next chapter.

Chapter 4

Search for the VBS in the 4l final state using the Run 2 data

4.1 Preface to the chapter

This chapter covers published results on the search for the VBS in the $ZZ \to 4l2j$ channel using the full Run 2 data and is a continuation of a previous study in the same channel that used 2016 data to extract the EWK signal [64]. The paper is a joined effort of the CMS diboson SM group.
The biggest challenge of the analysis is a small cross-section of the signal, being one of the smallest ever measured at the LHC at only $0.3\ fb$, approximately 30 times less than the irreducible background. Another feature of this channel is a large contribution from the QCD-induced production of the two Z bosons represented by diagrams containing, at least, 2 QCD vertices. This is the main background to the analysis.
However, unlike final states containing W bosons, this channel is characterized by a fully reconstructable final state. Because of this, it will be amongst the most important channels to separate the longitudinal polarization of the Z boson in the future. In addition, it is the most sensitive channel for studying certain anomalous quartic gauge couplings (aQGCs), specifically f_{T8} and f_{T_9}. Lastly, it had not yet been observed in CMS.

My contribution in the published paper is the development of the BDT classifier used as an alternative signal extraction method. This is described in section 4.6. In addition, I am the main contributor to deriving the limits on the anomalous quartic gauge couplings in the EFT approach. The procedure for deriving aQGCs is presented in section 4.7.

I begin the chapter by describing the data sets and Monte Carlo simulations used in the analysis. The following section defines event selection. In section 4.4 I define variables used for the signal extraction with BDT which I also use to show the agreement between the data and the simulation.
In the published paper the MELA discriminant was used as a main tool for the signal extraction and is discussed in section 4.5.1. In the same section I describe how the VBS significance and the cross-section in VBS and VBS+QCD fiducial regions were calculated. Section 4.8 will discuss the systematic uncertainties used in both the MELA and the BDT signal extraction approaches and also in the derivation of the limits on the aQGCs.
In section 4.9 I will present results on the VBS significance using both the MELA and the BDT approaches and compare the two. The results on the aQGCs are reported here as well. The key points of the chapter are summarized in section 4.10.

4.2 Monte Carlo simulations and data sets

4.2.1 Monte Carlo samples

Several Monte Carlo (MC) samples have been produced and are used in this analysis to optimize the event selection, evaluate signal efficiency and acceptance, optimize the search strategy for the VBS as well as for a search for anomalous quartic gauge couplings (aQGCs).

Signal

In this analysis the signal is defined as the purely electroweak (EWK) production of the two jets and the two leptonically decaying Z bosons. It was simulated at leading order (LO) using the *MadGraph5 aMCatNLO* (henceforth *MG5*) tool by requiring explicitly the number of QCD vertices to be zero:

$$generate\ pp > zzjj\ QCD = 0,\ z > l + l-$$

Z bosons are only allowed to decay into electrons and muons. This is performed using the *MadSpin* tool in order to preserve the spin correlations between the leptons. The resulting sample includes contributions from the SM Higgs boson produced in vector boson fusion (VBF) as well as from the interference with non-Higgs diagrams and diagrams featuring triboson production with one hadronically decaying W boson. The latter is supressed by requiring the dijet invariant mass, m_{jj} to be greater than 100 GeV.
An additional sample was produced using the *Phantom* tool which includes off-shell Z boson decays and was used to cross check the sample produced with *MG5*.

Irreducible backgrounds

The dominant, irreducible background in the analysis is the QCD-induced $pp \rightarrow ZZ$ production (henceforth qqZZ). This process was simulated at next-to-leading order (NLO) with up to two jets using MG5 and merged with parton showers using FxFx scheme.

$$generate\ pp > l + l - l + l -\ [QCD]\ @0$$

$$add\ process\ pp > l + l - l + l - j\ [QCD]\ @1$$

The idea behind the FxFx jet merging scheme is to remove the overlap between jets produced at matrix elements (ME) and those produced by parton showers (PS) and thus removing the double counting of jets [65, 66]. This is the nominal sample for the qqZZ background in this analysis.
In order to study the interference between the signal and the irreducible background, an additional sample was generated using MG5. It was shown, by comparing the event yields and distribution shapes between the signal sample and the interference sample, that the yield ratio is between 1% and 6%. This was taken into account in the analysis via a proper scaling.

An additional background to the signal is the gluon loop-induced ZZ production process (henceforth ggZZ). Although this process is suppressed by two additional strong couplings, it never the less contributes to inclusive ZZ production at around 10% level.
A dedicated sample was studied and produced with *MG5* [67] specially for this analysis [68]. The process is simulated

at LO with up to 2 jets modeled from matrix-element and matched to PS using the MLM matching scheme [69] for the first time:

$$generate\ gg > zz\ [noborn = QCD]$$

$$add\ process\ pp > zzj\ [noborn = QCD]$$

$$add\ process\ pp > zzjj\ [noborn = QCD]$$

The requirement in the square brackets instructs MG5 to only consider loop diagrams. One can see that, for the 0j sample, a gg initial state was used, while qq initial state was used for the $1j$ and $2j$ samples. In former, this is equivalent since there are no extra loop-induced diagrams included. However, it is important to use pp initial state for $1j$ and $2j$ samples in order to include the Initial State Radioation (ISR) processes. In this case a quark will first transform to a gluon through the ISR, after which it will be involved in the hard process. This results in significantly more diagrams from which only genuine loop-induced diagrams should be kept. This is achieved using the "diagram filter" designed specially for this purpose [70]. After the filter is applied, only genuine loop-induced diagrams survive. It must be emphasized that $1j$ and $2j$ diagrams are not simply $0j$ diagram with some ISR decoration. These processes include some new diagrams with different structures that can't evolve from the $0j$ sample. Some examples are shown in Figure 4.1 where jets are emitted directly from the loop.

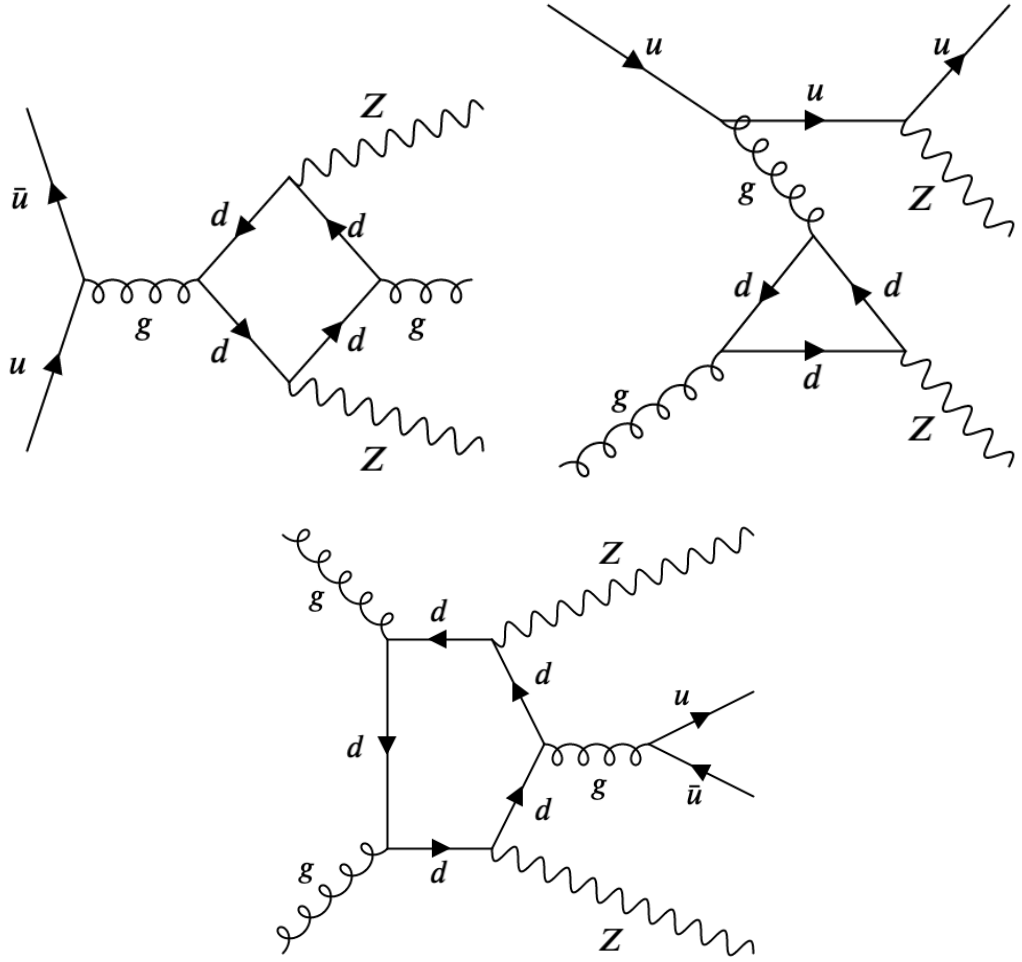


Figure 4.1: Example diagrams of loop induced ggZZ 1/2-jet process which can't evolve from the 0j sample

Although time consuming, this simulation covers the dijet phase space much better than the $0j$ sample where two jets are modeled from the PSs. However, since *MadSpin* generator can't decay particles generated in the loop-induced processes from the ME calculation, the decaying of the Z bosons is implemented in *Pythia8* such that spin correlations between the outgoing leptons are not included. The MLM matching scheme was applied to avoid the double counting when merging jets modeled with ME and those modeled by parton shower.

The dijet phase-space produced from the loop-induced process is expected to be more accurately modeled with this sample compared to an alternative approach using the MCFM generator [71]. The difference between the new *MG5* ggZZ production and the *MCFM* production is especially visible in the p_T spectrum of the two leading jets. This is shown in Figure 4.2 for different jet multiplicities. The difference is most notable in the softer p_T spectrum for the 0,1,2 jet merged sample (purple) produced with *MG5*. This has as a consequence a lower efficiency after applying the inclusive ZZjj selection (for details on event selection criteria see section 4.3). An additional effect the is the higher mass of the ZZ pair. The ggZZ background modeling described here, thus, gives the most accurate description of the dijet phase space so far.

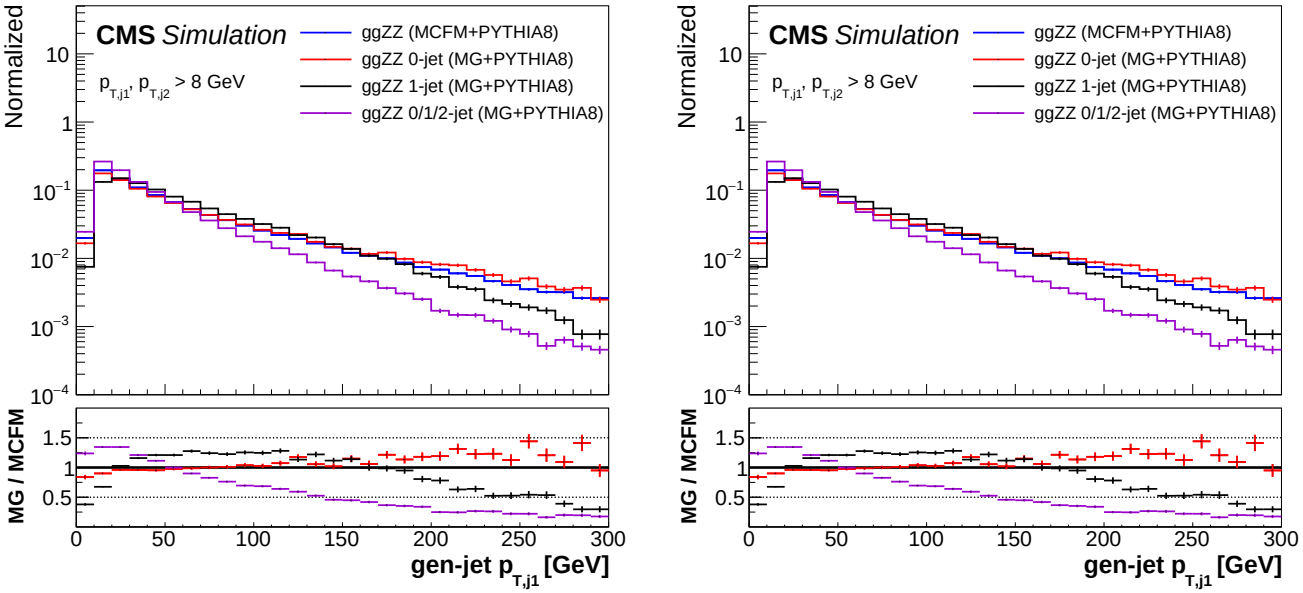


Figure 4.2: The p_T spectrum of the two leading jets in the QCD loop-induced samples generated with *MCFM* and the new, state-of-the-art samples with up to two jets merged with MLM matching scheme and generated for the first time for this analysis using the *MG5*. The figure is taken from Ref. [68]

In addition to the state-of-the-art gg sample, an additional sample was generated to validate the former. The simulation was done at LO with 1 jet using MG5 and the following syntax:

$$generate\ gg > zzj\ [noborn = QCD],\ z > l + l$$

Pythia8 was again used to perform the decay of the Z bosons and thus the correlation between the spin of the decay leptons is ignored. For this reason another sample was produced at LO using *MCFM* 7.0 tool [71].

Pythia8 package was used for parton showering and hadronization for all MC samples, with parameters set by the CUETP8M1 tune [72] for the 2016 and the CP5 tune [73] for 2017 and 2018 data-taking periods.
A NNPDF 3.0 parton distribution function (PDF) was used for all 2016 samples and NNPDF 3.1 for all 2017 and 2018 samples [74]. MC samples are reweighted with true number of interactions in each event to match the level of PU observed in data.

All simulated backgrounds are summarized in Table 4.1. The dijet mass for WZZ and ZZZ at generator level is required to be larger than 100 GeV event by event in order to avoid double counting with the signal sample.

Process	Generator	Cross-section [fb]	Remarks
signal samples for 2016, (2017 and 2018)			
$ZZ \to 4l\ +\ 2\ jets$	MadGraph (LO)	0.441 (427)	$m_{jj} > 100\ GeV$
$ZZ \to 4\mu\ +\ 2\ jets$	Phantom (LO)	0.418	used to cross-check MadGraph sample
$ZZ \to 4e\ +\ 2\ jets$	Phantom (LO)	0.418	used to cross-check MadGraph sample
$ZZ \to 2e2\mu\ +\ 2\ jets$	Phantom (LO)	0.836	used to cross-check MadGraph sample
irreducible background samples for 2016, 2017 and 2018			
$ZZ \to 4l\ +\ 0,1\ jets$	MadGraph (NLO)	1218	
$gg \to ZZ \to 4l\ +$ $0,1,2\ jets$	MadGraph (LO)	5.84	cross-section computed at $\mu = m_{ZZ}/2$
$gg \to ZZ \to 4l\ +$ $1\ jet$	MadGraph (LO)	4.45	used to cross-check nominal sample
$gg \to ZZ \to 4\mu$	MCFM (LO)	1.59	used to cross-check MG5 samples
$gg \to ZZ \to 4e$	MCFM (LO)	1.59	used to cross-check MG5 samples
$gg \to ZZ \to 2e2\mu$	MCFM (LO)	3.19	used to cross-check MG5 samples
minor background samples for 2016, 2017 and 2018			
$t\bar{t}Z \to 4l2\nu$	MadGraph	253	
$WWZ\ +\ jets$	MadGraph (NLO)	165.1	
$WZZ\ +\ jets$	MadGraph (NLO)	55.7	inclusive decays, $m_{jj} < 100\ GeV$
$ZZZ\ +\ jets$	MadGraph (NLO)	14.0	inclusive decays, $m_{jj} < 100\ GeV$

Table 4.1: List of signal and background samples used in the analysis for the 2016, 2017 and 2018 data-taking periods.

Although the qqZZ background is simulated at NLO, the cross-section has been computed at NNLO [75]. Thus, NNLO/NLO k-factors for the qqZZ process is applied to the MG5 sample as a function of $m(ZZ)$.
In addition, NLO EWK corrections dependent on the initial-state quark flavor and kinematics are applied to the qqZZ background in the region $m(ZZ) > 2m(Z)$ [76].
For the ggZZ background the NLO/LO (NNLO/NLO) k-factor of 1.53 (1.64) extracted from [77, 78] was applied.

Reducible background

The reducible background for the $ZZ \rightarrow 4l$ analysis, henceforth Z+X, comes from processes which contain one or more non-prompt leptons in the four-lepton final state. The main source of such leptons are non-isolated electrons and muons coming from the decays of the heavy-flavour mesons, mis-reconstructed jets usually coming from the light-flavour quarks, and photon conversions. Any such occurrence will be referred to as the "fake lepton".

The contribution from the Z+X background is minor, and is estimated by measuring the ratios of fake electrons and fake muons which also pass the final selection criteria over those which do pass the loose selection criteria. The selection criteria are discussed in section 4.3. These ratios, referred to as the *fake rates*, are used to extract the expected background yields in the signal region.
A detailed description of the procedure is not needed to follow the analysis presented in this chapter and is left out. However, an interested reader can find the detailed discussion on the measurement of fake rates elsewhere [49, 68].

4.2.2 Data samples

This analysis uses the data collected in 2016, 2017 and 2018 data-taking periods corresponding to an integrated luminosity of $137\ fb^{-1}$. Only the data that passed the quality certification by all detector subsystems, stored in so called golden JSON files, are used in the analysis. These are processed and stored in files formats that are easier to use in the analyses. One such format, known as the MINIAOD [79], was used here.

The analysis relies on five different primary data sets (PDs): DoubleEG, DoubleMu, MuonEG, SingleElectron, and SingleMuon. Each of these PDs combines a certain collections of HLT paths with exact requirements dependent on the data-taking period. Two primary data sets, DoubleEG and SingleElectron, were merged in 2018 into EGamma PD. Run periods used, together with reconstruction versions, are listed in Table 4.2.
The HLT paths used in the three data-taking periods are shown in Tables 4.3 - 4.5.

To avoid duplicate events from different primary data sets, events are taken:

- from DoubleEG
  - if events pass the diEle trigger (HLT EleXX EleYY CaloIdXX TrackIdXX IsoXX(DZ))
  - or if events pass the triEle trigger (HLT EleXX EleYY EleZZ CaloIdXX TrackIdXX)
- from DoubleMuon
  - if events pass the diMuon trigger (HLT MuXX TrkIsoVVL MuYY TrkIsoVVL)
  - or if events pass the triMuon trigger (HLT TripleMu XX YY ZZ)
  - and if events fail the diEle and triEle triggers
- from MuEG
  - if events pass the MuEle trigger (HLT MuXX TrkIsoXX EleYY CaloIdYY TrackIdYY IsoYY)
  - or if events pass MuDiEle trigger (HLT MuXX DiEleYY CaloIdYY TrackIdYY)
  - or if events pass DiMuEle trigger (HLT DiMuXX EleYY CaloIdYY TrackIdYY)
  - and if events fail diEle, triEle, diMuon and triMuon triggers

- from SingleElectron
  - if events pass the singleElectron trigger (HLT EleXX etaXX WPLoose/Tight(Gsf))
  - and if events fail all triggers above
- from SingleMuon
  - if events pass the singleMuon trigger (HLT IsoMuXX OR HLT IsoTkMuXX)
  - and if events fail all triggers above

where XX, YY and ZZ are year-dependent thresholds.

Primary data set	**Run and reconstruction version**
DoubleMuon DoubleEG MuonEG SingleMuon SingleElectron	Run2016B-17Jul2018-v1
	Run2016C-17Jul2018-v1
	Run2016D-17Jul2018-v1
	Run2016E-17Jul2018-v1
	Run2016F-17Jul2018-v1
	Run2016G-17Jul2018-v1
	Run2016H-17Jul2018-v1
DoubleMuon	Run2017B-31Mar2018-v1
DoubleEG	Run2017C-31Mar2018-v1
MuonEG	Run2017D-31Mar2018-v1
SingleMuon	Run2017E-31Mar2018-v1
SingleElectron	Run2017F-31Mar2018-v1
DoubleMuon	Run2018A-17Sep2018-v1
MuonEG	Run2018B-17Sep2018-v1
SingleMuon	Run2018C-17Sep2018-v1
EGamma	Run2018D-PromptReco-v2

Table 4.2: The list of data samples used in the analysis. All runs for each of the data streams are used, for a total of 76 primary data sets in the MINIAOD format.

HLT path	prescale	primary data set
HLT_Ele17_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	1	DoubleEG
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	1	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	1	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	1	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL	1	DoubleMuon
HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL	1	DoubleMuon
HLT_TripleMu_12_10_5	1	DoubleMuon
HLT_Mu8_TrkIsoVVL_Ele17_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu17_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu23_TrkIsoVVL_Ele8_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL	1	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL	1	MuonEG
HLT_Ele25_eta2p1_WPTight	1	SingleElectron
HLT_Ele27_WPTight	1	SingleElectron
HLT_Ele27_eta2p1_WPLoose_Gsf	1	SingleElectron
HLT_IsoMu20 OR HLT_IsoTkMu20	1	SingleMuon
HLT_IsoMu22 OR HLT_IsoTkMu22	1	SingleMuon

Table 4.3: HLT paths for 2016 data-taking period

HLT path	prescale	primary data set
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_*	1	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	1	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	1	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8	1	DoubleMuon
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8	1	DoubleMuon
HLT_TripleMu_12_10_5	1	DoubleMuon
HLT_TripleMu_10_5_5_D2	1	DoubleMuon
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	1	MuonEG
HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	1	MuonEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	1	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ	1	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL	1	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL_DZ	1	MuonEG
HLT_Ele35_WPTight_Gsf_v*	1	SingleElectron
HLT_Ele38_WPTight_Gsf_v*	1	SingleElectron
HLT_Ele40_WPTight_Gsf_v*	1	SingleElectron
HLT_IsoMu27	1	SingleMuon

Table 4.4: HLT paths for 2017 data-taking period

HLT path	prescale	primary data set
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_v*	1	EGamma
HLT_DoubleEle25_CaloIdL_MW_v*	1	EGamma
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8_v*	1	DoubleMuon
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_v*	1	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*	1	MuonEG
HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*	1	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ_v*	1	MuonEG
HLT_Ele32_WPTight_Gsf_v*	1	EGamma
HLT_IsoMu24_v*	1	SingleMuon

Table 4.5: HLT paths for 2018 data-taking period

4.3 Event selection

The final state in this analysis consists of at least two Z bosons decaying into pairs of oppositely charged leptons. The hallmark sign of the signal events are the two hadronic jets with large pseudo-rapidity gap between them. In order to maximize the measurement sensitivity, a set of selection criteria was used.
The objects reconstruction is based on the PF algorithm which uses information from all CMS subdetectors to identify individual particles within an event. These, so called, PF candidates are then classified as either electrons, muons, photons, neutral hadrons or charged hadrons. Higher-level objects such as jets and isolated leptons are created from PF candidates [80, 81].

Electrons

Reconstructed electrons with $p_T > 7\ GeV$ and $|\eta| < 2.5$ that also satisfy a loose primary vertex constraint defined by $|d_{xy}| < 0.5\ cm$ and $|d_z| < 1\ cm$, so called *loose electron*s, are considered for the analysis. Requirements on SIP parameter, presented in section 3.3.1, were imposed as well. In addition, leptons coming from the decaying Z bosons are required to be isolated as discussed in section 3.3.3. To account for the detector effects on electron momentum and energy, corrections were applied on MC simulations using the information from the data. $Z \to ee$ sample was used to match the reconstructed dielectron mass spectrum in data to the one in simulation. This was discussed in section 3.2.5. Eventual discrepancies between the data and MC samples is corrected as presented in section 3.4.1.
Those electrons that pass all presented requirements, so called *tight electrons*, are considered candidates from which a Z bosons can be built.

Muons

Loose muons are defined with $p_T > 5\ GeV$, $|\eta| < 2.4$, $|d_{xy}| < 0.5\ cm$ and $|d_z| < 1\ cm$. The same requirements on SIP parameter, as for electrons, are required.
Unlike for electrons, muon identification and isolation are done separately. Loose muons with $p_T < 200\ GeV$ are considered identified muons if they also pass the PF muon ID, while loose muons with $p_T > 200\ GeV$ are considered identified muons if they the PF ID or the Tracker High-p_T ID [68].
Muons are required to be isolated and this is done using the PF-based isolation. Muon isolation is defined by the parameter R_{iso} which measures activity in the cone of radius ΔR around the lepton and is defined as

$$R_{iso} = \left[\sum_{\substack{charged \\ hardons}} p_T + max\ (0, \sum_{\substack{neutral \\ hadrons}} E_T + \sum_{photons} E_T - \Delta\beta) \right] \Big/ p_T^l$$

where the sum runs over the charged and neutral hadrons and photons in the cone of radius ΔR around the lepton. The $\Delta\beta$ correction defined as $\Delta\beta = \frac{1}{2} \sum_{PU}^{chargedhad.} p_T$ gives an estimate of the energy deposit of neutral particles from the PU vertices and is used to remove the PU contribution for muons. The parameter ΔR is set to 0.3, and the isolation requirement is satisfied if $R_{iso} < 0.35$. Muon momentum scale is measured in data by fitting a CB function to the di-muon mass spectrum around the Z boson peak in the $Z \to \mu\mu$ control region.
Like for electrons, the discrepancy between the data and MC is cured by applying SFs obtained using the TnP.
Those muons that pass all presented requirements, so called *tight muons*, are considered candidates from which a Z bosons can be built

FSR recovery

The Final State Radiation (FSR) recovery algorithm was simplified since the Run 1, without degrading the performance. Since the effect of FSR on this analysis is small, the details on algorithm itself are omitted. An interested reader can find the full description elsewhere [49].

Jets

Jets are reconstructed from PF candidates using the $anti-k_t$ algorithm with a distance parameter $R = 0.4$ [82], after rejecting charged hadrons that are associated to a PU primary vertex. In order to be included in the analysis, all jets must have a corrected p_T larger than 30 GeV and should be within $|\eta| < 4.7$
In order to achieve a good reconstruction efficiency and to mitigate background and PU effects, loose ID criteria [will be defined in chapter 2] was applied on jets. In order to mitigate the PU contamination, a multivariate variable, the pileup jet ID (PUJetID), based on the compatibility of the associated tracks with the primary vertex and the topology of jet shape, was applied. Additionally, jets are cleaned from any tight lepton and FSR photons by requiring $\Delta R(j, l/\gamma) > 0.4$.

Since the detector response to particles is not linear, the energy of the reconstructed jets does not correspond to the true particle-level energy. For this reason, the reconstructed jet energy is corrected to take into account effects such as interactions with matter, PU, and detector response and response. These corrections are derived from simulations and are crosschecked by studying energy balance in dijet, multijet, γ + jet and leptonic Z/γ + jet events [20,83].

Unpredicted issues occurred during the three data-taking periods, which impact the quality of the reconstructed jets. In order to remedy the situation, additional requirements were imposed on jets. In 2018 it was noticed that a significant fraction of ECAL trigger primitives (TPs) in the forward region were wrongly associated with the previous bunch crossing. This was due to the degraded transparency of the ECAL crystals in the forward regions and progressed through 2016 and 2017. If the early fired L1 object has E_T above the treshold, a previous event will be sent to the HLT instead of the current event that will be rejected. This feature in the 2016 and 2017 data-taking periods is called L1 prefiring and was mitigated by calculating a probability that event didn't prefire and then applying this as a weight to the simulations. This was corrected in 2018 by a recalibration of the ECAL [55].

An increase in the ECAL noise in the 2017 data-taking period caused the appearance of the peaks (henceforth "horns") in the jet η distributions around $2.5 < |\eta_{jet}| < 3$. An effect of these horns on the analysis was tested by removing soft jets with $p_T < 50\ GeV$ in $2.65 < |\eta| < 3.139$ region. No significant impact was observed.

ZZ selection The four lepton candidates are build from the tight leptons discussed earlier. An additional lepton cleaning is performed by requiring the distance between reconstructed electron and muon, ΔR, be larger than 0.4. This removes fake electrons that arise from the muon track being wrongly matched to the electromagnetic cluster coming from an FSR emission of the muon.

A *Z candidate* is defined as the pair of the same-flavor, and opposite charge leptons (e^+e^- or $\mu^+\mu^-$) with dilepton invariant mass withing window $60\ GeV < m_{ll} < 120\ GeV$. The Z boson mass includes FSR photons if present.

A *ZZ candidate* is defined as a pair of non-overlaping Z candidates and satisfies the following requirements:

1. **Ghost removal**: $\Delta R(\eta, \phi) > 0.02$ between each of the four leptons.
2. **lepton p_T**: two out of the four selected leptons should satisfy $p_T(l_1) > 20\ GeV$ and $p_T(l_2) > 10\ GeV$.
3. **Z mass**: the mass of both Z_1 and Z_2 must be larger than $60\ GeV$ in order to comply with MC samples that do not describe the off-shell ZZ^* distributions.
4. **four-lepton invariant mass**: $m_{4l} > 180\ GeV$ in order to comply with MC samples that do not describe the off-shell ZZ^* distributions.
5. **QCD suppression**: regardless of flavor, all four opposite-sign pairs that can be built from the four leptons must satisfy $m_{ll} > 4\ GeV$. Selected FSR photons are not used in the calculation because a dilepton coming from QCD processes (e.g. J/Ψ) may have photons in vicinity (e.g. from π^0).
6. **"smart cut"**: defining Z_a and Z_b as the mass-sorted alternative pairing Z candidates (Z_a is the one with mass closest to the nominal Z mass), that satisfy $NOT(|m_{Za} - m_Z| < |m_{Z1} - m_Z|\ AND\ m_{Zb} < 12GeV)$. Here, the FSR photons are not included in calculation of the m_Z. This cut removes $4e$ and 4μ candidates where the alternative pairing looks like an on-shell Z boson accompanied by a low-mass lepton pair.

Only events containing at least one selected ZZ candidate are kept. If more ZZ pairs pass the selection requirements, a pair with the largest scalar p_T sum of the leptons constituting the Z_2 candidate is selected. This is because the false ZZ candidates are likely to be built from fake leptons which are more prominent at low p_T.
A ZZ pair that satisfies all requirements is selected with Z_1 being the one with higher p_T, and the Z_2 being the other one.

Inclusive and VBS selections

In order to select a VBS enriched part of the phase space, an additional set of requirements is imposed. At least two jets with $|\eta| < 4.7$ and $p_T > 30\ GeV$ are required in an event. In case more events are present in an event, the two with the highest p_T, referred to as the *tagging jets*, are taken. The tagging jets are required to have the invariant mass above 100 GeV in order to suppress hadronic WZ decays.
This set of requirements, on top of the ZZ selection, is referred to as the *inclusive selection* and was used to measure the signal significance, the total fiducial cross-sections and to impose limits on the aQGCs.

In addition, two more selections were defined to cross-check the signal extraction with the cut-based selection and to perform the measurement of the VBS and VBS+QCD cross-sections. A *loose VBS selection* requires, on

top of the ZZ selection, $m_{jj} > 400\ GeV$ and $|\Delta\eta| > 2.4$. A *tight VBS selection* requires $m_{jj} > 1\ TeV$ and $|\Delta\eta| > 2.4$ on top of the ZZ selection.

A control region, used to check the agreement between the data and MC, is defined by requiring events to pass a ZZjj inclusive selection, but to fail at least one condition of the loose VBS selection.
All selection criteria are summarized in Table 4.6.

lepton candidates	$p_T^e > 7GeV \qquad p_T^\mu > 5GeV$ $\|\eta\|^e < 2.5 \qquad \|\eta\|^\mu < 2.4$ $\|d_{xy}\| < 0.5\ cm$ $\|d_z\| < 1\ cm$ $\|SIP_{3D}\| < 4$ ID passed iso. in $ID \qquad R_{iso}^\mu < 0.35$
jet candidates	$p_T > 30\ GeV \qquad \|\eta\| < 4.7$ $\Delta R(j, l/\gamma) > 0.4$ ID passed L1 prefiring correction
Z candidate	tight lepton pair (e^+e^- or $\mu^+\mu^-$) $60\ GeV < m_{ll} < 120\ GeV$
ZZ selection	require pair of non-overlapping Z bosons $\Delta R(\eta, \phi) > 0.02$ between each of the four leptons $p_T(l_1) > 20\ GeV \qquad p_T(l_2) > 10\ GeV$ $m_{Z1} > 60\ GeV \qquad m_{Z2} > 60\ GeV$ $m_{4l} > 180\ GeV$ QCD suppression cut "smart" cut
Inclusive ZZjj selection	ZZ selection + $m_{jj} > 100GeV$
loose VBS selection	ZZ selection + $m_{jj} > 400GeV + \|\Delta\eta_{jj}\| > 2.4$
tight VBS selection	ZZ selection + $m_{jj} > 1\ TeV + \|\Delta\eta_{jj}\| > 2.4$

Table 4.6: Summary of the analysis selection criteria.

4.4 VBS observables

The smoking gun sign of VBS are the two hadronic jets separated by a large psudo-rapidity gap. Therefore, the most important kinematic variables describing a VBS process are the dijet invariant mass, m_{jj} and the difference in pseudo-rapidity between the two tagging jets, $\Delta\eta_{jj}$.

Variables $\eta^*(Z_1)$ and $\eta^*(Z_2)$, so called Zeppenfeld variables, were first introduced as a means of isolating events with no gluon emissions between the tagging jets in the vector boson fusion (VBF) processes [84]. Therefore, they measure activity between the two tagging jets.

Other variables used to isolate VBS are the ratio between the p_T of the tagging jet system and the scalar p_T sum of the tagging jets ($R^{jet}_{p_T}$) and the event balance ($R^{hard}_{p_T}$) defined as the transverse component of the vector sum of the Z bosons and leading jets momenta normalized to the scalar p_T sum of the same objects. The $qgtagger(j_i)$ variables quantify a probability that jets are originating from quarks rather than gluons.

A full list of variables used for signal extraction is shown in Table 4.7.

variable	definition
m_jj	invariant mass of the two leading jets
$\Delta\eta_{jj}$	pseudo-rapidity separation of the two leading jets
m_{4l}	invariant mass of the ZZ pair
$\eta^*(Z_1)$	direction of the Z_1 relative to the leading jets: $\eta^*(Z_1) = \eta(Z_1) - \frac{\eta(j_1)+\eta(j_2)}{2}$
$\eta^*(Z_2)$	direction of the Z_2 relative to the leading jets: $\eta^*(Z_2) = \eta(Z_2) - \frac{\eta(j_1)+\eta(j_2)}{2}$
$R^{hard}_{p_T}$	transverse component of the vector sum of the two leading jets and four leptons normalized to the scalar p_T sum of the same objects $R^{hard}_{p_T} = \frac{(\sum_{i=4l,\ 2j} \vec{V_i})_{transverse}}{\sum_{4l,\ 2j} p_T(i)}$
$R^{jet}_{p_T}$	transverse component of the vector sum of the two leading jets normalized to the scalar p_T sum of the same objects $R^{jet}_{p_T} = \frac{(\sum_{i=2j} \vec{V_i})_{transverse}}{\sum_{2j} p_T(i)}$
$p_T(j_1)$	transverse momentum of the leading jet
$p_T(j_2)$	transverse momentum of the second-leading jet
$y(j_1)$	rapidity of the leading jet: $y(j_1) = \frac{1}{2} ln \left[\frac{E(j_1)+p_L(j_1)}{E(j_1)-p_L(j_1)}\right]$
$y(j_2)$	rapidity of the second-leading jet: $y(j_2) = \frac{1}{2} ln \left[\frac{E(j_2)+p_L(j_2)}{E(j_2)-p_L(j_2)}\right]$
$\eta(j_1)$	pseudo-rapidity of the leading jet
$\eta(j_2)$	pseudo-rapidity of the second-leading jet
$\lvert\eta_{min}(j)\rvert$	smallest absolute value of the jet pseudo-rapidity
$\lvert\eta_{max}(j)\rvert$	largest absolute value of the jet pseudo-rapidity

$\sum \eta(j)$	sum of the pseudo-rapidity of selected jets
$\sum \lvert\eta(j)\rvert$	sum of the absolute value of the pseudo-rapidity of selected jets
$m_{jj}/\Delta\eta(jj)$	quotient of the invariant mass and the pseudo-rapidity gap of the two leading jet
$qgtagger(j_1)$	probability that the leading jet is coming from a quark rather than a gluon
$qgtagger(j_2)$	probability that the second-leading jet is coming from a quark rather than a gluon
$p_T(l_3)$	transverse momentum of the third-leading lepton
$\lvert\eta_{min}(lep)\rvert$	smallest absolute value of the lepton pseudo-rapidity
$\lvert\eta_{max}(lep)\rvert$	largest absolute value of the lepton pseudo-rapidity
$p_T(Z_1)$	transverse momntum of the Z_1
$p_T(Z_2)$	transverse momntum of the Z_2
$y(Z_1)$	rapidity of the Z_1: $y(Z_1) = \frac{1}{2} ln \left[\frac{E(Z_1)+p_L(Z_1)}{E(Z_1)-p_L(Z_1)}\right]$
$y(Z_2)$	rapidity of the Z_2: $y(Z_2) = \frac{1}{2} ln \left[\frac{E(Z_2)+p_L(Z_2)}{E(Z_2)-p_L(Z_2)}\right]$
$\Delta\phi(Z_1, Z_2)$	angular separation between the two Z bosons

Table 4.7: Set of 28 variables used to check the agreement between the data and MC.

Distributions of variables m_{jj} and $|\Delta\eta_{jj}|$, used to define the control region, are shown in Fig. 4.3 for all three data-taking periods and demonstrate a good agreement between the data and the simulation.
A good agreement between the data and the simulation is also observed in the signal region for a full set of variables used to extract the signal. This can be seen in Fig. 4.4 for the 2018 data-taking period with the baseline selection applied. All distributions for the three data-taking periods can be found in Appendix A.

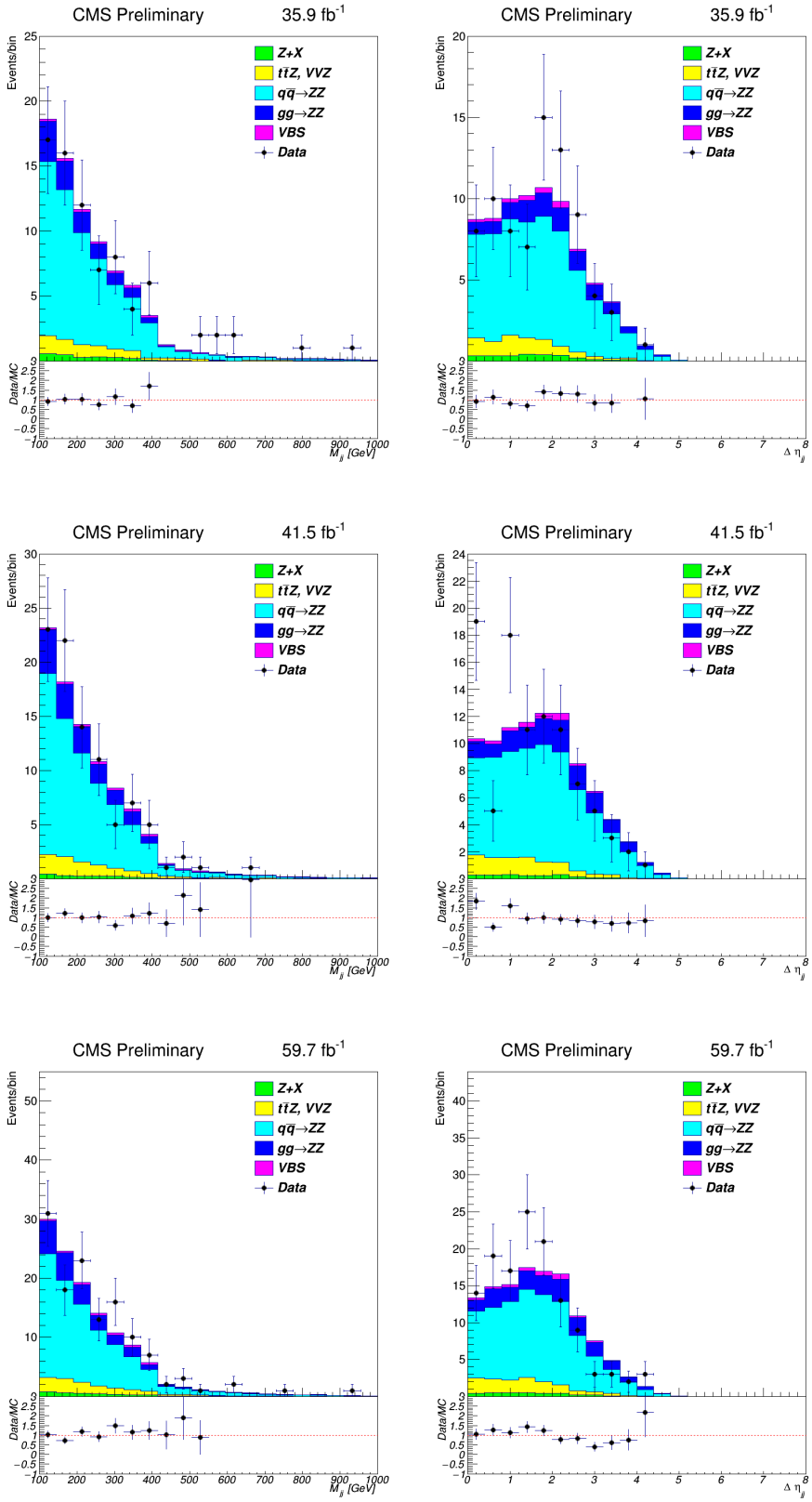
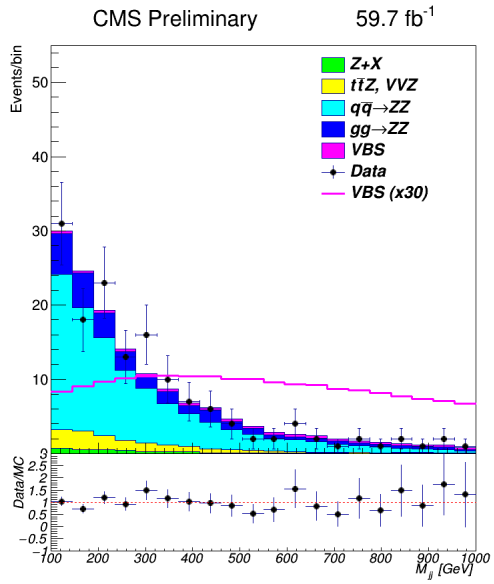
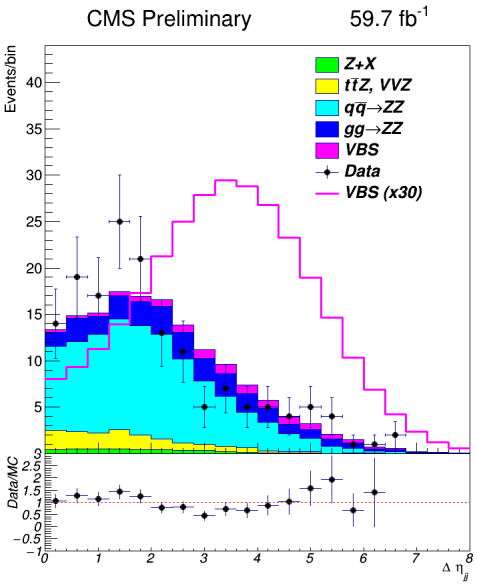
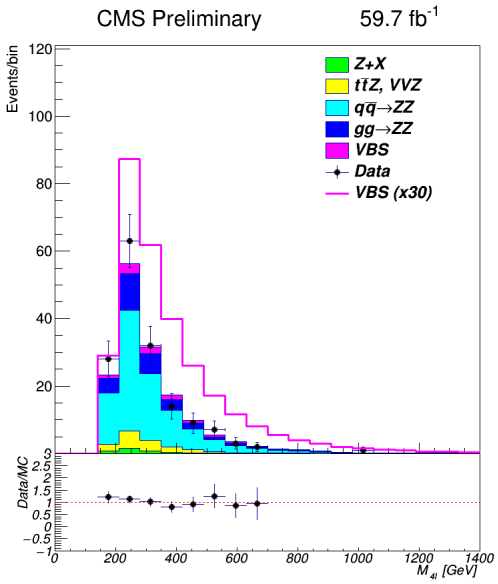
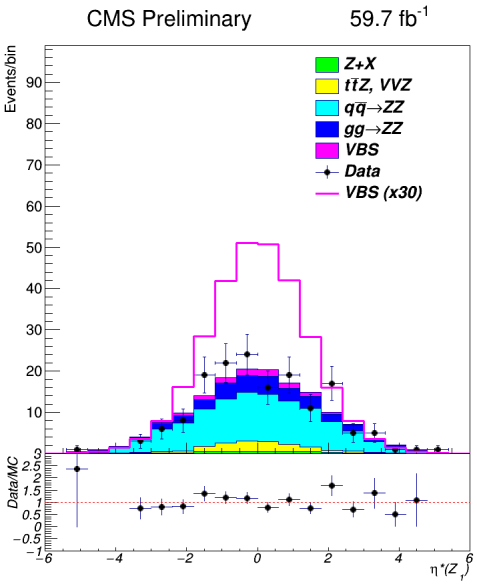


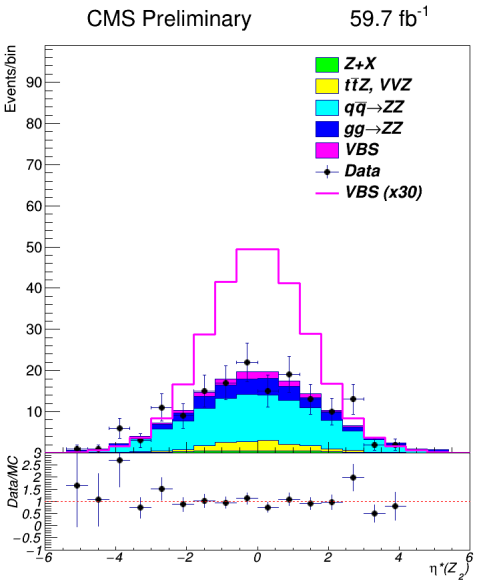
Figure 4.3: A comparison of data to background and signal estimations in 2016 (top row), 2017 (middle row) and 2018 (bottom row) samples in the control region.

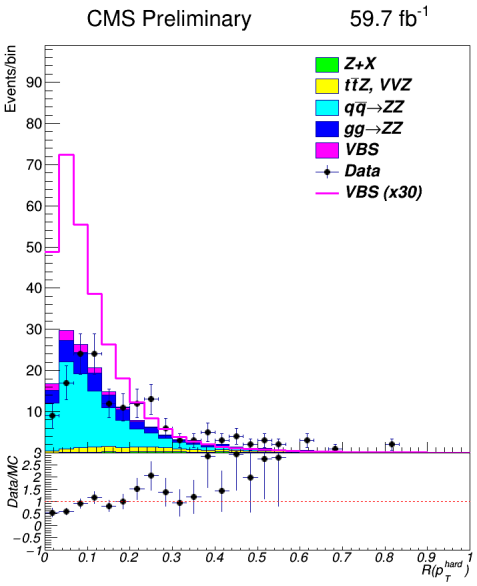

CMS Preliminary 59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
Mjj [GeV]

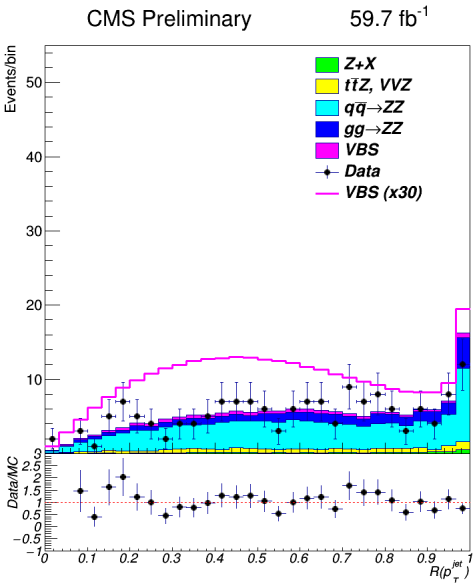

CMS Preliminary 59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
Δ ηjj

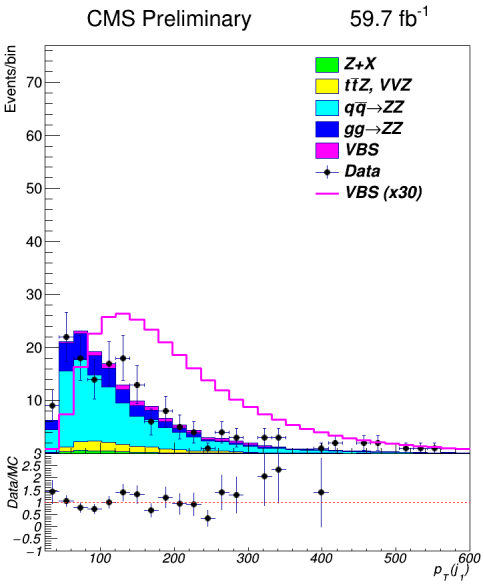

CMS Preliminary 59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
M4l [GeV]

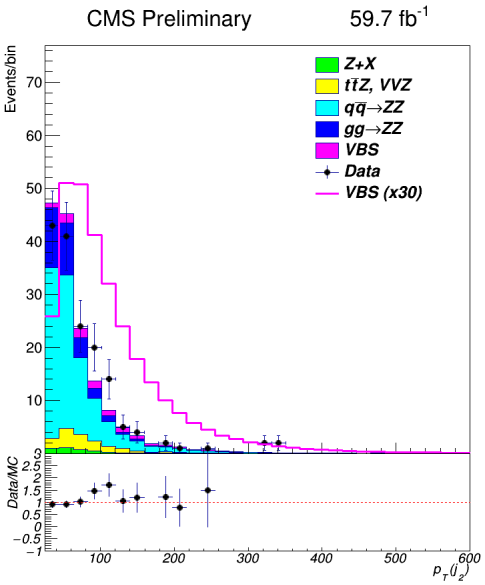

CMS Preliminary 59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η*(Z1)

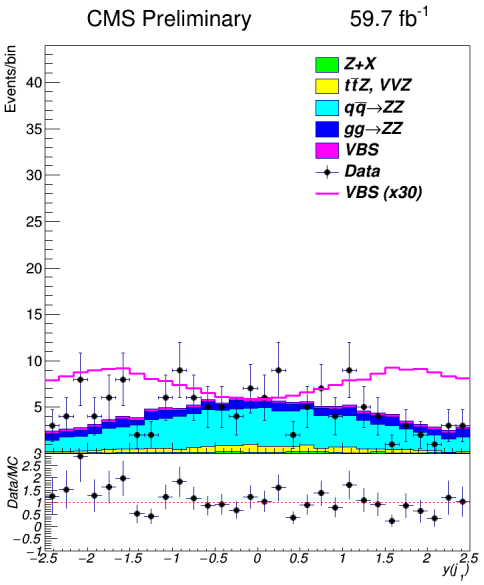

CMS Preliminary 59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η*(Z2)

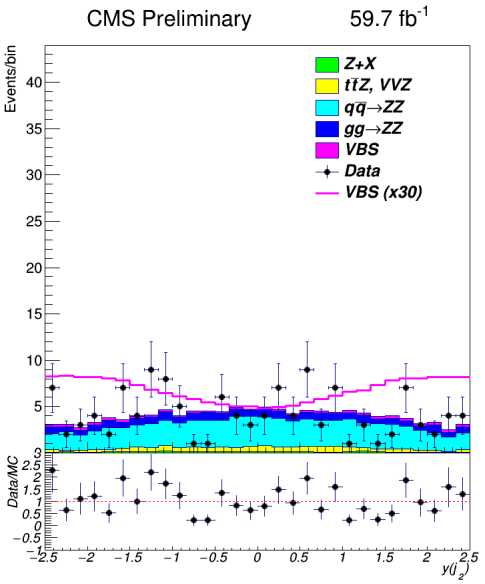

CMS Preliminary 59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(pT hard)

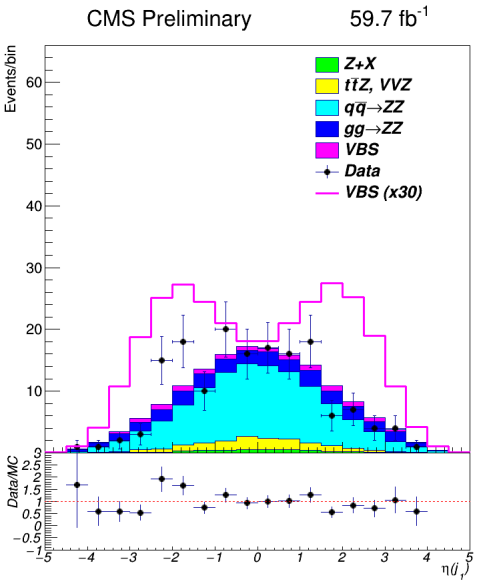

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(p_T^jet)

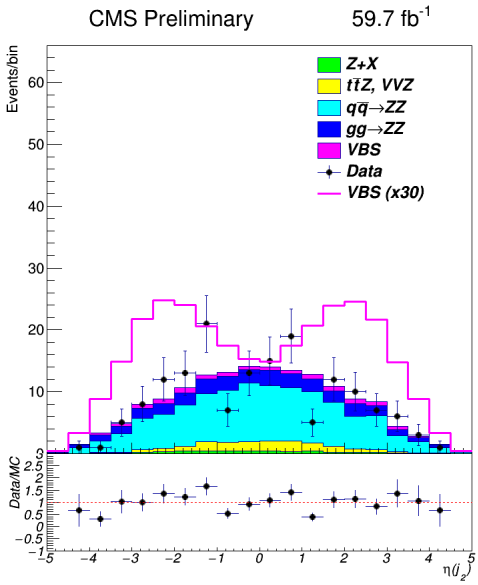

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
p_T(j_1)

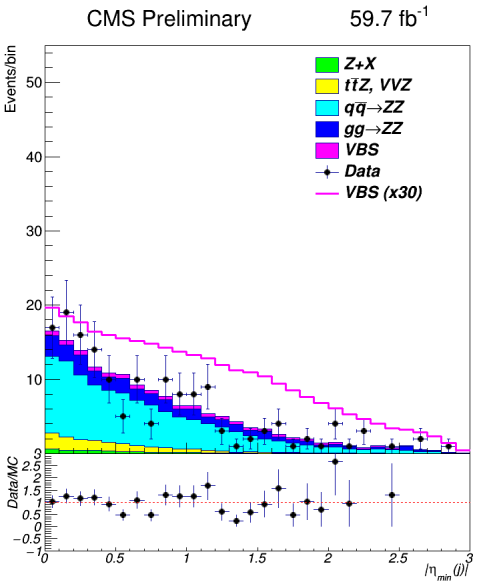

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
p_T(j_2)

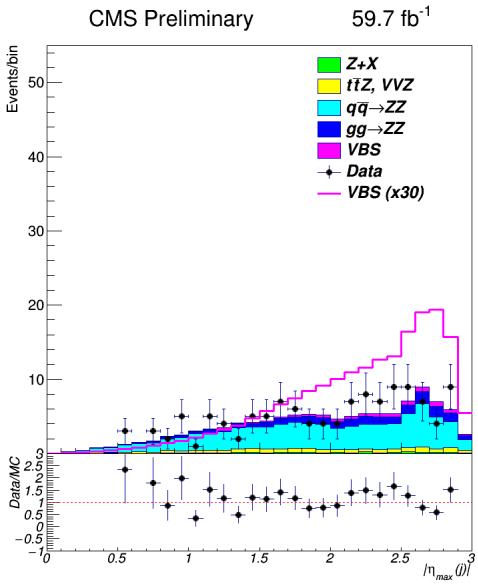

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j_1)

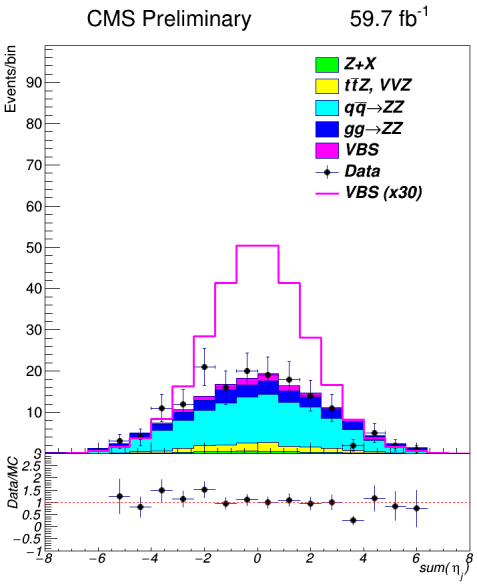

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j_2)

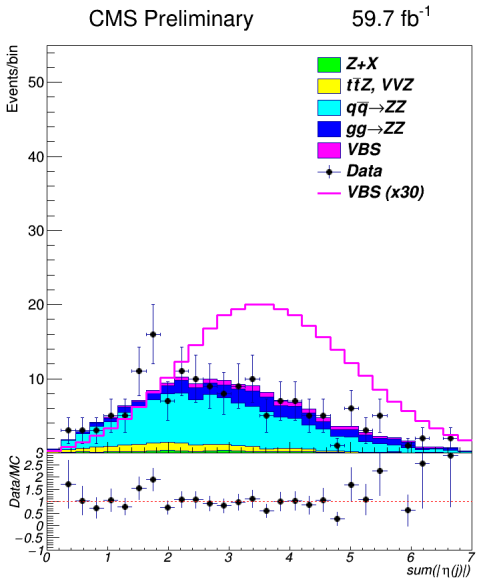

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j_1)

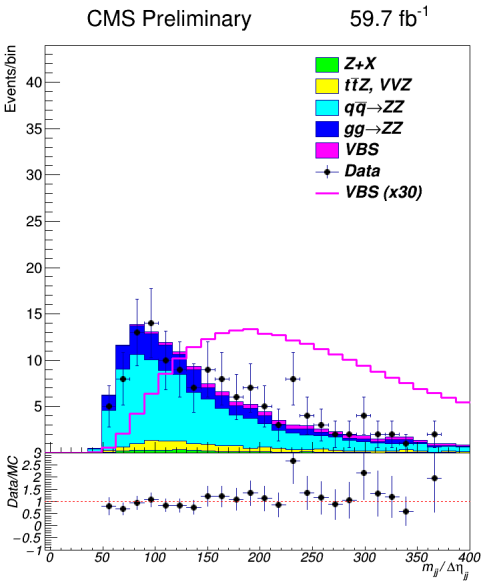

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j2)

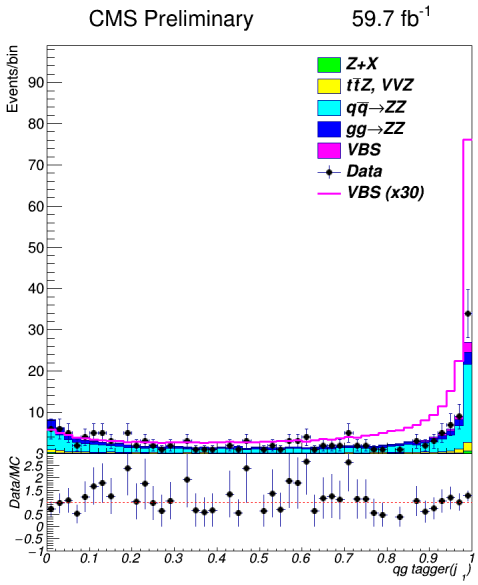

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmin(j)|

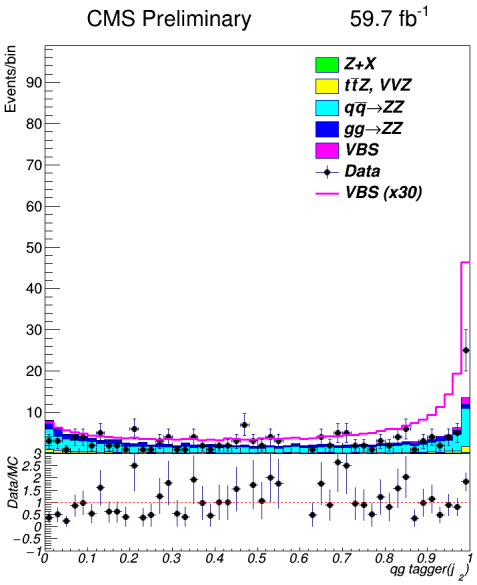

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmax(j)|

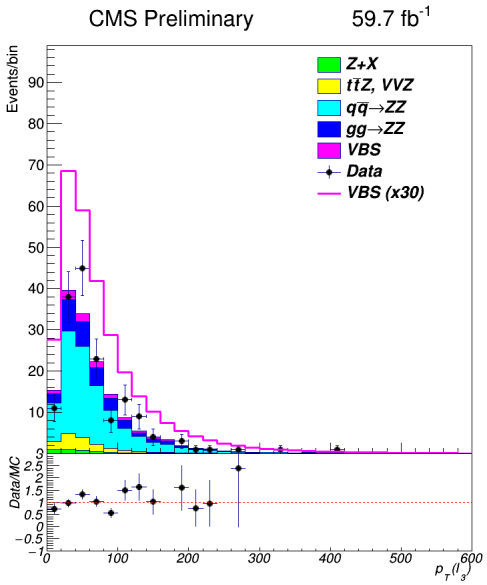

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(ηj)

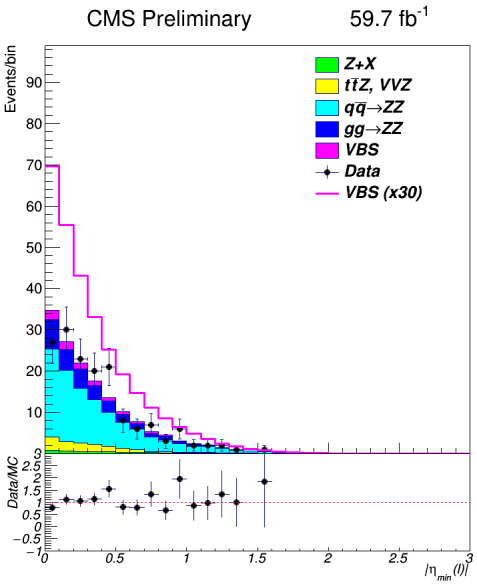

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(|η(j)|)

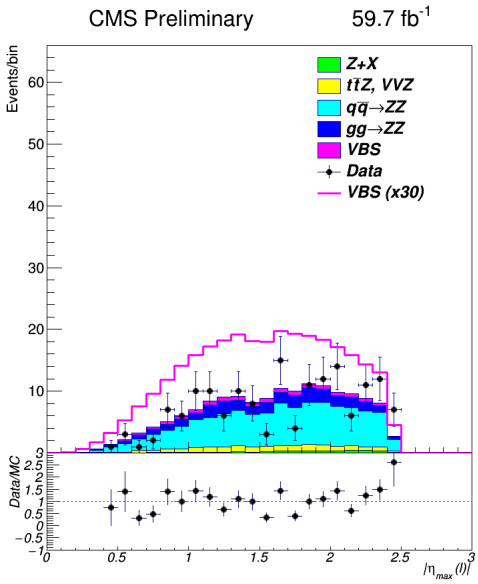

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
mjj/Δηjj

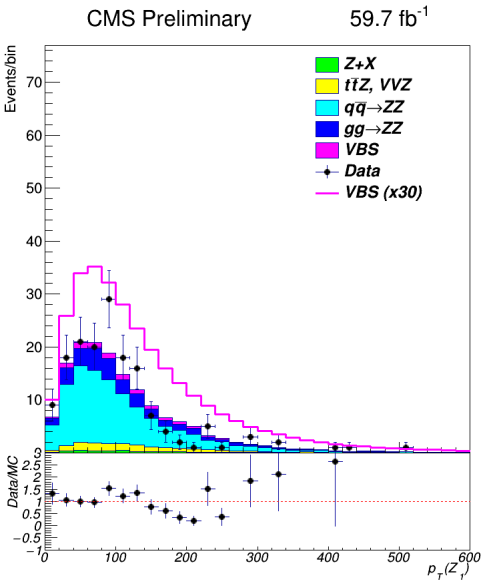

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
qg tagger(j 1)


CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
qg tagger(j 2)


CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(l3)


CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmin(l)|


CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmax(l)|


CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(Z1)

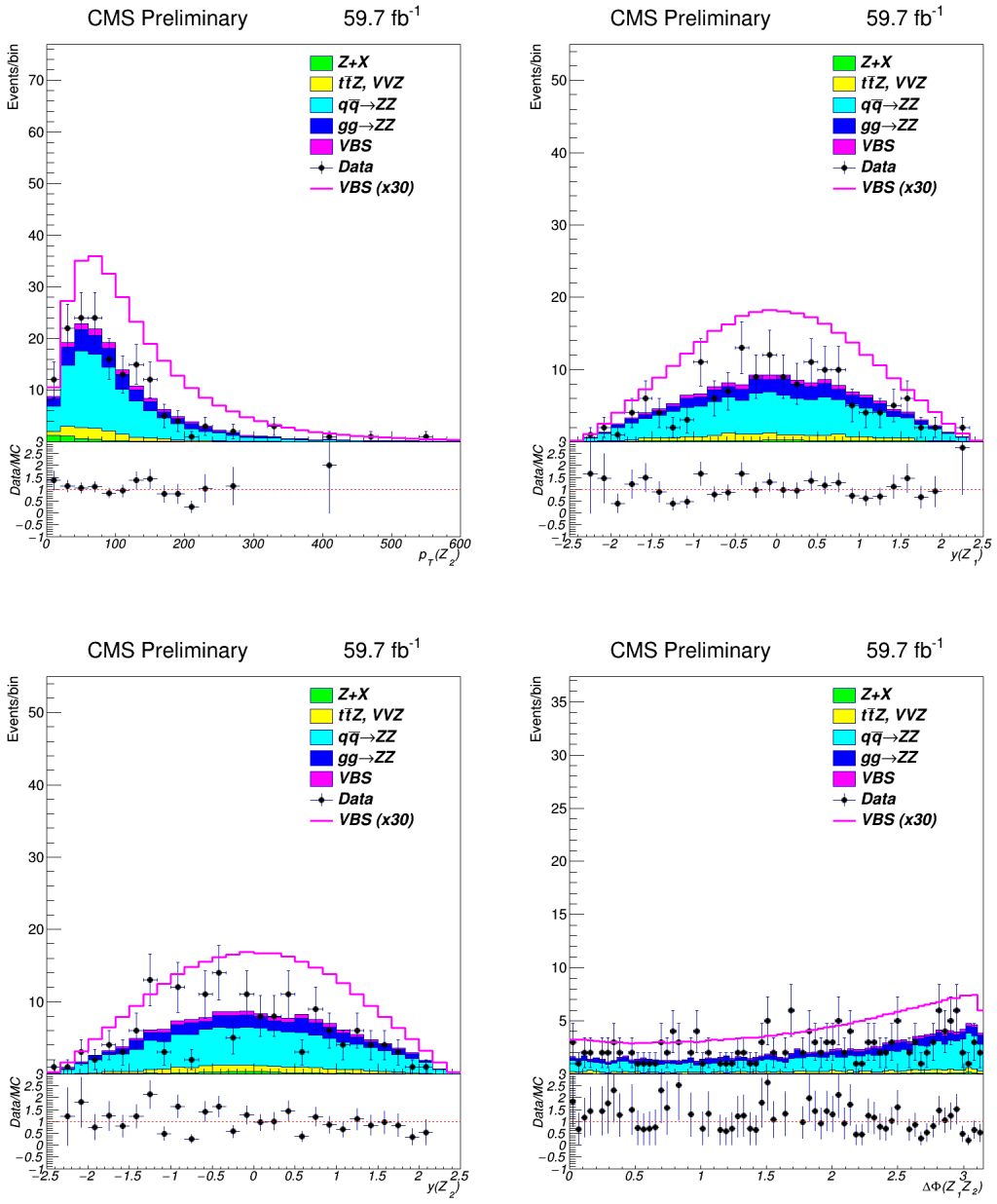


Figure 4.4: Comparison of data to background and signal estimations in 2018 samples used in the analysis. All 28 variables from Table 4.7 are shown.

4.5 Signal extraction and the cross-section measurement using the MELA discriminant

4.5.1 The MELA discriminant

In the published paper [85] on the search for the VBS in the $4l$ final state using the Run 2 data, the signal extraction approach was based on a kinematic discriminant that uses *MCFM* matrix elements for the EWK signal and the main $qqZZ$ background to describe process probabilities. This is the basis of the Matrix Element Likelihood Approach (MELA). AT the hearth of MELA lies the fact that the kinematics of the VBS $4l$ final state coming from the decay of vector bosons can be FULLY described by THE set of variables summarized in Table 4.8 and illustrated in Figure 4.5 [86–88].

variable	description
m_{4l}	invariant mass of the 4 final-state leptons
m_{Z_1}	invariant mass of the Z_1
m_{Z_2}	invariant mass of the Z_2
θ^*	angle between the Z_1 boson and the z axis
Φ	angle between the normal vectors of the decay planes of Z_1 and Z_2
Φ_1	angle between the beam axis and the plane of the Z_1 decay products in the $4l$ rest frame
θ_1	angle between Z_1 direction and momenta of the decay lepton in Z_1 rest frame
θ_2	angle between Z_2 direction and momenta of the decay lepton in Z_2 rest frame

Table 4.8: The set of eight variables needed to fully characterize $4l$ final state originating from the decay of Z bosons. All eight variables are used to construct the kinetic discriminant K_D.

From the set of mentioned variables, a kinematic discriminant, K_D is constructed:

$$K_D = \left[1 + c(m_{4l}) \cdot \frac{P_{QCD-JJ}(\vec{\Omega}^{4l+JJ}|m_{4l})}{P_{VBS+VVV}(\vec{\Omega}^{4l+JJ}|m_{4l})}\right]$$

In theabove expression, $P_{VBS+VVV}$ represents the probability, obtained from the *MCFM* matrix elemets, of an event coming from EWK processes. Similarly, P_{QCD-JJ} is the probability, obtained in the same way, that event originated from the QCD-induced production of the $4l2j$ final state. $\vec{\Omega}$ represents a set of invariant mass and angle variables from the Table 4.8. Finally, $c(m_{4l})$ is an m_{4l}-dependent constant that is used to bound the distribution in range [0, 1].

Figure 4.6 shows a good agreement, in the control region, of the K_D distribution between the data and the simulation for all three data-taking periods. Figure 4.7 shows the performance of the variable in discriminating EWK signal from the backgrounds. The EWK signal is visible in the region with large values of K_D. The baseline selection was applied. The figure also shows a good agreement between the data and the simulation in the VBS-enriched region.

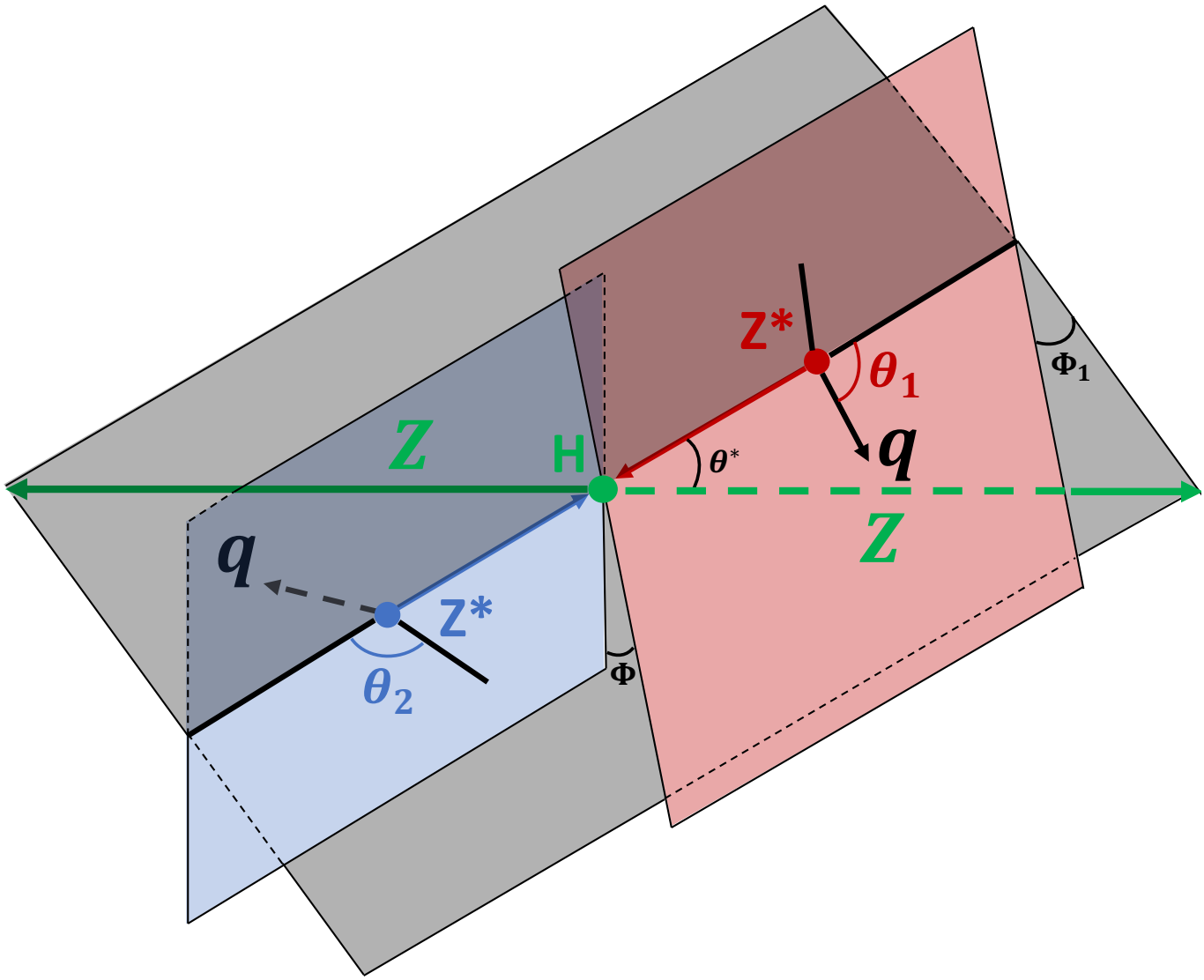


Figure 4.5: Illustration of angles defined in Table 4.8 used, together with the three invariant masses, to build the kinematic discriminant K_D.

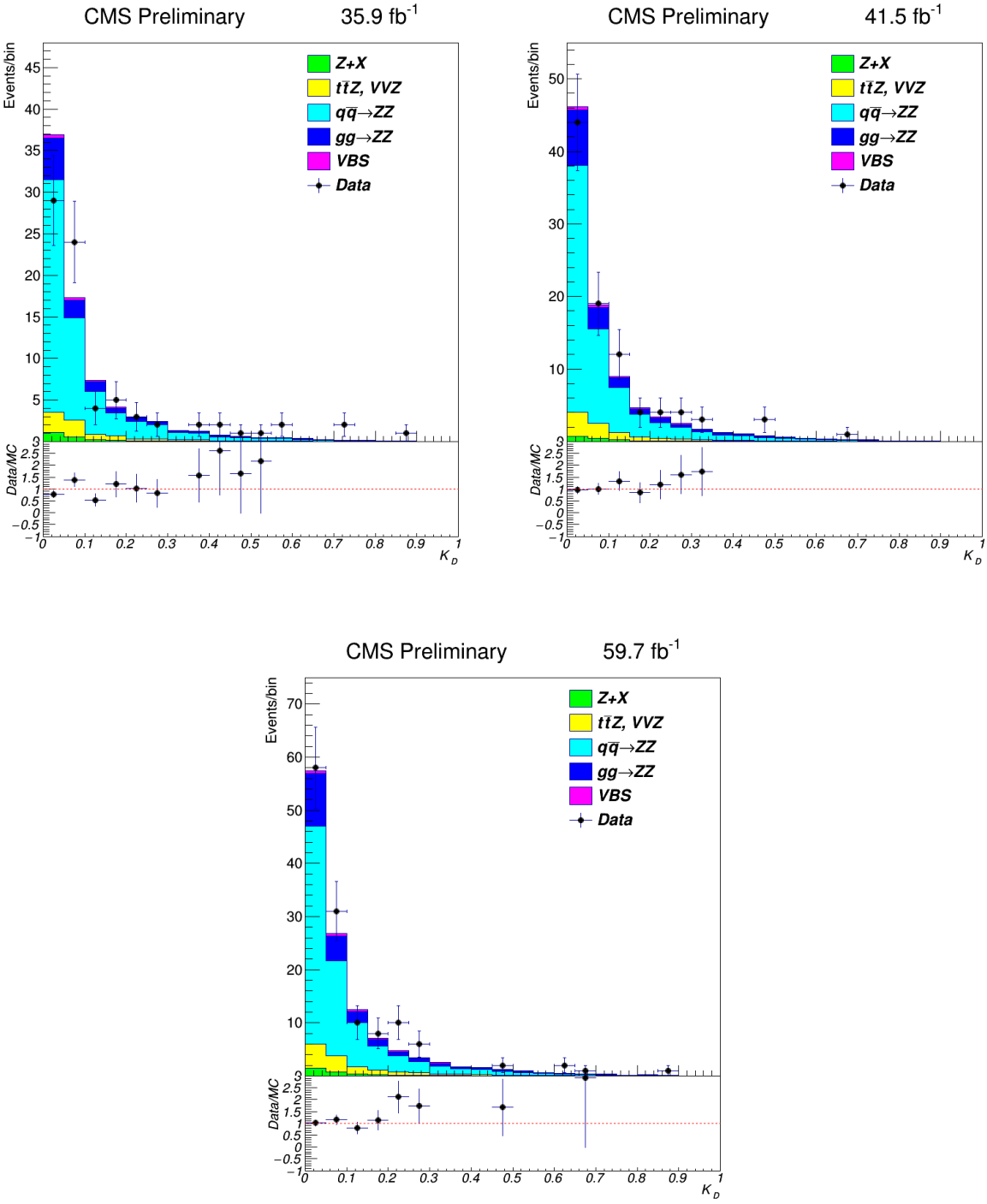


Figure 4.6: Comparison of data to background and signal estimations, in the control region, for the kinematic discriminant, K_D, in all three data-taking periods. A good agreement between the data and the simulation is observed.

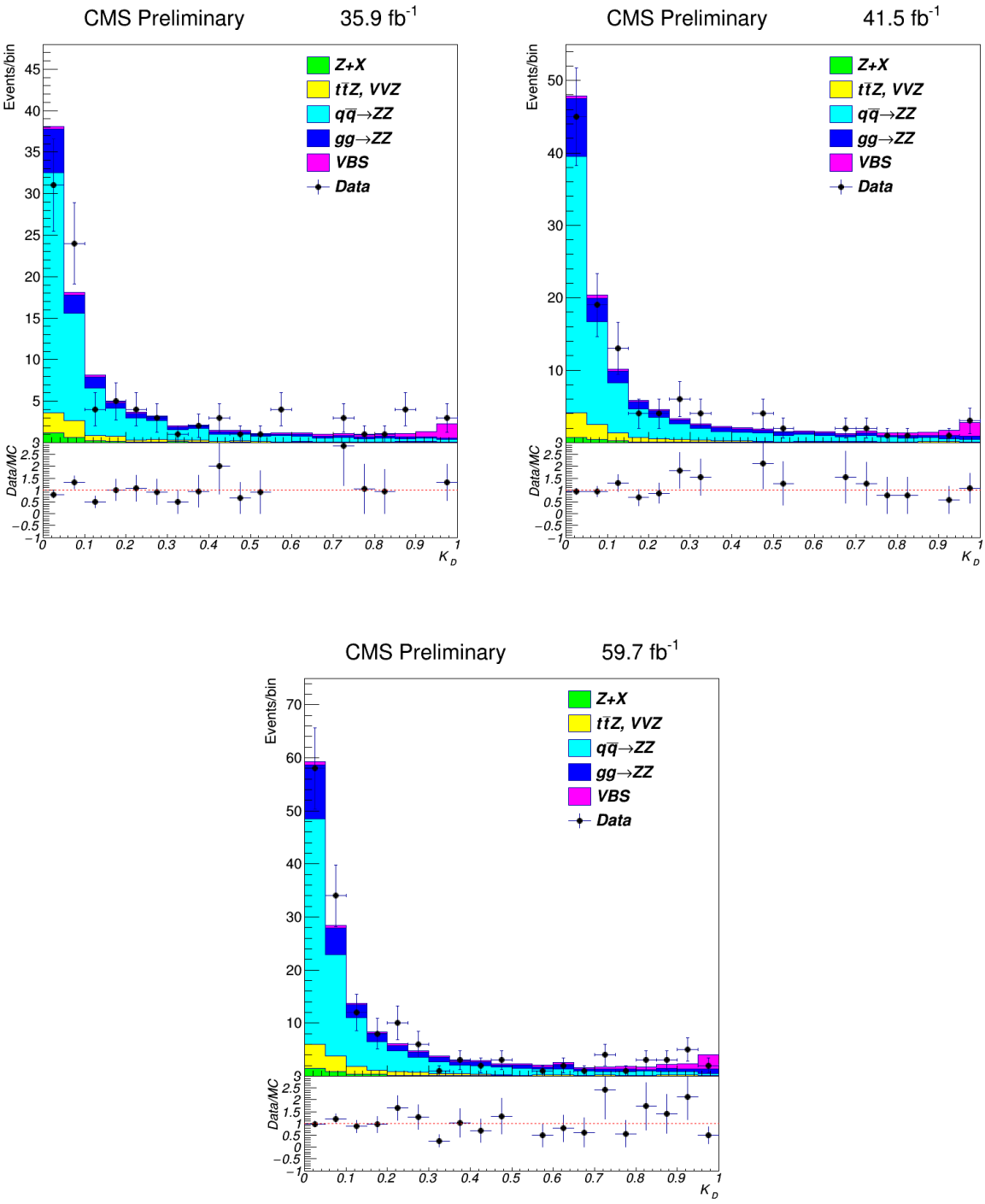


Figure 4.7: Comparison of data to background and signal estimations, in the signal region, for the kinematic discriminant, K_D, in all three data-taking periods. Plots shows the performance of the variable in discriminating between the signal and background distributions. The EWK signal is visible in the region with large values of K_D.

4.5.2 Significance and cross-section measurement

The expected and observed significance for the three data-taking periods, as well as the combined significance, was calculated using the "*combine*" tool. This tool was deigned to provide the user with command-line interface to fit a signal and background models to the data.

The previously defined matrix element discriminant, K_D, was used to produce a histogram to model each contribution of interest. These histograms, together with event yields of every process, were fed into the combine tool in the form of configuration files called datacards. All histograms were used as a template to perform a maximum likelihood fit to the observed data. This procedure was done for each year separately. The expected distributions for the signal and the irreducible backgrounds were obtained from the MC simulation. The reducible background was estimated from data.

In each of the three datacards, the systematic uncertainties from all the sources were specified and treated as nuisance parameters in the fit. The sources of the systematic uncertainties are discussed in section 4.8. In order to constrain the QCD-induced production from the background-dominated region of the K_D distribution, the shape and normalization of each contribution were allowed to vary up and down in fit. The signal significance for the integrated luminosity $\mathcal{L} = 137.1\ fb^{-1}$ was obtained by combining all three periods. This is done simply in *combine* by merging the individual datacards and performing a new fit.

The EWK and EWK+QCD cross-sections were estimated in the fiducial regions defined in Table 4.9. These were defined very closely to the selection criteria at the reco level. The same fit used to obtain the signal significance was also used to calculate the signal strength, μ, defined as the ratio of the measured cross-section to the SM expectation $\mu = \frac{\sigma}{\sigma_{SM}}$.

Since the K_D spectrum was optimized to separate the EWK signal from the backgrounds, the cross-section for the EWK component was obtained by exploiting the shape of the MELA discriminant. The procedure here was identical to the one used to obtain the EWK signal significance. On the other hand, fits that only use event counts in the three fiducial regions were used to obtain the EWK+QCD cross-section. This is possible because the EWK+QCD determination is, mostly, background-free.

The next section will describe an alternative signal extraction approach using boosted decision trees. The results for both approaches are discussed and compared in section 4.9.

Particle type	Selection
ZZjj inclusive	
Leptons	$p_T(l_1) > 20\ GeV$ $p_T(l_2) > 10\ GeV$ $p_T(l) > 5\ GeV$ $\|\eta(l)\| < 2.5$
Z and ZZ	$60 < m_{ll} < 120\ GeV$ $m_{4l} > 180\ GeV$
Jets	at least 2 $p_T(j) > 30\ GeV$ $\|\eta(j)\| < 4.7$ $m_{jj} > 100\ GeV$ $\Delta R(j,l) > 0.4$ for each $j,\ l$
VBS-enriched (loose)	
Leptons	same as ZZjj inclusive
Jets	ZZjj inclusive + $\|\Delta\eta_{jj}\| > 2.4$ $m_{jj} > 400\ GeV$
VBS-enriched (tight)	
Leptons	same as ZZjj inclusive
Jets	all above + $m_{jj} > 1\ TeV$

Table 4.9: Particle-level selections used to define the fiducial regions for EWK and EWK+QCD cross-sections

4.6 Signal extraction using Boosted Decision Trees

4.6.1 A Tool for MultiVariate Analysis (TMVA)

The signal extraction discussed in the following sections is based on a multivariate approach. For this, the *Toolkit for MultiVariate Analysis* (*TMVA*) was used. *TMVA* project started in 2005 with goal of building a consistent, feature-rich framework for the multivariate analysis (MVA). It provides a ROOT-integrated [89] environment for processing, parallel evaluation and application of classification and regression techniques. All MVA techniques implemented in the tool are based on supervised learning:

- Fisher
- Linear description (LD)
- Functional description analysis (FDA)
- Projective likelihood
- Cuts
- Probability density estimator—range search (PDE-RS)
- Probability density estimator—foam (PDE-foam)
- Neuronal network (MLP)
- Boosted decision trees (BDT)
- Support vector machine (SVM)
- Rule ensembles (RuleFit)

with each of the techniques implemented in C++/ROOT [90].
Apart from the techniques listed above, *TMVA* offers auxiliary tools such as parameter fitting and various data set transformations. It also provides training, testing and performance evaluation algorithms. Finally, it has implemented a Graphical User Interface (GUI) which enables users to easily obtain desired output plots [56].
The *TMVA* logic for both a classification and a regression problem is as follows

1. The data is fed to *TMVA* via ROOT TTrees or from ASCII file.
2. The user defines variables from the input file that will be used during the training and test phase.
3. If needed, selection cuts and event weights are defined. At this stage, *TMVA* gives to the user a convenient way to select desired preprocessing technique (normalisation, decorrelation, principal components analysis or gaussianisation).
4. The user chooses to do either classification or regression.
5. The desired MVA technique is selected.
6. The hyperparemeters are defined for the selected MVA technique.
7. The training is initiated on one part of the available data sample followed by the implementation of the training to the unknown set (i.e. the test set)

8. *TMVA* evaluates the chosen MVA method(s) and produces the result in various formats: Receiver Operating Characteristic (ROC) curve, curve of signal efficiencies and corresponding background rejection rates for each point on the ROC curve, signal significance, signal purity and a classifier distribution for the signal and background. Each value on the classifier distribution (henceforth the *cut value*) can be used to obtain a pair of (signal efficiency, background rejection) values (henceforth the *working point*).
9. *TMVA* stores the training result in the form of weights available in the "weight" file
10. Saved weight are used for the application of the training on individual signal and background samples

In this analysis *Boosted Decision Trees* (*BDT*) classifier was used to extract EWK signal from the backgrounds.

4.6.2 Introduction to Boosted Decision Trees

A decision tree is a supervised machine learning method that can be used in either classification or regression problems. In this thesis we are interested in labeling each event as either signal or background. Thus, decision trees are used here to solve a classification problem.
A decision tree is a data structure that consists of the root node, decision nodes, leaf nodes and branches. By definition, the root node is simultaneously a decision node. Every decision tree is built starting from the root node. From here, the data are split, using some conditions, into different sub-trees. The process is completed when every branch has only leaf nodes. One simple decision tree is shown in Figure 4.8. Some "buzzwords" used in decision trees theory are summarized in Table 4.10.
In order to understand how decision trees work, a simple example is prepared. The input data described by only two features, i.e. input variables, is shown in Figure 4.9. Every red ball represents a signal and every blue ball represents a background. A green ball represents a new data that will be classified once the decision tree is trained. It is not used in the forthcoming calculations.
The first step in the training of the decision tree is to load all training data in the tree. From here the root node is created. The idea behind every node is to consider all available features and select the one that does the best job in splitting the data into signal and background groups. For example, the data in Figure 4.9 can be split in two ways:

1. $x \leq -1$ or $x > -1$
2. $x \leq -4$ or $x > -4$

How does the tree decide which of the two lines it should use to split the data? This is done using the *Attribute Selection Measure* (*ASM*). The two examples of such tools are

- Gini index
- Information gain (IG)

Both give similar results, so only IG will be described here.
The IG is based on the minimization of the entropy obtained after the split. The entropy calculation is based on the information theory:

$$S = -\sum p_i \cdot log(p_i)$$

where p_i represents the probability of finding any class within a subgroup. The base of the logarithm can be arbitrarily chosen and is set to 2. In the beginning there are 10 red balls out of 20 balls in total (the green ball is not included in

the training) which gives the probability of selecting a red ball 50 %. The same argument holds for the blue balls. Thus, the entropy in the beginning has the maximal value

$$S = -\left[0.5 \cdot log(0.5) + 0.5 \cdot log(0.5)\right] = 1$$

The entropy of the region defined by $x \leq -1$ is then

$$S = -\left[\frac{4}{9} \cdot log\left(\frac{4}{9}\right) + \frac{5}{9} \cdot log\left(\frac{5}{9}\right)\right] = 0.99$$

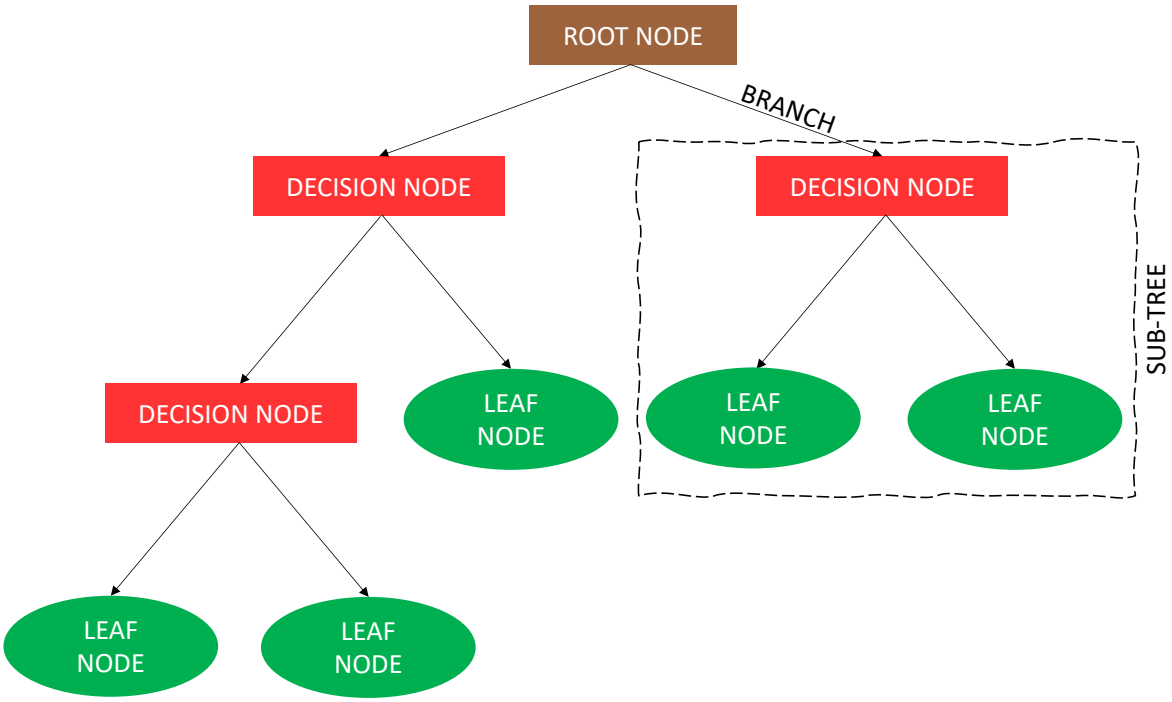


Figure 4.8: Illustration of the simple decision tree with a basic structure.

structure	definition
feature	input variable used in the training of the multivariate classifier
decision node	algorithm that splits the data depending on the value of the feature(s). Each decision node splits the structure and, therefore, creates an additional sub-tree
root node	the first decision node from which the decision tree is built
leaf node	ending node of a branch where all the data is classified as either signal or background. The branching of the tree ends at the leaf node.
branch	decision rule which creates a sub-tree.

Table 4.10: Definition of the basic decision tree structures

For the region defined by $x > -1$ this would be

$$S = -\left[\frac{6}{11} \cdot log\left(\frac{6}{11}\right) + \frac{5}{11} \cdot log\left(\frac{5}{11}\right)\right] = 0.99$$

Now one should calculate the IG from this split:

$$G = S_{parent} - \sum w_i \cdot S_{child}$$

where the factor w_i is the weight defined as the total number of balls in the region of interest divided by the total number of balls. Thus,

$$G_{-1} = 1 - \left[\frac{9}{20} \cdot 0.99 + \frac{11}{20} \cdot 0.99\right] = 0.01$$

For the region defined by $x \leq -4$ we have

$$S = -\left[\frac{1}{5} \cdot log\left(\frac{1}{5}\right) + \frac{4}{5} \cdot log\left(\frac{4}{5}\right)\right] = 0.72$$

and for the region defined by $x > -4$ we have

$$S = -\left[\frac{9}{15} \cdot log\left(\frac{9}{15}\right) + \frac{6}{15} \cdot log\left(\frac{6}{15}\right)\right] = 0.97$$

Thus, the IG for this split is

$$G_{-4} = 1 - \left[\frac{5}{20} \cdot 0.72 + \frac{15}{20} \cdot 0.97\right] = 0.09$$

From this it can be seen that splitting the data with line $X = -4$ results in a larger gain. For this reason the root node will split the data based on the condition $x \leq -4$ or $x > -4$. This is not to say that this is the best splitting option in this example. It was used merely for the demonstration purposes.

Although simple, this example describes exactly how decision tree splits the data using the available features. At every node, the tree will find the best feature using the *ASM* to split the data until nothing is left but leaves. Before the tree performance is tested on new data, a method of simplifying the tree by means of deleting unnecessary nodes, called pruning, is applied.

Next we want to test the performance of our tree by introducing new data (the green ball) and calculating how efficient the tree is in classifying it. The features of the ball will traverse through the entire tree, starting at the root node, until the leaf is reached. When this is done, our green ball will be classified as either a signal or a background.

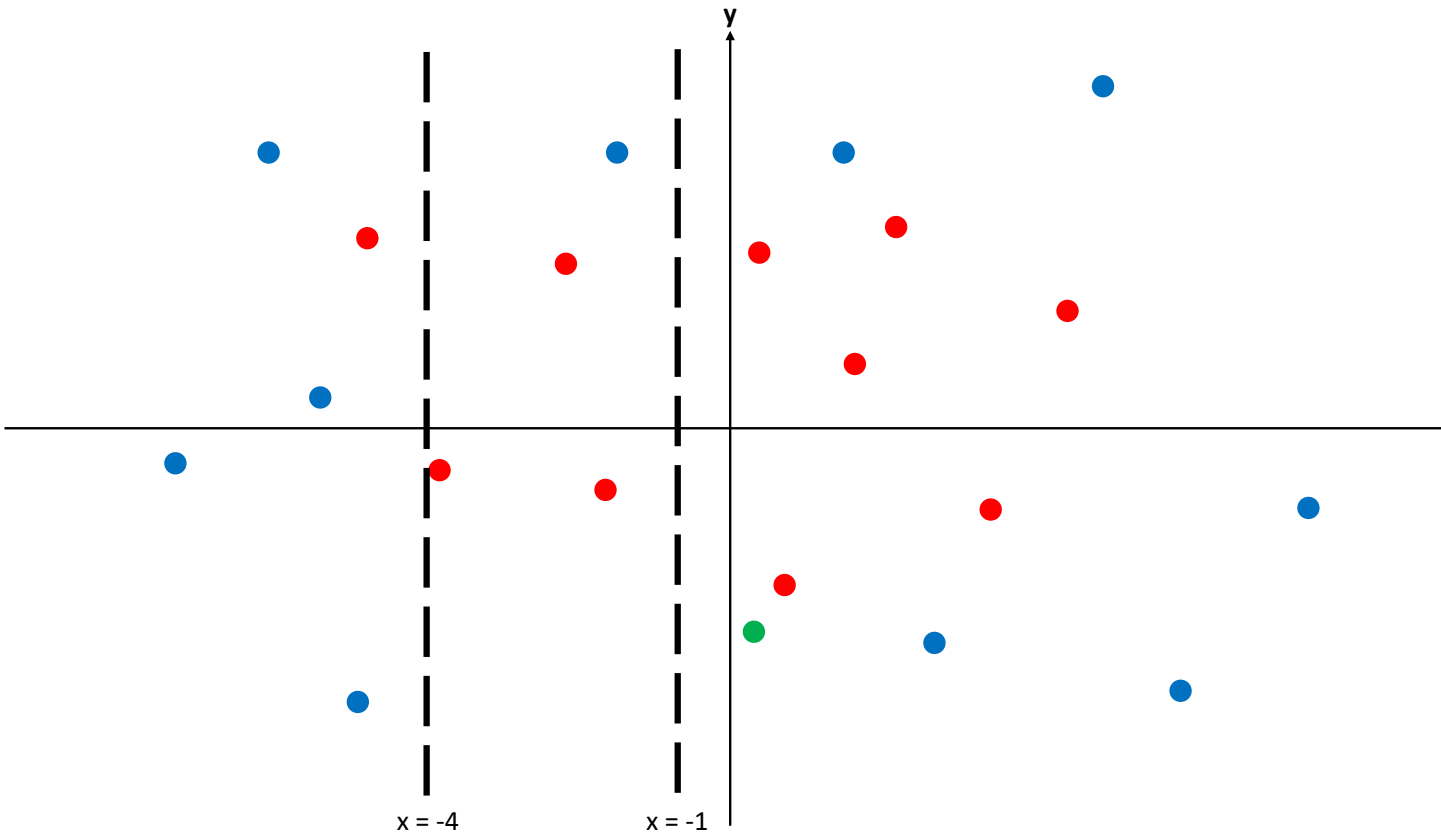


Figure 4.9: Illustration of a decision tree with a simple structure.

Boosted decision trees

The problem with using just a single decision tree, like in the example above, for doing a classification, or regression for that matter, is that a single decision tree has a tendency to overfit the data. This means that a decision tree is focusing on the noise in the data instead of the general behavior. This results in a poor performance in presence of a new data.

This problem is solved by means of boosting. Boosting is a method that relies on using many weak learners, instead of just one strong learner, to perform the task at hand. A weak learner is just a simple decision tree with a small number of leaf nodes. Boosting must not be confused with another method, called bagging, also used to combine several decision trees into one strong learner. The difference between bagging and boosting lies in the way information from many decision trees is combined into the final decision. In algorithms based on bagging, such as random forest algorithm, each tree is independent from the previous tree and the final decision is made by aggregating the predictions from all the trees.

The idea behind boosting lies in using the mistake of a previous tree to improve the prediction upon building another tree. Several boosting algorithms are available today, the most famous being AdaBost, Gradient boost and XGboost. *Gradient Boosted Decision Tree*, henceforth referred to as the *BDTG*, starts the classification training from some starting prediction. The residuals array, or simply errors array, is built next by calculating the prediction error with respect to each entry in the training set. This is calculated using some loss function. For a classification problem this can be achieved using logarithmic loss function, amongst others. These residuals go into the training of the second decision tree. In this way, the prediction error of the previous decision tree is passed on to the next decision tree. This

process is continued either until the maximum number of decision trees is reached, or there are no improvements from adding additional trees. At each step, the previous tree is modified by the new one. How large modifications are is defined by the parameter called the learning rate. If the learning rate is too small the *BDTG* algorithm will need a lot of time to converge. On the other hand, a too large learning rate may result in jumping around the minimum of the loss function and never reaching it. Small learning rate can be compensated by increasing the maximum number of the trees to be built. However, one must be careful because increasing the number of available tree also increases the probability of overtraining.

4.6.3 Algorithm setup for the signal extraction

The EWK signal extraction discussed in this chapter was based on two approaches:

1. a BDT classifier with gradient boosting (*BDTG*) that uses the first seven variables from the Table 4.7. Performance of these variables on separating the VBS contribution was presented in the previous study in this channel using 2016 data [64]. This approach is referred to as the *BDT7*

2. a BDTG classifier that uses all variables from the Table 4.7. This is referred to as the *BDT28* and it was used to check the signal significane gain when using additional variables.

Regardless of the approach used, the setup for the classifier training was the same.
The first step was to prepare the data to be inserted into the *TMVA* tool for the training of the classifier. For this, the baseline selection was applied and the data was stored in root files. Together with the data passing the baseline selection, weights were stored as well for each event. These incorporate L1 prefiring probability as well as the MC and PU weights, trigger efficiency, luminosity, cross-section, scale factors and K-factors for the qqZZ and ggZZ backgrunds. All weights were applied independently of the year with an exception of luminosity and L1 prefiring weights.

BDT classifier training

The EWK signal was trained only against the qqZZ background. Since the kinematics of the ggZZ events is rather similar to that of the qqZZ events, the gain of using it in the training would not be significant. Other backgrounds used in the analysis are minor and were not used in the training neither. While using available background samples in the training would increase separation slightly, at the same time, it would reduce robustness of the model. The result of the training, however, was applied to all samples.
The available signal and background data were equally split in the training set and the test set used to check the performance of the classifier on new data. The training and test samples were weighted, as previously discussed, in order to account for the difference in the distribution shapes between the different contributions. The hyperparameters used in the training are summarized in Table 4.11. It was checked that the training is stable under changes of hyperparameters.
After the training is completed, *TMVA* stores the result in "weight" files that is used to apply the training on the EWK signal and qqZZ, ggZZ, $t\bar{t}Z + VVZ$ and Z+X backgrounds. For each contribution, every event, correctly weighted, is passed through the BDT and it's BDT score is evaluated. This is used to produce a stacked BDT response histogram as shown in Figure 4.10. This is shown, for BDT7 and for the 2016 data, as an illustration here. The plot also shows the observed data together with the Data/MC on the bottom showing the level of agreement between the data and MC.

parameter	value	parameter meaning
NTrees	1000	number of trees
MinNodeSize	2.5	minimum percentage of training events required in a leaf node
Shrinkage	0.1	learning rate
nCuts	20	number of grid points used in finding optimal cut in node splitting
maxDepth	2	maximum allowed depth of the decision tree

Table 4.11: Hyperparameters used in the training of the *BDT7* and *BDT28* classifiers.

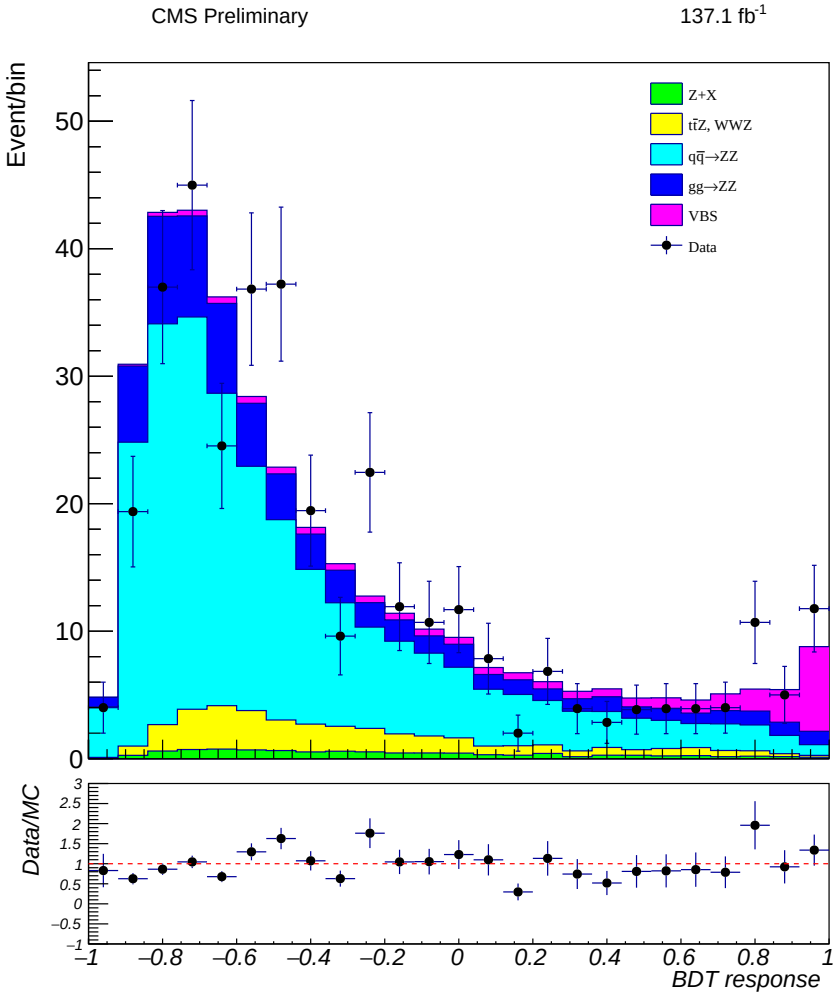


Figure 4.10: Top: BDT output distribution for each contribution, along with data, after *BDT7* training for the 2016 period. Each expected contribution is stacked on top of the previous one starting with the Z+X sample. Bottom: comparison between data and MC expectation.

As a final step, the expected and the observed signal significance is calculated. This is done using the "*combine*" tool which performs a maximum likelihood fit to the data. The expected significance is calculated by assuming an Asimov data set on top of the prediction. Distribution shapes and event yields for each contribution, together with systematic uncertainties, are provided in a file called the *datacard*. The systematic uncertainties are discussed in section 4.8. The "combine" tool also provides the user with an option to exclude systematic uncertainties when performing the fit.

4.6.4 Signal extraction using the BDT7

The distributions of input variables in the training of the BDT7 classifier are shown for the 2016 data-taking periods in Figure 4.11. The same distributions are shown in Figure 4.12 and Figure 4.11 for the 2017 and 2018 periods, respectively.

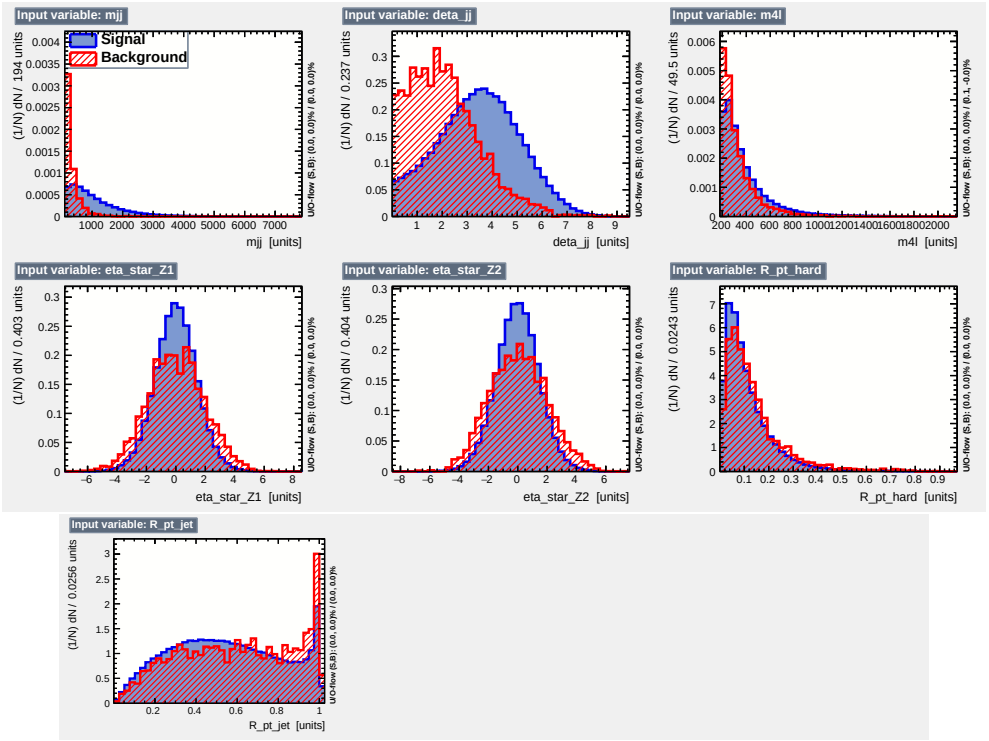


Figure 4.11: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT7* training for the 2016 period. [plot will be updated]

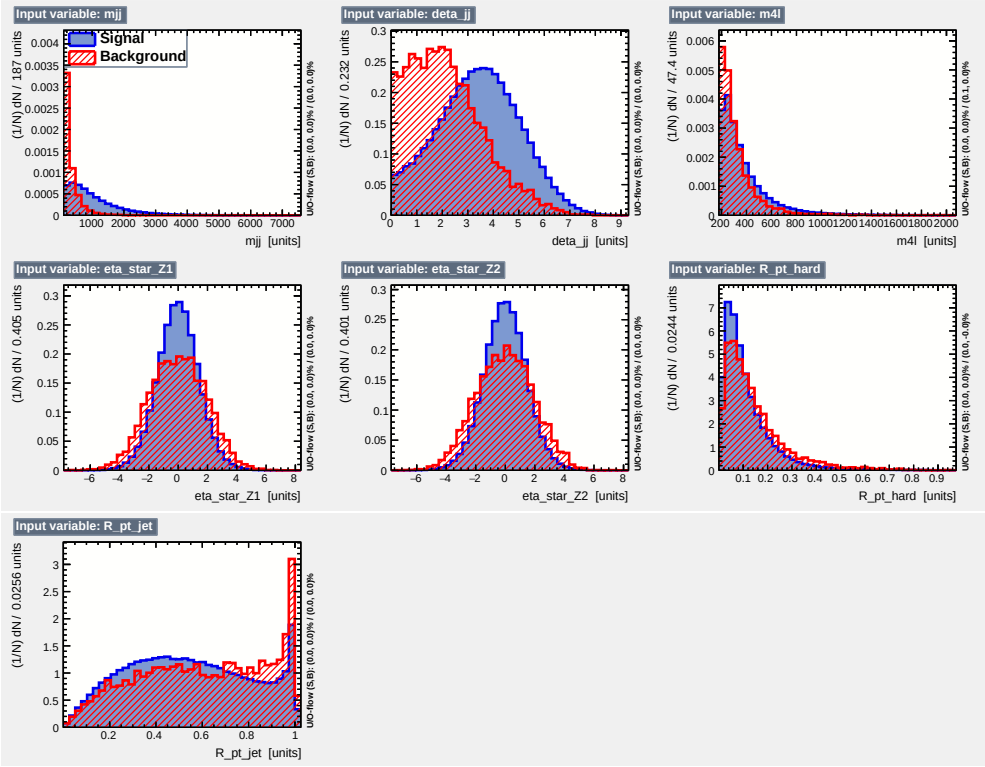


Figure 4.12: BDT input distributions for the EWK signal (in blue) and the QCD qq background (in red) to the *BDT7* training for the 2017 period. [plot will be updated]

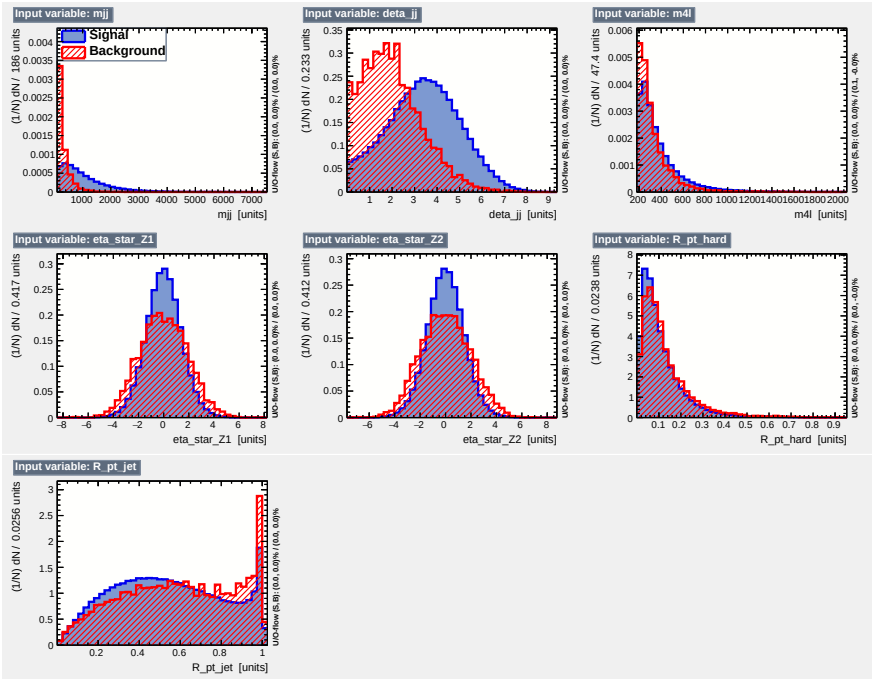


Figure 4.13: BDT input distributions for the EWK signal (in blue) and the QCD qq background (in red) to the *BDT7* training for the 2018 period. [plot will be updated]

The *BDT7* output distributions for the training and test samples together with the signal and background efficiency, purity and significance, for all three periods, are shown in Figures 4.14 - 4.16. Finally, Figure 4.17 shows the BDT

response distribution for all three data-taking periods, and for the three periods combined, where each contribution is stacked on the previous ones. The agreement between data and the MC prediction is within the uncertainties, which are large in the right tail of the distribution due to limited statistics in that region.

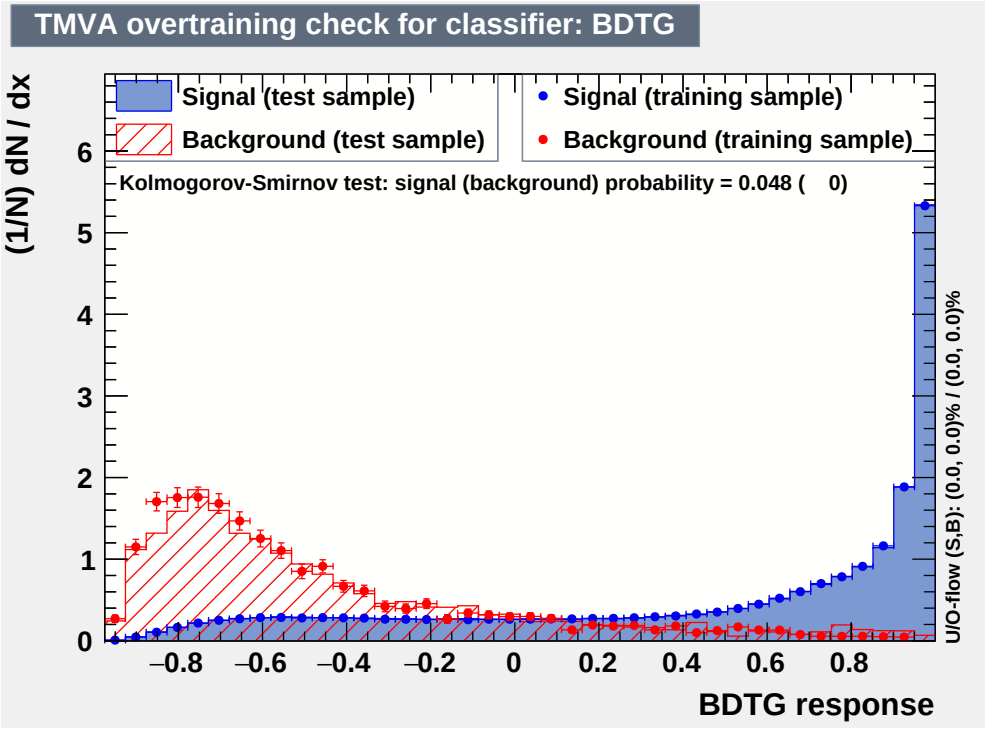


Figure 4.14: The BDT output distribution, together with the overtraining check, for the 2016 *BDT7* training.

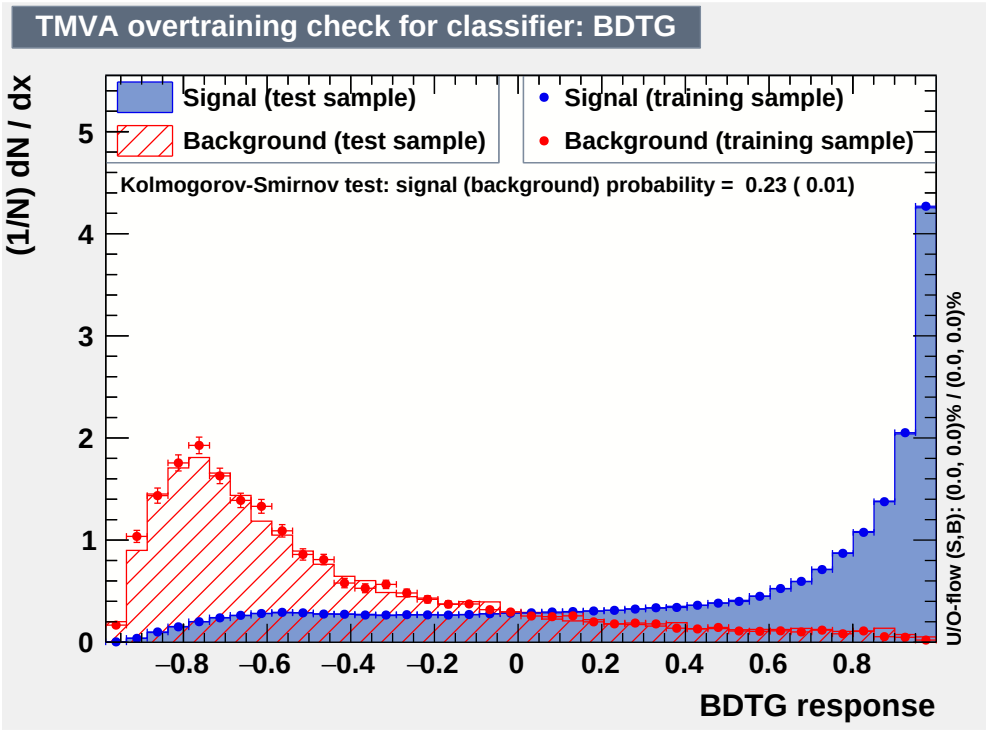


Figure 4.15: The BDT output distribution, together with the overtraining check, for the 2017 *BDT7* training.

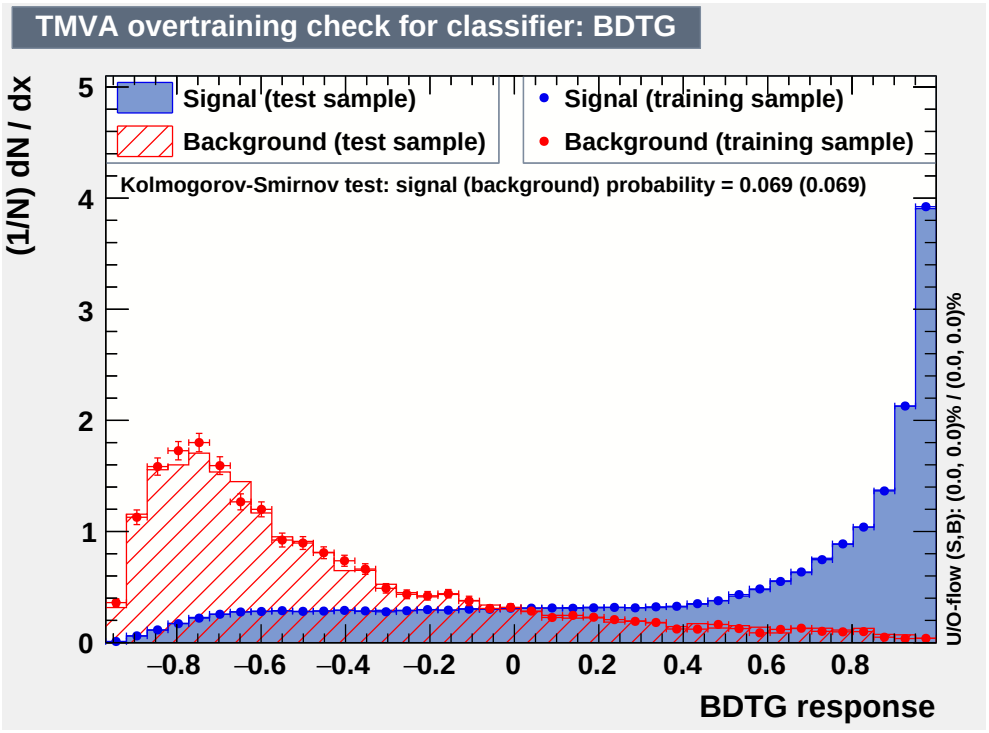


Figure 4.16: The BDT output distribution, together with the overtraining check, for the 2018 *BDT7* training.

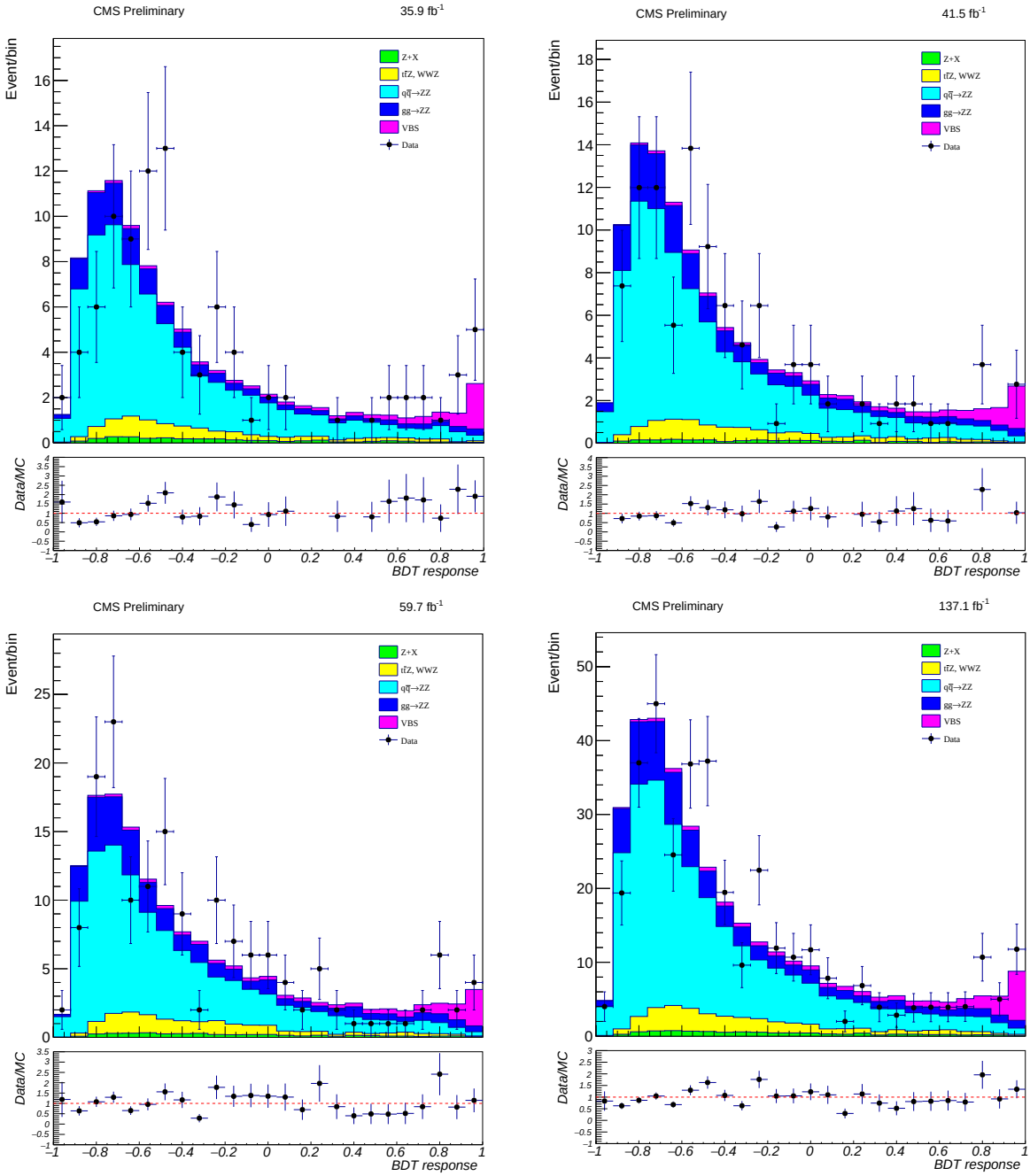
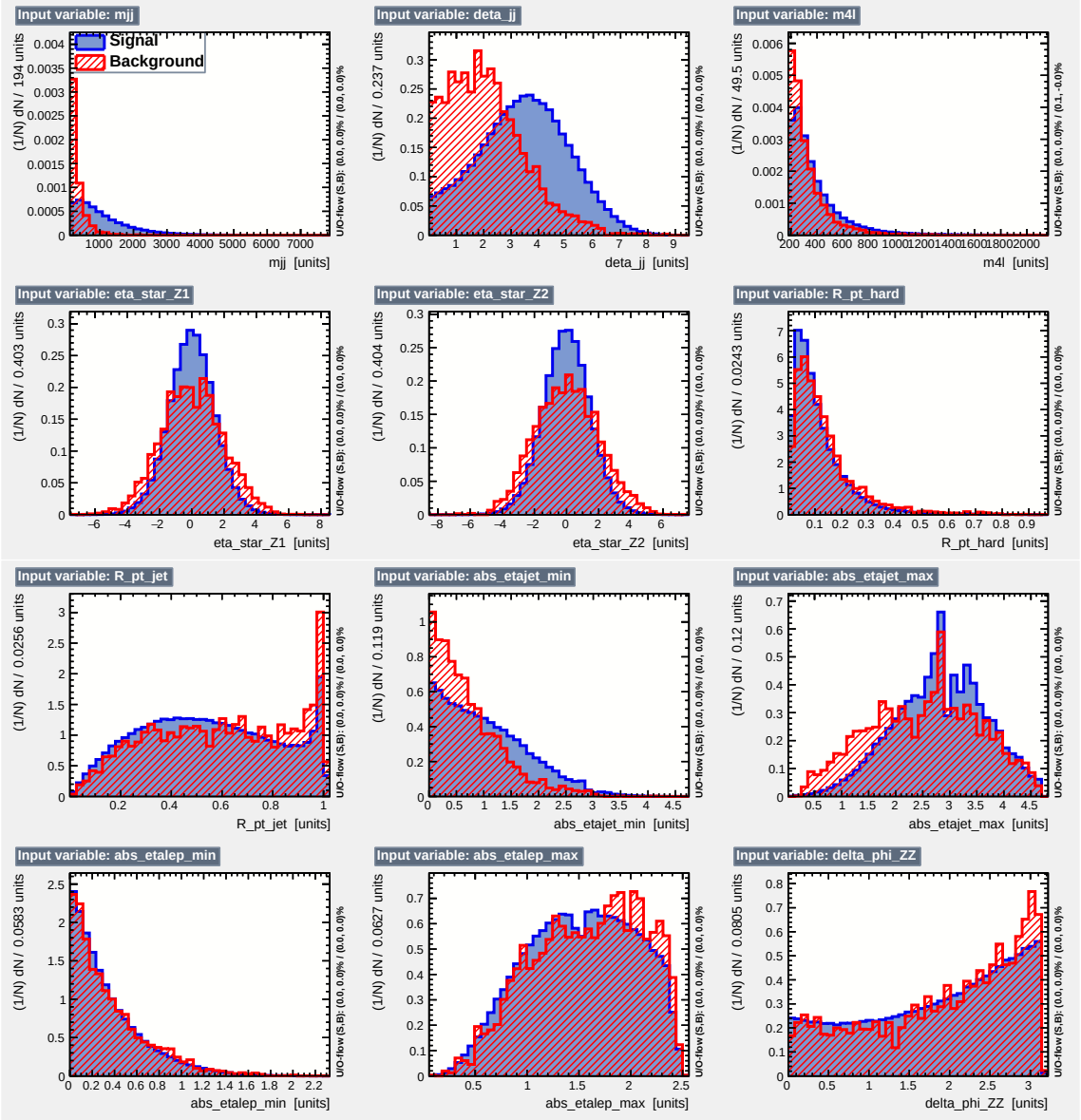


Figure 4.17: BDT output distribution for each contribution after the *BDT7* training for the 2016 (top-lef), 2017 (top-right) and 2018 (bottom-left) period together with the period-combined distribution (bottom-right). Each contribution is stacked on top of the previous one starting with the Z+X sample. Bottom: comparison between data and MC expectation.

4.6.5 Signal extraction using the BDT28

Input variables used in the training of the *BDT28* classifier are shown for 2016 data-taking periods in Figure 4.18. The same distributions are shown in Figure 4.19 and Figure 4.20 for the 2017 and 2018 periods respectively. Looking at new variables introduced to the BDT28 training one can notice a great separation power of variables $y(j_1)$, $y(j_2)$, $\eta(j_1)$ and $\eta(j_2)$. However, these are correlated with the $\Delta\eta_{jj}$ variable already present in the BDT7. The

same is true for the other jet variables as well. Variables $p_T(Z_1)$ and $p_T(Z_2)$ also show good separation power, but are correlated to m_{4l} and $R_{p_T}^{hard}$. For these reasons, one would not expect to gain a lot in terms of the BDT28 performance with respect to the BDT7.



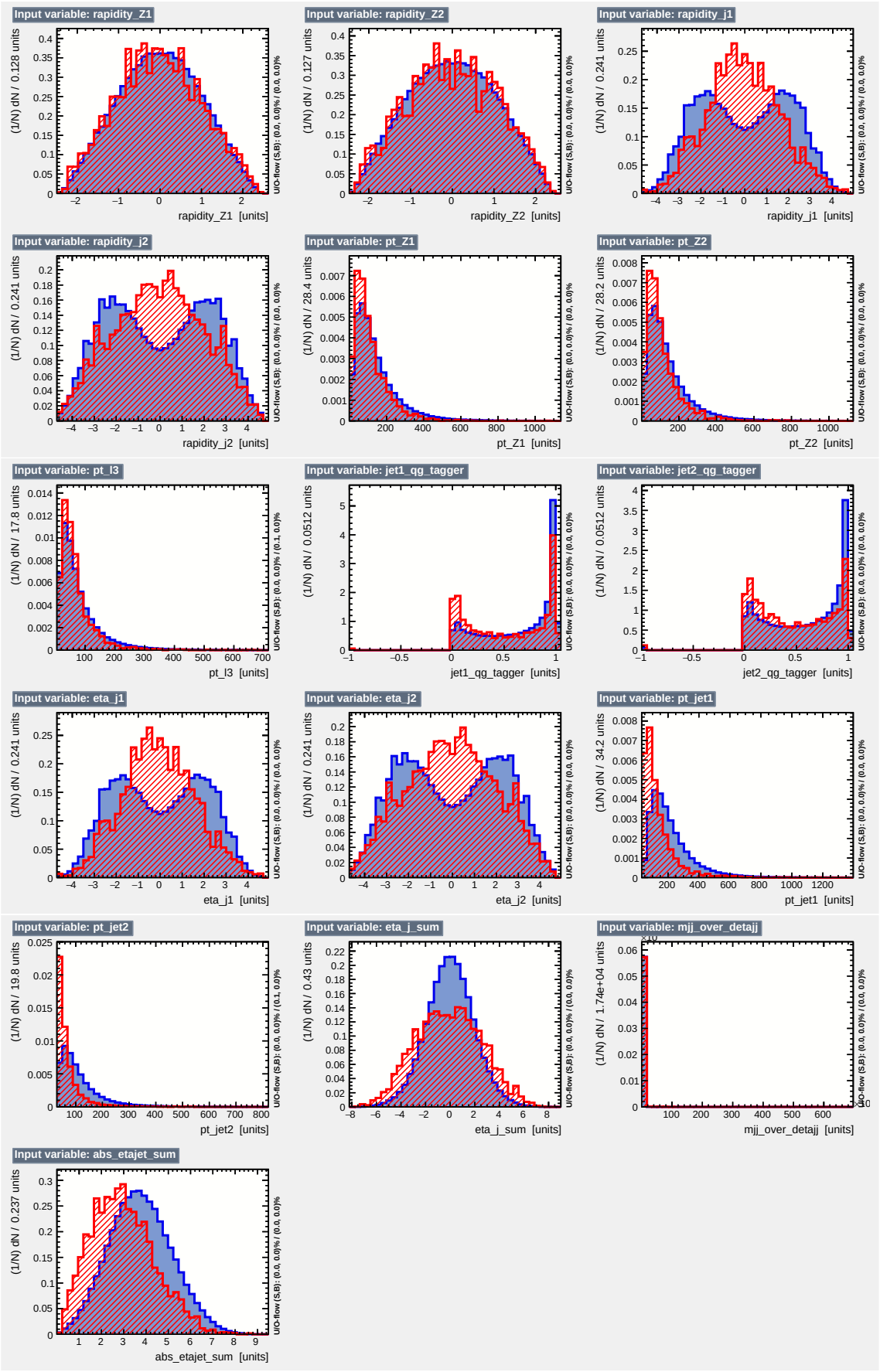
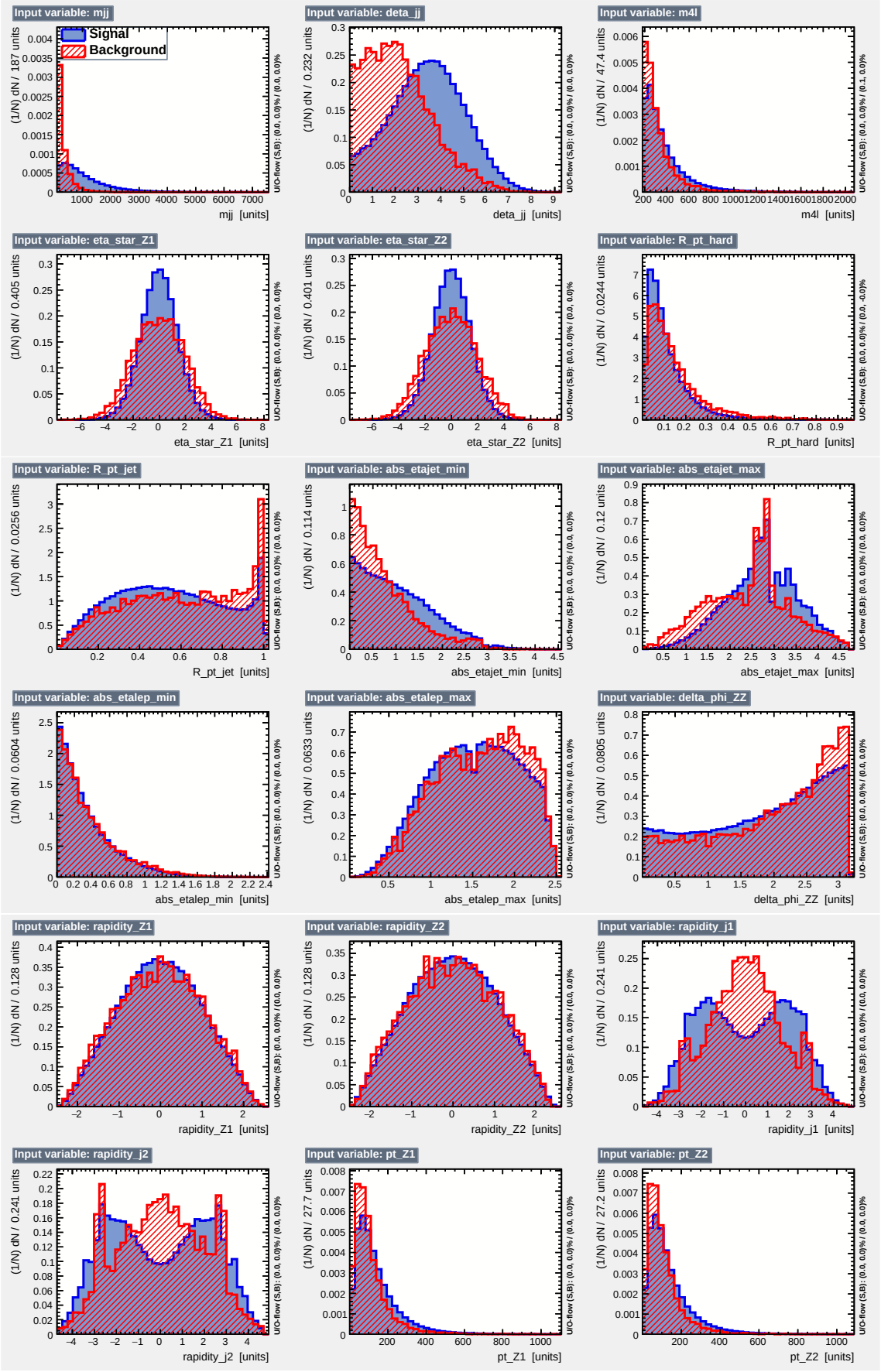


Figure 4.18: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT28* training for 2016 period. [plot will be updated]


Input variable: mjj
Signal
Background
(1/N) dN / 187 units
mjj [units]
Input variable: deta_jj
(1/N) dN / 0.232 units
deta_jj [units]
Input variable: m4l
(1/N) dN / 47.4 units
m4l [units]
Input variable: eta_star_Z1
(1/N) dN / 0.405 units
eta_star_Z1 [units]
Input variable: eta_star_Z2
(1/N) dN / 0.401 units
eta_star_Z2 [units]
Input variable: R_pt_hard
(1/N) dN / 0.0244 units
R_pt_hard [units]
Input variable: R_pt_jet
(1/N) dN / 0.0256 units
R_pt_jet [units]
Input variable: abs_etajet_min
(1/N) dN / 0.114 units
abs_etajet_min [units]
Input variable: abs_etajet_max
(1/N) dN / 0.12 units
abs_etajet_max [units]
Input variable: abs_etalep_min
(1/N) dN / 0.0604 units
abs_etalep_min [units]
Input variable: abs_etalep_max
(1/N) dN / 0.0633 units
abs_etalep_max [units]
Input variable: delta_phi_ZZ
(1/N) dN / 0.0805 units
delta_phi_ZZ [units]
Input variable: rapidity_Z1
(1/N) dN / 0.128 units
rapidity_Z1 [units]
Input variable: rapidity_Z2
(1/N) dN / 0.128 units
rapidity_Z2 [units]
Input variable: rapidity_j1
(1/N) dN / 0.241 units
rapidity_j1 [units]
Input variable: rapidity_j2
(1/N) dN / 0.241 units
rapidity_j2 [units]
Input variable: pt_Z1
(1/N) dN / 27.7 units
pt_Z1 [units]
Input variable: pt_Z2
(1/N) dN / 27.2 units
pt_Z2 [units]
UO-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%

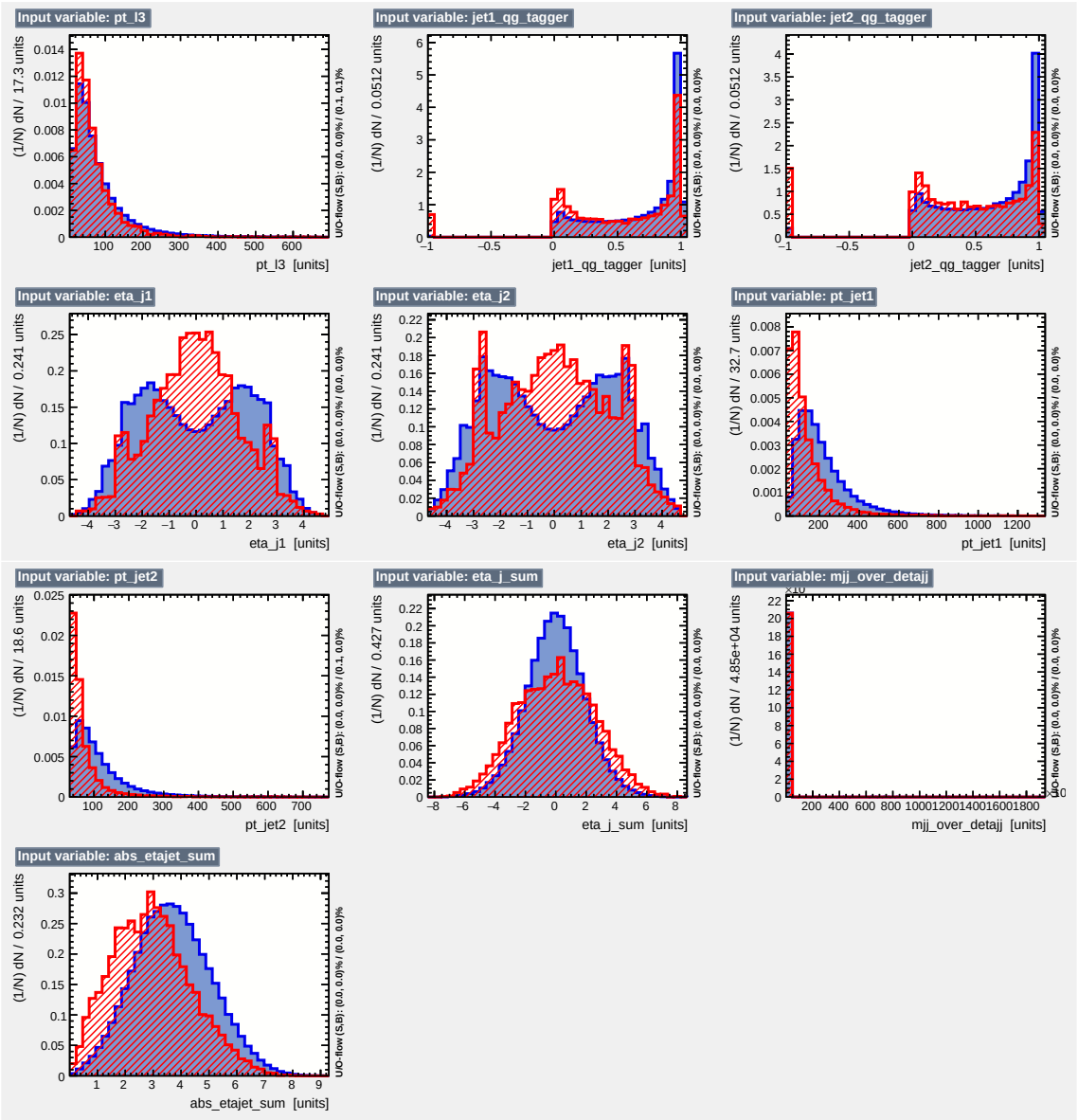
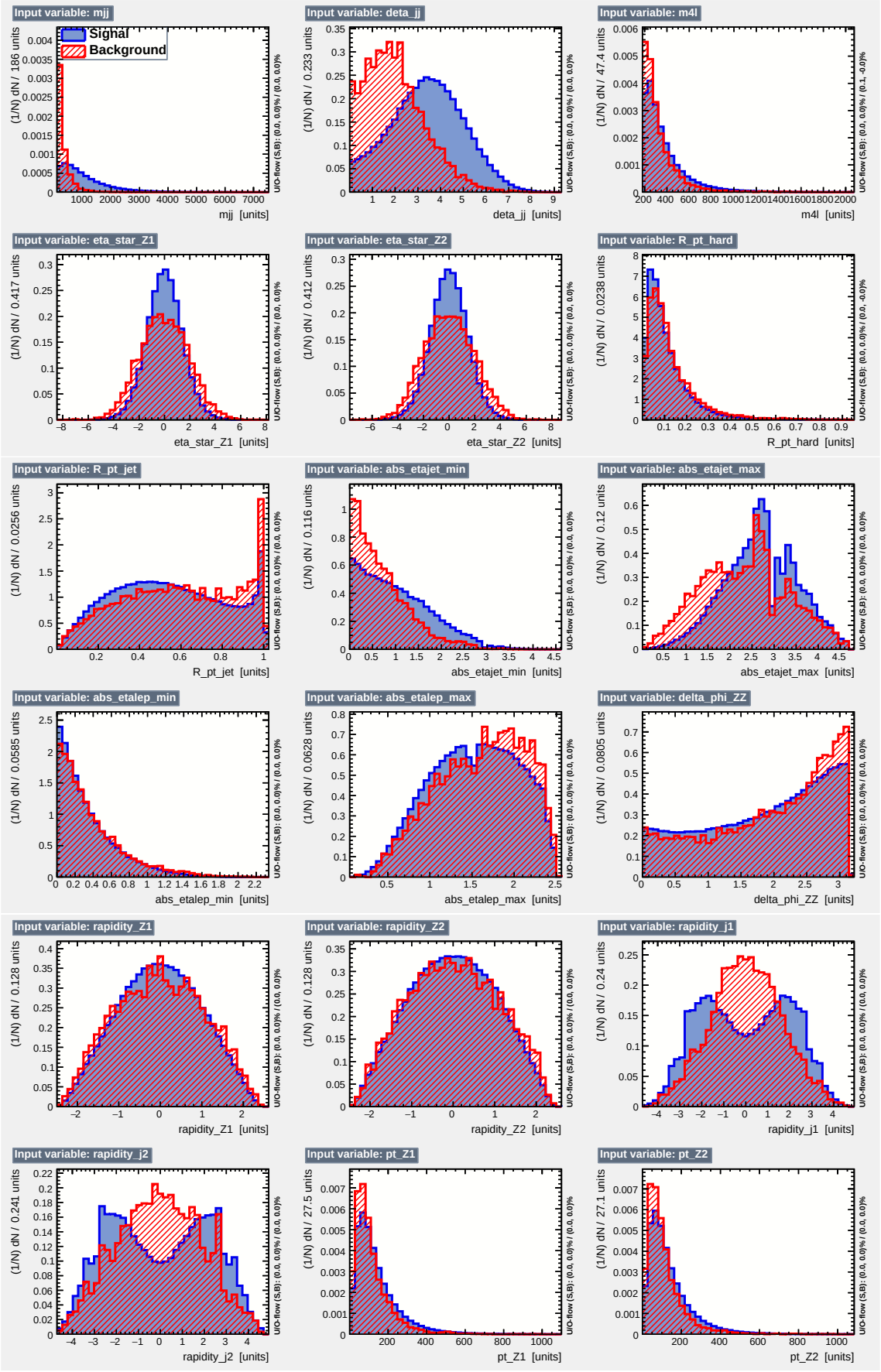


Figure 4.19: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT28* training for 2017 period. [plot will be updated]


Input variable: mjj
Signal
Background
(1/N) dN / 186 units
mjj [units]
Input variable: deta_jj
(1/N) dN / 0.233 units
deta_jj [units]
Input variable: m4l
(1/N) dN / 47.4 units
m4l [units]
Input variable: eta_star_Z1
(1/N) dN / 0.417 units
eta_star_Z1 [units]
Input variable: eta_star_Z2
(1/N) dN / 0.412 units
eta_star_Z2 [units]
Input variable: R_pt_hard
(1/N) dN / 0.0238 units
R_pt_hard [units]
Input variable: R_pt_jet
(1/N) dN / 0.0256 units
R_pt_jet [units]
Input variable: abs_etajet_min
(1/N) dN / 0.116 units
abs_etajet_min [units]
Input variable: abs_etajet_max
(1/N) dN / 0.12 units
abs_etajet_max [units]
Input variable: abs_etalep_min
(1/N) dN / 0.0585 units
abs_etalep_min [units]
Input variable: abs_etalep_max
(1/N) dN / 0.0628 units
abs_etalep_max [units]
Input variable: delta_phi_ZZ
(1/N) dN / 0.0805 units
delta_phi_ZZ [units]
Input variable: rapidity_Z1
(1/N) dN / 0.128 units
rapidity_Z1 [units]
Input variable: rapidity_Z2
(1/N) dN / 0.128 units
rapidity_Z2 [units]
Input variable: rapidity_j1
(1/N) dN / 0.24 units
rapidity_j1 [units]
Input variable: rapidity_j2
(1/N) dN / 0.241 units
rapidity_j2 [units]
Input variable: pt_Z1
(1/N) dN / 27.5 units
pt_Z1 [units]
Input variable: pt_Z2
(1/N) dN / 27.1 units
pt_Z2 [units]
UO-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%

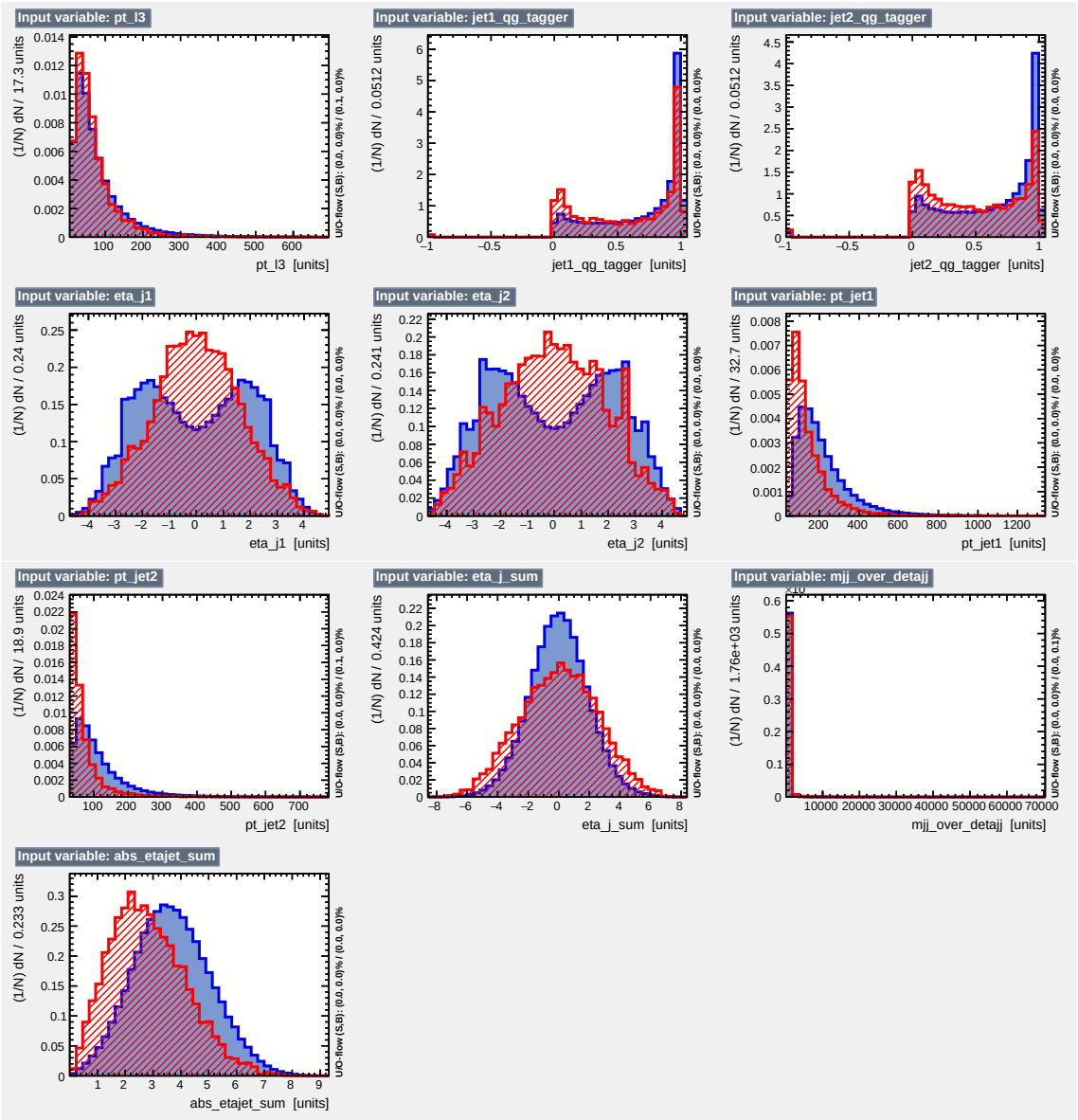


Figure 4.20: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT28* training for 2018 period. [plot will be updated]

The *BDT28* output distributions for the training and test samples together with the signal and background efficiency, purity and significance, for all three periods, are shown in Figures 4.21 - 4.23. Finally, Figure 4.24 shows the BDT response distribution for all three data-taking periods, and for the three periods combined, where each contribution is stacked on the previous ones. The agreement between data and the MC prediction is within the uncertainties.

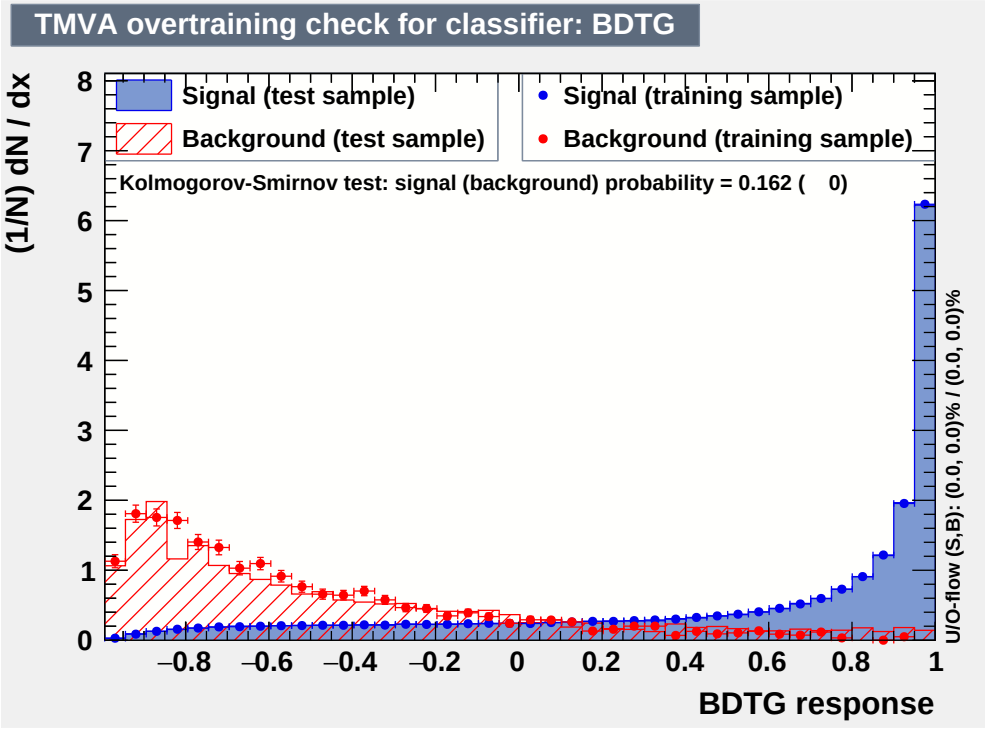


Figure 4.21: The BDT output distribution, together with the overtraining check, for the 2016 *BDT28* training.

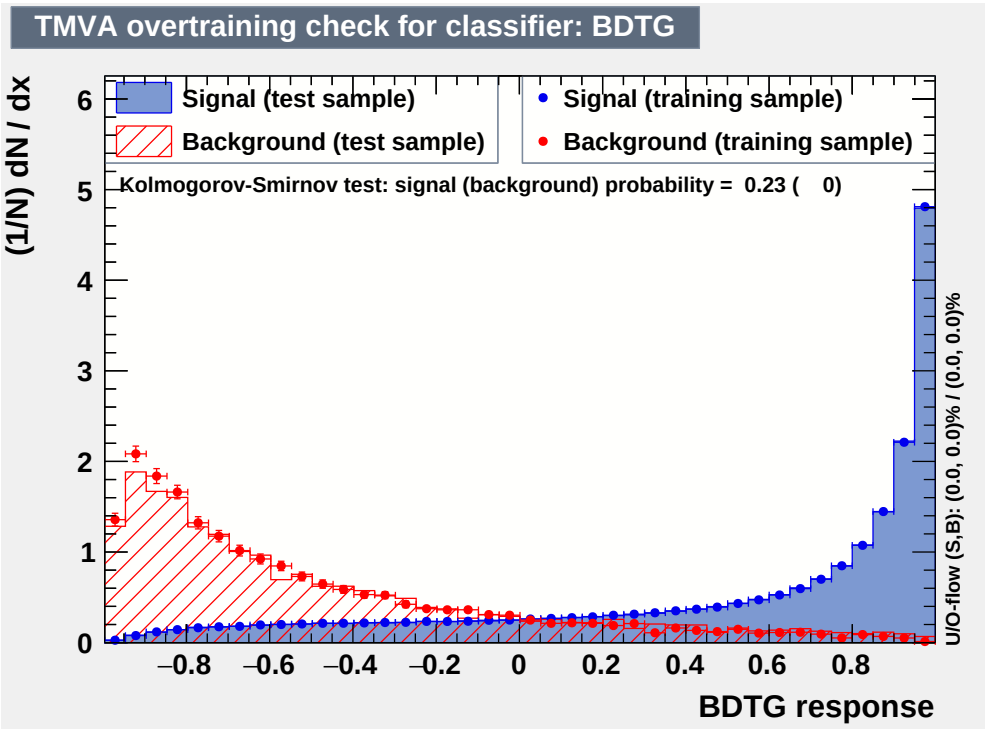


Figure 4.22: The BDT output distribution, together with the overtraining check, for the 2017 *BDT28* training.

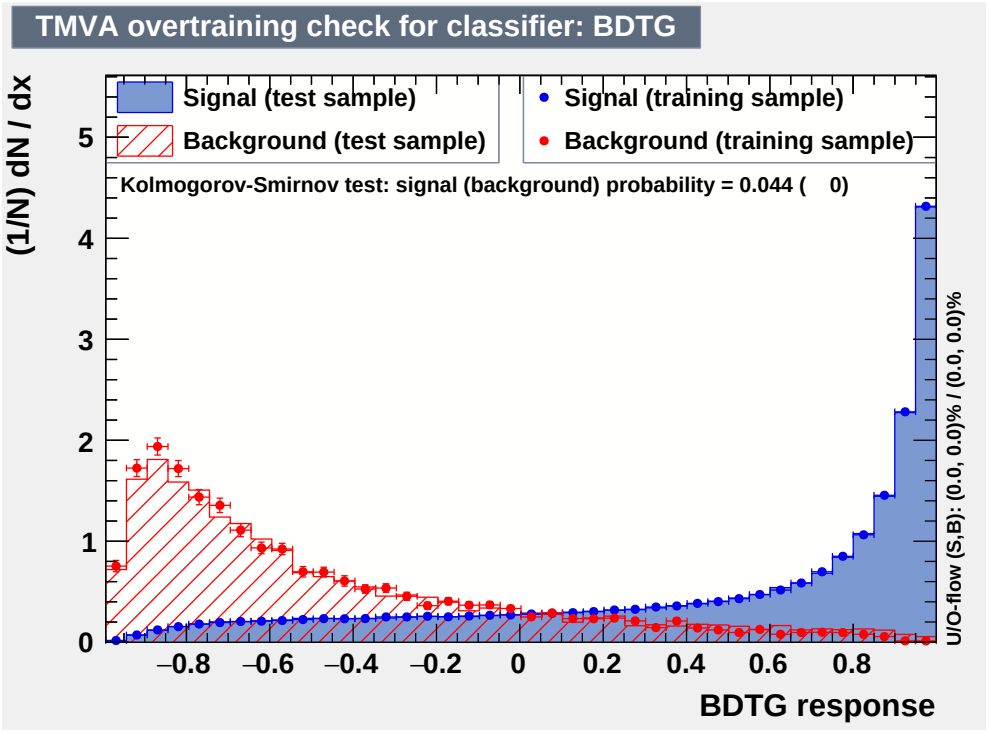


Figure 4.23: The BDT output distribution, together with the overtraining check, for the 2018 *BDT28* training.

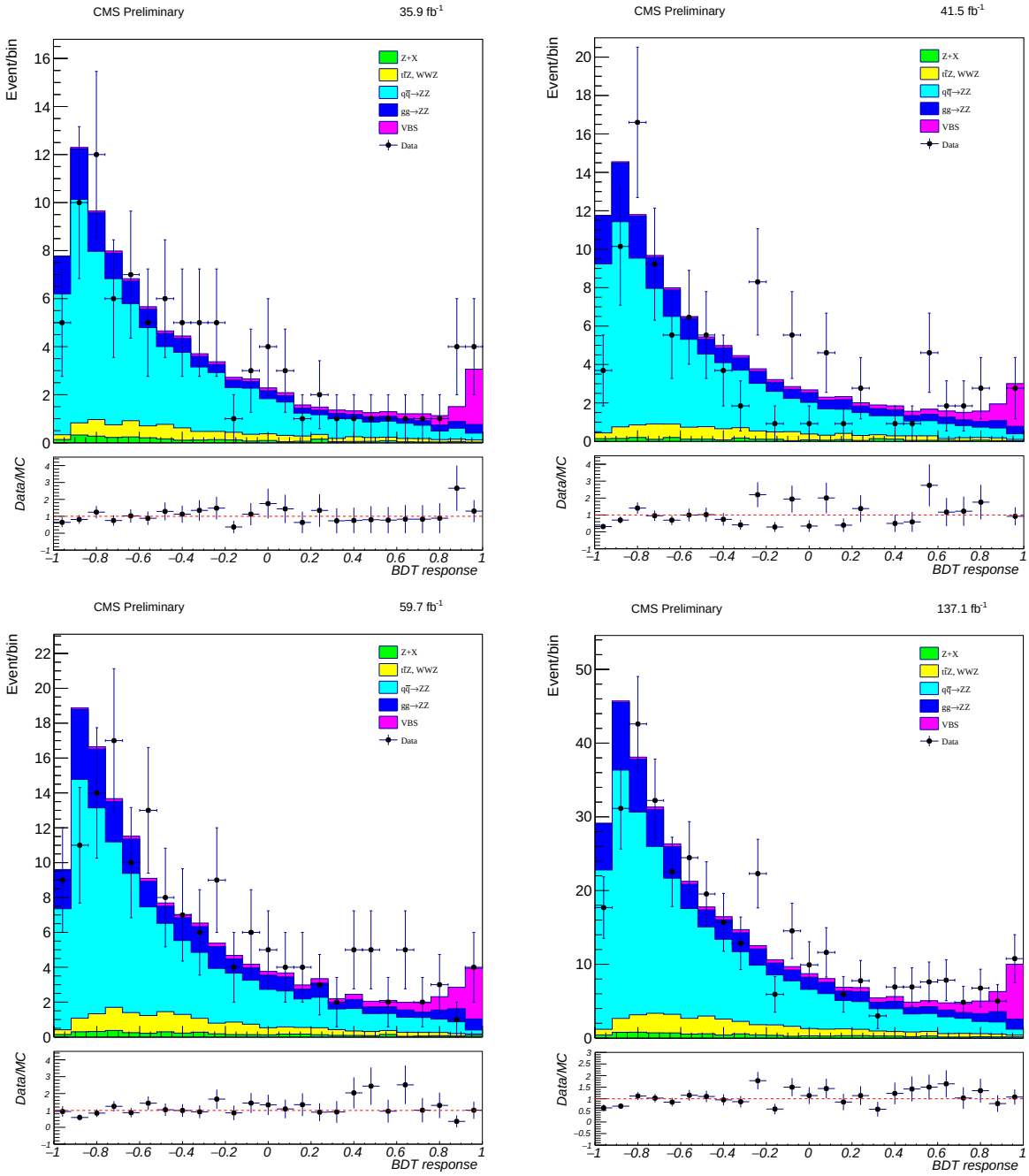


Figure 4.24: BDT output distribution for each contribution after the *BDT28* training for the 2016 (top-lef), 2017 (top-right) and 2018 (bottom-left) period together with the period-combined distribution (bottom-right). Each contribution is stacked on top of the previous one starting with the Z+X sample. Bottom: comparison between data and MC expectation.

4.7 Setting limits on anomalous quartic gauge couplings

Limits on anomalous quartic gauge couplings (aQGCs) are derived in the effective field theory framework where the 8-dimensional operators originate from the covariant derivatives of the Higgs doublet and the charged and neutral field strength tensors.The latter generates eight independent operators which correspond to the couplings of the transverse degrees of freedom, T_i, of the gauge fields [85].
The ZZjj channel exploited in this analysis is particularly sensitive to the neutral-current operators T_8 and T_9 as well as the charged-current operators T_0, T_1 and T_2 [14] which enhance the production cross-section at large values of m_{ZZ}.
Limits on the aQGC parameters f_{T_i}, corresponding to the Wilson coefficients of the operators, are derived based on the m_{4l} distribution following the previous analysis of the anomalous couplings in this channel [64]. The reason for choosing the m_{4l} distribution lies in the fact that the m_{4l} is Lorentz invariant and thus less sensitive to the higher-order corrections. This is crucial since the effect is dominant in the far tail of the distribution.

A dedicated MG sample was produced for the aQGC analysis:

$$generate\ p\ p > z\ z\ j\ j\ QED = 4\ QCD = 0\ NP = 1$$

The reweighting functionality of the *MG5* was used to obtain the expected distributions for different values of the couplings without needing to produce additional samples. The method uses event weights, w_{new}, to reweigh the nominal event sample to the alternative hypotheses of the coupling strength:

$$w_{new} = w_{old} \cdot \frac{|\mathcal{M}_{new}|^2}{|\mathcal{M}_{old}|^2}$$

where $\mathcal{M}_{new}$ and $\mathcal{M}_{old}$ are matrix element with the modified coupling strength and the nominal matrix element respectively. The ratio of the aQGC to SM yields was calculated for several discrete coupling values and then fit with a quadratic function. The result is a semi-analytic description of the expected m_{ZZ} distribution for every bin as a function of the aQGC couplings. This is shown for the operator $T8$ in Figure 4.25 for the last four bins of the m_{4l} distribution. The overflow is included in the last bin. It can be seen that the effect on yields is rising towards the tail of the distribution. The same plots, corresponding to the last bin of the m_{4l} distribution for the $T0$, $T1$, $T2$ and $T9$ operators, are shown in Figure 4.26.

Figure 4.27 shows the expected m_{4l} distribution for the SM, with post-fit normalizations, and the expected distribution for one aQGC scenario, as well as the observed distribution. The fit was performed in the same way as for the EWK signal significance calculation, i.e., using the "*combine*" tool. The test statistic is the log likelihood ratio with all systematic uncertainties profiled as nuisance parameters.
The 95% confidence level (CL) intervals were determined using the Wilk's theorem assuming that the likelihood approaches the χ^2-distribution with one degree of freedom.
The expected limits were obtained using the pre-fit yields for the background and the EWK signal. The observed limits for the combined data set, setting the other coupling to zero, were obtained using the post-fit yields for the background and the signal expectations.
Finally, the unitarity limits were calculated using both the *VBFNLO* package [91] and a theoretical approach as suggested recently [92].

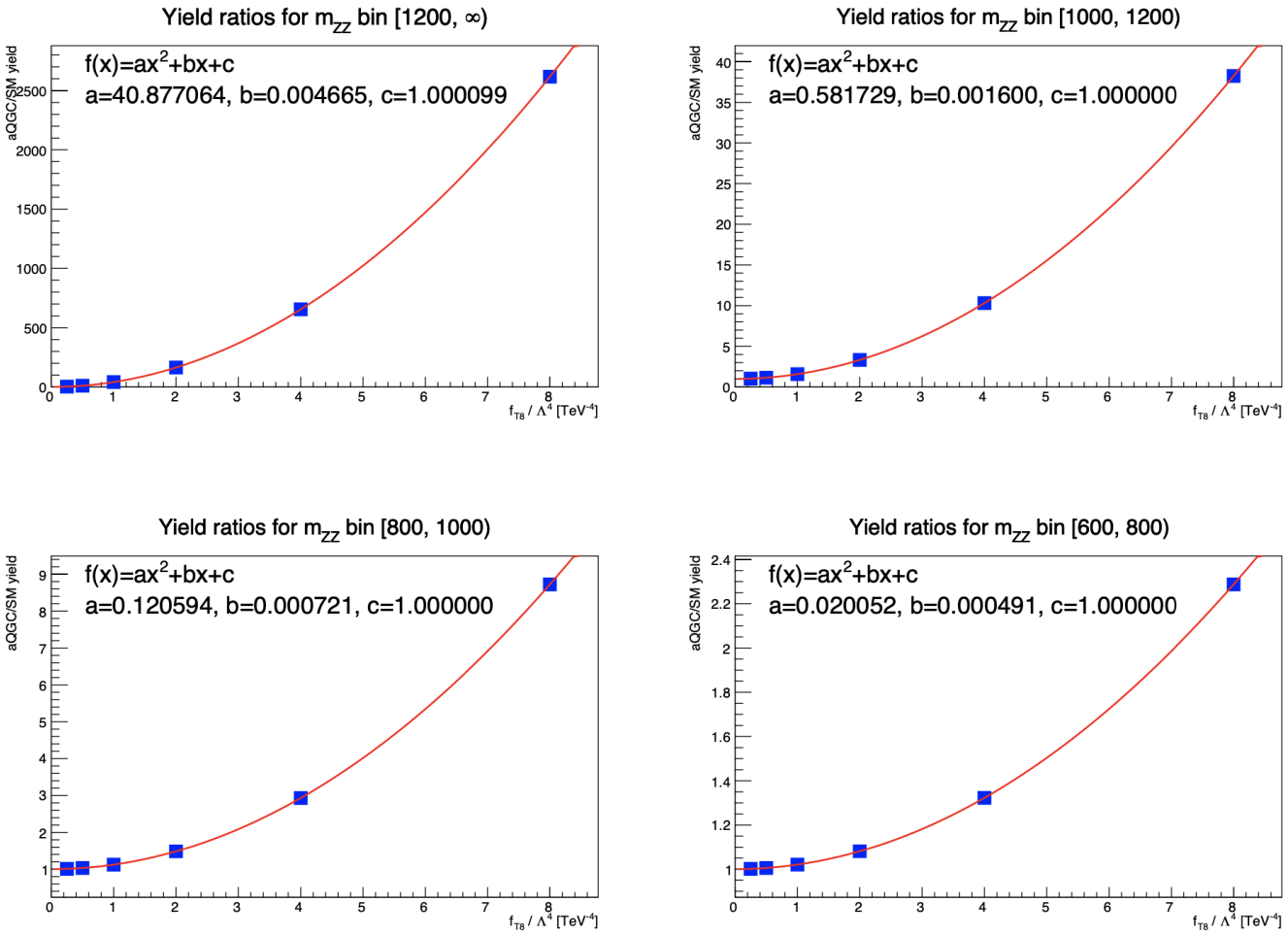


Figure 4.25: Yield ratios for a few values of the operator couplings, f_{T8}/Λ^4, obtained from the reweighing and the fitted quadratic interpolation for the most relevant mass bins used in the statistical analysis.

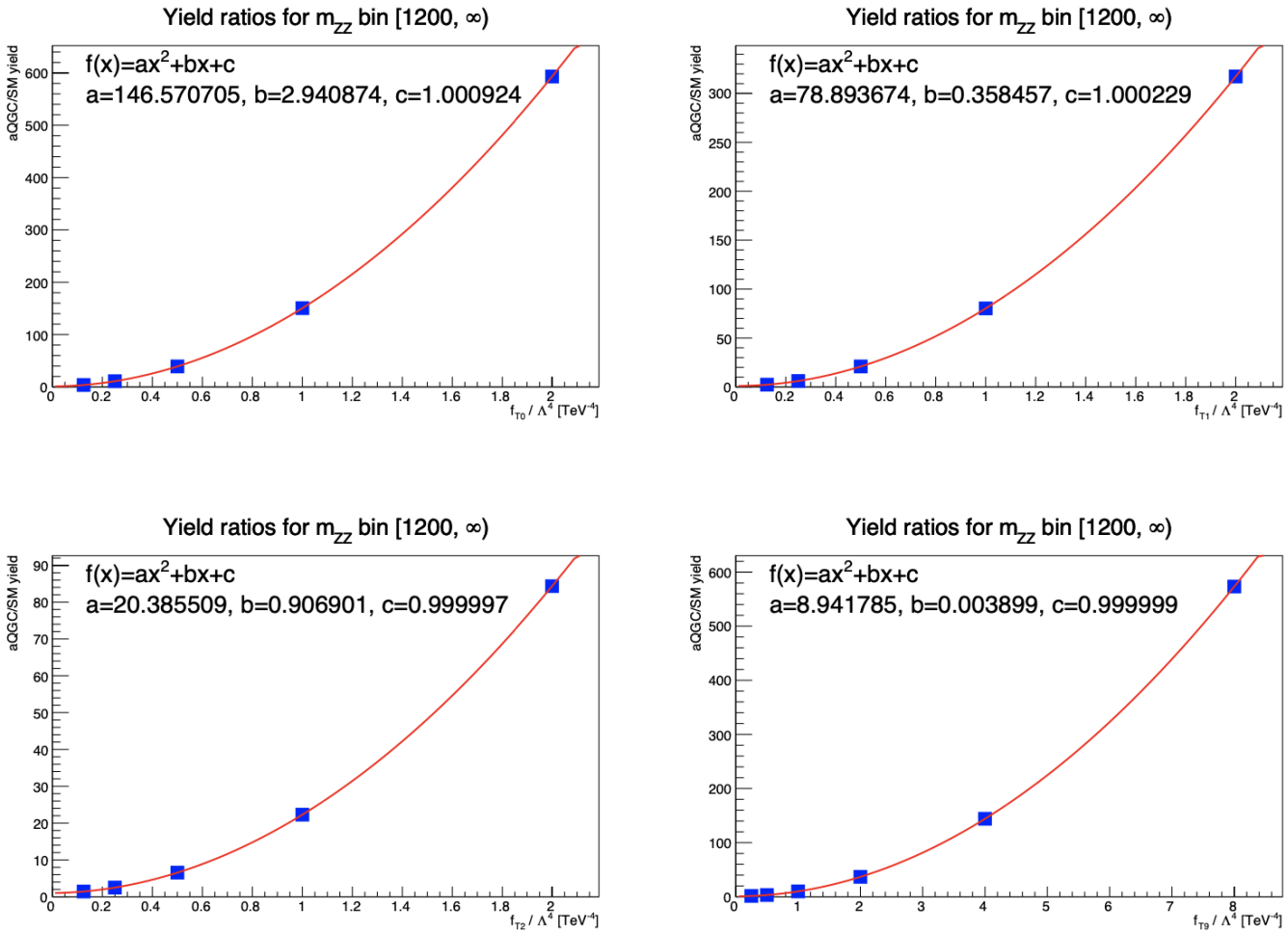


Figure 4.26: Yield ratios for a few values of the operator couplings obtained from the reweighing and the fitted quadratic interpolation for each of the mass bins used in the statistical analysis. The last bin of the m_{4l} distribution is shown for the f_{T0}/Λ^4 (top left), f_{T1}/Λ^4 (top right), f_{T2}/Λ^4 (bottom left) and f_{T9}/Λ^4 (bottom right) operators.

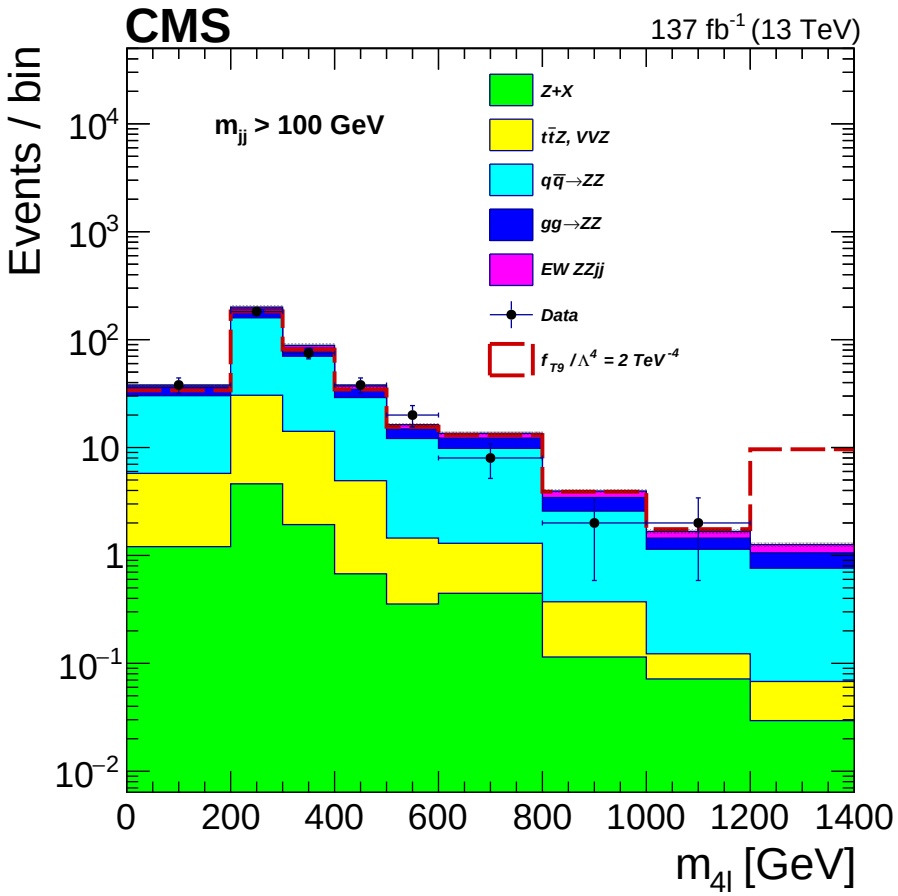


Figure 4.27: Postfit distributions of the four-lepton invariant mass for events sattisfying the ZZjj inclusive selection. Points represent the data, filled histograms the fitted signal and background contributions, and the gray band the uncertainties derived from the fit covariance matrix. As an example, the expected distribution for $f_{T9}/\Lambda^4 = 2\ TeV^{-4}$ is also shown.

4.8 Systematic uncertainties

Regardless of the signal extraction method used, the MELA discriminant, or the BDTs discussed in section 4.6, the same set of systematic uncertainties is applied. QCD scale and PDF uncertainties are originating from the incomplete theoretical description of the underlying physics. The rest described bellow come from imperfect description of the detector effects or simulation.

When calculating the cross-section of the desired process, one can find that, sometimes, it can't be explicitly done due to the divergences that appear. The nature of such infinities can be twofold:

1. ultraviolet (UV) divergences which arise due to the large momentum transfers in loop of Feynman diagrams representing the process amplitude
2. infrared (IR) divergences that can arise either because massless particle radiates another massless particle or because a virtual or real particle reaches zero momentum

In order to solve the UV divergences, the renormalization scale, μ_R, is introduced. Consequently, the running coupling constant, α_s, becomes a function of parameter μ_R.
If the IR divergences appear because of a massless particle being radiated by another massless particle, they can be cured by introducing a factorization scale, μ_F. Consequently, the parton distribution functions (PDFs) and the fragmentation functions, which defines the evolution of the collision fragments, become a function of μ_F [93].
QCD scale uncertainties are estimated using the common procedure of varying the normalization and factorization scales up and down by a factor two (excluding the extreme cases) with respect to the nominal value. Unlike for the EWK signal where the uncertainty is shape-dependent, a constant uncertainty, between 9% and 14%, is used for qqZZ and ggZZ backgrounds.

Uncertainties related to the choice of the PDFs and the strong coupling constant α_s are evaluated from the variations of the respective eigenvalues set [74]. Although different PDFs were used for different data-taking periods, the associated uncertainties are very similar. A constant uncertainty, between 3.3% and 6.6%, was used for different samples [68, 85].

The uncertainty in the LHC integrated luminosity is taken from [94] and is 2.3-2.5%. Since the correlated component amongst years is small, and because the overall effect of systematic uncertainties in the measurements is also small, the uncertainty in the luminosity between the years is assumed to be uncorrelated.

The uncertainty in the data-driven reducible background estimate is dominated by the statistical uncertainties because of the limited number of events in the control regions and ranges from 33% to 45% depending on the final state.

Processes estimated from the simulation are limited by the statistics of the MC sample. This is taken as a source of the shape-dependent, year-uncorrelated systematic uncertainty. For the cut-and-count analyses (i.e. calculation of the EWK and EWK+QCD cross-sections, and a derivation of the limits on the aQGCs) integrated uncertainties of the MC sample were used, while for the template analysis (i.e. signal extraction using MELA) the $autoMCstats$ feature of the "combine" tool was used to obtain the shape-dependent uncertainty profile. For the calculation of limits on aQGCs, the uncertainties were enlarged because the sensitivity comes from the high-m_{ZZ} bins only.

Uncertainties coming from the trigger and lepton reconstruction and selection range from 2.5% to 9% depending on the final state and those coming from the PU reweighting range between 0.2% and 2.7% depending on the sample and year [95].

The jet energy scale (JES) uncertainty ranges from 4.9 - 11.4% for the QCD qqZZ background and 0.7% - 1.2% for the EWK signal. The jet energy resolution (JER) ranges from 2.2% - 6.3% and 0.2% - 0.4% for QCD qqZZ background and EWK signal respectively [20].

L1 prefiring weight variations range from 0.6% to 3.0% depending on the sample.

Systematic uncertainties are summarized in Table 4.12.

Systematic source	qqZZ	ggZZ	VBS	Z+X	Shape	Years correlated
QCD scales [%]	10 - 12	9 - 14	6	-	+	+
PDF + α_s [%]	3.2	5	6.6	-		+
Lepton trigger, reco, sel. [%]	2.5 - 9	2.5 - 9	2.5 - 9	-		+
L1 prefiring [%]	0.6 - 1.0	0.6	1.8 - 3.0	-	+	
Luminosity [%]	2.3 - 2.5	2.3 - 2.5	2.3 - 2.5	-		
JES [%]	4.9 - 11.4	3.6 - 10.2	0.7 - 1.2	-	+	
JER [%]	2.2 - 6.3	1.0 - 2.2	0.2 - 0.4	-		
MC samples [%]	2.5-4.2 (11-28)	3.2 (17-22)	$\ll 1$	-	+	
Pileup [%]	0.2 - 2.6	0.4 - 2.7	0.3 - 1.7	-		
Reducible background [%]	-	-	-	33 - 45		

Table 4.12: Systematic uncertainties on the signal and background yields. Minor backgrounds, for which the systematics is dominated by the MC sample size (19% - 24%), are not shown. The numbers in parentheses refer to the uncertainties used in the derivation of limits on aQGCs.

4.9 Results

Table 4.13 shows the expected and observed event yields for the ZZjj inclusive selection as well as the two VBS-enriched regions. A very good agreement between the predicted and measured event yields is reported for all data-taking periods.

Measured cross-sections and the corresponding SM predictions in the three fiducial regions obtained using the MELA discriminant for both EWK and EWK+QCD are summarized in Table 4.14. The same table shows the measured and expected EWK signal strength. Total uncertainty is quoted for all the measurements with statistical only separated in parentheses. SM predictions were extracted from the generated events in the MC samples used in the analysis including the K-factors where applicable. For the EWK ZZjj inclusive region, in addition to the higher-order calculations at NLO in QCD [96, 97] and theoretical predictions at LO in QCD, NLO EWK corrections [98] were included. Uncertainties in all all SM predictions come from variations of the factorization and renormalization scales. $PDF + \alpha_s$ variation uncertainties are summed in quadrature, except from the prediction from [98].

Year	EWK signal	Z+X	$q\bar{q} \to ZZjj$	$gg \to ZZjj$	$t\bar{t}Z + VVZ$	Tot. predict.	Data
ZZjj inclusive							
2016 ($35.9\ fb^{-1}$)	6.3 ± 0.07	2.8 ± 1.1	65.6 ± 9.5	13.5 ± 2.0	8.4 ± 2.2	96 ± 13	95
2017 ($41.5\ fb^{-1}$)	7.4 ± 0.8	2.4 ± 0.9	77.7 ± 11.2	20.3 ± 3.0	9.6 ± 2.5	117 ± 15	111
2018 ($59.7\ fb^{-1}$)	10.4 ± 1.1	4.1 ± 1.6	98.1 ± 14.2	29.1 ± 4.3	14.2 ± 3.8	156 ± 20	159
All ($137.1\ fb^{-1}$)	24.1 ± 2.5	9.4 ± 3.6	241.5 ± 34.9	62.9 ± 9.3	32.2 ± 8.5	370 ± 48	365
VBS signal-enriched (loose)							
2016 ($35.9\ fb^{-1}$)	4.2 ± 0.4	0.4 ± 0.2	9.7 ± 1.4	3.2 ± 0.5	1.1 ± 0.3	18.7 ± 2.3	21
2017 ($41.5\ fb^{-1}$)	4.9 ± 0.5	0.5 ± 0.2	13.5 ± 1.9	5.5 ± 0.8	1.2 ± 0.3	25.5 ± 3.1	17
2018 ($59.7\ fb^{-1}$)	6.9 ± 0.7	0.8 ± 0.3	14.9 ± 2.2	8.3 ± 1.2	1.7 ± 0.5	32.6 ± 3.9	30
All ($137.1\ fb^{-1}$)	16.0 ± 1.7	1.6 ± 0.6	38.1 ± 5.5	17.0 ± 2.5	4.1 ± 1.1	76.8 ± 9.3	68
VBS signal-enriched (tight)							
2016 ($35.9\ fb^{-1}$)	2.4 ± 0.3	0.10 ± 0.04	1.3 ± 0.2	0.7 ± 0.1	0.24 ± 0.06	4.8 ± 0.5	4
2017 ($41.5\ fb^{-1}$)	2.7 ± 0.3	0.05 ± 0.02	1.9 ± 0.3	1.2 ± 0.2	0.14 ± 0.04	6.0 ± 0.7	3
2018 ($59.7\ fb^{-1}$)	3.9 ± 0.4	0.17 ± 0.06	2.0 ± 0.3	1.5 ± 0.2	0.30 ± 0.08	7.8 ± 0.9	10
All ($137.1\ fb^{-1}$)	9.0 ± 1.0	0.32 ± 0.12	5.3 ± 0.8	3.3 ± 0.5	0.68 ± 0.18	18.6 ± 2.1	17

Table 4.13: Predicted signal and background yields with total uncertainties, and the observed number of events for the ZZjj inclusive selection as well as the VBS loose and tight signal-enriched selections. Integrated luminosities per data set are reported in parentheses.

	SM σ $[fb]$	**Measured σ $[fb]$**	μ_{exp}	μ_{obs}
ZZjj inclusive				
EWK	LO: $0.275 \pm 0.021_{th.}$ NLO QCD: $0.278 \pm 0.017_{th.}$ NLO EWK: $0.242^{+0.015_{th.}}_{-0.013_{th.}}$	$0.33^{+0.11\ (+0.04)}_{-0.10\ (-0.03)}$	$1.00^{+0.43\ (+0.39)}_{-0.36\ (-0.34)}$	$1.21^{+0.47}_{-0.40}$
EWK+QCD	$5.35 \pm 0.51_{th.}$	$5.29^{+0.31\ (+0.46)}_{-0.30\ (-0.46)}$	$1.00^{+0.13\ (+0.06)}_{-0.12\ (-0.06)}$	$0.99^{+0.13}_{-0.12}$
VBS signal-enriched (loose)				
EWK	LO: $0.186 \pm 0.015_{th.}$ NLO QCD: $0.197 \pm 0.013_{th.}$	$0.200^{+0.078\ (+0.023)}_{-0.067\ (-0.013)}$	$1.00^{+0.45\ (+0.40)}_{-0.38\ (-0.35)}$	$1.08^{+0.47}_{-0.38}$
EWK+QCD	$1.21 \pm 0.09_{th.}$	$1.00^{+0.12\ (+0.06)}_{-0.11\ (-0.05)}$	$1.00^{+0.16\ (+0.13)}_{-0.15\ (-0.12)}$	$0.83^{+0.15}_{-0.13}$
VBS signal-enriched (tight)				
EWK	LO: $0.104 \pm 0.008_{th.}$ NLO QCD: $0.108 \pm 0.007_{th.}$	$0.09^{+0.04\ (+0.02)}_{-0.03\ (-0.02)}$	$1.00^{+0.52\ (+0.50)}_{-0.44\ (-0.41)}$	$0.87^{+0.48}_{-0.39}$
EWK+QCD	$0.221 \pm 0.014_{th.}$	$0.20^{+0.05\ (+0.02)}_{-0.04\ (-0.02)}$	$1.00^{+0.42\ (+0.40)}_{-0.34\ (-0.32)}$	$0.92^{+0.39}_{-0.32}$

Table 4.14: SM cross-sections in the three fiducial regions together with the fitted value of the signal strength. Total uncertainty is quoted for all measurements with the statistical only contribution in parentheses. The theory uncertainty for the expected SM cross-section is also quoted. For the EWK ZZjj inclusive region, NLO EWK corrections are quoted in addition to the higher-order calculations at NLO in QCD and theoretical predictions at LO in QCD.

The significance of the EWK signal using the MELA classifier was obtained by calculating the probability of the background-only hypothesis (*p*-value) as the tail integral of the test statistic evaluated at $\mu_{EWK} = 0$ under the asymptotic approximation [99]. The background-only hypothesis was excluded with $4.0\ \sigma$ ($3.5\ \sigma$ expected).

The expected significance using the two BDTs was calculated for the separated data-taking periods as well as for the combined period. The results are summarized in Table 4.15. An expected significance of $3.9\ \sigma$ (stat. only) and $3.8\ \sigma$ (stat. + sys.) is reported for the combined period using the *BDT7*. The value of $3.8\ \sigma$ obtained using the *BDT7* classifier is comparable to the MELA result. Similar performance of the *BDT7* and MELA is also confirmed by comparing the ROC curves in Figure 4.28.

In order to asses the potentioal gain of using 28 variables in the BDT training, the EWK signal significance was calculated for the *BDT28* as well. For the *BDT28*, an expected significance of $4.0\ \sigma$ (stat. only) and $3.9\ \sigma$ (stat. + sys.) is reported for the combined period. A small increase in sensitivity is obtained at the expense of a loss of model robustness. This shows that the *BDT7* is capable of capturing and exploiting the kinematical difference between signal and background without a need of additional variables.
The observed signal significance for the three periods, as well as for the combined period, for the *BDT7* and *BDT28* is also reported in Table 4.15. An upward fluctuation in the data can be seen from the bottom right plots in Figure 4.17 and Figure 4.24. This is reflected in the increase of observed signal significance for the combined period in both *BDT7* and *BDT28*.

Possible gain in sensitivity was also looked for by using for the training events that passed the VBS loose selection instead of the ZZjj baseline. This was done for the 2016 data-taking period with the *BDT7* training and a negligible increase (< 0.4%) in the signal sensitivity was observed while, at the same time, loosing some signal

events.

Year	Exp. significance [σ]	Obs. significance [σ]
BDT7		
2016 ($35.9\ fb^{-1}$)	2.07 (2.12)	4.08 (4.05)
2017 ($41.5\ fb^{-1}$)	2.8 (2.14)	1.79 (1.69)
2017 ($59.7\ fb^{-1}$)	2.44 (2.53)	2.90 (3.12)
All ($137.1\ fb^{-1}$)	3.77 (3.93)	5.09 (5.19)
BDT28		
2016 ($35.9\ fb^{-1}$)	2.13 (2.18)	3.24 (3.29)
2017 ($41.5\ fb^{-1}$)	2.14 (2.20)	2.02 (1.91)
2017 ($59.7\ fb^{-1}$)	2.46 (2.55)	2.85 (3.08)
All ($137.1\ fb^{-1}$)	3.85 (4.01)	4.69 (4.81)

Table 4.15: Expected and observed EWK signal significance for the three data-taking periods as well as the combined period. Both results for the BDT7 and BDT28 are reported. Result with only statistical uncertainties are shown in the parentheses.

The expected and observed lower and upper 95% CL limits on the couplings of the charged-current operators T_0, T_1 and T_2 as well as of the neutral-current operators T_8 and T_9 are shown in Table 4.16. Results with only statistical uncertainties included are shown in parentheses. The unitarity limits obtained using both the *VBFNLO* package and the approach suggested in the recent publication [92] are also shown. These were the most stringent limits, at the time, on the neutral-current operators T8 and T9. A recent study by CMS collaboration in the $Z\gamma$ channel provided even further improvements on these measurements [100].

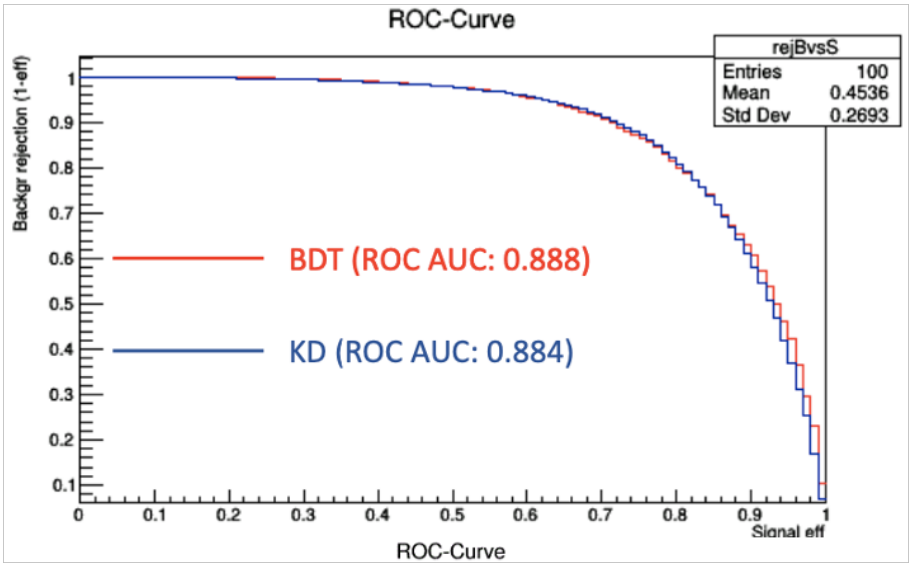


Figure 4.28: Performance of the *BDT7* compared to the MELA using the ROC curve and area under curve (AUC).

Coupling	Exp. lower	Exp. upper	Obs. lower	Obs. upper	Unit. limit (VBFNLO)	Unit. limit (Eboli)
f_{T_0}/Λ^4	-0.37	0.35	-0.24	0.22	2.9	2.4
f_{T_1}/Λ^4	-0.49	0.49	-0.31	0.31	2.7	2.6
f_{T_2}/Λ^4	-0.98	0.95	-0.63	0.59	2.8	2.5
f_{T_8}/Λ^4	-0.68	0.68	-0.43	0.43	1.8	1.8
f_{T_9}/Λ^4	-1.46	1.46	-0.92	0.92	1.8	1.8

Table 4.16: Observed and expected lower and upper 95 % CL limits on the coupling of the quartic tensor operators T_0, T_1 and T_2 as well as the neutral-current operators T_8 and T_9. The unitarity limits are also reported. All couplings are expressed in TeV^{-4} while the unitarity limits are expressed in TeV. Results are obtained using the postfit distributions.

4.10 Summary

A search for VBS in the $qq \rightarrow ZZ \rightarrow 4ljj$ channel using CMS data from the full Run 2 was presented in this chapter. Because of the fully reconstructable final state, this channel is amongst the most sensitive for the extraction of the longitudinal component of the ZZ scattering, thus providing a better insight into the scalar sector of the Standard Model. Since the channel is sensitive to the neutral-current operators, it enables to probe into the anomalous quartic gauge coupling phenomena and provides a tool for the exploration of physics the beyond the SM.
In order to provide the best possible description of signal and background processes, a special care was given to the MC simulations. This is especially true for the QCD loop-induced background which was simulated using *MG5* with up to two hadronic jets modeled at the matrix element and matched to parton shower using the MLM matching scheme for the first time.

The Matrix Element Likelihood Approach (MELA) discriminant was used to calculate the EWK and EWK+QCD cross-sections in three fiducial regions defined to be as close as possible to the reco-level selection. The measurements were done using the MELA distribution as a base for a maximum likelihood fit to the observed data and a cut-and-count approach for the EWK and EWK+QCD cross-section measurements, respectively. The Electroweak (EWK) signal strength measurement in the three regions was reported as well.
The EWK signal was measured with background-only hypothesis rejected with significance of 4.0 (3.5 expected) standard deviations.

A Boosted Decision Tree (BDT) classifier was used as an alternative signal extraction method in order to gauge possible gain in the sensitivity, with respect to MELA. The nominal BDT classifier used seven input variables to extract the EWK signal from the main QCD-induced background and is referred to as the *BDT7*. An additional BDT was built using the a of 28 variables, referred to as the *BDT28*, in order to asses possible gain when using a larger set of variables.
The shape of the BDT classifier was used as the template for the maximum likelihood fit. The background-only hypothesis was rejected using the *BDT7* with significance of 3.77 (expected) standard deviations for the combined data-taking period. This can be compared to the significance of 3.5 standard deviations obtained using MELA. It shows that MELA is able to capture the full kinematics of the event and a small gain in the significance (< 6%) was not deemed enough to change the methodology. Observed EWK signal significance of 5.1 standard deviations using the *BDT7* is reported.
The observed (expected) significance using BDT28 was found to be 4.69 (3.85) standard deviations. This shows only a marginal gain in sensitivity ($\approx 2\%$) was achieved compared to BDT7.

The expected and observed lower and upper 95% CL limits on the anomalous quartic gauge couplings for the charged-current operators T_0, T_1 and T_2 as well as the neutral-current operators T_8 and T_9 are also reported. The limits obtained for the neutral-current operators T_8 and T_9 and discussed in this chapter were the tightest bounds available for these couplings at a time.

Chapter 5

Prospective studies for the High-Lumi and the High-Energy LHC

5.1 Preface to the chapter

The previous chapter showed that the Run 2 data have opened the door for the measurement of the VBS processes with two Z bosons accompanied by the two jets coming from EWK vertices. However, the measurement of the individual vector boson polarizations remains out of reach because of the low cross-section of these processes. At the same time, it is the longitudinal polarization of vector bosons that is directly connected to the EWSB and the Higgs mechanism. In 2018, I did a study, using the 13 TeV LHC data [101], to project the measurement sensitivity of the longitudinal polarization of the Z bosons for the HL- and HE-LHC conditions. This was done by simply scaling the measured yields with luminosity and cross-section expected at future LHC conditions. This provided a motivation to simulate a detailed kinematics at 14 and 27 TeV and do a more in-depth analysis. This analysis is presented in here.

In the beginning sections, the reader will familiarize himself with the MC simulations of the signal and background processes that were prepared for this analysis. Event selection is defined in section 5.3. Additionally, I studied the effect of extended acceptance for electrons that is foreseen for the detector upgrade at HL-LHC phase that will introduce the High Granular Calorimeter Nose in front of HF (referred to as the HF nose). [will be discussed in chapter 2]. Since no additional treatment for the HF nose upgrade was needed in the analysis, it is only referred to in the last section when results are discussed.
Section 5.4.1 will describe the lepton-jet cleaning algorithm that I designed in order to remove lepton duplicates from the jet collection. Origin and the effect of these on the analysis are also discussed.
I studied the effect of parton showers and PU on the selection of the leading and the subleading jets. This is presented in sections 5.4.2 and 5.4.3.
In order to maximise the signal sensitivity measurement, I designed two signal extraction algorithms: the combined-background BDT and the 2D BDT. This will be covered in section 5.6.1.
Next, the kinematics for the signal and the background processes and the application of the signal extraction techniques on 14 TeV and 27 TeV samples will be shown.
Results are presented in section 5.7 followed by the summary of the key points discussed in the chapter.

5.2 Simulations of the signal and backgrounds

The first step in the analysis is the simulation of the hard processes of interest. This was done with MadGraph5_aMC@NLO (henceforth MG5) package for all EWK processes and irreducible QCD $pp \to ZZ$ background. For the gluon loop-induced QCD background this was done using the MCFM (Monte Carlo for FeMtobarn processes) tool. MG5 is a fully automated and publicly available framework born from merging features of MadGraph5 and aMC@NLO tools. The framework is capable of computing tree-level as well as one-loop amplitudes for arbitrary processes. Below the surface, MG5 is a meta-code tool written in Python that utilizes Python, C++ and Fortran to simulate the desired process. In essence, a user defines a theory model and a set of process-independent building blocks. Subsequent steps are done by the tool automatically [102].

MCFM is a parton-level integrator tool developed by Fermilab that allows the calculation of any infra-red finite quantity up to NLO in α_s. The event generation in MCFM is done with the help of the general-purpose integration algorithm VEGAS. The integration routine consists of producing several iterations of sets of events and a grid optimization after each iteration for faster convergence. Currently, more than 300 processes are included in the tool [103–107].

The next step in MC simulation is the parton showering and hadronization of the outgoing particles and the simulation of the detector effects. The former is done using the PYTHIA8 framework and the later using the DELPHES tool.

PYTHIA8 is a standalone tool used to generate events in the high-energy collisions. However, in this analysis, it has been used in conjoncture with MG5 through the usage of Les Houches Event (LHEF) files [108]. This is a standard file format used in high energy physics to store process and event information obtained from event generators. The matrix element calculation is done by MG5, and the output is stored in the standard LHEF format. This is then used by PYTHIA to simulate the parton showering and hadronization [109].

Beforementioned MG5 and PYTHIA8 tools deal with the event production based on purely theoretical considerations. However, to do a proper analysis, one cannot dismiss the importance of the interaction between matter and radiation with the detector. Whenever such analysis requires a high level of accuracy, these interactions are simulated using the GEANT4 package. It is important to note that, although this tool provides the most sophisticated simulation of the detector effects, it is also very complex and time consuming. For this analysis such level of precision is not required. Thus, detector effects were simulated using the DELPHES tool which was designed by the LHC collaborations to be two to three orders of magnitude faster than GEANT4. This is done by propagating particles emerging from hard processes to the calorimeters in the uniform magnetic field parallel to the beam direction. The energies and momenta of long-lived particles are smeared to match the detector response. To take into account the CMS measurement efficiencies in different η regions, an efficiency parametrization from the full detector simulation is used. All these effects are stored in configuration files that must be forwarded to Delphes at runtime. Standard configuration files used by the CMS collaboration were used for generation of all processes. CMS_PhaseII_0PU_v02.tcl was used for 0 PU, while CMS_PhaseII_200PU_v03.tcl was used for 200 PU samples.

In Delphes, it is assumed that electrons and photons leave all their energy in the electromagnetic calorimeter (ECAL) and forward calorimeters (FCAL). At the same time, neutral and charged hadrons leave all their energy in the hadron calorimeter (HCAL) and FCAL. Finally, the sharing of particle energy between two or more neighboring cells, in case the particle hits a cell near its edge, is not implemented.

Electrons and muons are identified in Delphes with no fake rates. For both, the efficiency is exactly zero outside the tracker acceptance. Both final electrons and final muons are obtained by smearing their 4-momentum.

In the analysis, the final states are dominated by jets. As such, it is important to identify them correctly. It is possible in Delphes to produce jets by starting from different collections. These can be generated jets, calorimeter jets or particle-flow jets. Generated jets are obtained by clustering generator-level (henceforth gen level) particles after parton shower and hadronization. Calorimeter jets are reconstructed by using calorimeter towers which are overlayed collections of cells from ECAL and HCAL. Energy-flow jets are obtained by combining the information from

particle-flow tracks and particle-flow towers. Particle-flow tracks are reconstructed tracks from the ECAL and HCAL originating from the charged hadrons.

At last, there are six different jet clustering algorithms in Delphes that can be used to reconstruct jets: CDF jet clusters, CDF MidPoint, Seedless Infrared Safe Cone, Longitudinally invariant kt jet, Cambridge/Aachen jet and Anti-kt jet algorithm. In this analysis the Anti-kt jet algorithm was used [110, 111]. In simple terms, the Anti-kt algorithm can be understood as the following. One can assume that within an event there is a number of well-separated hard particles with transverse momenta $kt1$, $kt2$, $kt3$... and many soft particles. If there are no other hard particles closer than $2R$ around a given hard particle, then all soft particles in circle of radius R will be clustered together with the hard particle resulting in a perfect conical jet. If there are two hard particles with distance $R < d_{ij} < 2R$ between them, then there will be two hard jets. If $k_{t1} \gg k_{t2}$ then only jet 1 will be conical since the second jet will miss part that overlaps with the first jet. If $k_{t1} = k_{t2}$ than neither of the jets will be conical with the overlapping part being equally divided between the two [112].

5.2.1 Simulations of the EWK signal

In this analysis, the signal is the purely electroweak production of the two longitudinally polarized, leptonically decaying Z bosons accompanied by two hadronic jets originating from electroweak vertices. In the rest of the text, this process will be referred to as simply LL. It was simulated at the LO by explicitly requiring that the number of QCD vertices be zero:

$$generate\ pp > z\{0\}z\{0\}jj\ QCD = 0,\ z > l + l-$$

Samples for both HL-LHC and HE-LHC configurations were simulated by requiring 7 TeV and 13.5 TeV beam energy respectively.

Important parameter to be set is the parton distribution function (PDF). A PDF is defined as the probability of finding a parton within a proton with a given fraction of the total proton energy. In the MC simulation of the signal, cteq6l1 PDF set was used [113].

In addition, 10 GeV and 3 GeV cuts were imposed on the p_T of the jets and leptons respectively. Cut on pseudo-rapidity for both jets and leptons have been left open to enable a study of the effect of the future hadronic nose (henceforth HF nose) upgrade on the measurement sensitivity. Finally, a cut of 100 GeV is imposed on the di-jet system mass to suppress the tri-boson contribution. Samples with and without parton showering were simulated at both 14 TeV and 27 TeV to check the effect of parton showering on the tagging jets. In addition, zero PU samples were produced to check the effect of PU. In the end, 200 PU samples with parton showering included were used to obtain the signal significance.

5.2.2 Simulations of the EWK backgrounds

The EWK background in this analysis is the purely EWK production of the two leptonically decaying Z bosons accompanied by two hadronic jets originating from electroweak vertices where at least one Z boson has transverse polarization. These processes will be referred to as LT and TT in the following chapters.

$$generate\ pp > z\{0\}z\{T\}jj\ QCD = 0,\ z > l + l - \qquad (LT\ and\ TL\ polarisation)$$

and

$$generate\ pp > z\{0\}z\{T\}jj\ QCD = 0,\ z > l + l - \qquad (TT\ polarisation)$$

Generator level cuts are identical to those used in the signal simulations. The cross-sections of EWK samples used in the analysis are given in Table 5.1

	EWK LL 14 TeV	**EWK LT 14 TeV**	**EWK TT 14 TeV**	**EWK LL 27 TeV**	**EWK LT 27 TeV**	**EWK TT 27 TeV**
$\sigma[fb]$	0.033	0.189	0.317	0.115	0.669	1.142

Table 5.1: Cross-sections for all EWK processes after the gen level at HL-LHC and HE-LHC.

From the table one can see that the LL contribution is only $\approx 6\%$ of the total at 14 TeV which is also the case at 27 TeV. The cross-section of each contribution rises by a factor ≈ 3.5 when going from 14 TeV to 27 TeV. Examples of Feynman diagrams showing the EWK production of two Z bosons and 2 hadronic jets can be seen in Figure 5.1. Figure also shows interference diagram with the Higgs boson which ensures the unitarization of the theory.

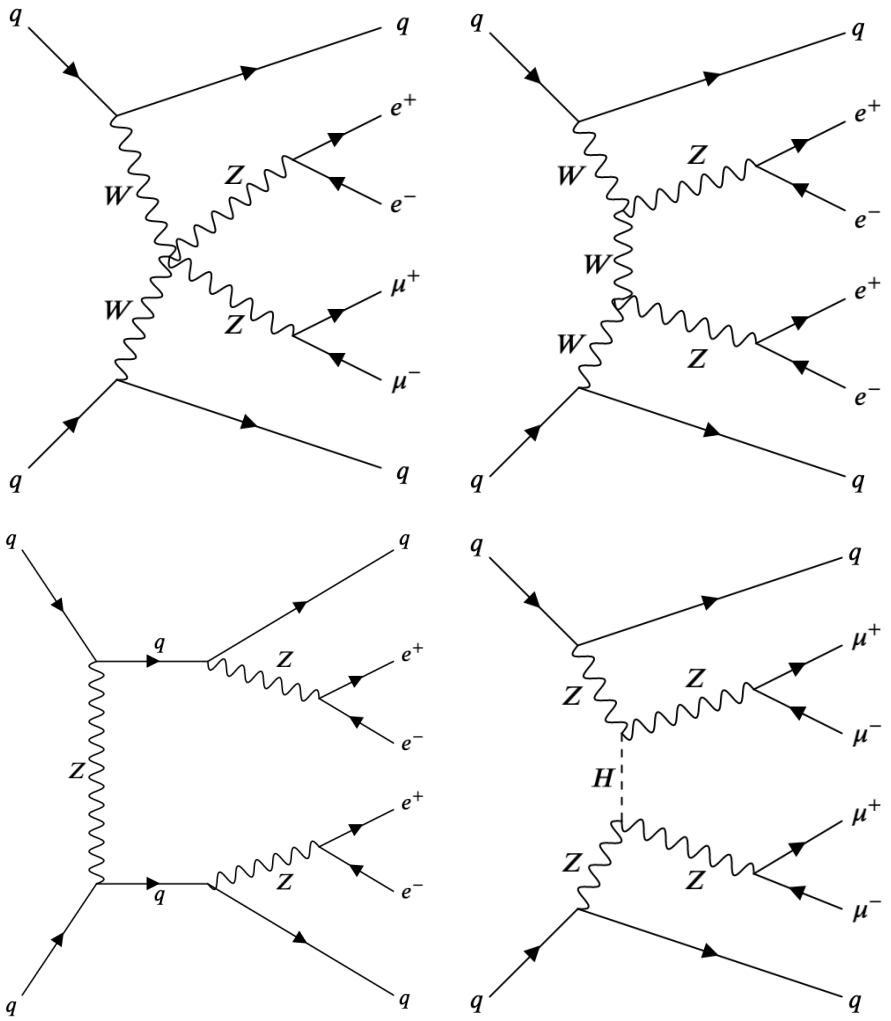


Figure 5.1: Example diagrams for the EWK production of two jets and two Z bosons decaying leptonically. The interference of the bottom-right diagram featuring the Higgs boson exchange with the processes depicted in the top row ensures the unitarization of the theory. [to Claude: discuss comment in anotated document]

5.2.3 Simulations of the QCD backgrounds

The dominant background for this analysis is QCD-induced $pp \to ZZ$ process with up to 2 [to Claude: why comment "at least" in annotated doc?] extra parton emissions coming from the QCD vertices. This is an irreducible background since the final state is identical to that coming from the signal process. Henceforth, the main QCD background will be referred to as qq background.

As for the EWK processes, the qq background was simulated at HL- and HE-LHC conditions with zero PU as well as 200 PU and with the parton showering included as well as without it. In addition, 1,2-jet samples were simulated at LO and 1-jet sample was simulated at the NLO. All the samples were simulated with MG5 with the following syntax:

$$generate\ pp > zzj\ QCD = 1\ QED = 2,\ z > l + l - \qquad (1j@LO)$$

$$generate\ pp > zzjj\ QCD = 2\ QED = 2,\ z > l + l - \qquad (2j@LO)$$

$$generate\ pp > zz > l + l - l + l - j\ [QCD] \qquad (1j@NLO)$$

The same set of generator level cuts was used for the LO samples. The jet p_T was set to 10 GeV, while jet η was set to 5. For the leptons, the p_T cut was set to 3 GeV, while η was left opened. The di-jet mass of the 2j@LO sample was set to 100 GeV. For the 1j@NLO sample the jet p_T cut was set to 15 GeV. Other cuts were later defined at the reco level.

The LO samples were used to assess the effect of PU on the leading and the subleading jets in case of single jet and two jets produced at ME. The 1j@NLO sample with parton shower included is the nominal sample and was used in the analysis. To simulate the 1j@NLO sample without the parton shower, one must specify this in the main Delphes configuration file:

$$PartonLevel : ISR = off$$

$$PartonLevel : FSR = off$$

Like for the EWK samples, cteq6l1 PDF was used for the LO samples. For the NLO sample, NN23NLO PDF was used.

Finally, there is also a gluon loop-induced ZZ production simulated at LO and hence denoted gg background. Although it contributes only at around 10% level with respect to the main background, it is nevertheless included to obtain better projections of the signal sensitivity. It is simulated at both HL- and HE-LHC conditions using the MCFM package. To simulate a desired process in MCFM, one must define a process number in the configuration card. This was set to 132 which corresponds to a LO production of the $gg \to ZZ$ processes with 4 leptons in the final state. In order to faster simulate a gg contribution, only $2e2\mu$ final state was included, therefore omitting the $4e$ and 4μ final states. Effectively, only half of the phase space is simulated this way which is reflected in the event counts. To counter this, expected event counts are doubled before performing a multivariate analysis. For this process, NN2.3NL PDF set was used. In the previous chapter, the state-of-the-art gg simulation was used. This was not done here since that level of precision was not needed.

Examples of the QCD background diagrams are shown in Figure 5.2. The cross-sections from the QCD samples used in the analysis are given in Table 5.2. At 14 (27) TeV, the cross-section of the qq background is ≈ 14 (≈ 11) times larger than the gg cross-section. The increase in cross-section when moving to 27 TeV is more pronounced in the gg background (≈ 3 times) than in the gg background (2.3 times).

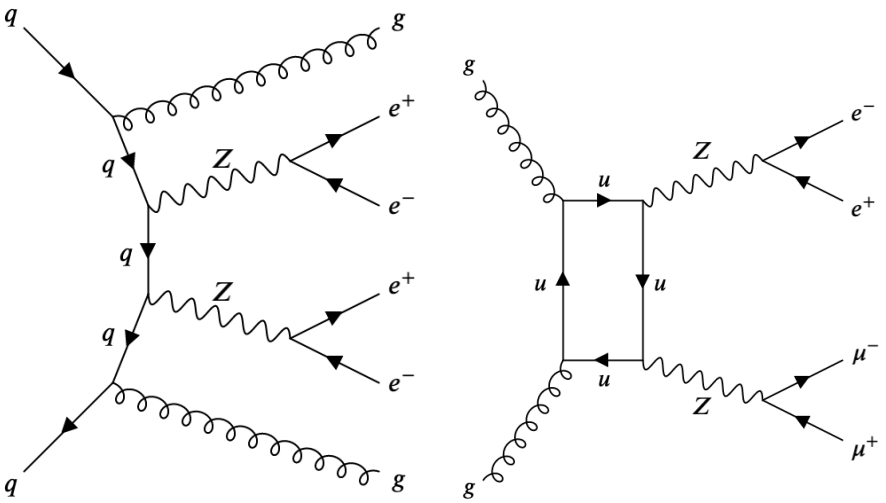


Figure 5.2: Example diagrams of the QCD induced production of two Z bosons in the fully leptonic decay channel. The left figure shows the irreducible background with 2 jets in the final state. The right figure shows loop-induced production of the gg background.

	QCD qq 14 TeV	**QCD gg 14 TeV**	**QCD qq 27 TeV**	**QCD gg 27 TeV**
$\sigma[fb]$	49.6	3.57	116	10.9

Table 5.2: Cross-sections from the QCD qq 1j@NLO and QCD gg@LO samples for HL-LHC and HE-LHC energies.

5.3 Event selection

The $ZZ \to 4l2j$ channel is a good candidate for the study of EWSB mechanism due to the clean final state that can be fully reconstructed. For this reason, it is also a great fit for studying the scattering of longitudinal vector bosons. However, a small cross-section of the LL, compared to background, process makes it challenging to measure. In order to suppress the background as effectively as possible, a set of efficient selection criteria has to be put in place. This was done in several steps:

1. In the analysis we require isolated objects. In the Delphes framework an object, such as an electron or a muon, is said to be isolated if the activity in a cone of radius R around the lepton direction is small enough. This is precisely defined with the variable I as

$$I = \frac{\sum_{i \neq P}^{\Delta R(i)<R,\ p_T(i)>p_T^{min}} p_T(i)}{p_T(P)}$$

   where the nominator sums over the p_T of all particles that are in the cone around the particle of interest, P, and the denominator is the p_T of the particle P. Values of $I > I_{min}$ indicate that the particle is isolated. The parameters $R,\ p_T^{min}$ and I_{min} are set to 0.5, 0.1 GeV and 0.1 respectively.
   For isolated electrons (muons) , the p_T is required to be above 7 (5) GeV, while the $|\eta|$ is required to be less than 3 (2.8). Additionally, the extended acceptance for electrons was considered for which the $|\eta|$ acceptance for electrons was increased to 4.

2. With both HL- and HE-LHC conditions, PU is expected to affect the analyses and the CMS collaboration has been working on a number of algorithms to mitigate its effects. One such technique, the charged-hadron subtraction (CHS), was designed to remove charged particles coming from pileup vertices from the reconstructed objects and was used to treat leptons in this analysis. This method is not efficient enough when the PU contributions come from neutral hadrons. For this reason, a new PU mitigation approach, the pileup per particle identification (PUPPI), was devised. This technique was built on top of the CHS algorithm, and it estimates the probability that the neutral particle comes from the PU. It then scales the energy of such particles based on the calculated probability [114]. PUPPI algorithm was used to reduce the effect of PU on jets.
   In this analysis, cuts were set to 25 GeV for the jet p_T and 4.7 for the jet $|\eta|$

3. Final state leptons are coming from the decay of Z bosons. Each Z boson candidate is reconstructed from a pair of oppositely charged electrons or muons with a dilepton mass in the window $60\ GeV < m_{ll} < 120\ GeV$. Each event is required to have a pair of non-overlapping Z bosons where the leading Z boson is chosen as the one with the highest p_T.

This set of requirements is referred to as the *ZZ selection*. For the analysis, two regions of interest are defined by additional selections:

- a baseline selection is built on top of the ZZ selection by requiring the $m_{jj} > 100GeV$.
- a VBS selection is defined by requiring $m_{jj} > 400GeV$ and $|\Delta\eta_{jj}| > 2.4$

A summary of the selection criteria used in the analysis is shown in Table 5.3

lepton candidates	$p_T^e > 7GeV$ $p_T^\mu > 5GeV$ $\|\eta\|^e < 3\ (4)$ $\|\eta\|^\mu < 2.8$
jet candidates	at least two jets in the event with $p_T > 25\ GeV$ $\|\eta\| < 4.7$
ZZ selection	lepton pair (e^+e^- or $\mu^+\mu^-$) with $60\ GeV < m_{ll} < 120\ GeV$ pair of non-overlapping Z bosons Z_1 defined as the one with the highest p_T Z_2 defined as the one with the next-to-highest p_T
baseline selection	ZZ selection + $m_{jj} > 100GeV$
VBS selection	ZZ selection + $m_{jj} > 400GeV + \|\Delta\eta_{jj}\| > 2.4$

Table 5.3: Summary of the selection criteria used in the analysis. The number in parentheses for electron p_T is referring to the extended acceptance for electrons.

The efficiencies, defined for each contribution as the number of events passing the selection over the number of generated events, after the ZZ selection, baseline selection and VBS selection are reported in Table 5.4.

	ZZ selection		**Baseline selection**		**VBS selection**	
	14 TeV	27 TeV	14 TeV	27 TeV	14 TeV	27 TeV
Z_LZ_L efficiency [%]	51.5	44.2	44.3	38.4	30.3	27.5
Z_LZ_T efficiency [%]	53.9	47.2	47.8	42.5	31.1	29.8
Z_TZ_T efficiency [%]	59.0	52.6	52.6	47.8	32.7	32.3
qq efficiency [%]	44.9	36.6	9.80	11.1	1.40	1.90
gg efficiency [%]	42.1	40.9	13.7	16.7	3.50	4.80

Table 5.4: Signal and background efficiencies for the ZZ selection, baseline selection and VBS selection.

The table is discussed in section 5.5 where distributions for the different polarizations as well as for the qq and gg backgrounds are shown.

[I chose not to comment on the reasons why the numbers in the table are what they are. If I want to back up my claims, I have to show distributions. I believe having some plots showing kinematics here and then again some plots showing kinematics 2 sections later would be less nicer and would result in a forest of plots all around the place. In this way I tell reader that I will actually discuss this table later and then show all the kinematic distribution in one place and discuss it there.]

5.4 Cleaning of lepton-jets and effect of parton showering and pileup on the leading and subleading jets

5.4.1 Lepton-jet cleaning

The signal events in this analysis are characterized by two hadronic jets with high pseudo-rapidity gap between them. It is thus imperative to build the analysis that will be as effective as possible in identifying such jets.
However, it was found that Delphes populates the jet collection with objects previously reconstructed as leptons and stored in the lepton collections. If untreated, this will lead to double counting of objects in an event and wrong interpretation of analysis results. Leptons that are found in the jet collection will be referred to as the *lepton-jets*.

When comparing the η spectrum of leptons and two leading jets, one would expect to see clearly distinctive distributions. However, the top two rows in Figure 5.3 show that the η spectrum of leptons and jets is similar. In addition, one would expect a large pseudo-rapidity gap between the tagging jets in the signal sample which is not the case as can be seen on the bottom plot. This points to the lepton contamination of the jet collection used in the analysis. To check this, events were examined before the baseline selection. Example of one such event is given in Table 5.5. Along four leptons, this event has five jets. However, it can be seen that e_1 and j_1 are the same object stored once in the electron collection and once in the jet collection. The same is true for e_2 and j_3 and μ_1 and j_2. By looking in the LHEF file, it can be confirmed that these are the same final state particles. The small discrepancy in the kinematics seen in Table 5.5 comes from the smearing of energy and momentum in Delphes. The detector response is different for leptons and jets and thus the applied smearing is also different.

One major issue with lepton-jets is that analysis is sensitive to kinematic distributions and this information is used to extract the signal. In addition, event selection requires at least two jets in the event with a p_T of at least 25 GeV and a dijet mass of at least 100 GeV. When lepton-jets are removed from the event discussed in Table 5.5, the event does not pass the selection. Therefore, without removing lepton-jets, the final event counts will be wrong.
Finally, if the lepton is wrongly identified as a jet, a softer jet candidate coming from PU will have a lower probability of becoming a leading or subleading jet and thus the effect of PU will be underestimated.

Several lepton-jet cleaning algorithms were tested before the most efficient one was found:

1. Loop over each object in the jet collection.
2. For each jet, loop over every object in the electron collection. Calculate the distance, ΔR_{jl}, between the lepton and the jet.
3. Remove the jet closest to electron if $\Delta R_{jl} < 0.1$
4. Apply steps 1-3 also for objects in the muon collection.

The performance of the lepton-jet cleaning algorithm was thoroughly checked by going through dozens of events one-by-one and checking whether the algorithm removed fake jets while leaving others untouched. Next, all important kinematic variables were plotted to make sure that lepton-jets were removed. The same set of kinematic variables shown in 5.3 is shown in Figure 5.4 after applying lepton-jet cleaning algorithm. The distributions show the expected difference between the lepton and the jet kinematics, as well as the expected pseudo-rapidity separation between the tagging jets.

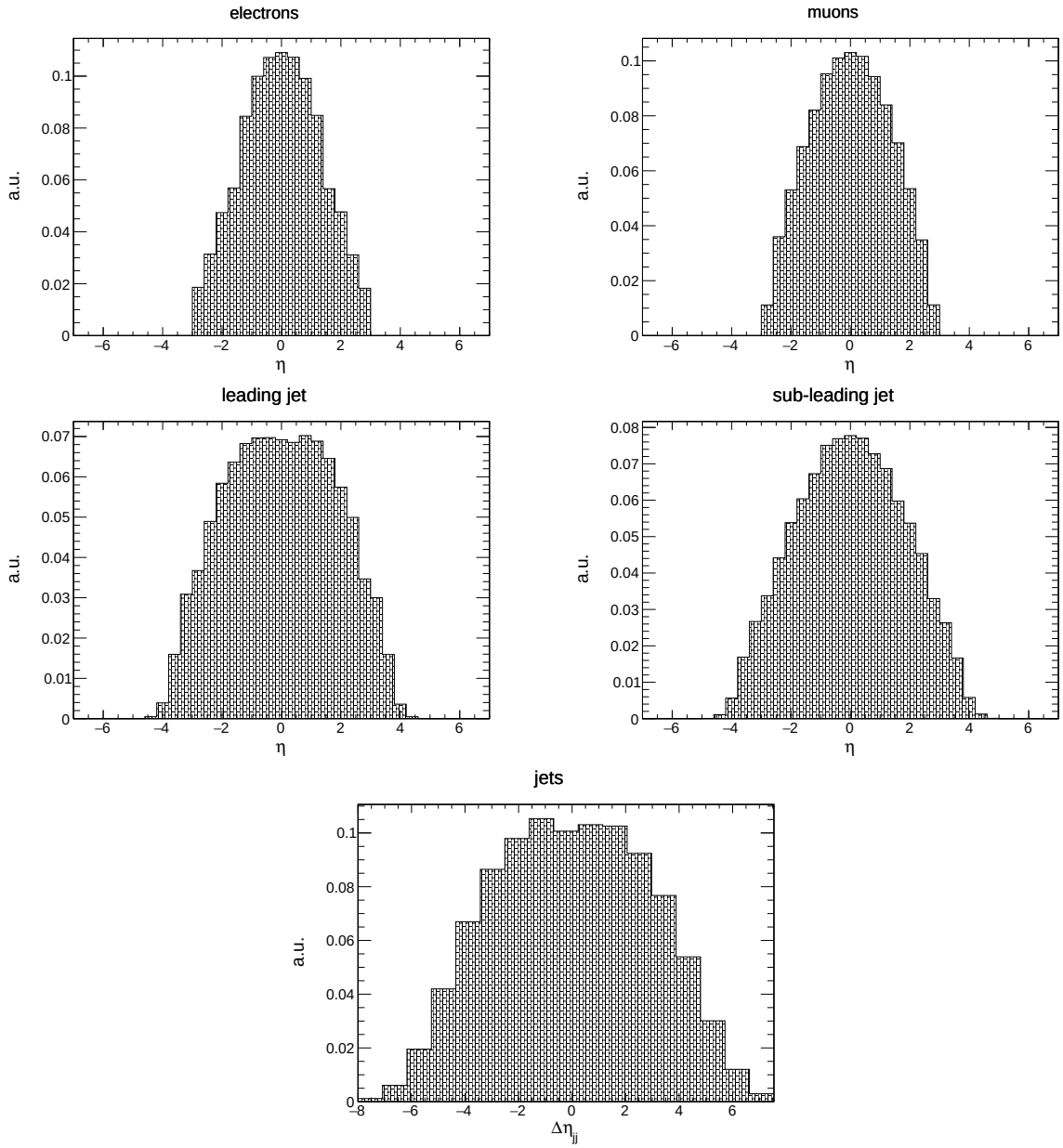


Figure 5.3: Top row: pseudo-rapidity spectrum of leptons. Middle row: pseudo-rapidity spectrum of the two leading jets. Bottom: pseudo-rapidity difference between the two leading jets. Distributions for signal sample are shown, obtained before implementing the lepton-jet cleaning algorithm. Samples are simulated at 14 TeV c.o.m. energy and the baseline selection was applied.

	e_1	e_2	μ_1	μ_2	j_1	j_2	j_3	j_4	j_5
p_T **[GeV]**	121.7	20.6	81.7	14.9	125	85	20.6	18.6	16.8
η	-0.43	-0.29	-0.80	1.29	-0.42	-0.80	-0.29	2.52	-1.64
ϕ	2.99	0.63	-0.31	-0.91	2.98	-0.32	0.63	-1.39	1.37

Table 5.5: An example of values of the two jet kinematic variables in single event before the baseline selection.

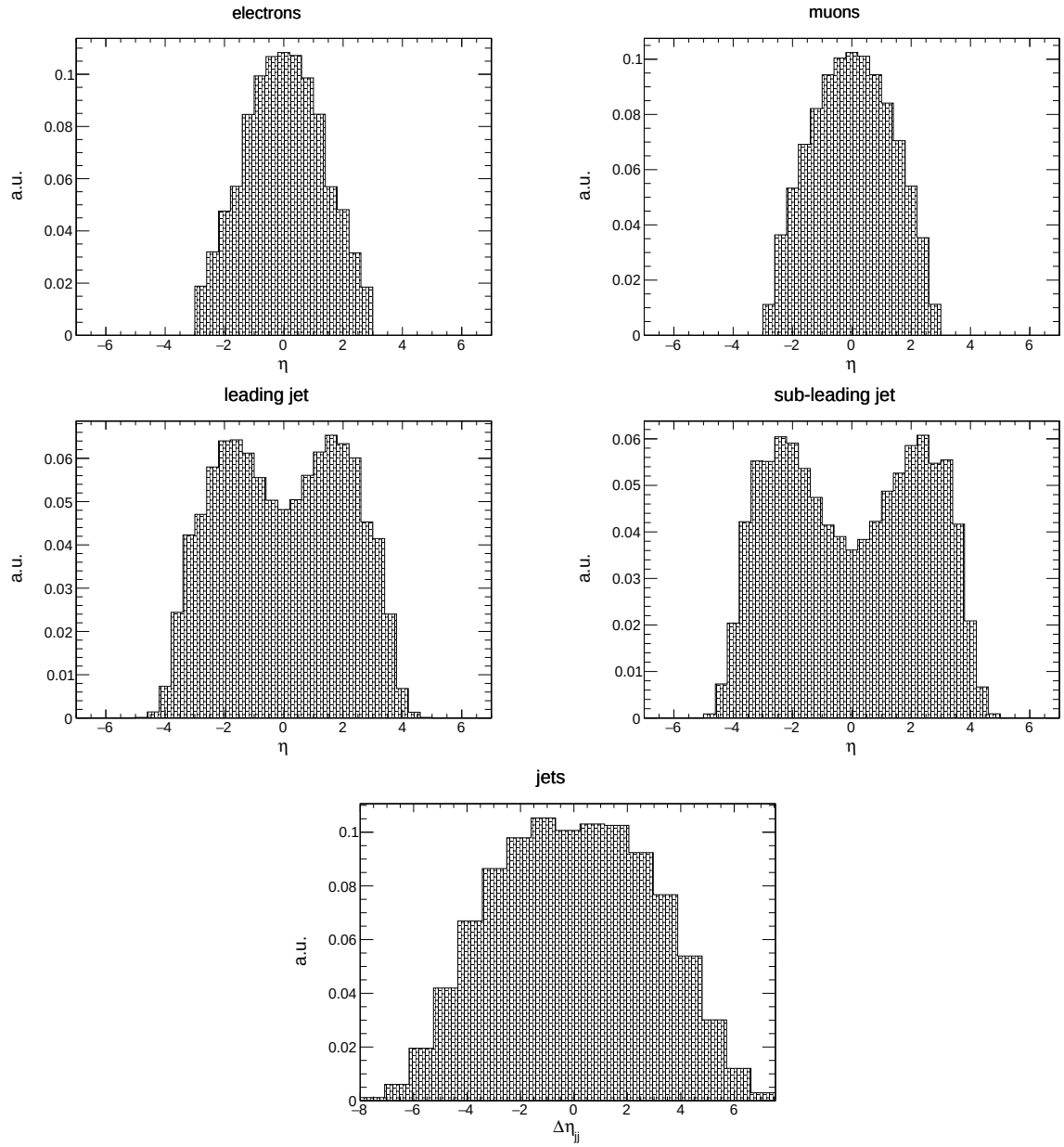


Figure 5.4: Top row: pseudo-rapidity spectrum of leptons. Middle row: pseudo-rapidity spectrum of the two leading jets. Bottom: pseudo-rapidity difference between the two leading jets. Distributions for signal sample are shown, obtained after implementing the lepton-jet cleaning algorithm. Samples are simulated at 14 TeV c.o.m. energy with the baseline selection applied.

5.4.2 Effect of parton showering on the leading and subleading jets

The parton showering is introduced via initial state radiation (ISR) and final state radiation (FSR) simulations which are modelled with differential equations that give the probability of emitting radiation as the parton shower evolves with time. For the FSR, this is done by replacing a mother particle with two daughter particles at each branching. Contrary to the FSR where the parton shower evolves forwards in physical time, the ISR is simulated by starting from a hard scattering partons and successively reconstructing prior branchings in rising sequence of parton energies. In other words, the ISR evolution is modelled backwards in physical time [109, 115].
This section shows the effects of parton showering on the choice of the leading and the subleading jets. For this, zero-PU (henceforth PU0) samples were used so to prevent from mixing the effects of PU and parton showers. The effect is shown for the VBS signal and the main QCD background. The study was done in several steps:

1. Run Delphes twice to obtain
   - sample with parton showering switched off (henceforth no-showering sample)
   - sample with parton showering switched on (henceforth showering sample)
2. record events that pass the baseline selection in the showering sample.
3. record events from the non-showering sample that were also recorded in the step 2. This ensures that the same events are being compared.
4. Compare the two leading jets from the step 2 to the two leading jets from the step 3 and check
   - how often only the leading jet was changed by the parton showering
   - how often only the subleading jet was changed by the parton showering
   - how often either of the two jets were changed by the parton showering
   - how often both the leading and the subleading jets were changed by the parton showering
   - if the leading and the subleading jets simply swapped places, or a new jet was intoduced

The primary effect of parton showering is the increase of jet multiplicity within the event. This is shown in Figure 5.5 that compares the number of jets within the same events for non-showering and showering samples of the VBS signal and the main QCD background.

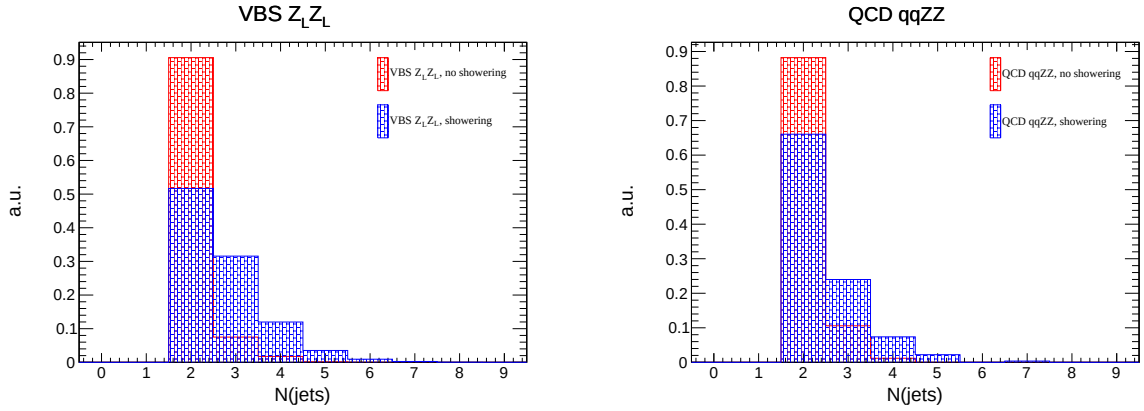


Figure 5.5: The effect of parton showering on the number of jets within the same event for the LL signal (left) and the qq background (right).

In addition, parton showering can change leading (subleading) jet. The leading (subleading) jet is said to be changed by parton showering if its distance, ΔR, to the leading (subleading) jet after parton showering is greater than 0.5. This is illustrated in Fig. 5.6. In the first case, both the leading jet (blue marble) and the subleading jet (red marble) remained the same after parton showering. In the second case, the parton showering caused the leading jet to be replaced by a new jet (green marble). In the third case, the parton showering caused the subleading to jet be replaced by a new jet (green marble). The fourth case depicts two possible scenarios in which both leading jets are changed by parton showering. In the first scenario, the two jets simply swapped places. However, in the second scenario, both leading jets have been replaced by new jets.

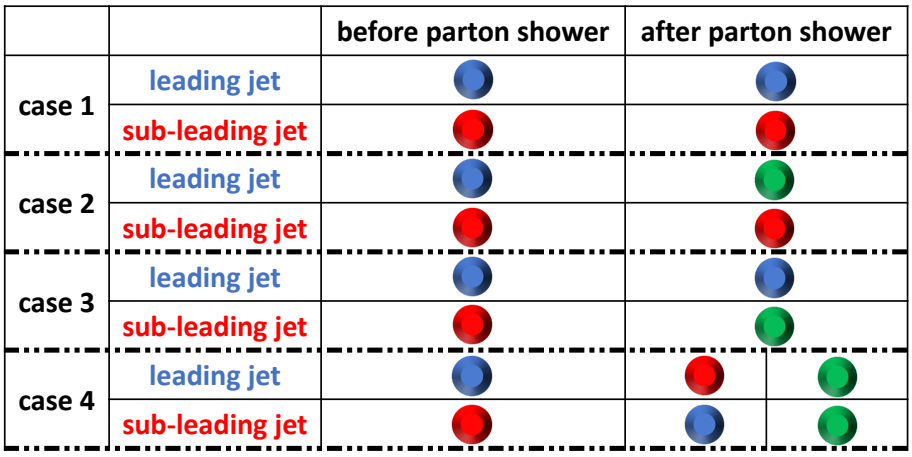


Figure 5.6: An illustration of the effect of parton showering on the two leading jets in an event. Parton showering can either simply swap the two leading jets, or it can introduce a new jet (green marble).

This effect is summarized for the VBS signal and the main QCD background in Table 5.6. In 1.5 % (0.8 %) events parton showering caused the leading jet in the signal (main background) sample to be replaced by a new jet coming from parton showers. This happened to the subleading jet in 16 % (29 %) events. On the other hand, in 23 % (21 %) of events parton showering changed both leading jets! However, in 69 % (56 %) of those events the two jets were simply swapped. Either of the two jets were changed in 40 % (51 %) events. The results indicate that parton showering significantly affects the selection of the leading and the subleading jets.

VBS signal		**QCD qq**	
jets changed [%]	jet replaced	jets changed [%]	jet replaced
1.5	only first	0.8	only first
16	only second	29	only second
23	both	21	both
40	any	51	any

Table 5.6: The left-hand side of the table shows how often jets coming from parton showers interchange or replace tagging jets. The right-hand side of the table shows the same for the leading jets of the main QCD background.

Comparing the cross-section weighted event counts for VBS processes after the baseline selection for the non-showering and showering samples in Table 5.7 one can see that, for the VBS signal, the difference is bellow 10 %. The effect of parton showers on the VBS background is similar. To further show that parton showering is under control, a set of lepton and jet plots for VBS signal is shown on Figure 5.7.

	number of cross-section weighted events after the baseline selection	
	non-showering samples	**showering sample**
LL	2.56	2.79
LT	6.90	7.38
TT	13.8	14.6

Table 5.7: Number of cross-section weighted events for the VBS processes after the baseline selection at 14 TeV. Both non-showering and showering samples were produced from the same gen level output so that the effect of parton showering can be isolated and quantified.

5.4.3 Effect of pileup on the leading and subleading jets

In 2018, LHC has reported a mean PU of 32 at 13 TeV c.o.m. energy. This number is expected to be around 200 for the HL-LHC at 14 TeV. PU makes physical analyses more difficult by adding a large background noise and, therefore, must be treated carefully. The study of PU effects was also done in several steps:

1. Run Delphes twice to obtain
   - sample without pileup (henceforth PU0 sample)
   - sample with 200 pileup (henceforth PU200 sample)
2. record events that pass the baseline selection in PU200 sample.
3. record events from the PU0 sample that were also recorded in step 2. This ensures that the same events are being compared.
4. Compare the two leading jets from the step 2 to the two leading jets from the step 3 and check
   - how often only the leading jet was changed by PU
   - how often only the subleading jet was changed by PU
   - how often either of the two jets were changed by PU
   - how often both the leading and the subleading jets were changed by PU
   - if the leading and the subleading jets simply swapped places, or a new jet was introduced

As for the previous study, the leading (subleading) jet is said to be changed by PU if its distance, ΔR, to the leading (subleading) jet after PU is greater than 0.5.
Table 5.8 summarizes the effect of PU on the leading and the subleading jets for the VBS signal and the qq background. In 0.3 % (1.1 %) events PU caused the leading jet in the signal (main background) sample to be replaced by a new jet coming from PU. This happened to the subleading jet in 9 % (15 %) events. In 11 % (12 %) of events PU changed both leading jets! However, in 81 % (67 %) of those events, the two jets were simply swapped.

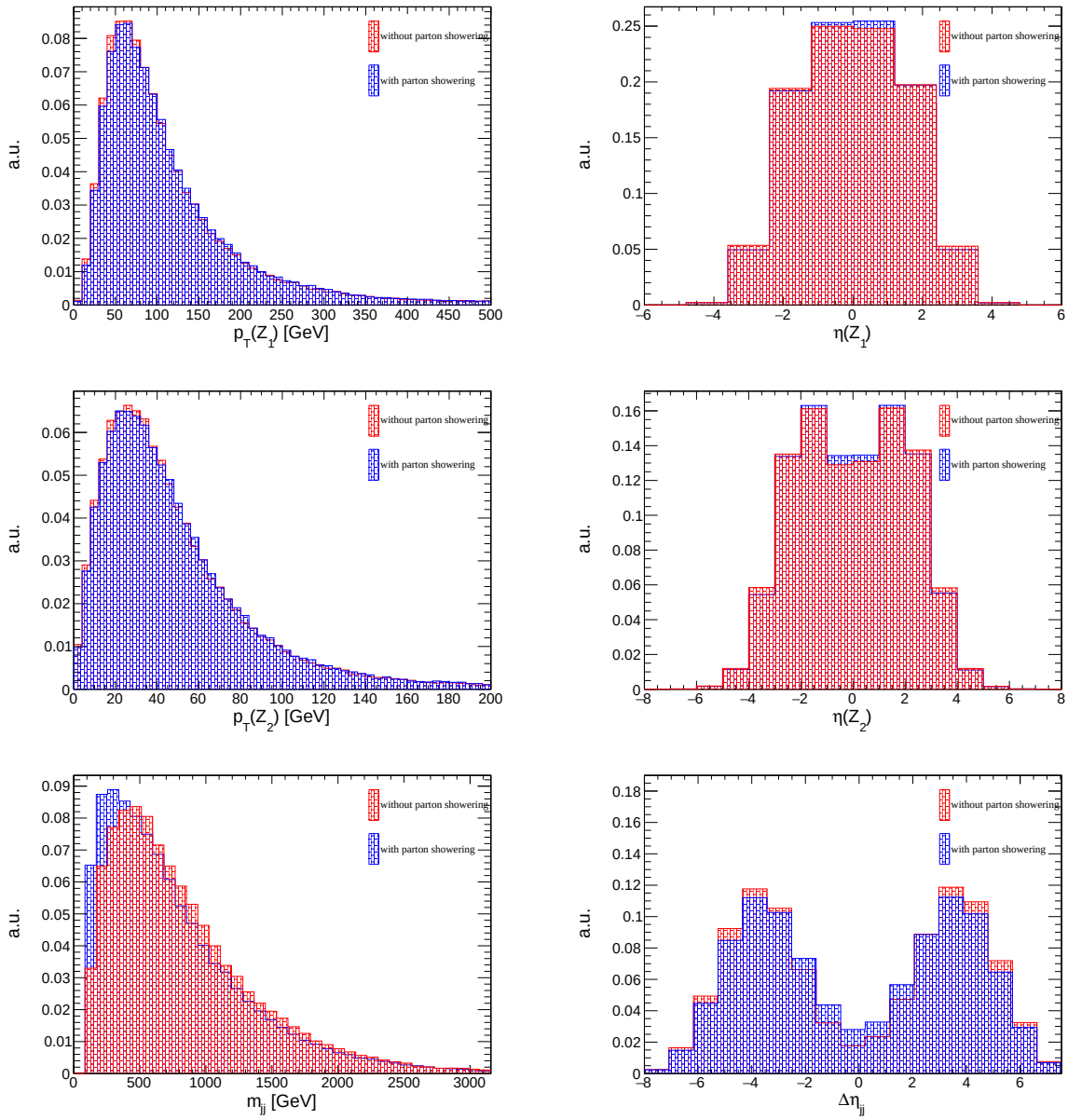


Figure 5.7: Lepton and jet kinematic distributions for the non-showering and showering samples after the baseline selection at 14 TeV. Distributions for the VBS signal are shown.

Either jet was changed in 20 % (29 %) of events. This result is especially significant for the VBS signal where the tagging jets are replaced by the PU jets in around 10 % events. Although this effect is not extreme, it is sizeable.

VBS signal		**QCD qq**	
jets changed [%]	which jet replaced	jets changed [%]	which jet replaced
0.3	only first	1.1	only first
9	only second	15	only second
11	both	12	both
20	any	29	any

Table 5.8: The left-hand side of the table shows how often jets coming from PU interchange or replace tagging jets. The right-hand side of the table shows the same for the leading jets of the main QCD background. Both samples are simulated with parton showers included.

Another interesting PU feature can be seen by looking at the jet η distribution for the LL signal and the qq background shown on Figure 5.8. The left-hand side plots show the pseudo-rapidity and the pseudo-rapidity separation between the two tagging jets in the signal sample, while the right-hand side plots show the same distributions for the two leading jets in the background sample. The effect of PU on the shape of jet distributions is especially pronounced in the qq sample.

Distributions of both the leading and the subleading jets show two horns in $3 < |\eta| < 4$ region. The low statistics of the 1j@NLO PU0 sample makes this harder to see. For this reason, the right-hand side distributions are also shown in Figure 5.9 using the 1j@LO high-statistics sample. This feature is more pronounced in the η distribution of the subleading jets of the background samples compared to the signal sample because the subleading jet in the background sample is generally softer then the second tagging jet of the signal sample and is more affected by PU. One can recall from the previous chapter that horns were observed, in both data and the simulation, in the 2017 data-taking period and it was traced back to the noisy crystals. Here, the PU represents the noise in the analysis that causes horns to appear. Importantly, it was shown that these horns have a small impact on the analysis.

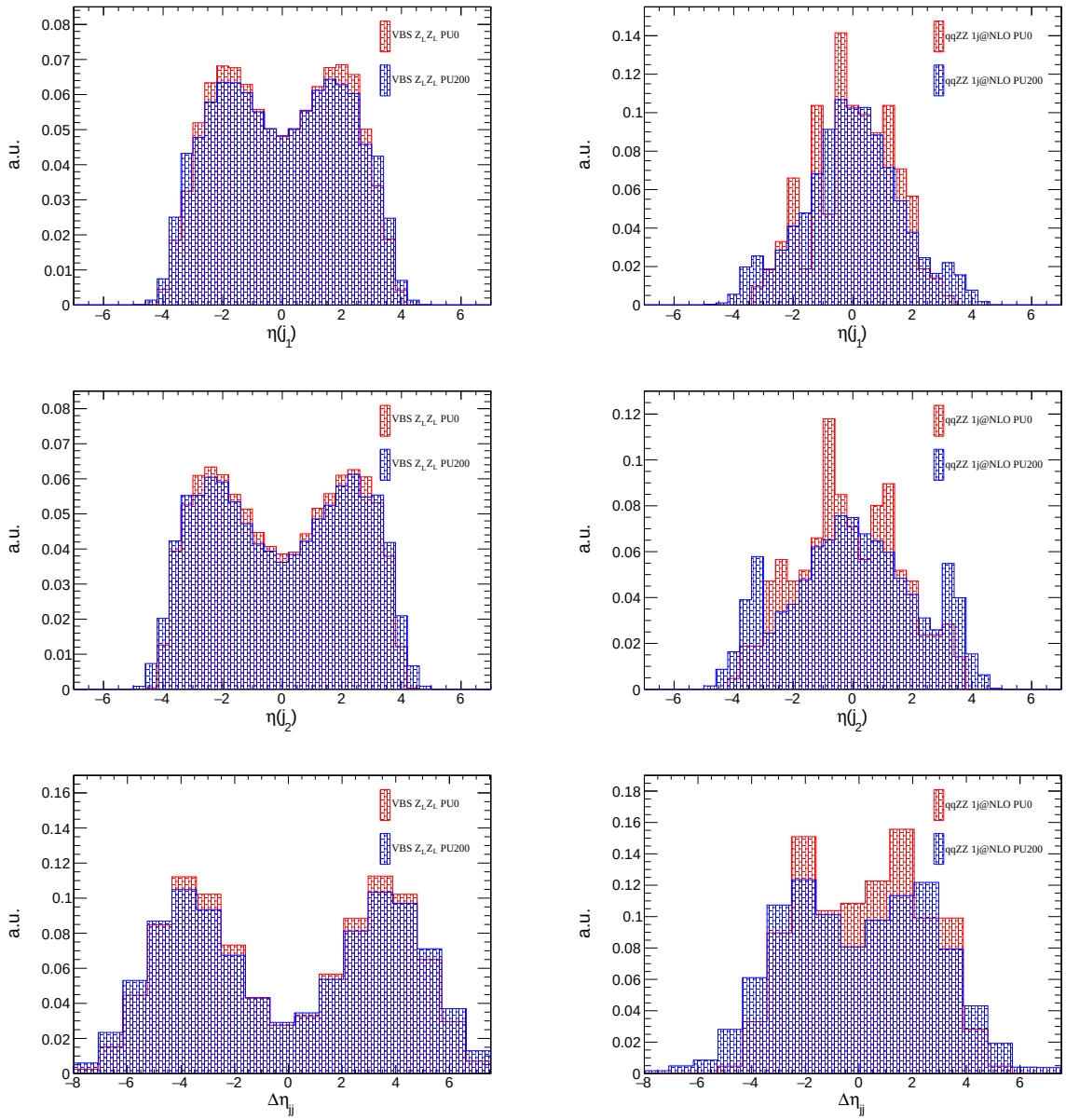


Figure 5.8: The left-hand side plots shows the effect of pileup on the pseudo-rapidity for the two tagging jets in the LL signal samples as well as the pseudo-rapidity gap between them. The right-hand side shows the same distributions for the QCD qq1j@NLO background. All samples were produced at 14 TeV with parton showers included in the simulations.

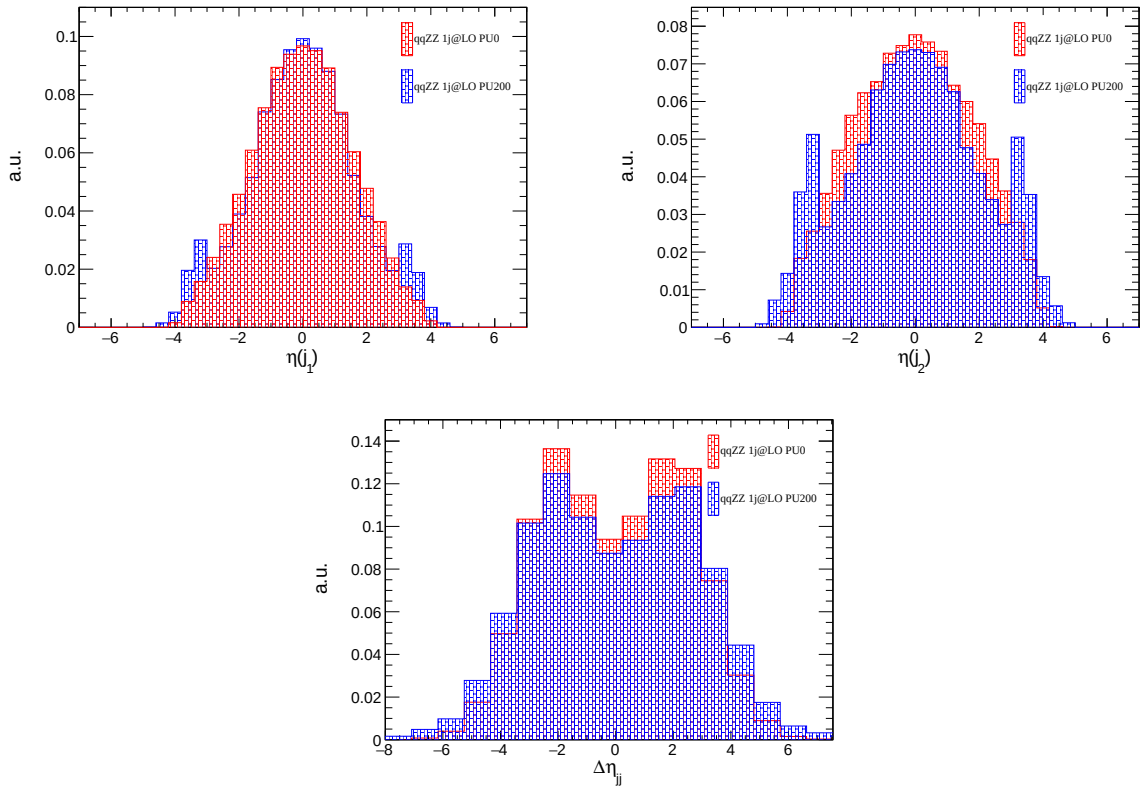


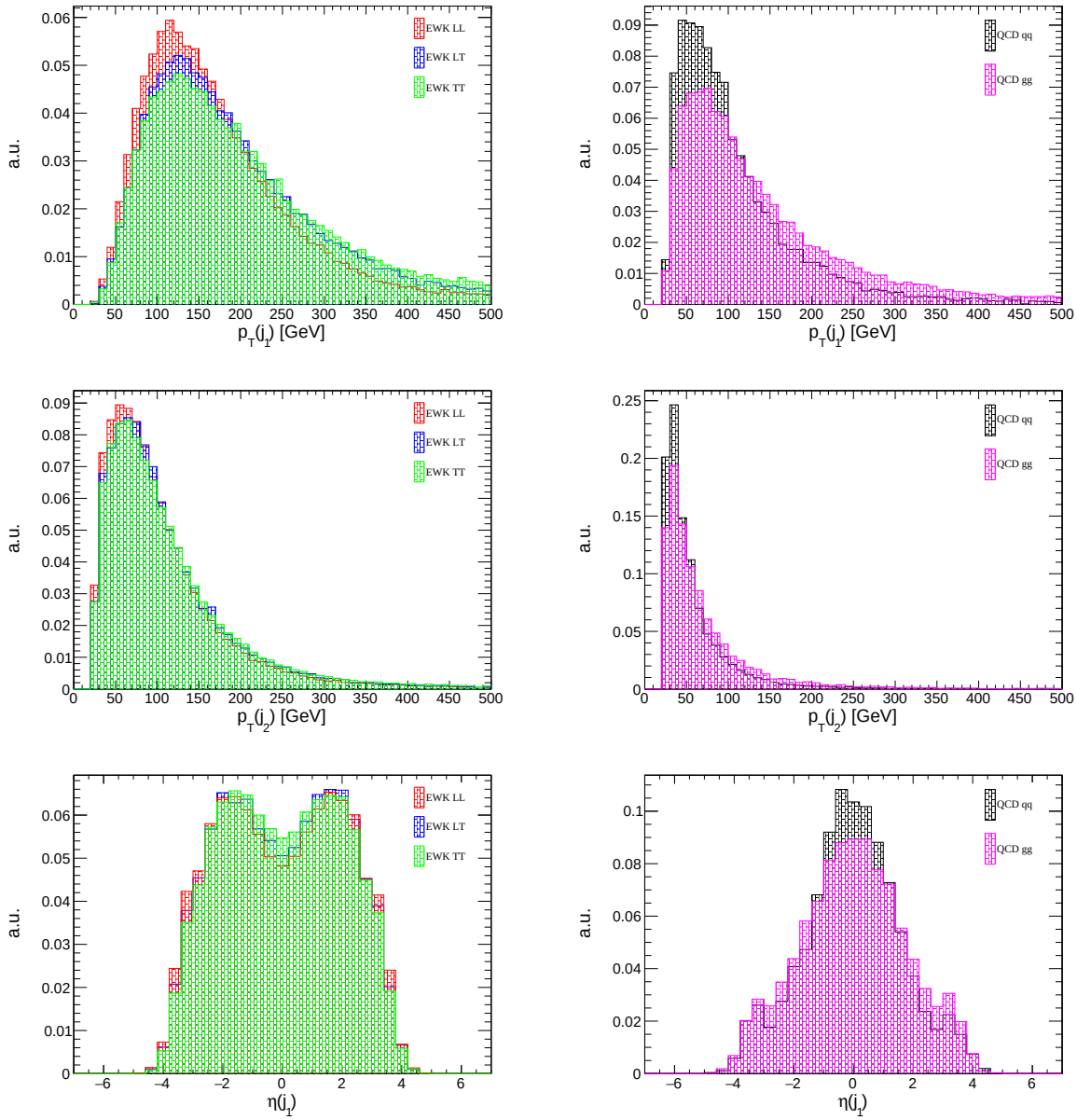
Figure 5.9: The effect of PU on the pseudo-rapidity for the two tagging jets in the QCD qq1j@LO samples as well as the pseudo-rapidity gap between them. Distributions are shown as a supplement for the right-hand plots of the Figure 5.8 since the LO samples were produced with higher statistics. All samples are produced at 14 TeV c.o.m. energy with parton showers included in the simulations.

5.5 Kinematics at 14 and 27 TeV

It was shown in section 5.3 (see Table 5.4) that the QCD background is more affected by the ZZ selection than the VBS contributions. This is mainly due to the requirements on the jets which have harder p_T spectrum for the latter. This is shown on Fig. 5.10. The same figure shows the η distribution of the two leading jets for all contributions. It can be seen that the tagging jets in the LL signal have somewhat softer p_T spectrum and are more forward than the LT and TT backgrounds.

The two leading jets in the VBS samples have a harder m_{jj} spectrum compared to the leading jets in the qq and gg samples. This is reflected in the large drop in efficiency for the QCD samples after the baseline selection ($m_{jj} > 100\ GeV$). This is shown in the top row of Fig. 5.11.

Finally, the VBS selection exploits the fact that the leading jets in the VBS samples have larger pseudo-rapidity separation compared to the leading jets of the qq and gg samples. This is shown in the bottom row of Fig.5.11.



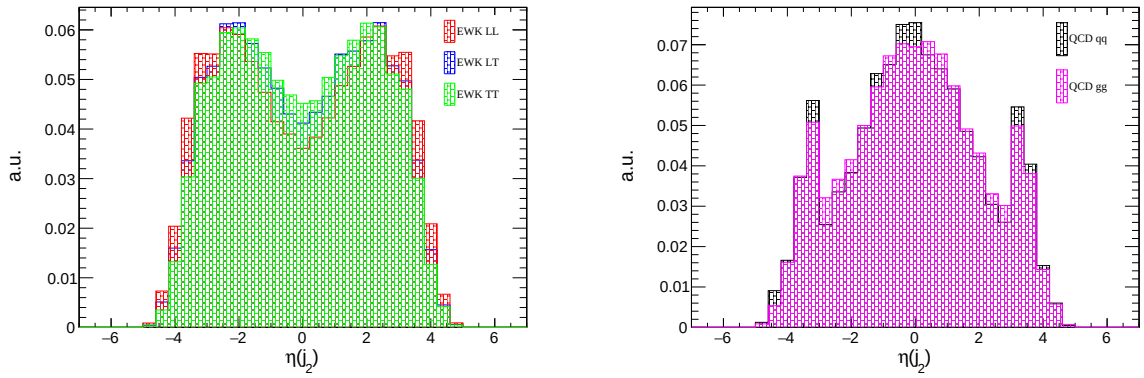


Figure 5.10: Transverse momentum and pseudo-rapidity of the leading jets for the VBS (left) and QCD (right) processes at 14 TeV. The baseline selection was applied.

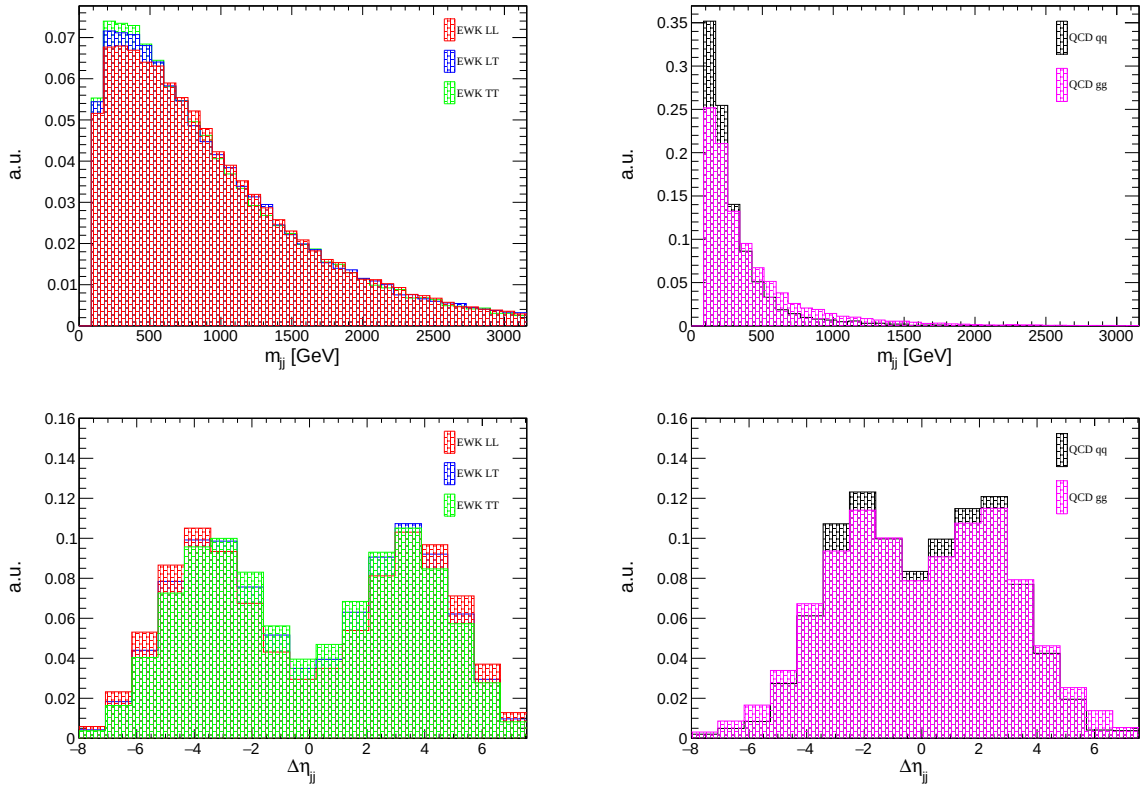


Figure 5.11: Dijet mass and pseudo-rapidity separation for the VBS (left) and QCD (right) processes at 14 TeV. The baseline selection was applied.

The same set of distributions is shown in Figs. 5.12 and 5.13 for 27 TeV. A bigger loss in the efficiency for the VBS contributions at 27 TeV comes mostly from the jet kinematics that shows harder m_{jj} spectrum with more forward jets at 27 TeV compared to 14 TeV.

All distributions at 14 and 27 TeV, along with 14 and 27 TeV distributions overlaid for easier comparison, are shown in Appendix B.

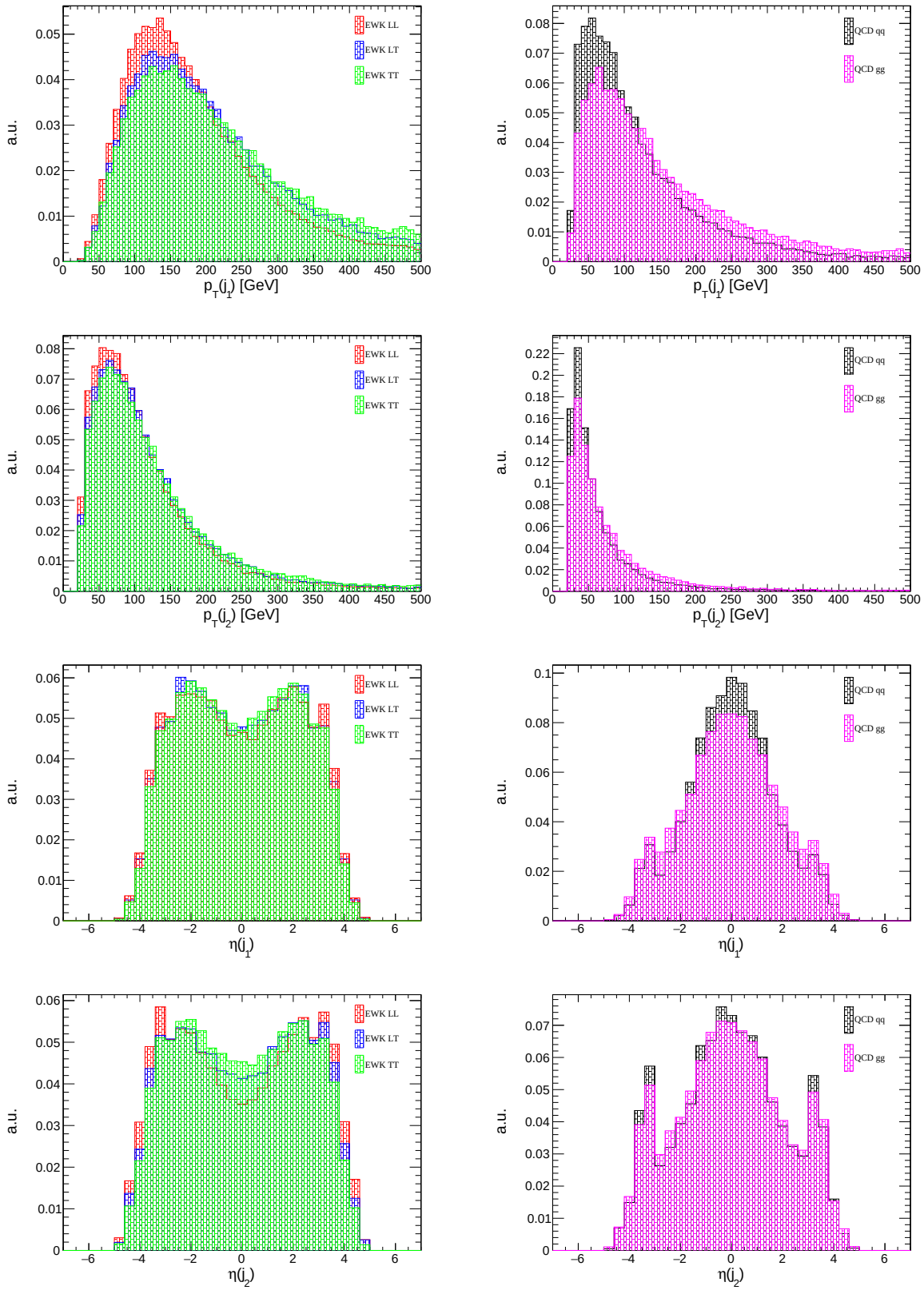


Figure 5.12: Transverse momentum and pseudo-rapidity of the leading jets for the VBS (left) and QCD (right) processes at 27 TeV. The baseline selection was applied.

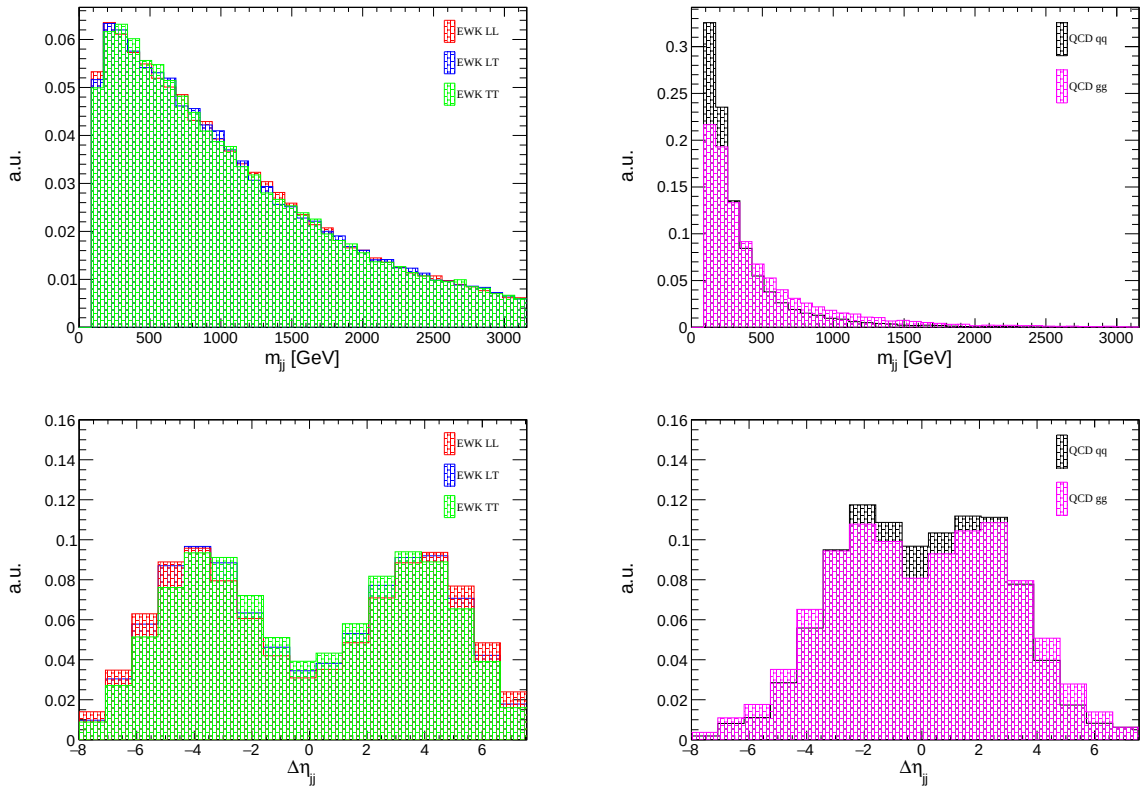


Figure 5.13: Dijet mass and pseudo-rapidity separation for the VBS (left) and QCD (right) processes at 27 TeV. The baseline selection was applied.

The emphasis of this chapter is on the extraction of the longitudinal polarization from the LT and TT polarizations and from the QCD backgrounds. The set of variables used to extract the LL signal is summarized in Table 5.9. The first seven variables were shown in the previous chapter to separate well the VBS contribution from the QCD.

Along with p_T and η of the two Z bosons, variables $\theta^*(Z_i)$, defined as a decay angle between the negatively charged lepton in the Z_i rest frame with respect and the momentum direction of the Z_i in the laboratory frame, were found to separate well the LL signal from the LT and TT backgrounds. Angles $\theta^*(Z_i)$ are illustrated in Fig. 5.14. Fig. 5.15 shows the distribution of the last six variables from Table 5.9 used to separate LL signal from the LT and TT backgrounds. It can be shown [116] that, when calculating decay rates, the matrix elements for transverse and longitudinal polarizations of vector bosons are

$$|M_-|^2 \approx (1 + cos\theta^*)^2 \qquad |M_+|^2 \approx (1 - cos\theta^*)^2 \qquad |M_L| \approx sin^2\theta^*$$

where $|M_-|$ and $|M_+|$ correspond to the left and right helicity states of the transverse polarization, respectively. From here, one would expect a very different angular distribution for transverse and longitudinal polarizations. This is exactly shown in the bottom row of Fig. 5.15.

As a consequence, the p_T spectrum of the longitudinally polarized Z bosons is softer than the p_T spectrum of the transversely polarized Z bosons and the longitudinal component is produced at larger η values.

The same set of plots for 27 TeV is shown in Fig. 5.16

variable	definition
m_{jj}	invariant mass of the two leading jets
$\Delta\eta_{jj}$	pseudo-rapidity separation between the two leading jets
m_{4l}	invariant mass of the ZZ pair
$\eta^*(Z_1)$	η direction of the Z_1 relative to the leading jets: $\eta^*(Z_1) = \eta(Z_1) - \frac{\eta(j_1)+\eta(j_2)}{2}$
$\eta^*(Z_2)$	η direction of the Z_2 relative to the leading jets: $\eta^*(Z_2) = \eta(Z_2) - \frac{\eta(j_1)+\eta(j_2)}{2}$
$R_{p_T}^{hard}$	module of the transverse component of the vector sum of the two leading jets and four leptons in the event normalized to the scalar p_T sum of the same objects $R_{p_T}^{hard} = \frac{\lvert(\sum_{i=4l,\ 2j} \vec{V}_i)_{transverse}\rvert}{\sum_{4l,\ 2j} p_T(i)}$
$R_{p_T}^{jet}$	module of the transverse component of the vector sum of the two leading jets and four leptons in the event normalized to the scalar p_T sum of the same objects $R_{p_T}^{jet} = \frac{\lvert(\sum_{i=2j} \vec{V}_i)_{transverse}\rvert}{\sum_{2j} p_T(i)}$
$p_T(Z_1)$	transverse momentum of the Z_1
$\eta(Z_1)$	pseudo-rapidity of the Z_1
$p_T(Z_2)$	transverse momentum of the Z_2
$\eta(Z_2)$	pseudo-rapidity of the Z_2
$cos\theta^*(Z_1)$	decay angle between the negatively charged lepton in the Z_1 rest frame with respect and the momentum direction of the Z_1 in the laboratory frame
$cos\theta^*(Z_2)$	decay angle between the negatively charged lepton in the Z_2 rest frame with respect and the momentum direction of the Z_2 in the laboratory frame

Table 5.9: Set of 13 variables used to separate the LL signal from the LT and TT polarizations and from the QCD backgrounds.

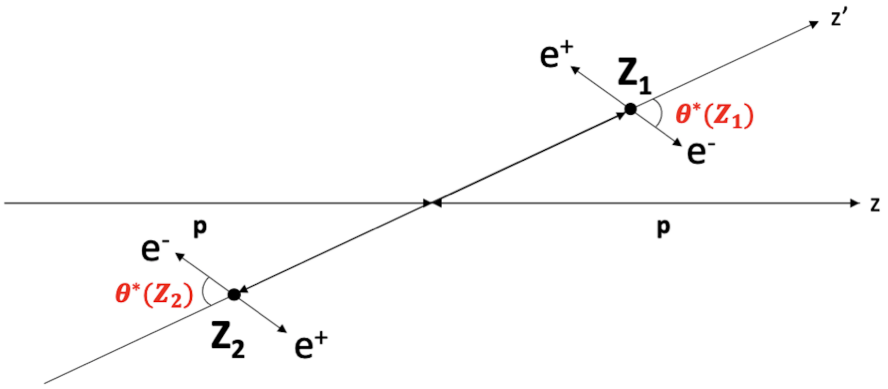


Figure 5.14: An illustration of angles $\theta^*(Z_1)$ and $\theta^*(Z_2)$ defined in Table 5.9

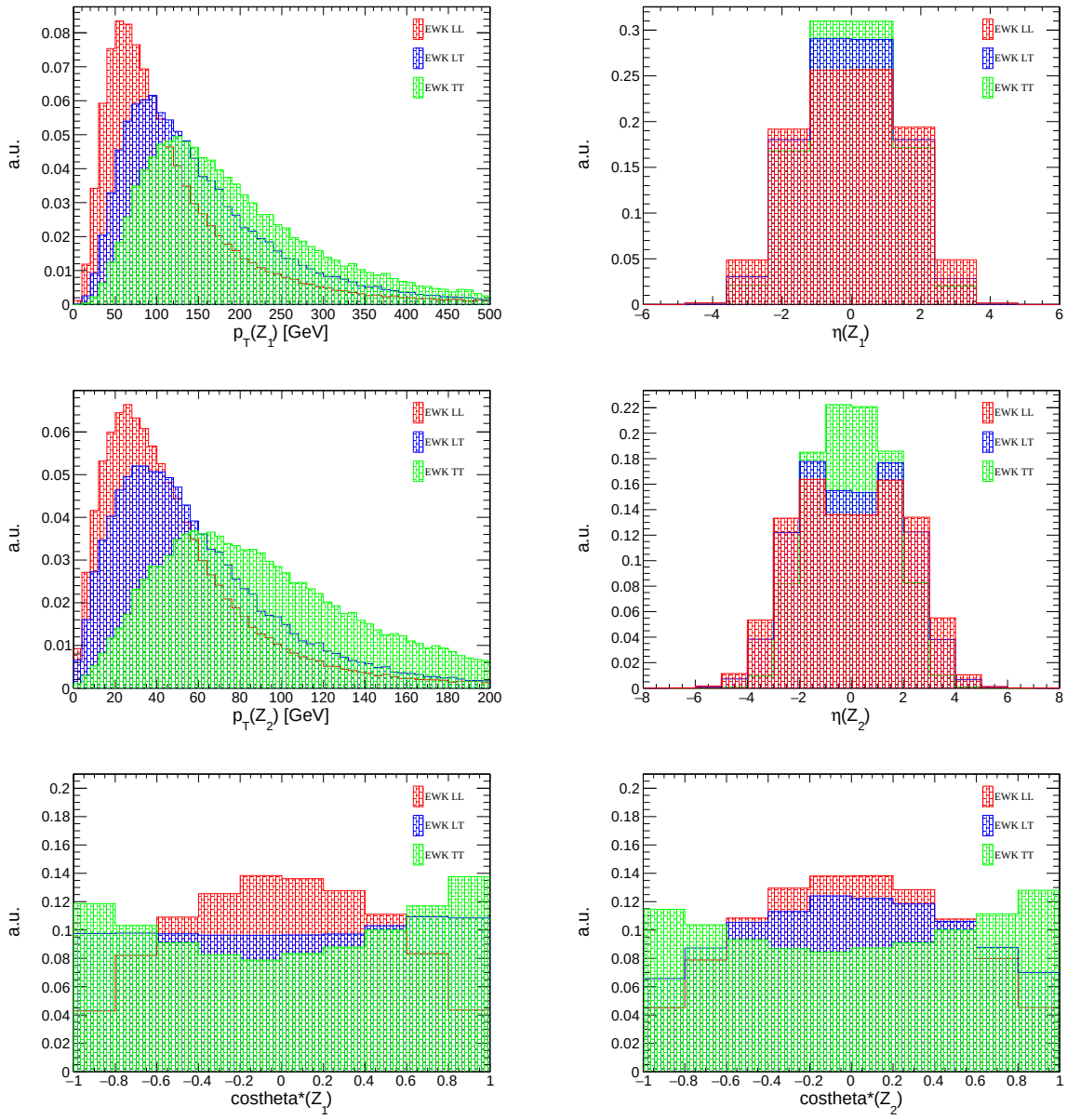


Figure 5.15: Distributions of the six variables from Table 5.9 used to extract the LL signal from LT and TT backgrounds. Plots for 14 TeV are shown, and the baseline selection was applied.

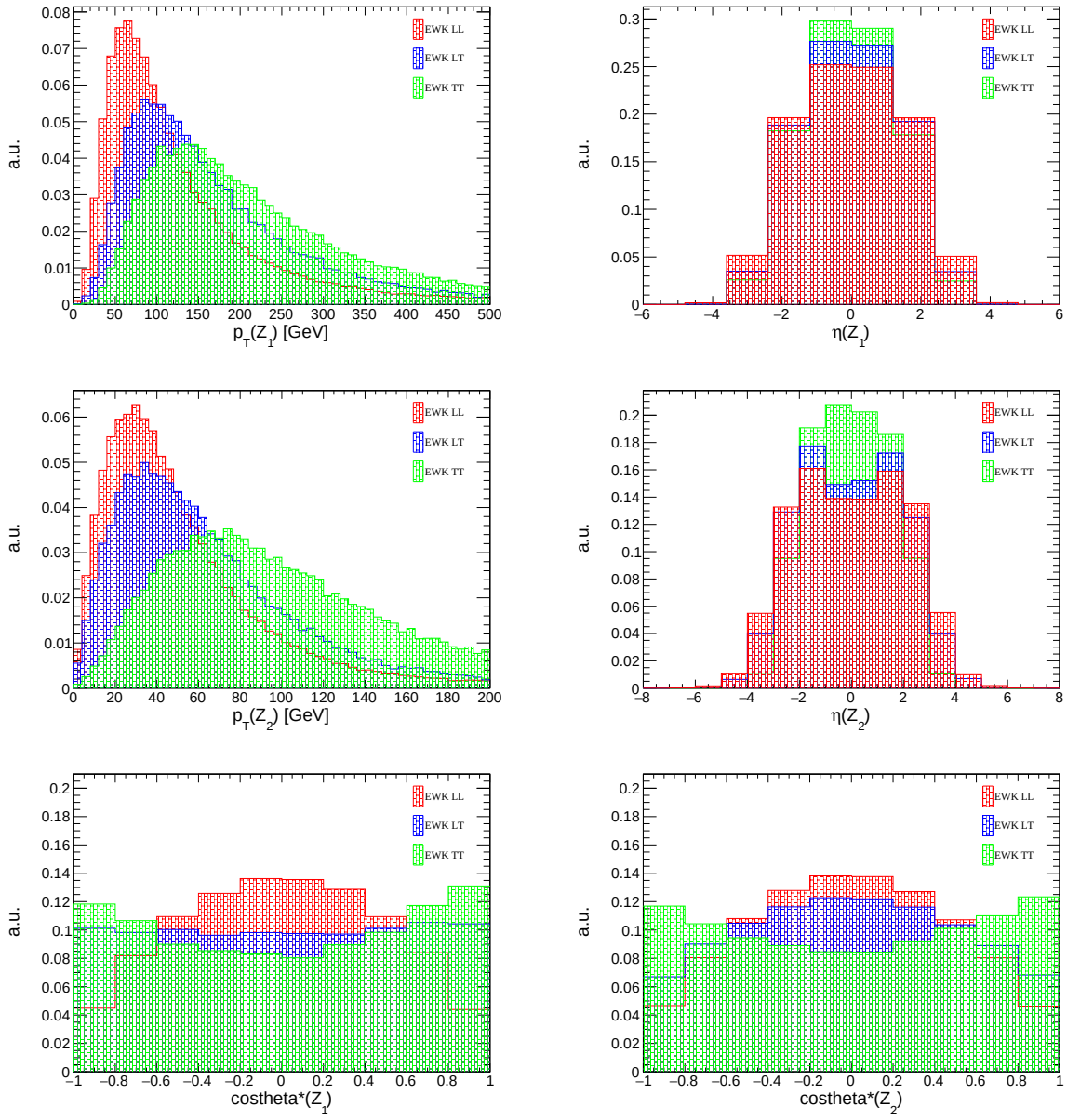


Figure 5.16: Distributions of the six variables from Table 5.9 used to extract the LL signal from LT and TT backgrounds. Plots for 27 TeV are shown, and the baseline selection was applied.

5.6 Signal extraction using a BDT and signal significance measurements

5.6.1 The combined-background BDT and the 2D BDT methods for signal extraction

Because of the low cross-section of the LL signal, it is important to devise a signal extraction method that will keep as much signal events as possible while maximally reducing the background. Unfortunately, none of the signal distributions alone is discriminating enough to accomplish such a task. For this reason, a more complex method must be used. Two such methods were studied in order to obtain the maximum signal sensitivity:

1. Combined-background BDT
   - Train the BDT classifier on the events that pass the baseline selection to discriminate the LL signal from the mixture of all backgrounds
2. 2D BDT
   - QCD BDT: Train the BDT classifier on the events that pass the baseline selection to discriminate the LL signal from the qq background.
   - VBS BDT: Train the BDT classifier on the events that pass the baseline selection to discriminate the LL signal from the mixture of LT and TT backgrounds.

Regardless of the method used, the same set of 13 variables shown in Table 5.9 was used to train the BDT. For both methods, the gradient boosting was used with the hyperparameters listed in Table 5.10.

parameter	value	parameter meaning
NTrees	1000	number of trees in the forest
MinNodeSize	2.5	minimum percentage of training events required in a leaf node
Shrinkage	0.1	learning rate
nCuts	20	number of grid points used in finding optimal cut in node splitting
maxDepth	2	maximum allowed depth of the decision tree

Table 5.10: Hyperparameters used in the BDT training for the combined-background BDT and the 2D BDT at both 14 TeV and 27 TeV.

The combined-background BDT

In the combined-background BDT method, the first step, after selecting the appropriate set of variables and hyperparameters, is to properly weight each contribution. Failing to do so would result in the suboptimal performance of the BDT which is sensitive to the shape of the input distributions. The weights used in the training are shown in Table 5.11.

	EWK LL weight	EWK LT weight	EWK TT weight	QCD qq weight
14 TeV	0.000661	0.003799	0.006376	1
27 TeV	0.000988	0.005726	0.0097774	1

Table 5.11: Weights used when filling the training and test trees in the combined-background BDT and the 2D BDT. The weights are obtained by dividing each sample's cross-section by the cross-section of the qq sample.

Although different, the kinematics of the loop-induced background is close to that of the main QCD background. In addition, the gg simulation used in this analysis is not state-of-the-art and using it in the BDT training at this stage would not gain much. For these reasons, the gg kinematics was not used in any training in this analysis. However, the result of the BDT training is always applied on the gg sample which is included in the calculation of the signal significance. Thus, in the combined-background BDT approach, the properly weighted LL signal is trained against the weighted mixture of the VBS backgrounds and the main QCD background.

For the demonstration purposes, an example of the BDT output distribution for the combined-background BDT is shown in Figure 5.17. There is no sign of overtraining. To find the WP that maximizes the significance, $S/\sqrt{B}$, the cut values on the BDT output distribution that give the signal efficiencies in the range [10 %, 70 %] are calculated. Then, for the given signal efficiency, the efficiency of each background is calculated followed by the calculation of the signal significance.

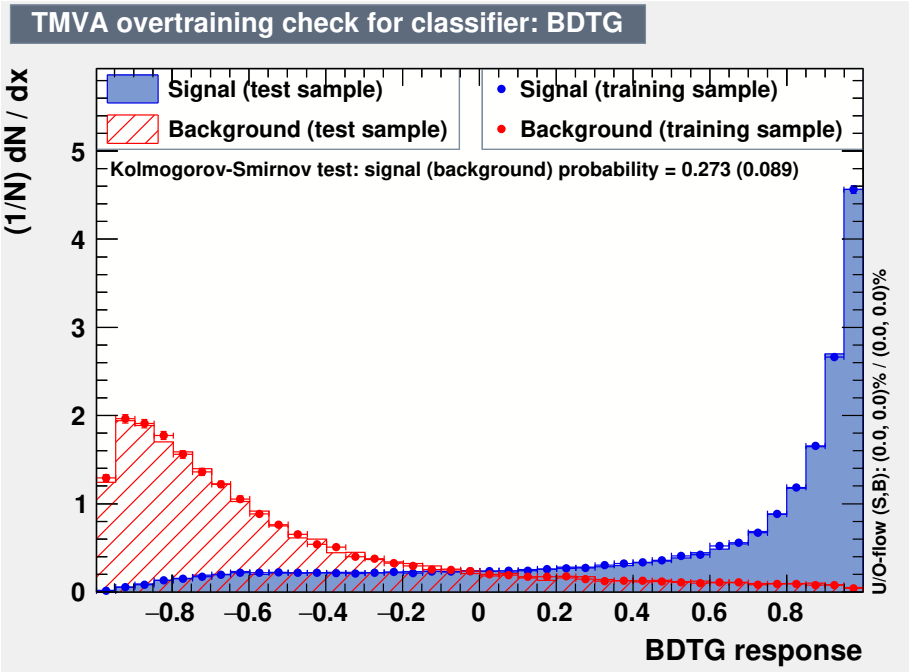


Figure 5.17: An example of the BDT output distribution of the training and test samples for the combined-background BDT approach.

The 2D BDT

The procedure of LL signal extraction with the 2D BDT method is illustrated in Fig. 5.18.
Both the QCD BDT and the VBS BDT are trained in parallel on the same set of events that passed the baseline selection (reffered to as the "original samples" in the rest of this section). The same set of weights, as used in the combined-baackground BDT, was also used here. For the demonstration purposes, an example of the BDT output

distributions obtained after training the QCD BDT and the VBS BDT is shown in Figure 5.19.

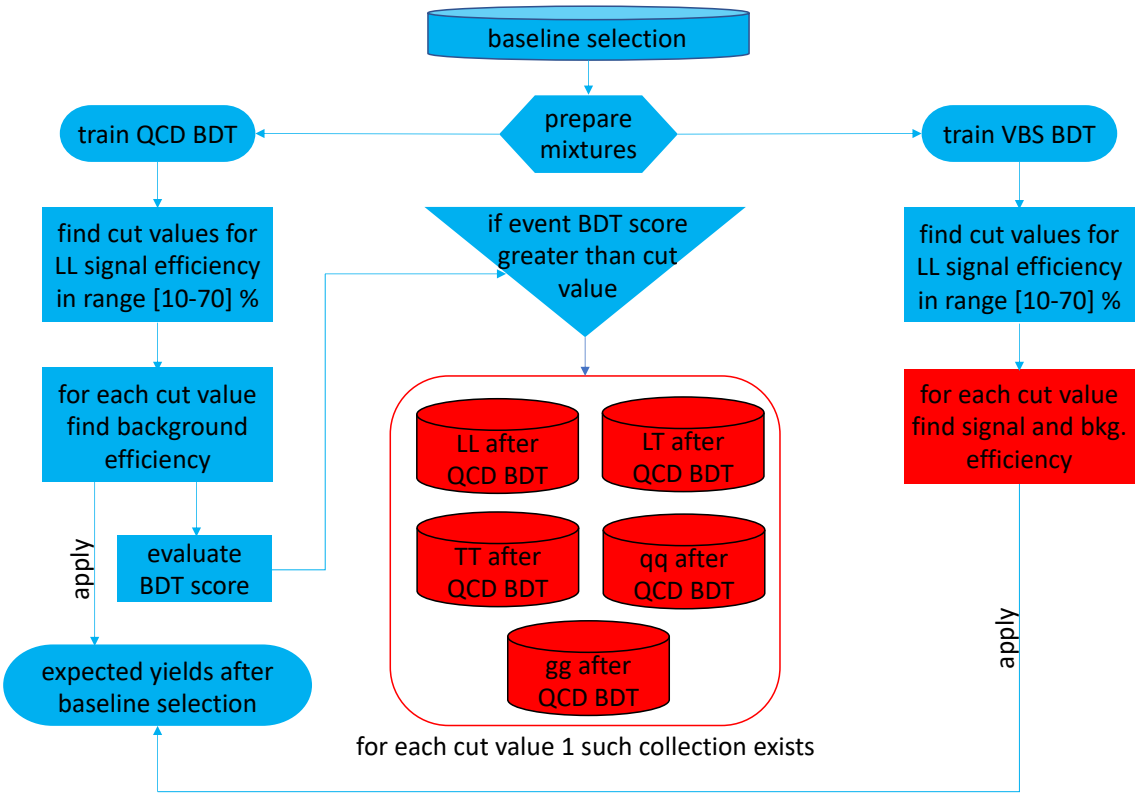


Figure 5.18: Illustration of the 2D BDT signal extraction approach. Blue color marks data sets obtained after the baseline selection, as well as any operation applied on them. Red color marks new data sets obtained after applying the QCD BDT training, as well as any operation applied on them.

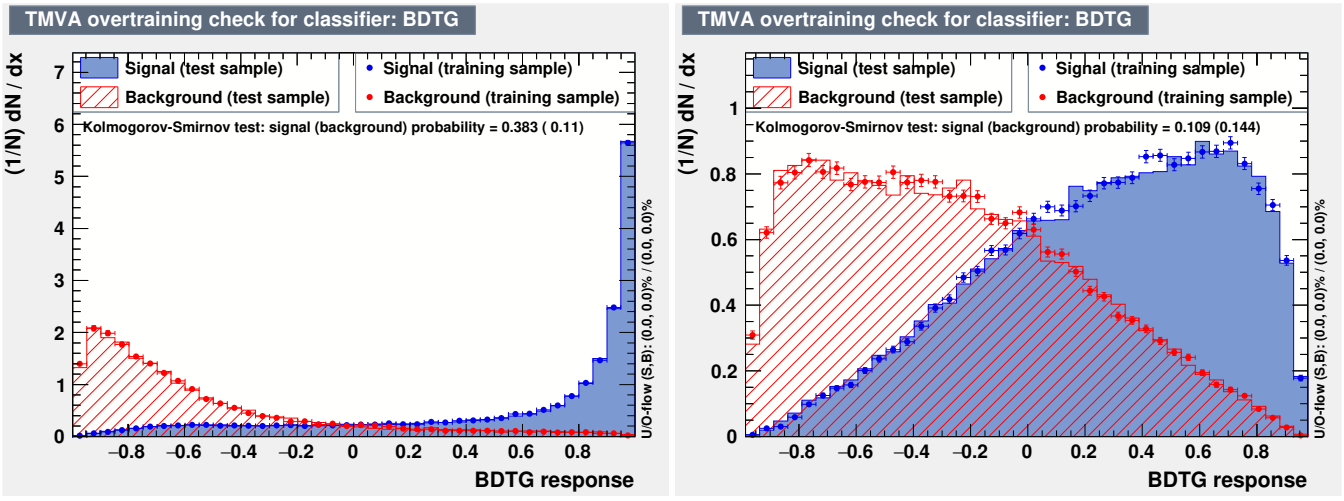


Figure 5.19: An example of the BDT output distributions of the training and test samples for the 2D BDT approach. The left-hand side plot top row shows the result of the QCD BDT training. The right-hand side plot shows the result of the VBS BDT training.

The cut value on the QCD BDT output distribution is chosen so that 10% signal efficiency is obtained. The training of the QCD BDT is now applied on the original samples from which one calculates the efficiency of each background contribution. The efficiencies of the signal and each of the backgrounds after the QCD BDT are used later.
At the same time, the BDT score is calculated for each event in the original samples. If the BDT score is greater than the cut value, an event is stored in a new sample (red discs in the illustration).

In parallel, the cut values on the VBS BDT output distribution are chosen so that signal efficiencies in range [10 %, 70 %] are obtained. For the sake of clarity, let's consider only a single cut value. This cut value is used to calculate the signal and background efficiencies, after the VBS BDT, in the new samples (red discs in the illustration).

To obtain the expected yields after the QCD BDT and the VBS BDT training, one applies both QCD BDT and VBS BDT efficiencies to the expected yields after the baseline selection. Now one can calculate the signal significance as $\frac{S}{\sqrt{B}}$. This procedure is repeated for other cut values in the VBS BDT output distribution

Finally, one will now choose several other cut values on the QCD BDT output distribution and repeat everything. This results in an array of VBS BDT signal and background efficiencies for each cut value in the QCD BDT output distribution, hence the name 2D BDT.
Because of the large cross-section of the QCD background, the WP for the QCD BDT must be chosen such that it heavily suppresses the QCD contribution. The VBS QCD will further reduce the QCD contribution, but at the expense of also reducing the LL signal.

5.6.2 Signal extraction and significance measurements at 14 TeV

Table 5.12 shows the number of generated events for all VBS and QCD process included in the analysis. For all contributions, the unweighted number of generated events is reported. In addition, for the VBS processes and the qq background, number of events, weighted by the process cross-section, is quoted as well. The events in the gg production are not weighted by the cross-section but by unity. Very few events, less than 0.3 %, have weight larger than 1. This is due to the setup of the computation grids where a balance between the precision and the time consumption was required. Such events have been rejected in the analysis and thus the unweighted number of events is slightly larger than the weighted number of events.

The expected number of events for the VBS and QCD qq contributions at luminosity L, expressed in inverse femtobarns, was calculated using the formula below:

$$N_{expected}^{L[fb^{-1}]} = \frac{N_{weighted}^{selection}}{N_{unweighted}^{generated}} \cdot 1000 \cdot L$$

The expected number of events for the VBS and QCD gg contributions at the luminosity L, expressed in inverse femtobarns, was calculated using the formula below:

$$N_{expected}^{L[fb^{-1}]} = 2 \cdot \sigma_{gen} \cdot \frac{N_{weighted}^{selection}}{N_{weighted}^{generated}} \cdot L \ ,$$

where the σ_{gen} is the cross-section of the generated sample expressed in femtobarns. The factor 2 amortizes the fact that only $2e2\mu$ final state was simulated for the gg contribution, thus neglecting the $4e$ and 4μ final states and therefore also neglecting half of the available phase space.

	VBS LL	VBS LT	VBS TT	QCD qq	QCD gg
unweighted events	227731	94345	100000	502500	109731
weighted events	7.4	17.7	31.6	24942	109729
ZZ selection	3.8	9.5	18.7	11199	46189
baseline selection	3.3	8.5	16.6	2440	15080
VBS selection	2.3	5.5	10.3	353	3798
	expected yields at 14 TeV c.o.m. energy				
baseline selection	43	269	499	14569	2941
VBS selection	30	175	310	2106	741

Table 5.12: Top: unweighted and weighted number of generated events. For VBS and QCD qq processes events are weighted by the process cross-section. Middle: weighted number of events after the selection. Bottom: expected number of events at HL-LHC and for $3000\ fb^{-1}$.

Combined-background BDT

Distributions for the LL signal and the combined background, normalized to unit area, are shown in Figure 5.20. The BDT output distributions for the training and test samples for the combined-background BDT are shown on Figure 5.21. The BDT output distribution shows no signs of overtraining.

The top part of Table 5.13 shows the expected yields for the LL signal and all backgrounds after the baseline selection at 14 TeV for $3000\ fb^{-1}$. The bottom part shows the cut value chosen from the BDT output distribution together with the corresponding signal efficiency. For each signal efficiency, the efficiency of all contributions is reported. Table 5.14 shows expected yields corresponding to the efficiencies shown in Table 5.13 together with the signal significance for each WP.

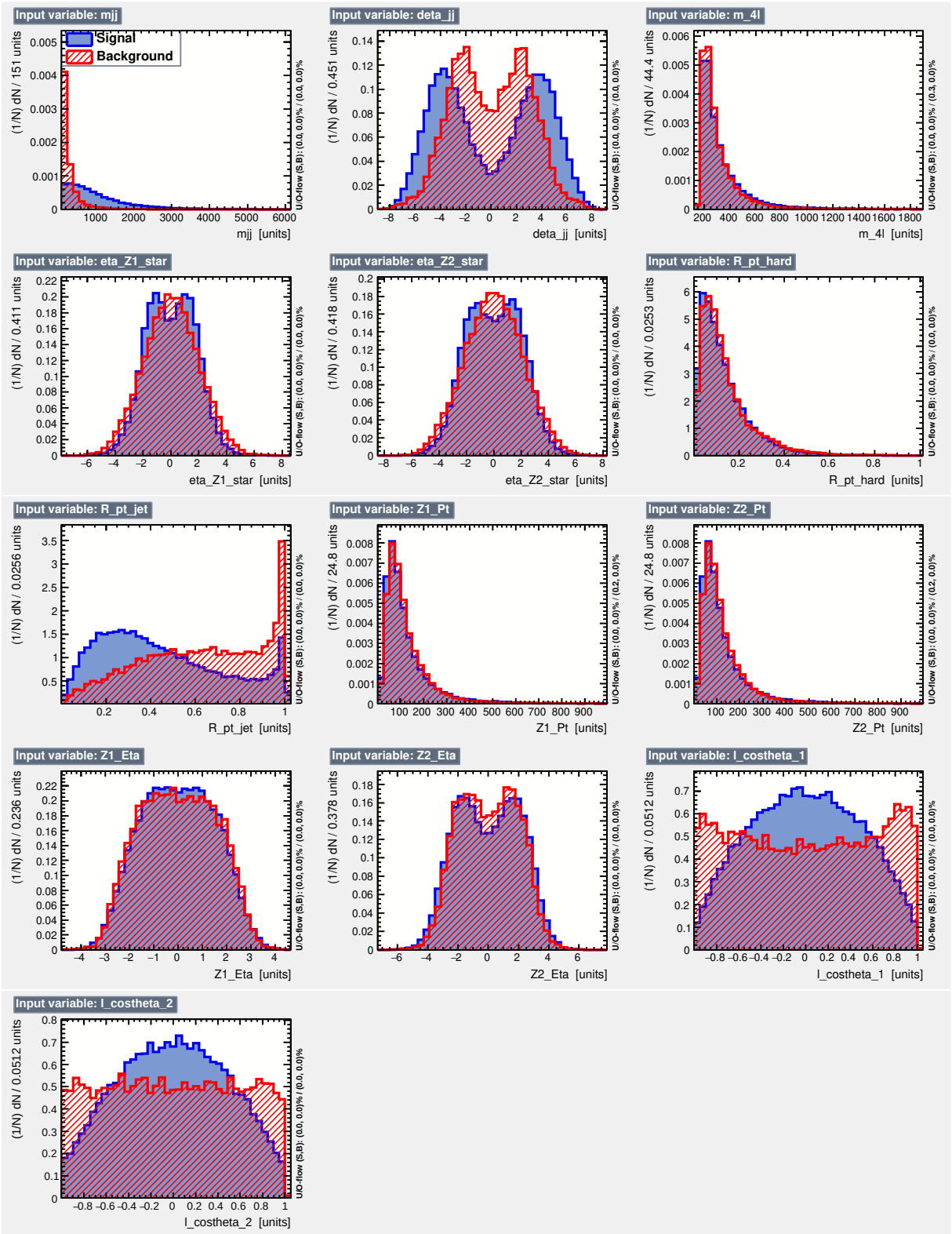


Figure 5.20: Input variables for the combined-background BDT training at 14 TeV. The LL signal is shown in blue and the mixture of backgrounds in red. [to be updated with nicer plot]

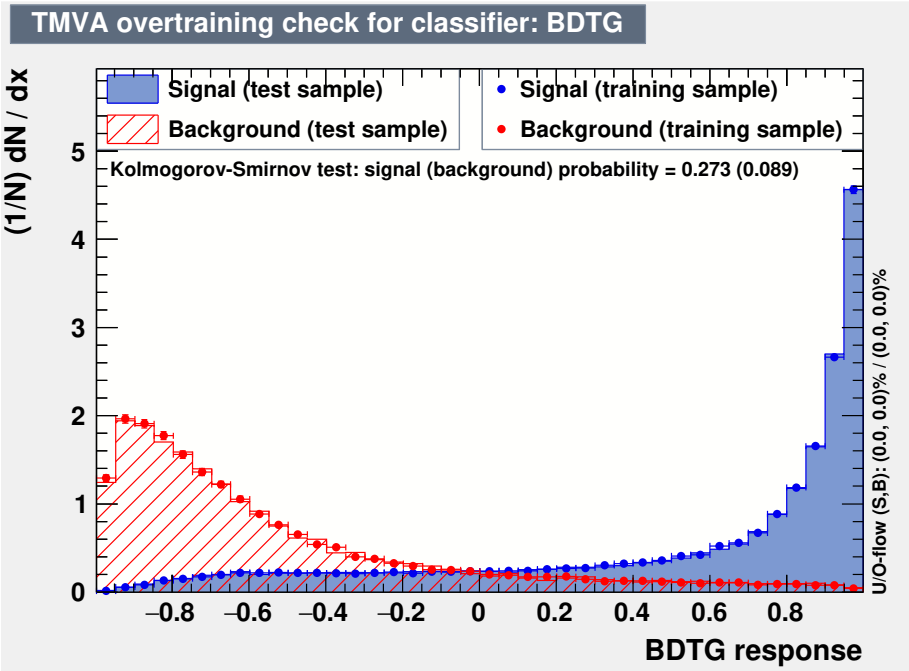


Figure 5.21: The BDT output distributions of the LL signal (in blue) and the mixture of backgrounds (in red) for the combined-background BDT training at 14 TeV.

	LL	LT	TT	qq	gg
expected yields after the baseline selection	43	269	499	15569	2941
signal efficiency [%]	**LL [%]**	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
45 (0.845)	45.0	33.5	22.7	0.70	3.30
40 (0.879)	40.0	28.5	18.3	0.50	2.30
35 (0.905)	35.0	23.9	14.5	0.30	1.70
30 (0.925)	30.0	19.4	11.1	0.20	1.20
20 (0.957)	20.0	11.3	5.60	0.06	0.50
15 (0. 969)	15.0	7.70	3.40	0.03	0.20

Table 5.13: Top: expected yields for the LL signal and all backgrounds after the baseline selection. Bottom: signal efficiencies and corresponding efficiencies for all contribution. Cut values corresponding to the signal efficiencies are shown in the parentheses. Results correspond to the combined-background BDT training at 14 TeV and for $3000\ fb^{-1}$.

	Number of events					
signal efficiency [%]	**LL**	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
45 (0.845)	19.4	90.0	113	108	97.1	0.96
40 (0.879)	17.2	76.6	91.5	81.2	69.0	0.96
35 (0.905)	15.0	64.3	72.6	48.2	49.5	0.98
30 (0.925)	12.9	52.1	55.6	29.4	35.5	0.98
20 (0.957)	8.60	30.4	28.0	9.20	13.8	0.95
15 (0.969)	6.40	20.7	16.7	5.00	6.00	0.93

Table 5.14: Expected yields for all contributions corresponding to efficiencies reported in Table 5.13. Cut values corresponding to the signal efficiencies are shown in the parentheses. Results correspond to the combined-background BDT training at 14 TeV and for $3000\ fb^{-1}$.

2D BDT

Figures 5.22 and 5.23 show the input variables for the QCD BDT and the VBS BDT. The QCD BDT output distribution for the training and test samples is shown in the top row of Figure 5.24. The bottom row shows the same distributions for the VBS BDT training. No overtraining is observed for either case.

Table 5.15 shows the efficiencies of all contributions after the VBS BDT training for the fixed QCD BDT signal efficiency of 20%. Table 5.16 shows the expected yields corresponding to the efficiencies quoted in Table 5.15 together with the signal significance for each WP. Scanning of the 2D BDT significance space was performed to find the optimal working points for both QCD BDT and VBS BDT training. This is shown in Figure 5.25.

A detailed discussion on the performance of the 2D BDT compared to the combined-background BDT at 14 TeV and for $3000\ fb^{-1}$ is presented in section 5.7.

[There was a discussion to comment on the differences between the combined-background BDT and the 2D BDT after I present tables. However, I decided not to do it. This goes with Chapter 4 where I decided not to discuss the MELA results since it would break the symmetry of the entire document. For MELA results we concluded that it is acceptable to, instead, tell the reader that results will be discussed in section XXX. In this way I don't break the symmetry of the document, but, at the same time, state that I didn't just leave the tables without any discussion. I will do that also for the 27 TeV case. The results are anyways right in the next section.]

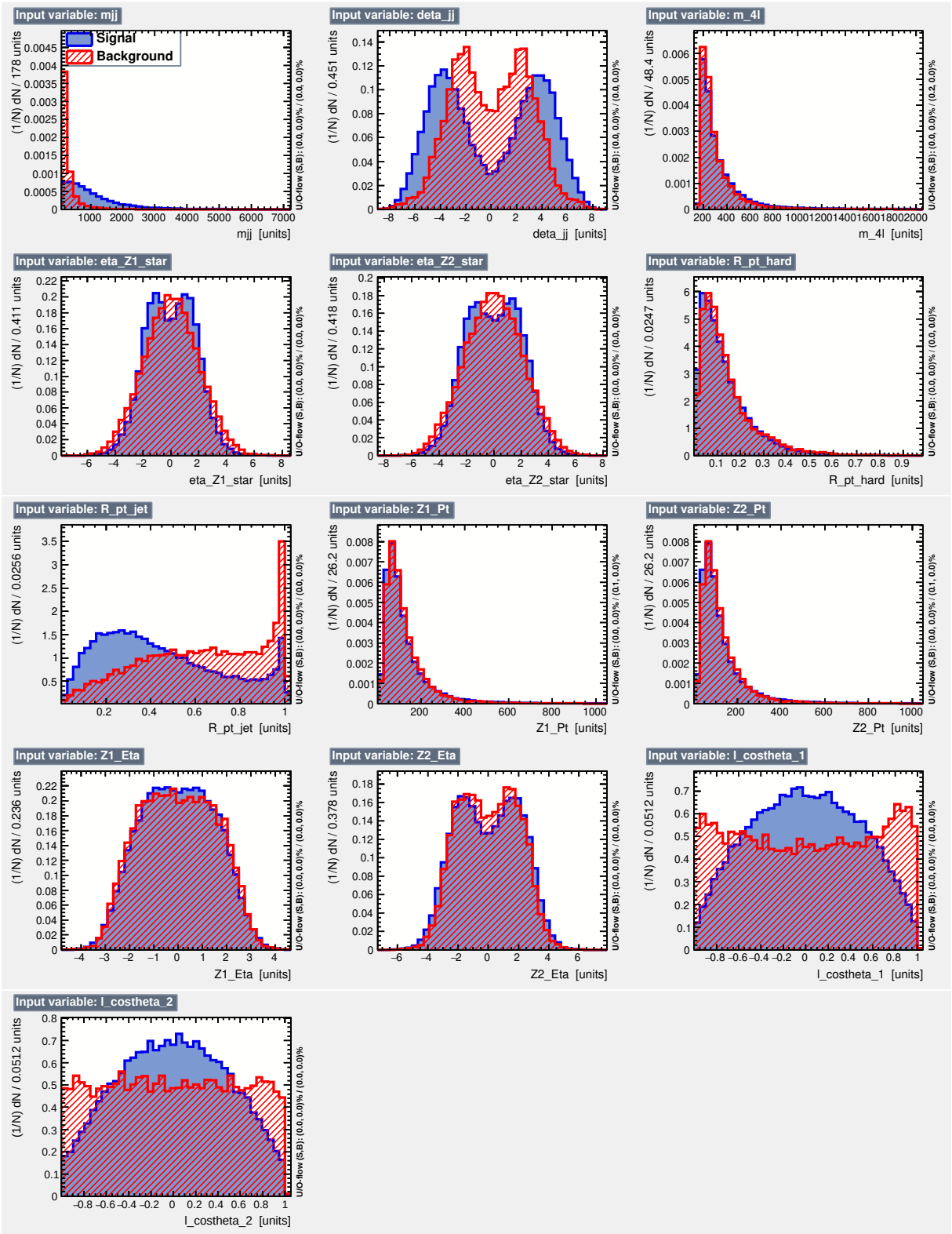


Figure 5.22: Input variables for the QCD BDT training at 14 TeV. The LL signal is shown in in blue and the qq background is shown in red. [to be updated with nicer plot]

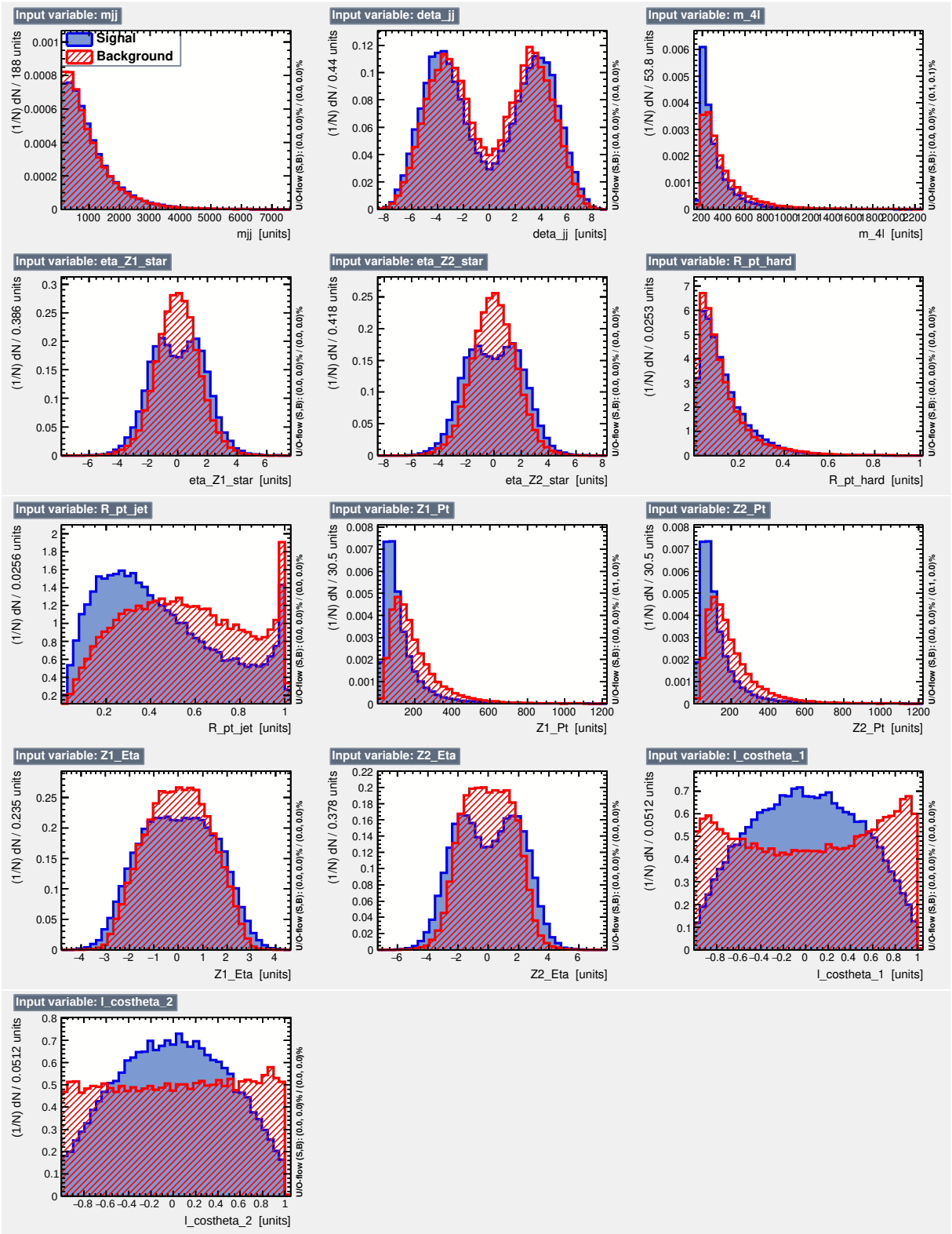


Figure 5.23: Input variables for the VBS BDT training at 14 TeV. The LL signal is shown in in blue and the mixture of the LT and TT backgrounds is shown in red. [to be updated with nicer plot]

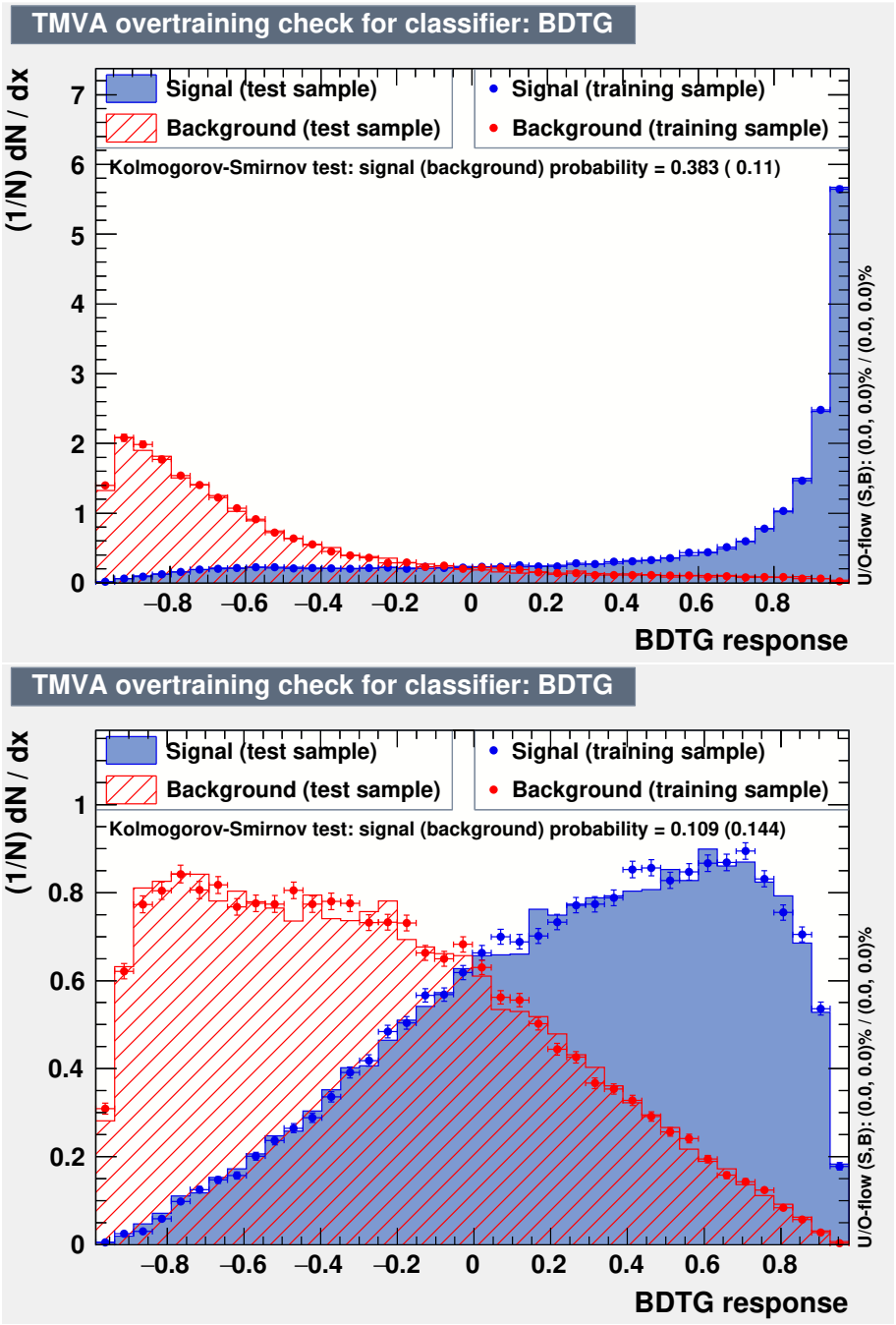


Figure 5.24: Top: the QCD BDT output distribution of the LL signal (in blue) and the qq background (in red) for the 2D BDT training at 14 TeV. Bottom: the VBS BDT output distribution of the LL signal (in blue) and the mixture of the LT and TT backgrounds (in red).

	LL [%]	LT [%]	TT [%]	qq [%]	gg [%]
efficiencies after the QCD BDT ($\epsilon_{sig} = 20\ \%$)	20	13.3	8.03	0.0766	0.59
VBS BDT signal efficiencies [%]	**efficiencies after VBS BDT for QCD BDT $\epsilon_{signal} = 20\ \%$**				
	LL [%]	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
50 % (0.314)	72.7	38.2	15.0	34.6	23.6
45 % (0.377)	68.1	33.3	12.2	23.1	20.2
40 % (0.438)	63.4	29.1	9.25	19.2	12.4
35 % (0.498)	58.1	24.8	7.07	11.5	11.2
30 % (0.558)	52.6	20.8	5.25	7.69	10.1
20 % (0.673)	39.9	12.8	3.08	7.69	5.62

Table 5.15: Top: efficiencies of the LL signal and all backgrounds after the QCD BDT. The QCD BDT signal efficiency is fixed to 20 %. Bottom: signal efficiencies and corresponding background efficiencies after the VBS BDT for the 20 % QCD BDT signal efficiency. Several signal efficiencies, corresponding to the working points in the bottom-left plot in Figure 5.24, were scanned to find the maximum signal significance. Cut values corresponding to the signal efficiencies are shown in the parentheses. Results are shown for the 2D BDT training at 14 TeV and for $3000\ fb^{-1}$.

	LL	LT	TT	qq	gg	$S/\sqrt{B}$
expected yields after the QCD BDT ($\epsilon_{sig} = 20\ \%$)	8.6	35.8	40.1	11.9	17.4	0.84
VBS BDT signal efficiencies [%]	**expected yields after 2D BDT for QCD BDT $\epsilon_{signal} = 20\ \%$**					
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
50 % (0.314)	6.30	13.7	6.00	4.10	4.10	1.18
45 % (0.377)	5.90	11.9	4.90	2.80	3.50	1.22
40 % (0.438)	5.50	10.4	3.70	2.30	2.10	1.27
35 % (0.498)	5.00	8.90	2.80	1.40	2.00	1.29
30 % (0.558)	4.50	7.40	2.10	0.90	1.80	1.29
20 % (0.673)	3.40	4.60	1.20	0.90	1.00	1.24

Table 5.16: Expected yields for all contributions corresponding to efficiencies reported in Table 5.15. Cut values corresponding to the signal efficiencies are shown in the parentheses. Results are shown for the 2D BDT training at 14 TeV and for $3000\ fb^{-1}$.

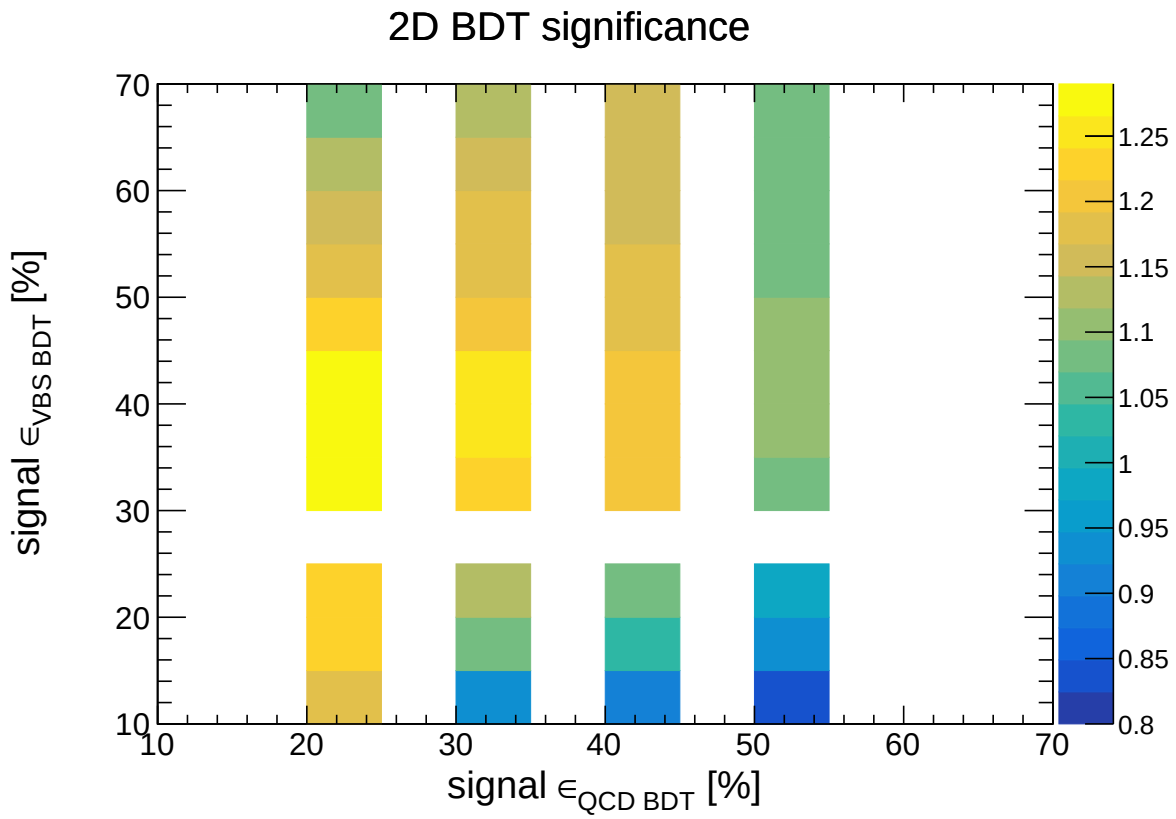


Figure 5.25: The 2D BDT significance plane used to scan for the optimal WP of the QCD BDT and the VBS BDT. Results at 14 TeV and for $3000\ fb^{-1}$ are shown. [plot will be updated to show all bins]

5.6.3 Signal extraction and significance measurements at 27 TeV

Table 5.17 shows the number of generated events for all VBS and QCD process included in the analysis. Number of weighted events after the selection together with the expected yields at 27 TeV is also reported.

	VBS LL	**VBS LT**	**VBS TT**	**QCD qq**	**QCD gg**
unweighted events	227731	94345	100000	502500	109731
weighted events	7.45	17.7	31.6	24941	109729
ZZ selection	3.83	9.55	18.7	11199	46189
baseline selection	3.30	8.46	16.6	2440	15080
VBS selection	2.25	5.51	10.3	353	3798
	expected yields at 14 TeV c.o.m. energy				
baseline selection	43	269	499	14569	2941
VBS selection	30	175	310	2106	741

Table 5.17: Top: unweighted and weighted number of generated events. For the VBS and QCD qq processes events are weighted by the process cross-section. Middle: weighted number of events after the selection. Bottom: expected number of events at HE-LHC and for $15000\ fb^{-1}$.

Combined-background BDT

Distributions for the LL signal and the combined background, normalized to unit area, are shown in Figure 5.26. The BDT output distributions for the training and test samples for the combined-background BDT are shown in Figure 5.27. The BDT output distribution shows no signs of overtraining.

The top part of Table 5.18 shows the expected yields for the LL signal and all backgrounds after the baseline selection at 27 TeV for $15000\ fb^{-1}$. The bottom part shows the cut value chosen from the BDT output distribution together with the corresponding signal efficiency. For each signal efficiency, the efficiency of all contributions is reported. Table 5.19 shows expected yields corresponding to the efficiencies shown in Table 5.18 together with the signal significance for each WP.

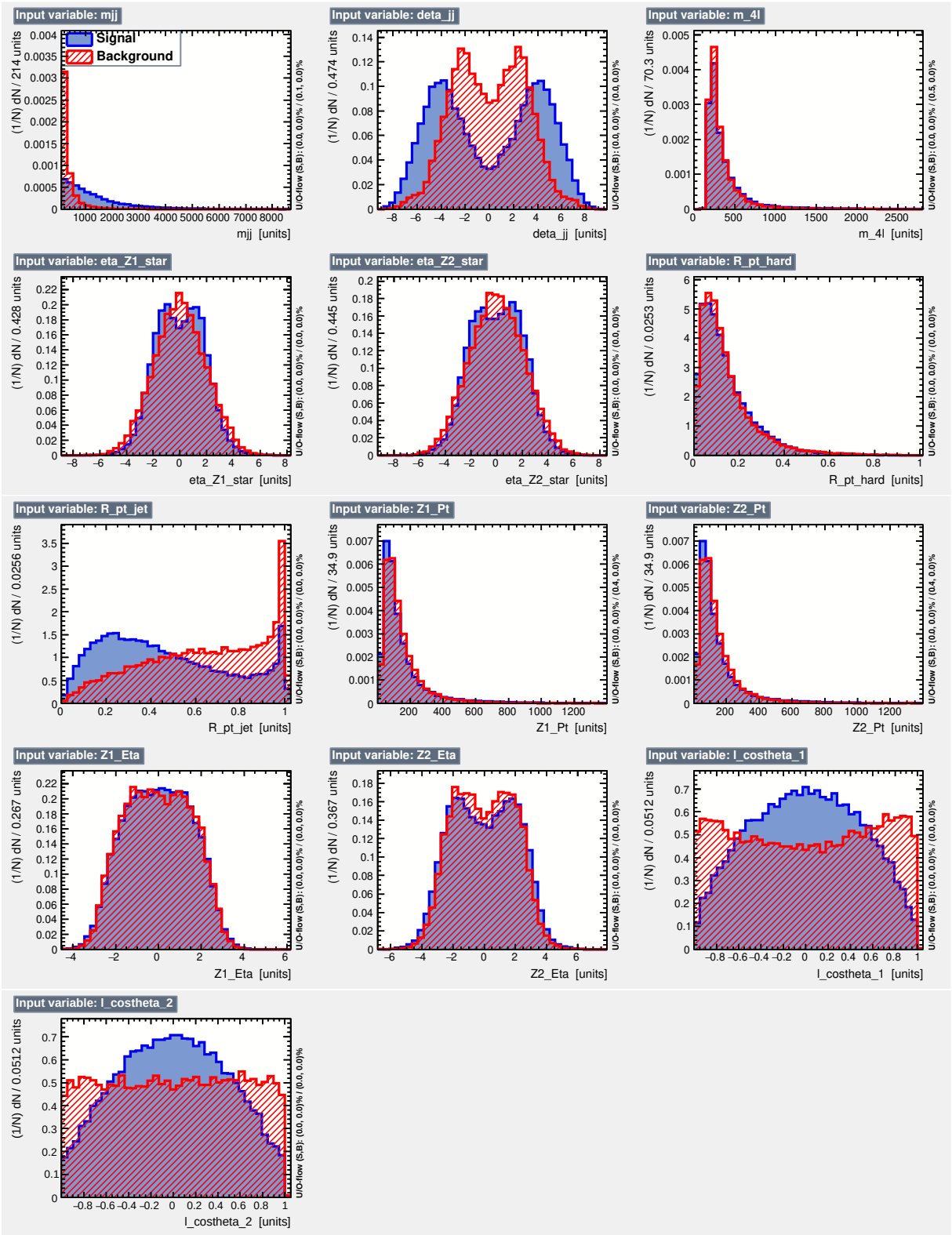


Figure 5.26: Input variables for the combined-background BDT training at 27 TeV. The LL signal is shown in blue and the mixutre of backgrounds in red. [to be updated with nicer plot]

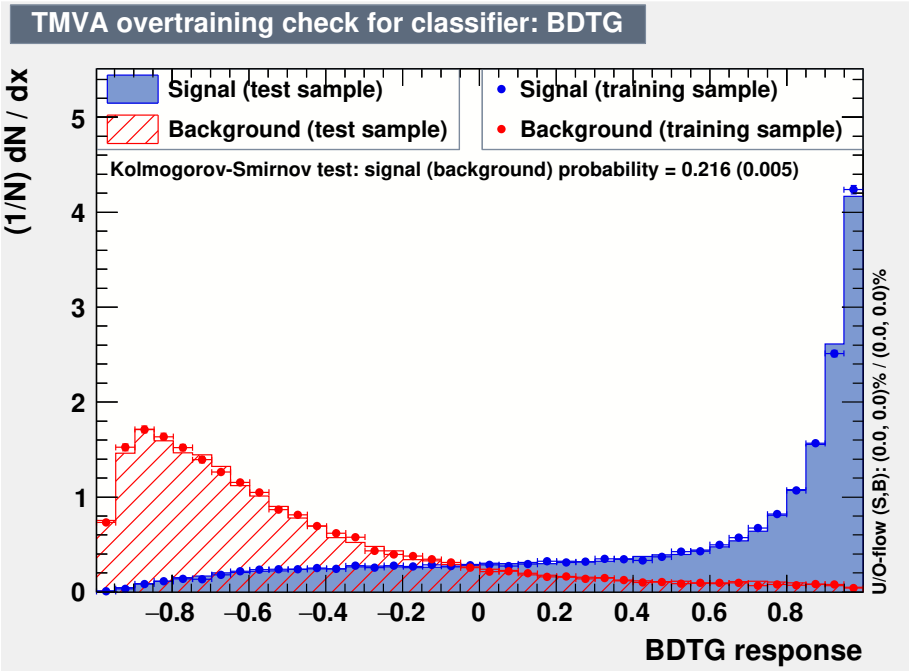


Figure 5.27: The BDT output distributions of the LL signal (in blue) and the mixture of backgrounds (in red) for the combined-background BDT training at 27 TeV.

	LL	LT	TT	qq	gg
expected yields after the baseline selection	664	4241	8178	187731	54503
signal efficiency [%]	**LL [%]**	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
45 (0.819)	45.0	36.6	27.6	0.86	3.43
40 (0.861)	40.0	31.4	22.7	0.59	2.43
35 (0.893)	35.0	26.3	18.4	0.38	1.76
30 (0.917)	30.0	21.5	14.4	0.23	1.19
20 (0.952)	20.0	13.1	7.67	0.09	0.52
15 (0.965)	15.0	9.32	5.03	0.04	0.27

Table 5.18: The: expected yields for the LL signal and all backgrounds after the baseline selection. Bottom: signal efficiencies and corresponding efficiencies for all contributions. Cut values corresponding to the signal efficiencies are shown in the parentheses. Results are shown for the combined-background BDT training at 27 TeV and for $15000\ fb^{-1}$.

	Number of events					
signal efficiency [%]	**LL**	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
45 (0.819)	299	1553	2256	1615	1868	3.50
40 (0.861)	266	1332	1857	1115	1322	3.54
35 (0.893)	232	1116	1505	705	961	3.55
30 (0.917)	199	914	1174	440	648	3.53
20 (0.952)	133	557	627	172	284	3.28
15 (0.965)	99.6	395	411	70.9	146	3.11

Table 5.19: Expected yields for all contributions corresponding to efficiencies reported in Table 5.18. Cut values corresponding to the signal efficiencies are shown in the parentheses. Results are shown for the combined-background BDT training at 27 TeV and for $15000\ fb^{-1}$.

2D BDT

Figures 5.28 and 5.29 show the input variables for the QCD BDT and the VBS BDT. The QCD BDT output distribution for the training and test samples is shown in the top row of Figure 5.30. The bottom row shows the same distributions for the VBS BDT training. No overtraining is observed for either case.

Table 5.20 shows the efficiencies of all contributions after the VBS BDT training for the fixed QCD BDT signal efficiency of 40%. Table 5.21 shows the expected yields corresponding to the efficiencies quoted in Table 5.20 together with the signal significance for each WP. Once more, scanning of the 2D BDT significance space was performed to find the optimal working points for both QCD BDT and VBS BDT training. This is shown in Figure 5.31.

A detailed discussion on the performance of the 2D BDT compared to the combined-background BDT at 27 TeV and for $15000\ fb^{-1}$ is presented in the next section.

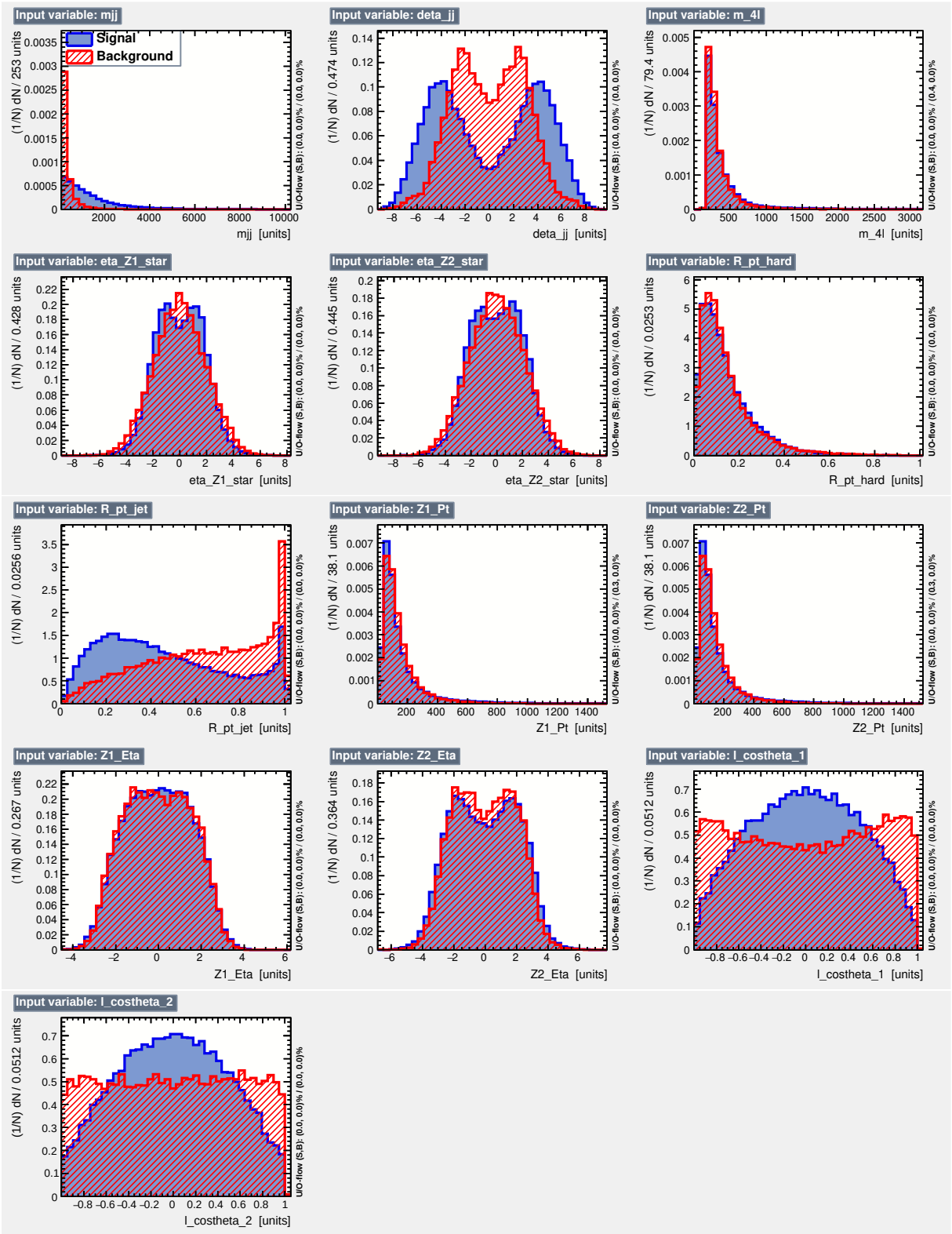


Figure 5.28: Input variables for the QCD BDT training at 27 TeV. The LL signal is shown in blue and the qq background is shown in red. [to be updated with nicer plot]

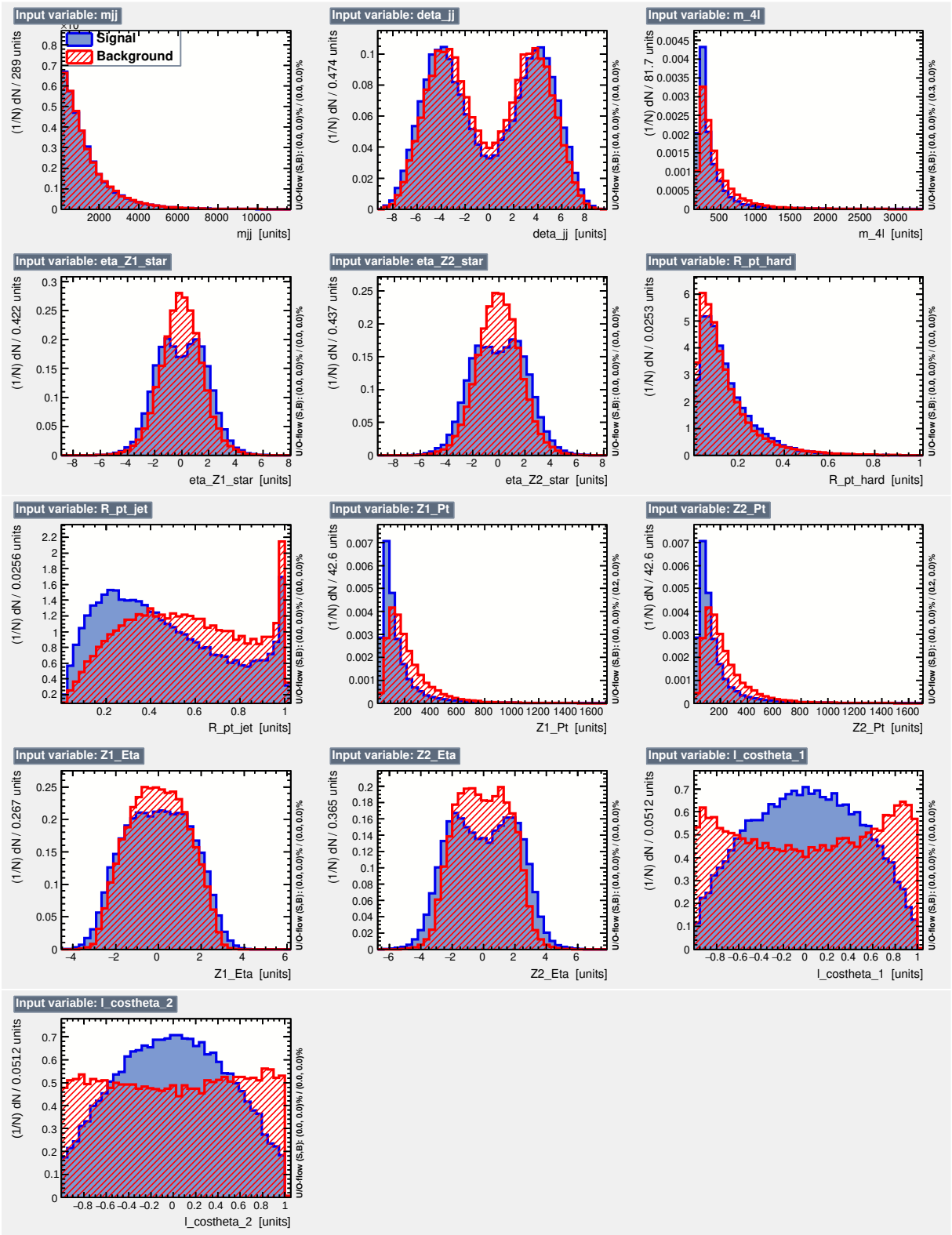


Figure 5.29: Input variables for the VBS BDT training at 27 TeV. The LL signal is shown in blue and the mixture of the LT and TT backgrounds is shown in red. [to be updated with nicer plot]

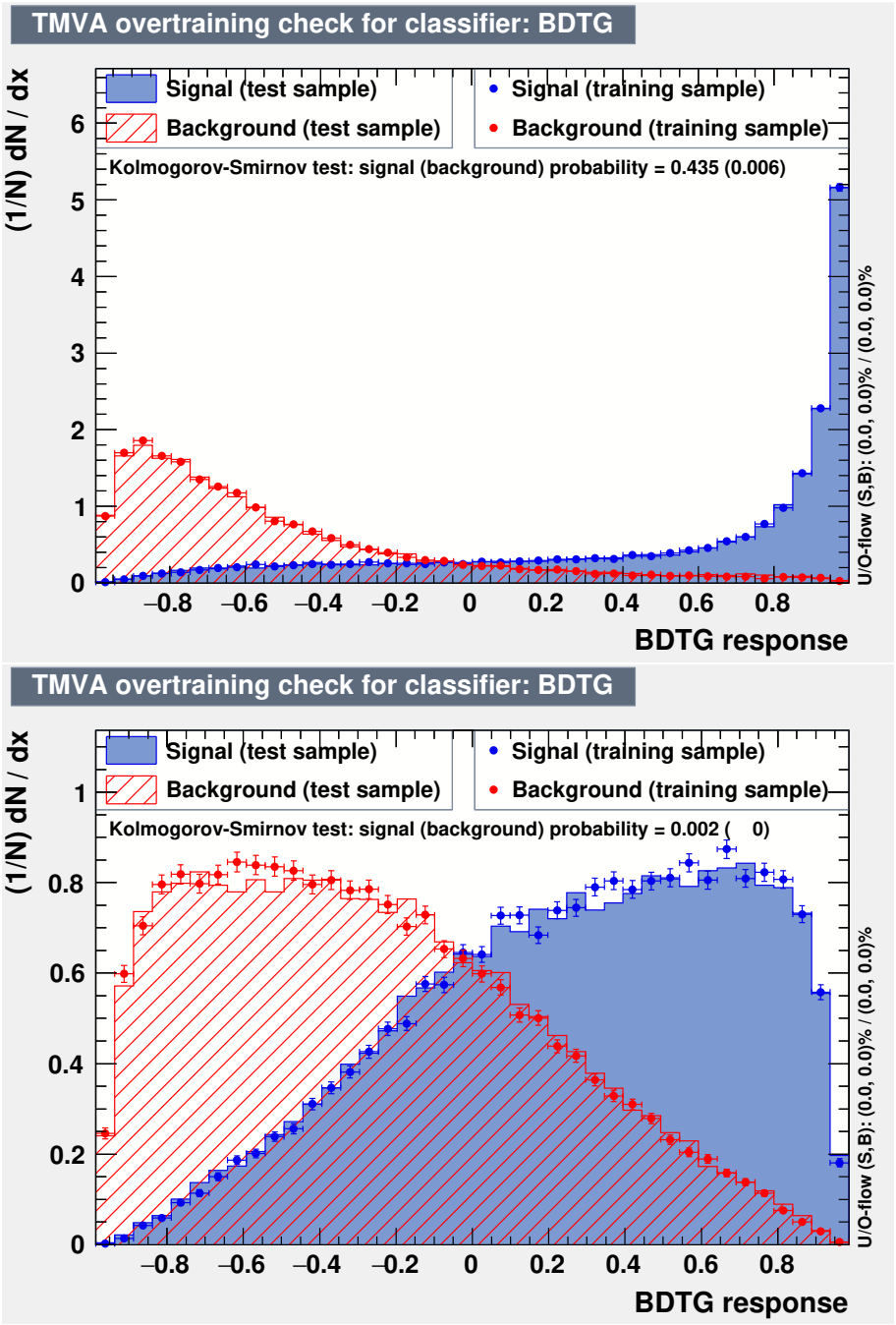


Figure 5.30: Top: the QCD BDT output distribution of the LL signal (in blue) and the qq background (in red) for the 2D BDT training at 27 TeV. Bottom: the VBS BDT output distribution of the LL signal (in blue) and the mixture of the LT and TT backgrounds (in red).

	LL [%]	LT [%]	TT [%]	qq [%]	gg [%]
efficiencies after the QCD BDT ($\epsilon_{sig} = 40\ \%$)	40	33.1521	25.8272	0.560157	2.74
VBS BDT signal efficiencies [%]	**efficiencies after VBS BDT for QCD BDT** $\epsilon_{signal} = 40\ \%$				
	LL [%]	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
55 % (0.243732)	68.0	35.3	13.8	46.8	25.4
50 % (0.308986)	63.1	30.5	10.9	41.5	19.2
45 % (0.374275)	58.1	25.8	8.58	34.0	14.4
40 % (0.438395)	52.9	21.2	6.79	30.1	11.4
35 % (0.499922)	47.3	17.2	5.05	25.5	8.40
30 % (0.561874)	41.6	13.7	3.75	21.3	6.20

Table 5.20: Top: efficiencies of the LL signal and all backgrounds after the QCD BDT. The QCD BDT signal efficiency is fixed to 40 %. Bottom: signal efficiencies and corresponding background efficiencies after the VBS BDT for the 40 % QCD BDT signal efficiency. Several signal efficiencies, corresponding to the working points in the bottom-left plot in Figure 5.30, were scanned to find the maximum signal significance. Cut values corresponding to the signal efficiencies are shown in the parentheses. Results are obtained at 27 TeV and for $15000\ fb^{-1}$.

	LL	LT	TT	qq	gg	$S/\sqrt{B}$
expected yields after the QCD BDT ($\epsilon_{sig} = 40\ \%$)	266	1406	2112	1052	1492	3.41
VBS BDT signal efficiencies [%]	**expected yields after 2D BDT for QCD BDT** $\epsilon_{signal} = 40\ \%$					
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
55 % (0.243732)	180.5	496.4	292.0	492.2	379.0	4.43
50 % (0.308986)	167.5	428.8	230.2	436.3	286.5	4.51
45 % (0.374275)	154.4	362.4	181.1	358.0	214.9	4.62
40 % (0.438395)	140.5	298.5	143.3	317.0	170.1	4.61
35 % (0.499922)	125.7	241.5	106.6	268.5	125.3	4.62
30 % (0.561874)	110.5	192.3	79.20	223.7	92.50	4.56

Table 5.21: Expected yields for all contributions corresponding to efficiencies quoted in Table 5.20. Cut values corresponding to the signal efficiencies are shown in the parentheses. Results are shown for the 2D BDT training at 27 TeV and for $15000\ fb^{-1}$.

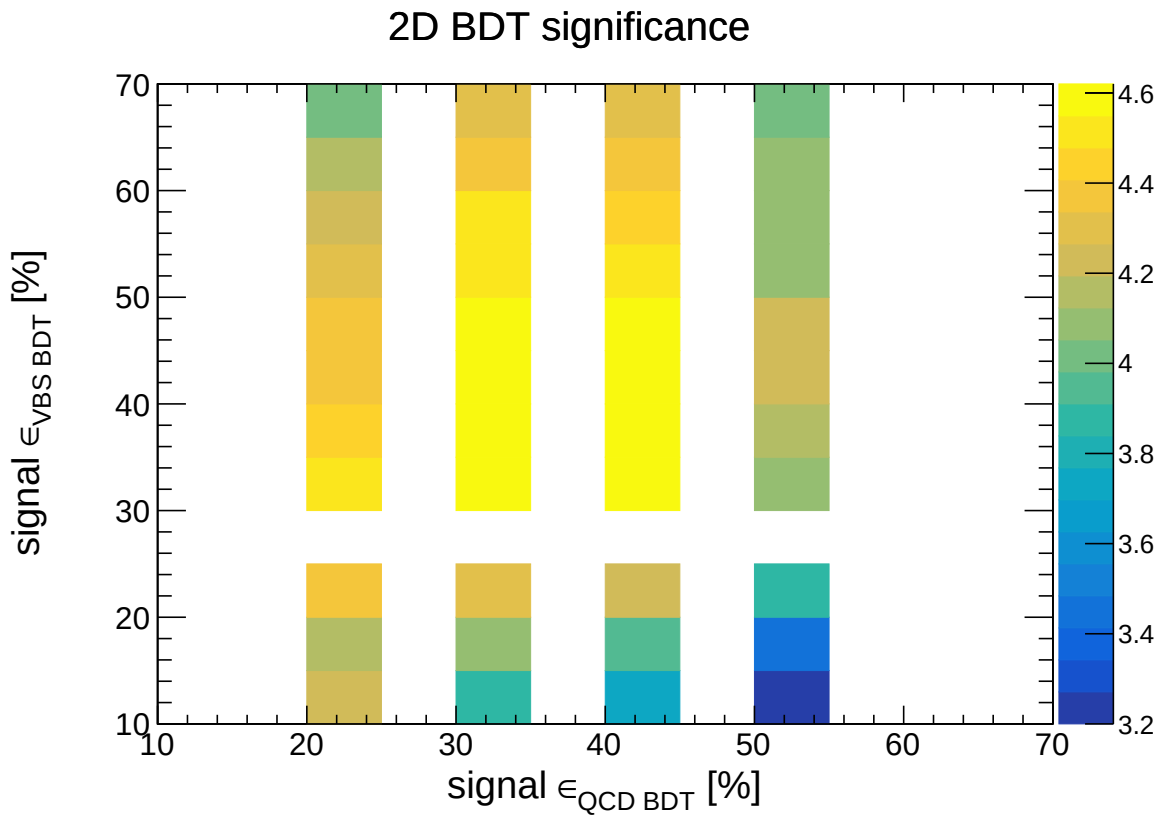


Figure 5.31: The 2D BDT significance plane used to scan for the optimal WP of the QCD BDT and the VBS BDT. Results at 14 TeV and for $3000\ fb^{-1}$ are shown. [plot will be updated to show all bins]

5.7 Results

The significance of the LL signal at 14 and 27 TeV obtained using the combined-background BDT and the 2D BDT method is shown in Table 5.22. Events after the baseline selection were used as the foundation for the multivariate analysis. Signal significance with the inclusion of the HF nose upgrade and the gain when moving from the combined-background BDT to the 2D BDT approach is also reported.

Using the combined-background BDT at 14 TeV, the confidence on the LL signal measurement is expected to reach 1σ level. This can be improved using the 2D BDT method with the significance of the LL signal reaching 1.3σ. Extending the η acceptance for electrons up to 4 is expected to increase the significance to 1.4σ using the 2D BDT approach. The gain when moving from the combined-background BDT to the 2D BDT at 14 TeV is 31.6 % if the HF nose option is not included and 33.7 % with the HF nose included.

At 27 TeV, the LL signal is expected to be measured at 4.6σ confidence level using the 2D BDT method. This is an improvement of 30.1 % with respect to the simpler combined-background BDT. Most of the gain at 27 TeV, with respect to 14 TeV, comes from an increased luminosity which enables harder suppression of the QCD background. Importance of the HF nose upgrade is especially noticeable at HE-LHC where the 5.4σ significance on the VBS LL measurement is expected if the 2D BDT approach is employed. This is due to the more forward kinematics at 27 TeV, with respect to 14 TeV.

	event counts at HL-LHC without the HF nose upgrade						
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$	**gain**
Combined-background BDT with $\epsilon_{signa} = 35\ \%$	15.1	64.3	72.6	48.2	49.5	0.98	31.6 %
2D BDT with $\epsilon_{signal}^{QCD\ BDT} = 20\ \%$ & $\epsilon_{signal}^{VBS\ BDT} = 35\ \%$	5.00	8.90	2.80	1.40	2.00	1.29	
	event counts at HL-LHC with the HF nose upgrade						
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$	**gain**
Combined-background BDT with $\epsilon_{signa} = 35\ \%$	16.1	68.2	77.1	47.9	48.6	1.04	33.7 %
2D BDT with $\epsilon_{signal}^{QCD\ BDT} = 20\ \%$ & $\epsilon_{signal}^{VBS\ BDT} = 35\ \%$	4.90	7.50	2.30	0.90	1.80	1.39	
	event counts at HE-LHC without the HF nose upgrade						
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$	**gain**
Combined-background BDT with $\epsilon_{signa} = 35\ \%$	232.4	1116	1501	704.8	960.9	3.55	30.1 %
2D BDT with $\epsilon_{signal}^{QCD\ BDT} = 40\ \%$ & $\epsilon_{signal}^{VBS\ BDT} = 45\ \%$	154.4	362.4	181.1	258.0	214.9	4.62	
	event counts at HE-LHC with the HF nose upgrade						
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$	**gain**
Combined-background BDT with $\epsilon_{signa} = 40\ \%$	293.2	1414	1888	1140.5	1310	3.87	38.2 %
2D BDT with $\epsilon_{signal}^{QCD\ BDT} = 30\ \%$ & $\epsilon_{signal}^{VBS\ BDT} = 40\ \%$	123.2	224.2	83.60	158.4	63.60	5.35	

Table 5.22: Event counts and corresponding signal significances for the combined-background BDT and the 2D BDT training at 14 and 27 TeV. Presented working points for both BDT training approaches give the most sensitive LL measurement. The gain using the 2D BDT compared to the simple combined-background BDT is also reported. Table shows both the results without the HF nose upgrade and with the HF nose upgrade included.

5.8 Summary

In the previous chapter it was shown that CMS is entering the measurement era for the VBS ZZ channel using the data from the LHC Run 2. Still, an important piece of the puzzle, i.e., the measurement of the longitudinal component of the Z boson, is missing. Measuring the longitudinal component of the Z boson in the clean VBS $ZZ \to 4l2j$ channel will enable to probe the scalar sector of the Standard Model and deepen our understanding of the EWSB mechanism. The final goal of this study was to optimize the signal extraction method and measure the LL signal significance for the HL-LHC and the HE-LHC.

Signal and background processes were simulated using MG5, MCFM, Delphes and Pythia8 tools at 7 and 13.5 TeV beam energies. Special care was given to jets since they dominate the final state and define the signal. The effect of parton showers on the leading jets was studied to make sure they are not affecting the identification of the tagging jets and thus making analysis unstable. With an increased luminosity, at HL- and HE-LHC conditions, the importance of pileup will increase as well. Thus, the effect of pile-up on the leading jets was studied as well. Both parton showers and pile-up were found to affect the leading jets at 10 % level.

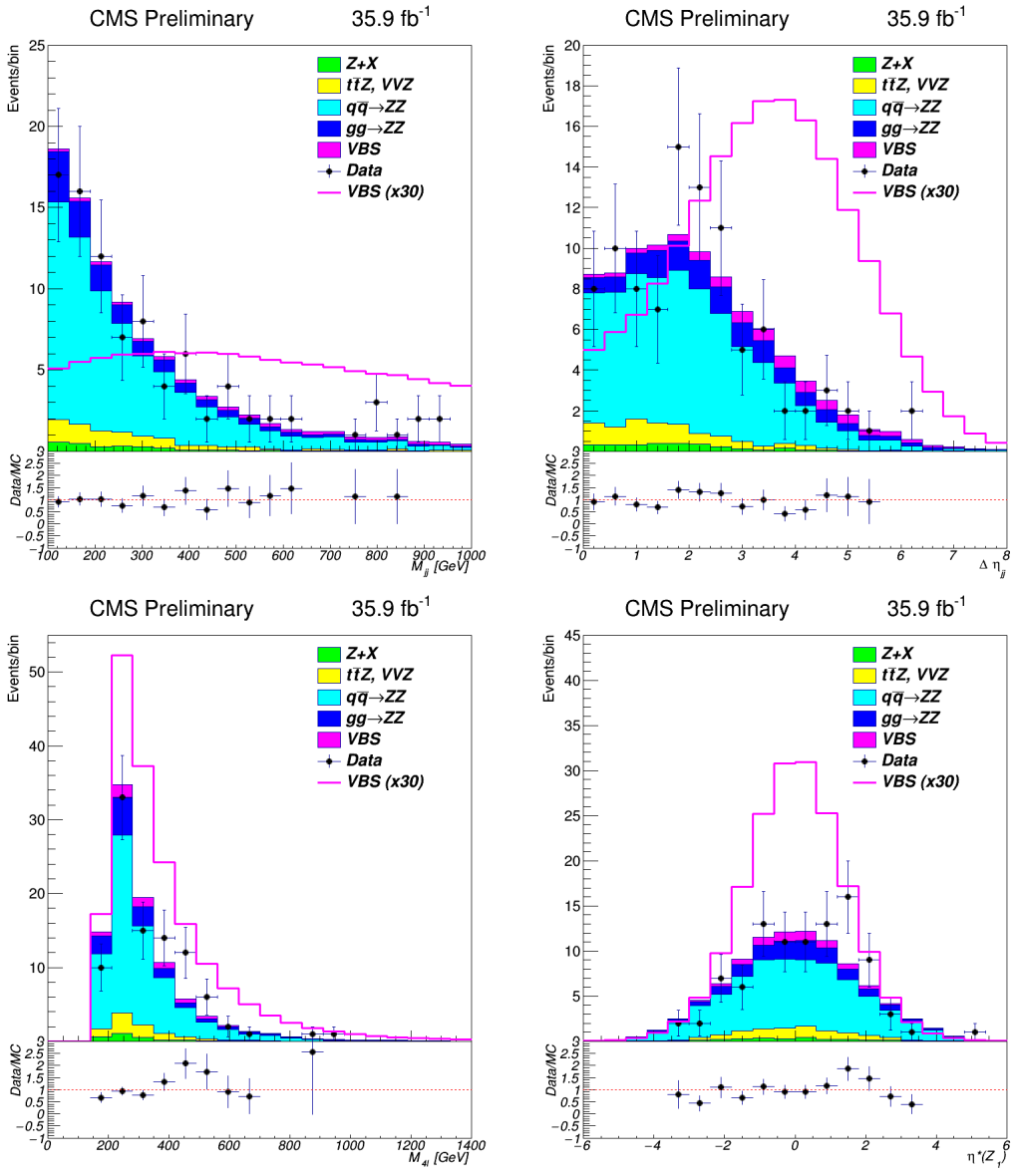
Since no single kinematic variable is discriminating enough to separate individual polarizations, multivariate approach was devised. Two such approaches were tested on both HL- and HE-LHC samples. The simpler of the two, the combined-background BDT, trained the LL signal against the proper mixture of VBS and QCD backgrounds to find the WP that maximizes signal sensitivity. The second approach exploits the difference in the kinematics of VBS and QCD processes to simultaneously train the BDT to separate the LL signal from the qq background and the LL signal from the mixture of the LT and TT backgrounds. This approach is referred to as the 2D BDT and is superior amongst the two. It was shown that as much as 160 % (120 %) can be gained in terms of the signal significance measurement by exploiting the 2D BDT at 14 TeV (27 TeV) compared to the simple cut-and-count approach. Without the HF nose upgrade, the LL signal is expected to be measured at 1.3σ (4.6σ) confidence level at 14 (27) TeV. Extending the lepton acceptance from $\eta = 3$ to $\eta = 4$, which corresponds to the HF nose upgrade, will increase the signal significance to 1.4σ (5.4σ) at 14 (27) TeV. These prospective studies show the great potential of the HE-LHC in observing the longitudinal component of the Z boson in the VBS $ZZ \to 4l2j$ channel.

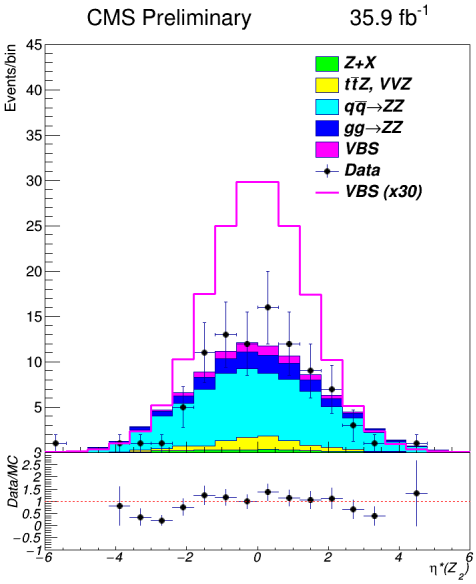
Chapter 6

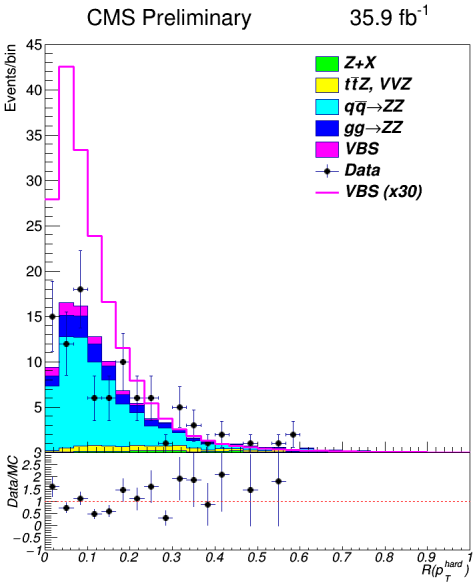
Summary

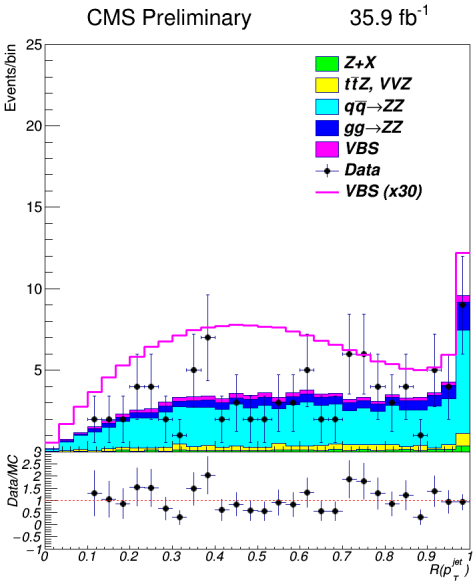
Appendix A: Supporting plots for the analysis presented in chapter 4

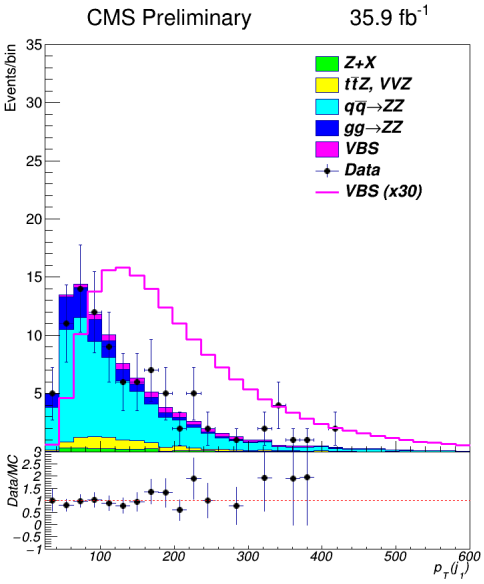
Figs. A.1 and A.2 show the distributions, defined in Table 4.7, for the 2016 and 2017 data-taking periods used to extract the VBS signal.

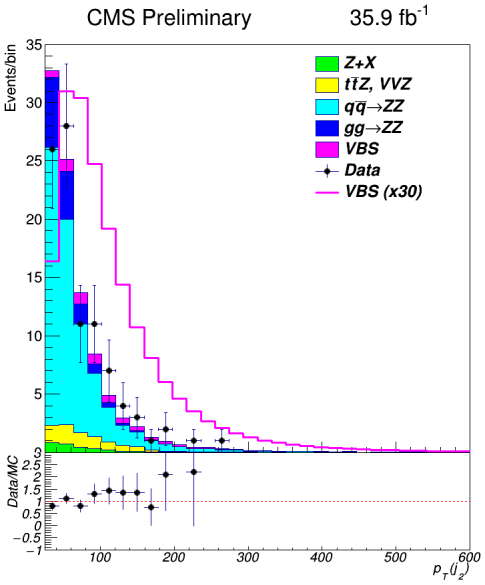


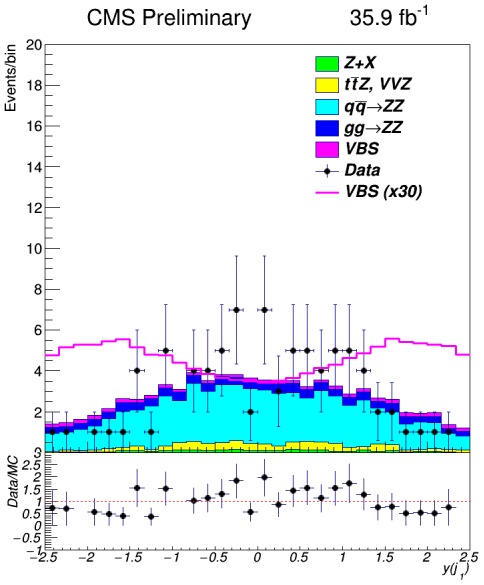

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η*(Z2)

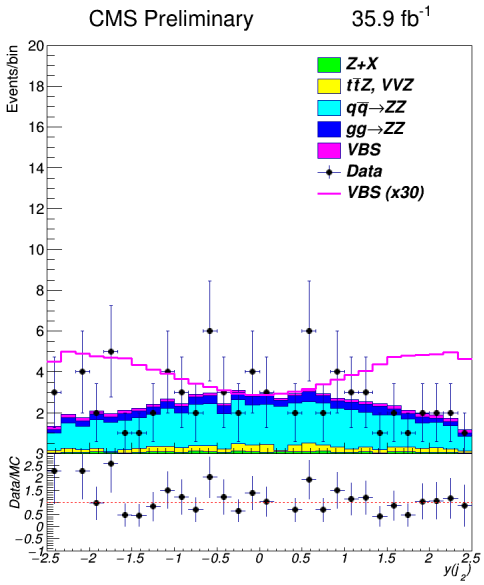

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(pT hard)

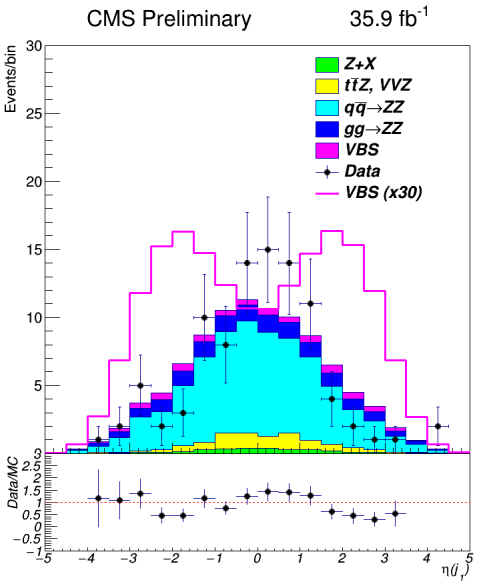

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(pT jet)

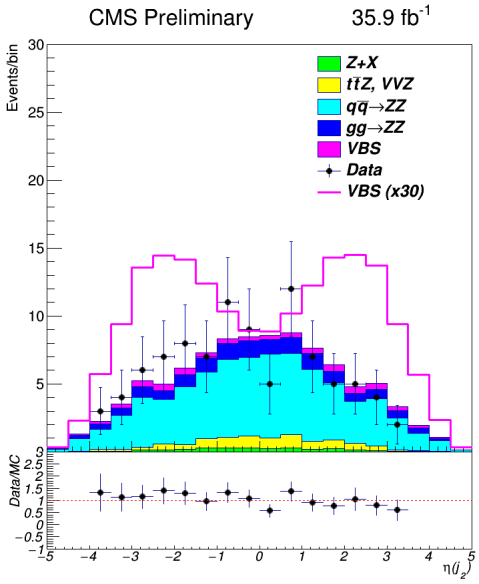

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(j1)

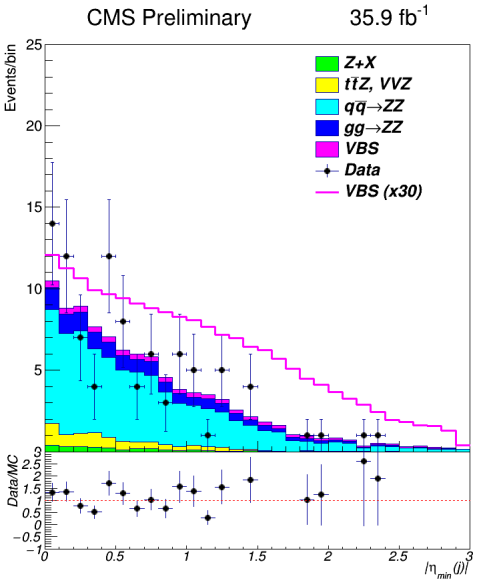

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(j2)

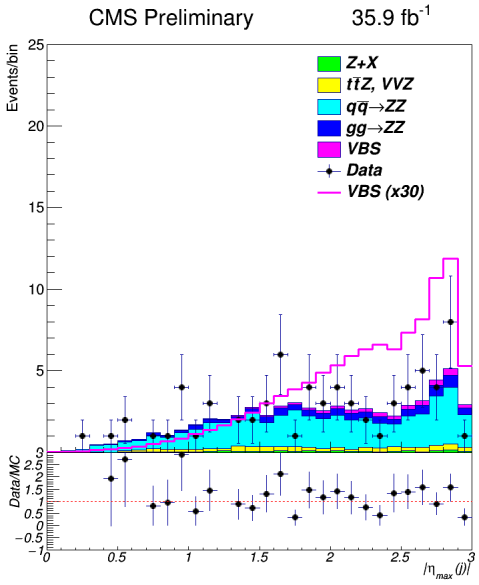

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j1)

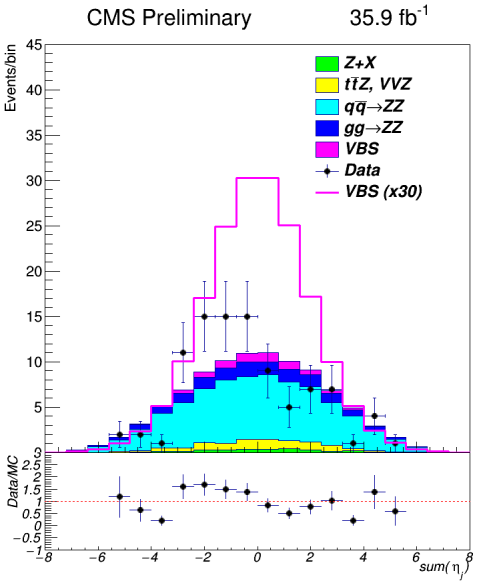

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j2)

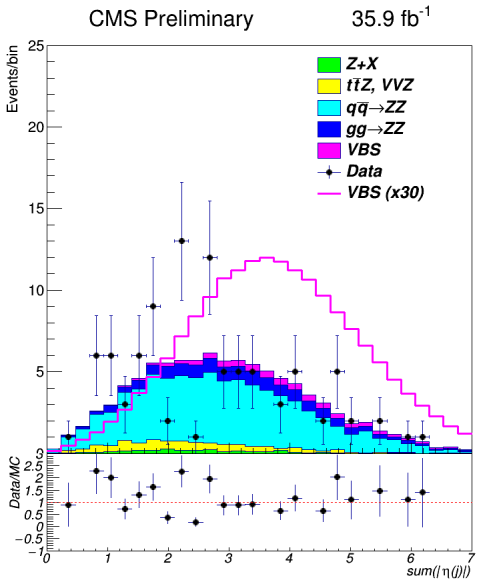

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j1)

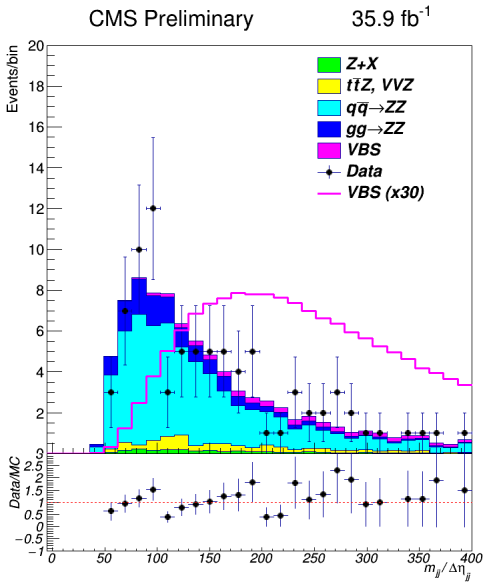

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j2)

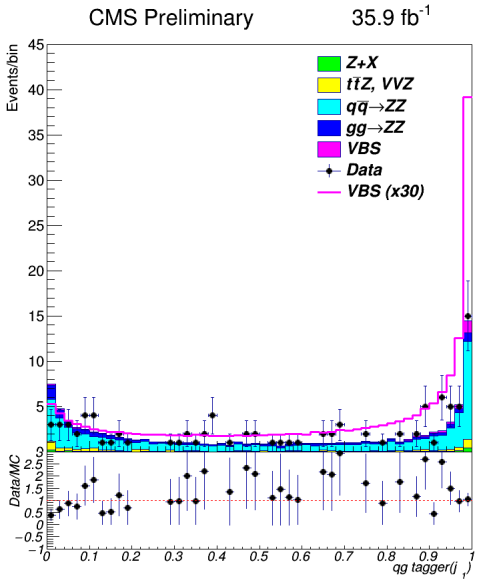

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmin(j)|

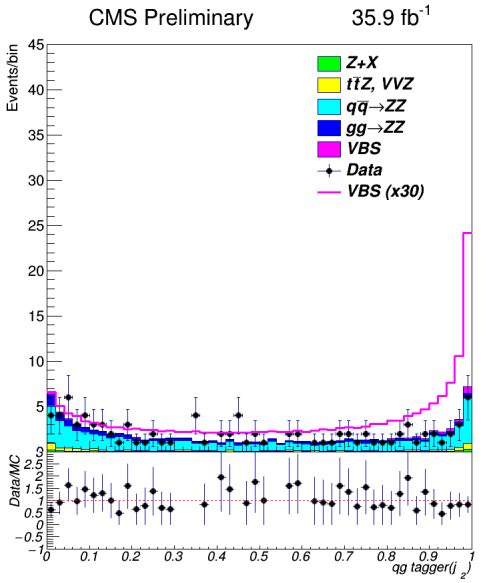

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmax(j)|

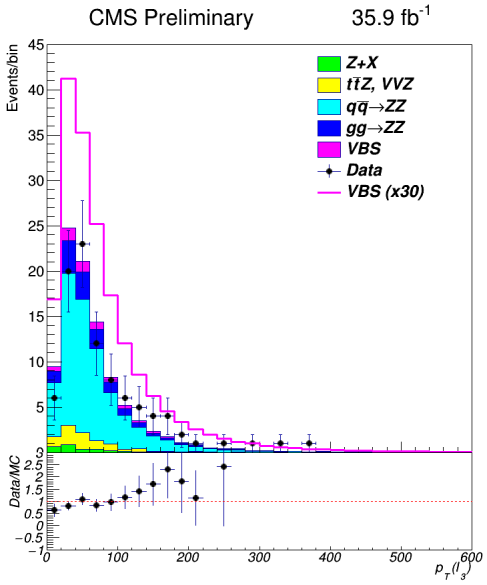

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(ηj)

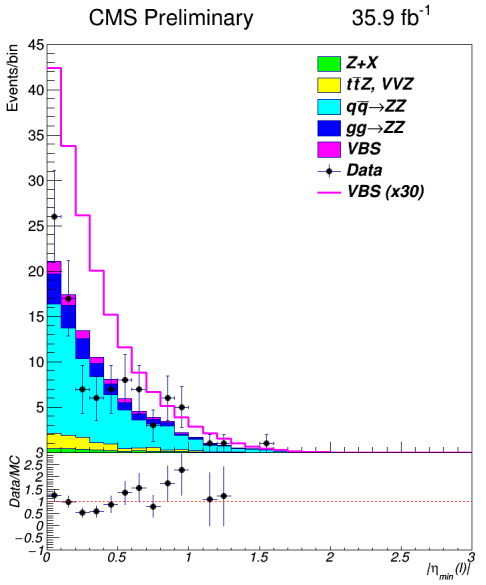

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(|η(j)|)


CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
mjj/Δηjj


CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
qg tagger(j 1)


CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
qg tagger(j 2)


CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(l3)


CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmin(l)|

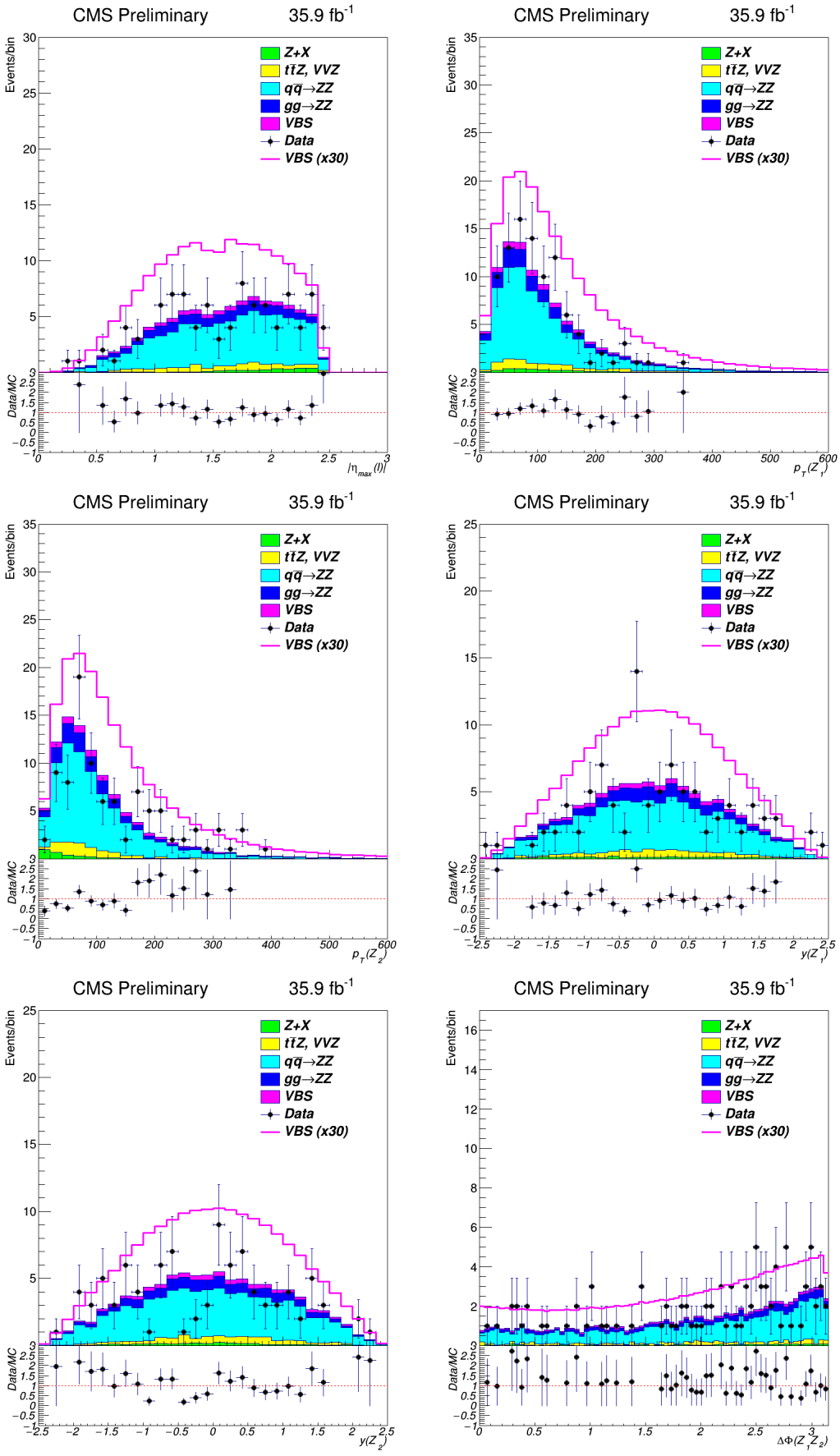
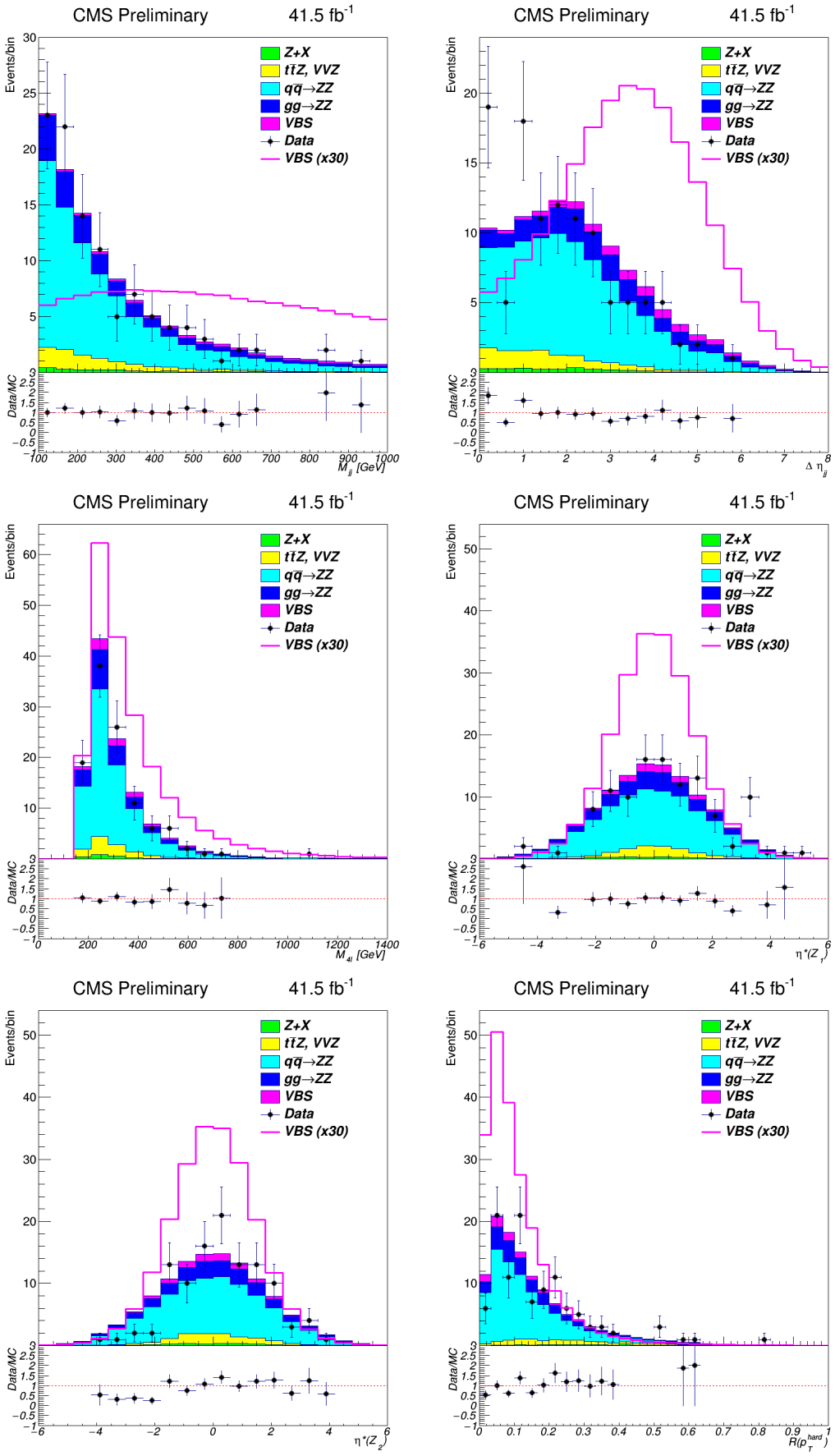
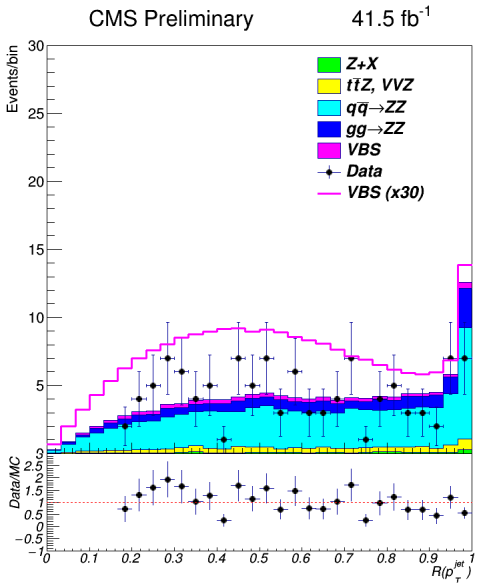
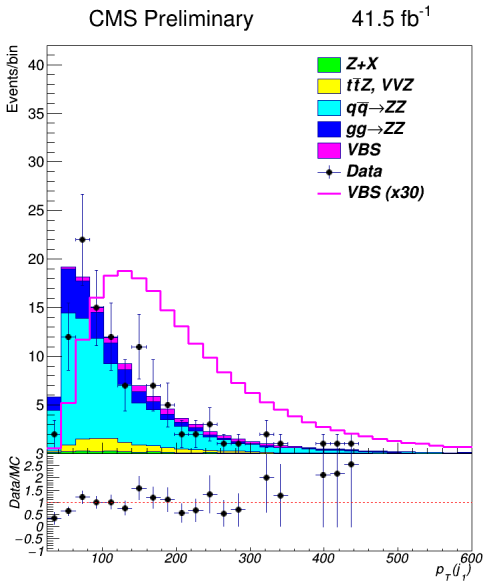
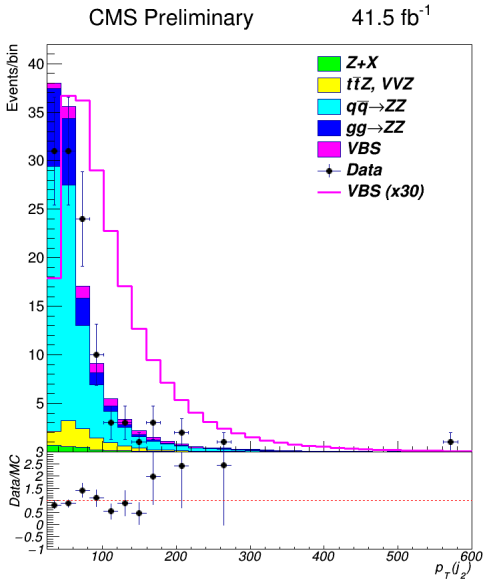


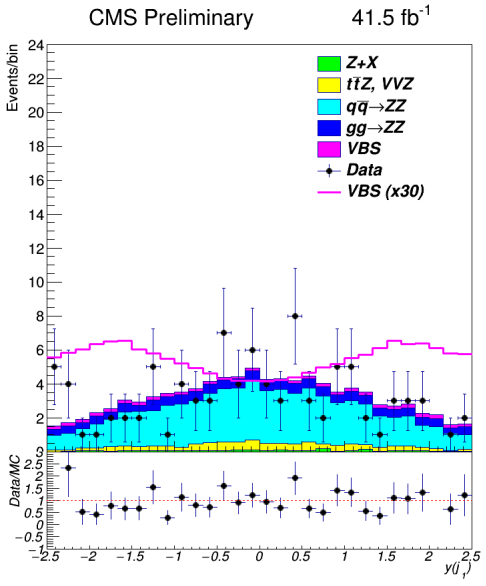
Figure A.1: Comparison of data to background and signal estimations in 2016 samples used in the analysis.

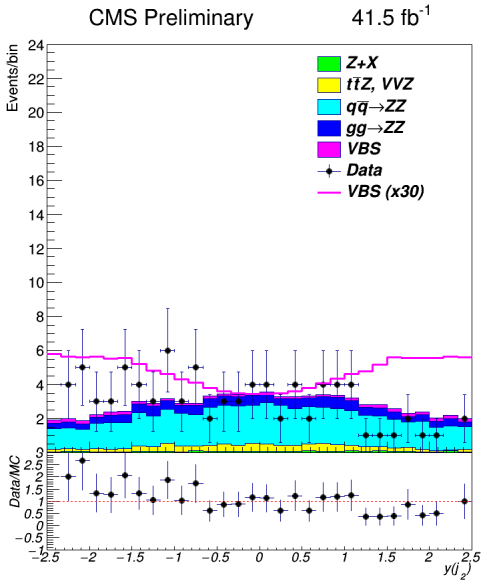

CMS Preliminary 41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
M_jj [GeV]
Δ η_jj
M_4l [GeV]
η*(Z_1)
η*(Z_2)
R(p_T^hard)

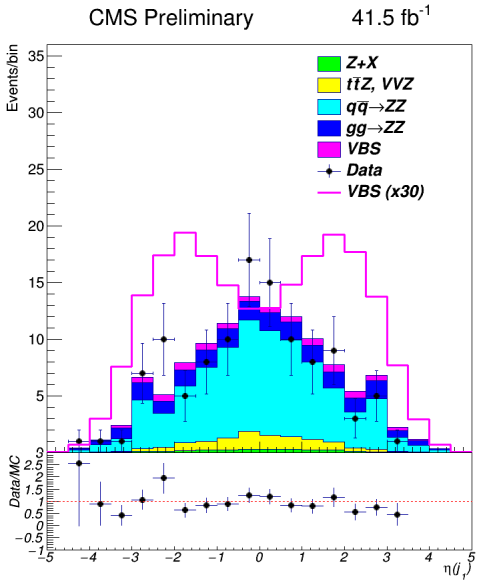

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(p_T^jet)

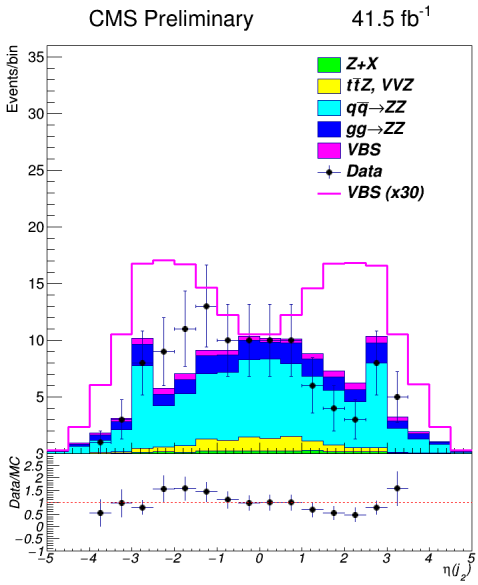

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
p_T(j_1)

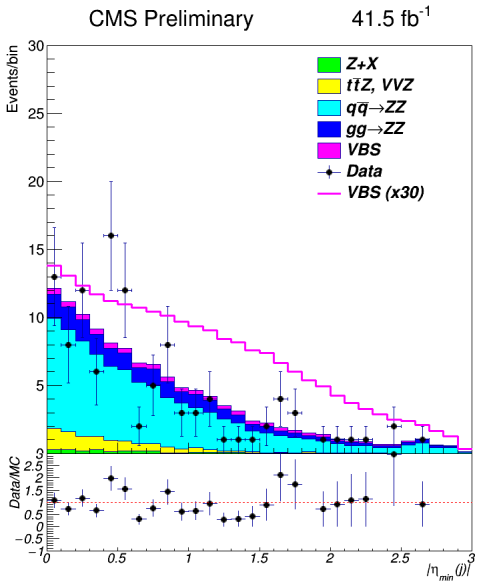

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
p_T(j_2)

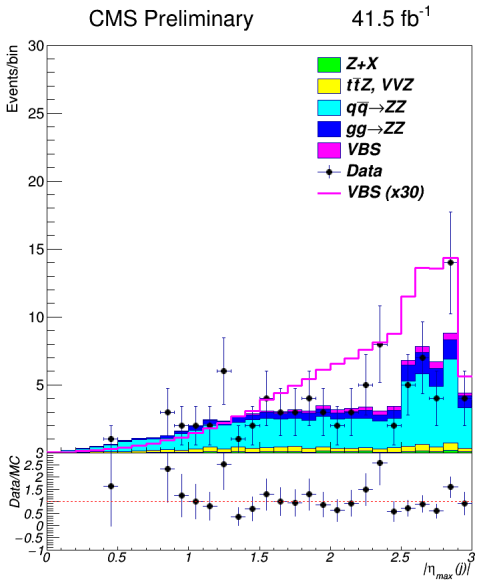

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j_1)

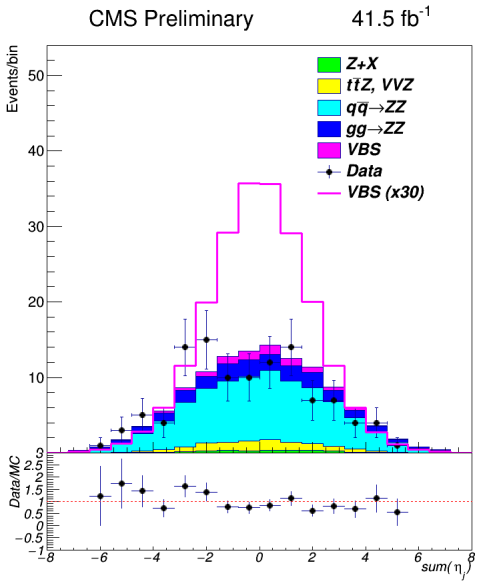

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j_2)

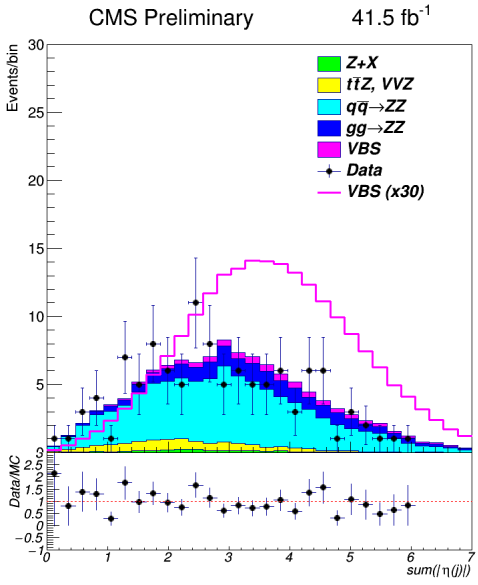

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j_1)

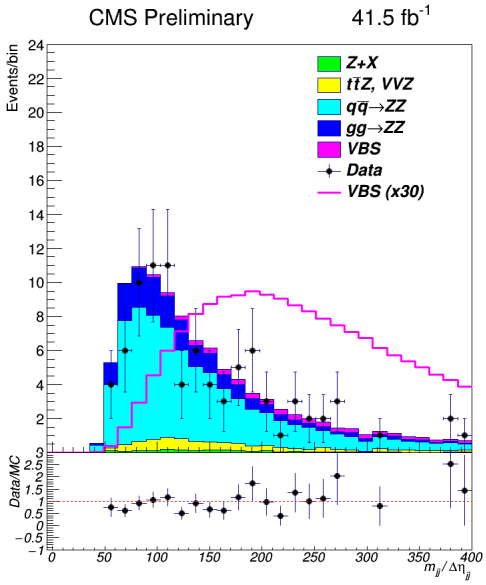

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j2)

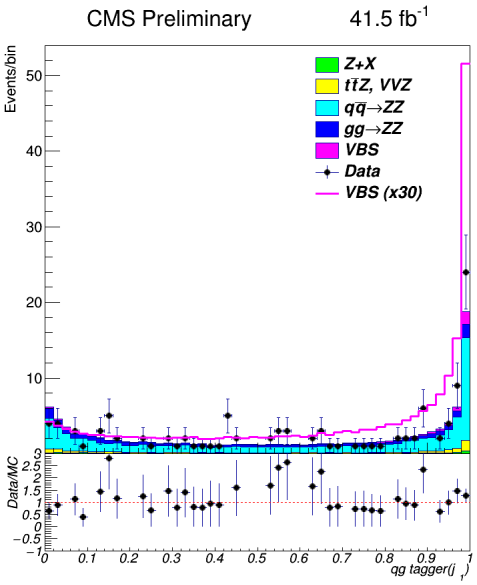

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmin(j)|

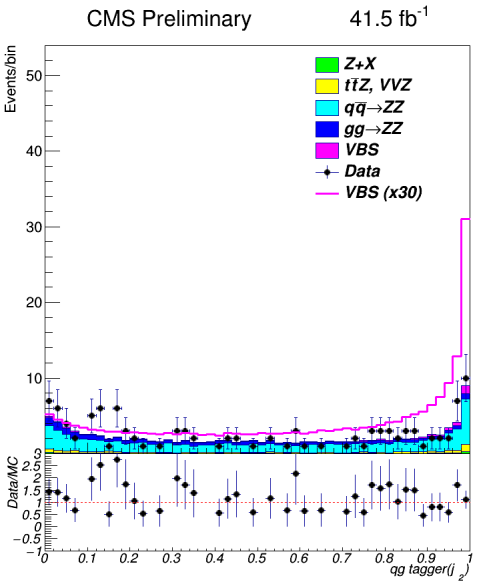

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmax(j)|

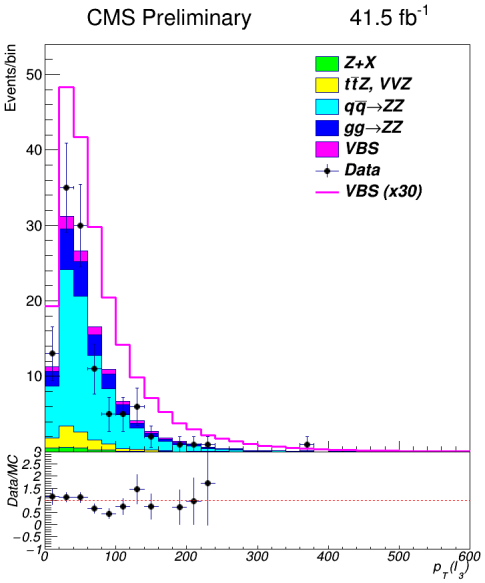

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(ηj)

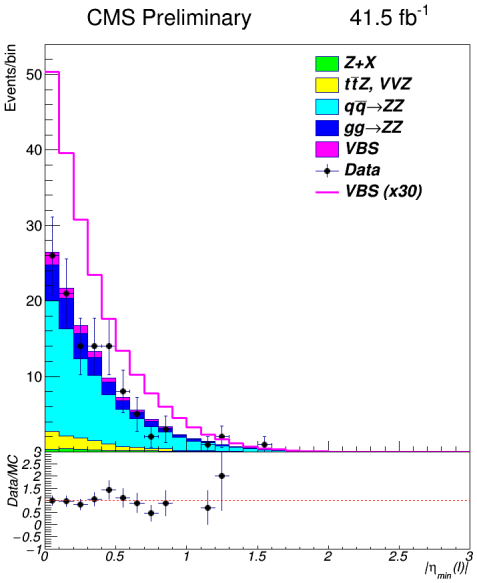

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(|η(j)|)

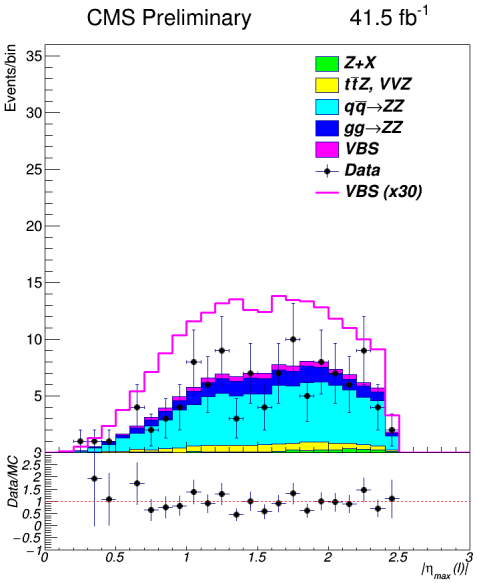

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
mjj/Δηjj

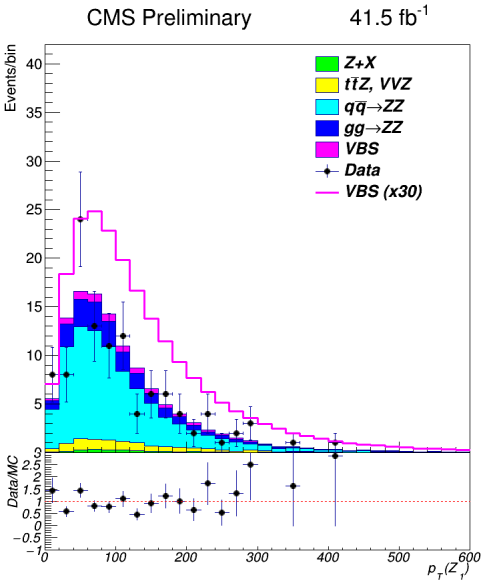

CMS Preliminary
41.5 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
qg tagger(j 1)


CMS Preliminary
41.5 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
qg tagger(j 2)


CMS Preliminary
41.5 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
pT(l3)


CMS Preliminary
41.5 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
|ηmin(l)|


CMS Preliminary
41.5 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
|ηmax(l)|


CMS Preliminary
41.5 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
pT(Z1)

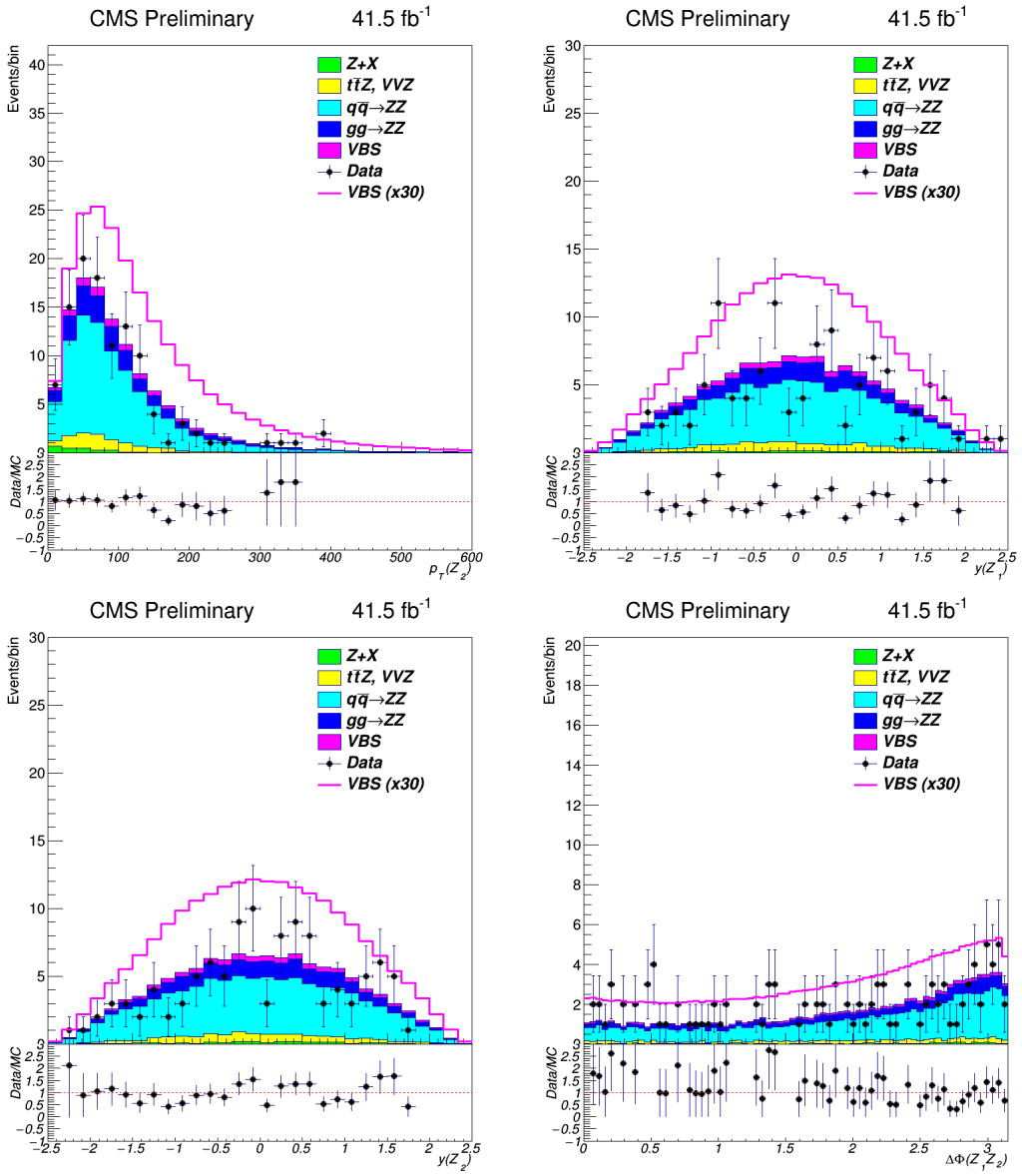
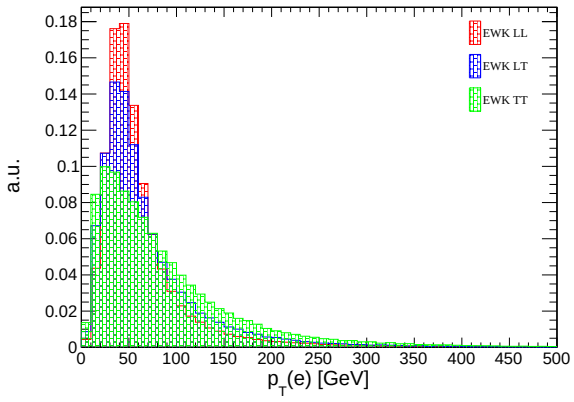
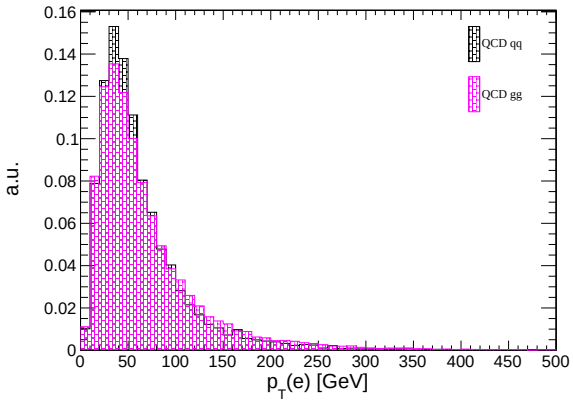


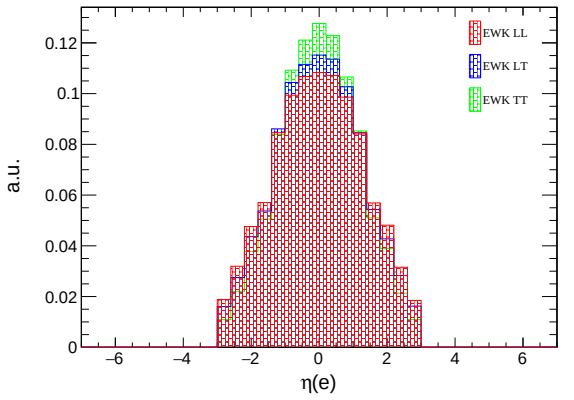
Figure A.2: Comparison of data to background and signal estimations in 2017 samples used in the analysis.

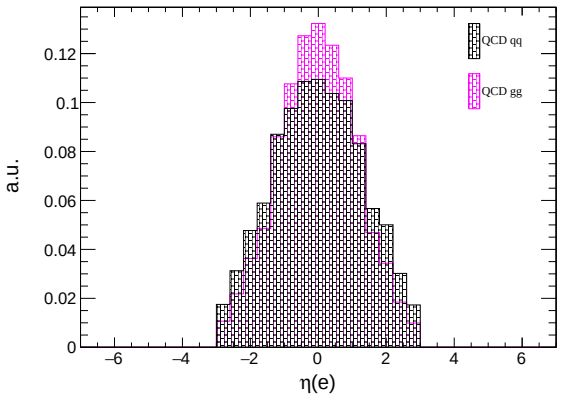
Appendix B: Supporting plots for the analysis presented in chapter 5

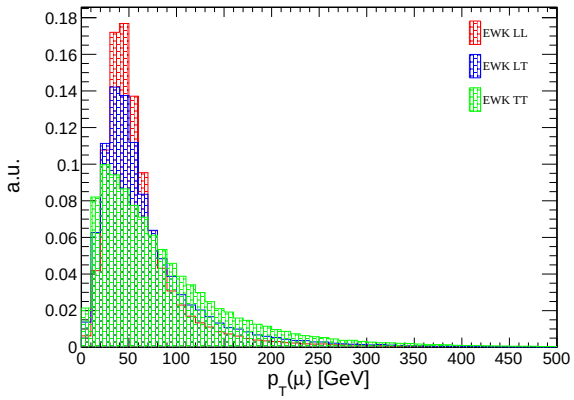
Figs. B.1 and B.2 show distributions at 14 and 27 TeV of all variables used in the analysis presented in chapter 5. Figs. B.3-B.7 compare the kinematics at 14 and 27 TeV for the LL, LT, TT, qq and gg contributions, separately.

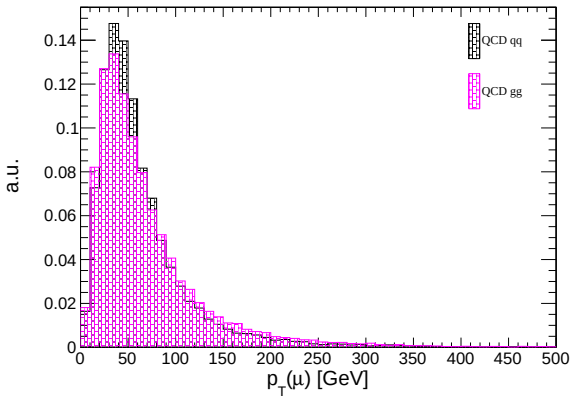


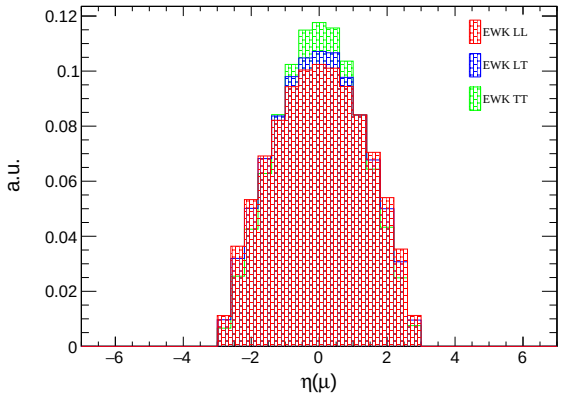


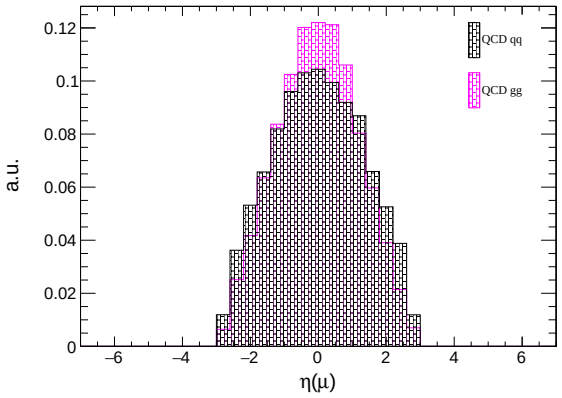


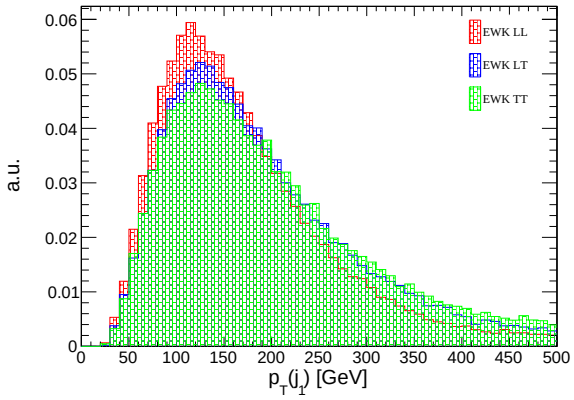


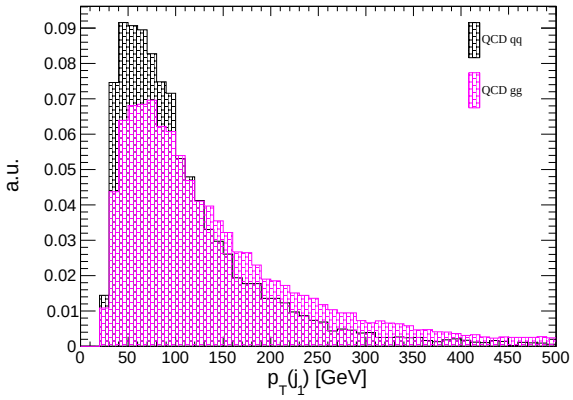


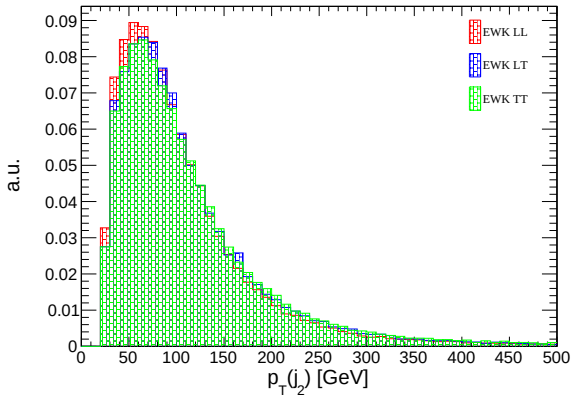


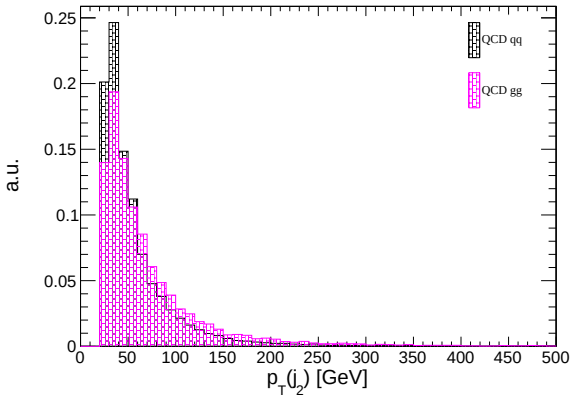

EWK LL
EWK LT
EWK TT
a.u.
η(μ)

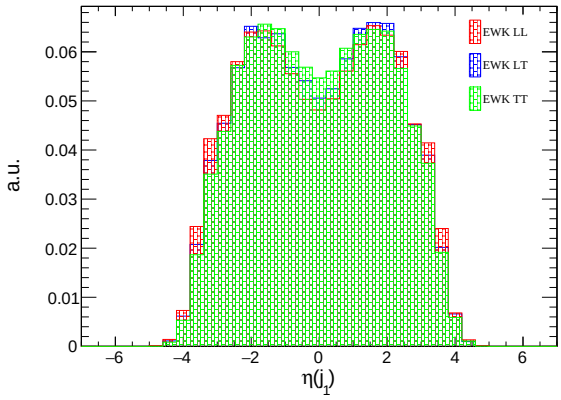

QCD qq
QCD gg
a.u.
η(μ)

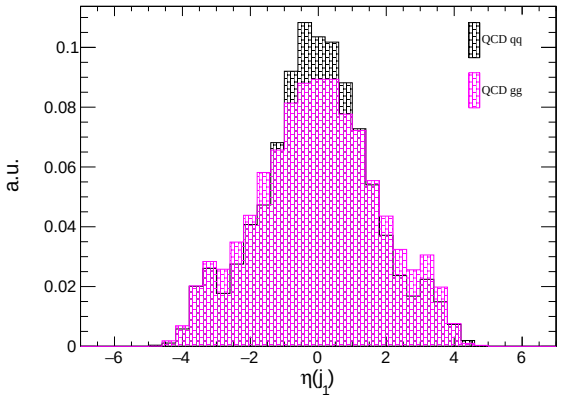

EWK LL
EWK LT
EWK TT
a.u.
p_T(j_1) [GeV]

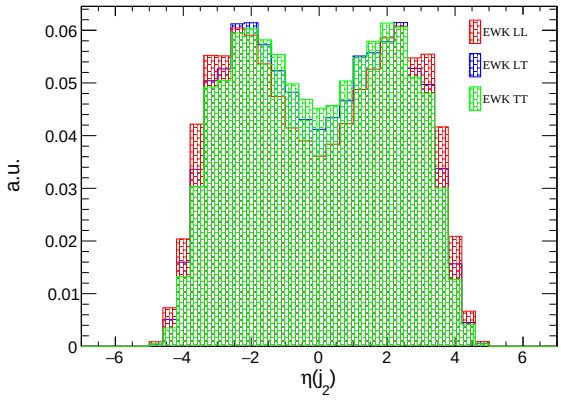

QCD qq
QCD gg
a.u.
p_T(j_1) [GeV]

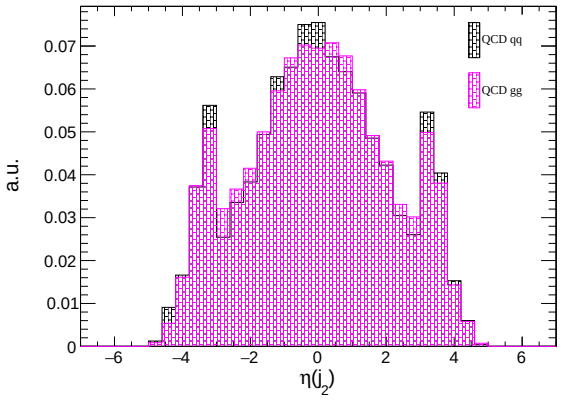

EWK LL
EWK LT
EWK TT
a.u.
p_T(j_2) [GeV]

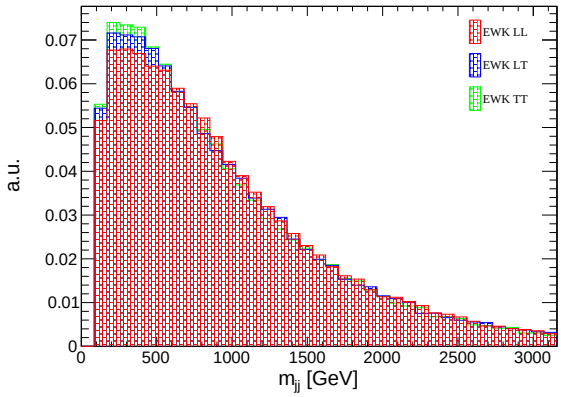

QCD qq
QCD gg
a.u.
p_T(j_2) [GeV]

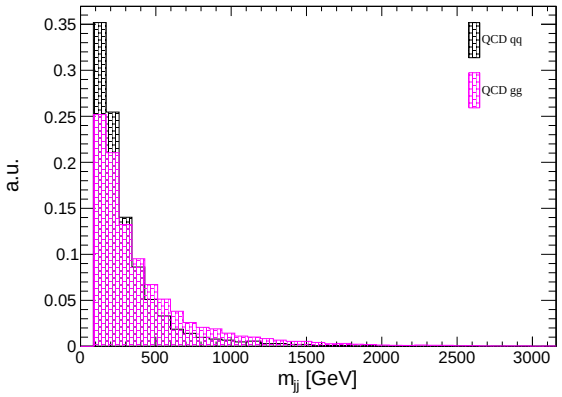

EWK LL
EWK LT
EWK TT
a.u.
η(j_1)

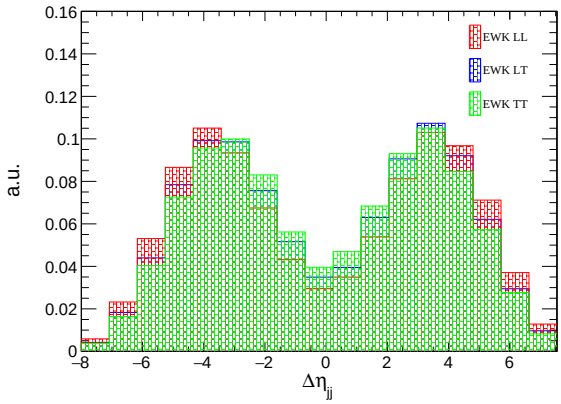

QCD qq
QCD gg
a.u.
η(j_1)

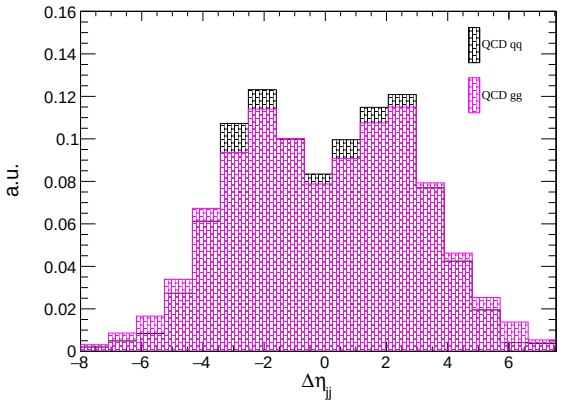

EWK LL
EWK LT
EWK TT
a.u.
$\eta(j_2)$

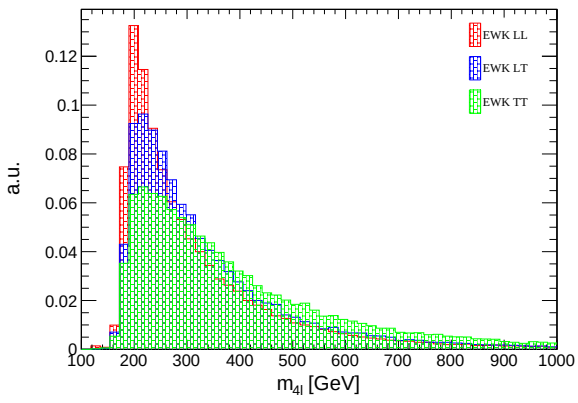

QCD qq
QCD gg
a.u.
$\eta(j_2)$

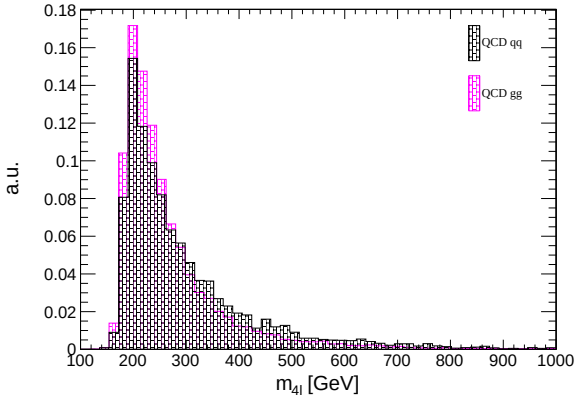

EWK LL
EWK LT
EWK TT
a.u.
m_{jj} [GeV]

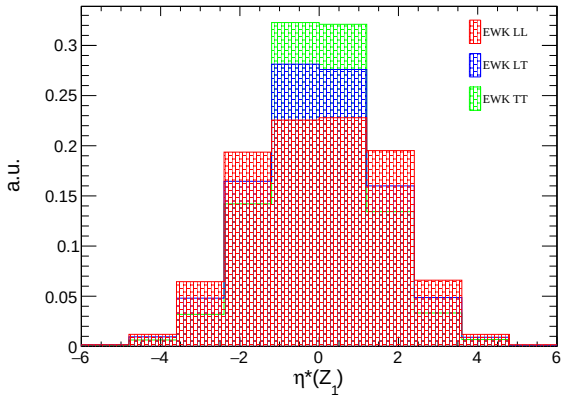

QCD qq
QCD gg
a.u.
m_{jj} [GeV]

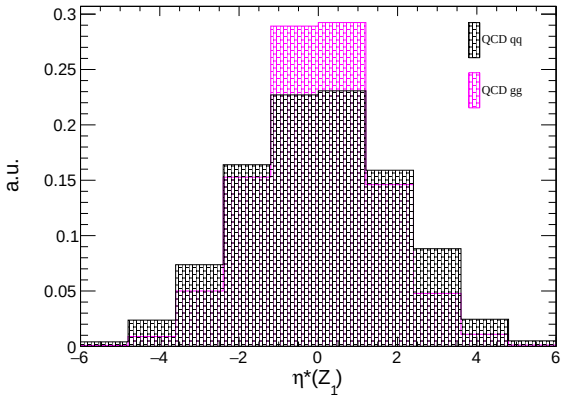

EWK LL
EWK LT
EWK TT
a.u.
$\Delta\eta_{jj}$

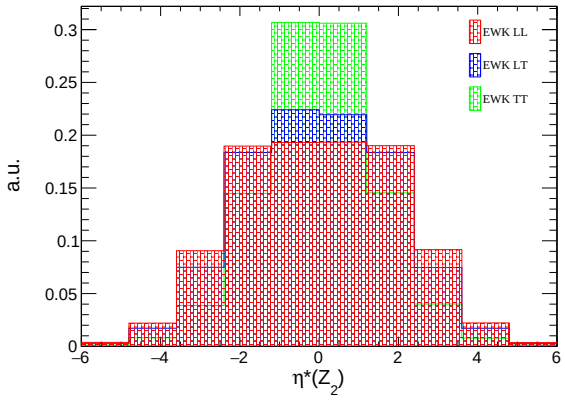

QCD qq
QCD gg
a.u.
$\Delta\eta_{jj}$

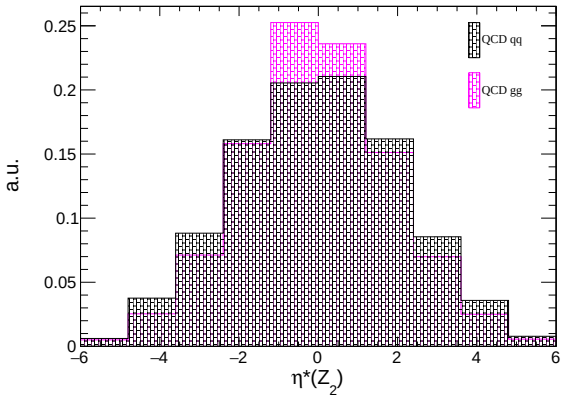

EWK LL
EWK LT
EWK TT
a.u.
m_{4l} [GeV]

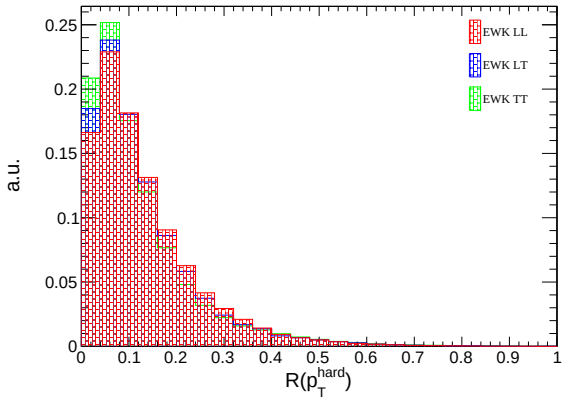

QCD qq
QCD gg
a.u.
m_{4l} [GeV]

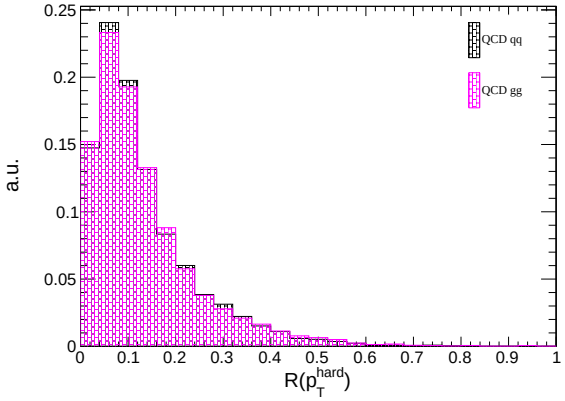

EWK LL
EWK LT
EWK TT
a.u.
η*(Z1)

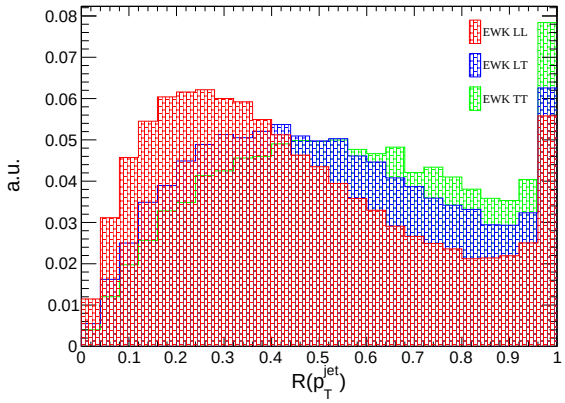

QCD qq
QCD gg
a.u.
η*(Z1)

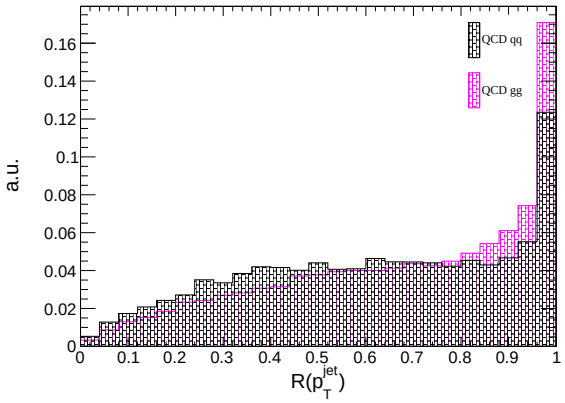

EWK LL
EWK LT
EWK TT
a.u.
η*(Z2)

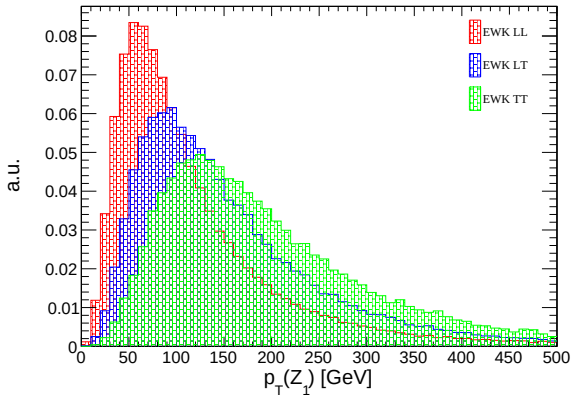

QCD qq
QCD gg
a.u.
η*(Z2)

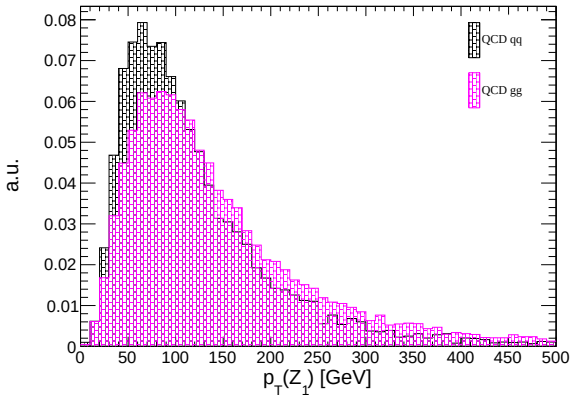

EWK LL
EWK LT
EWK TT
a.u.
R(pThard)

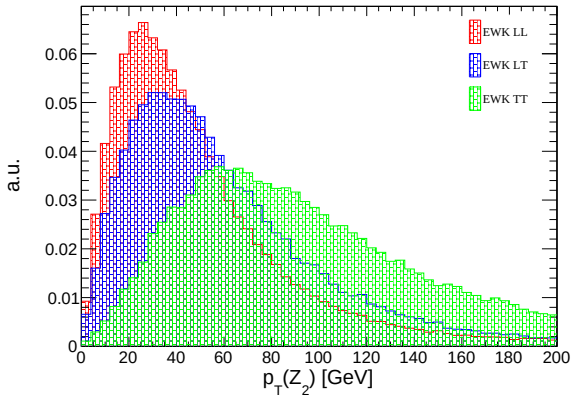

QCD qq
QCD gg
a.u.
R(pThard)

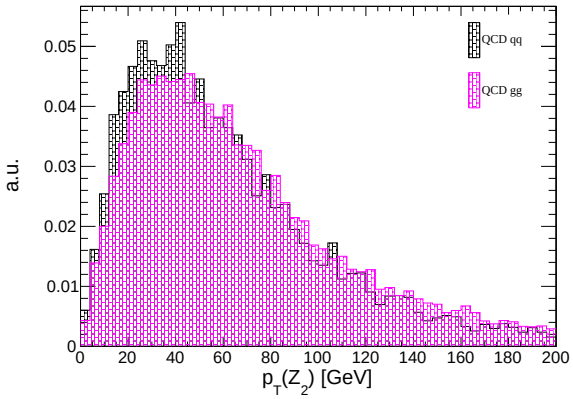

EWK LL
EWK LT
EWK TT
a.u.
R(pTjet)

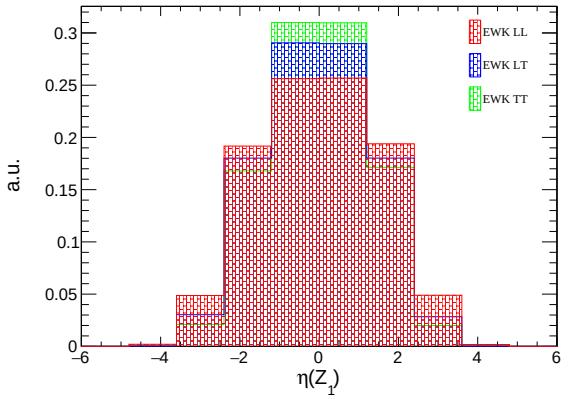

QCD qq
QCD gg
a.u.
R(pTjet)

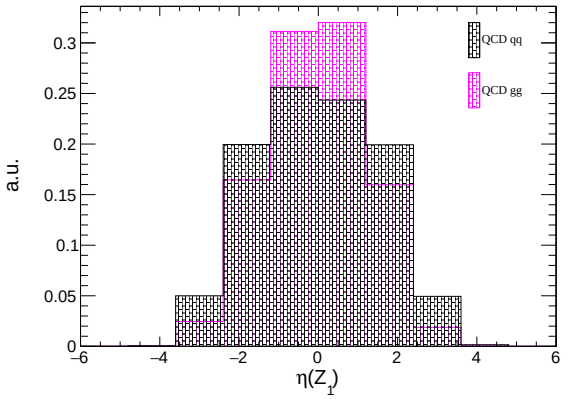

EWK LL
EWK LT
EWK TT
a.u.
$p_T(Z_1)$ [GeV]

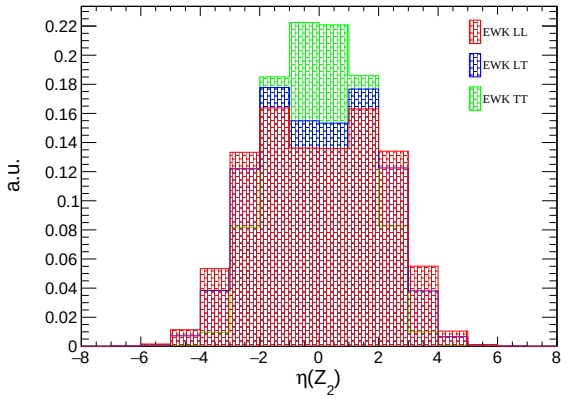

QCD qq
QCD gg
a.u.
$p_T(Z_1)$ [GeV]

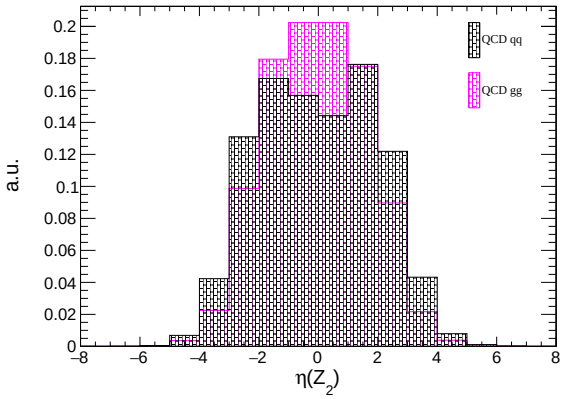

EWK LL
EWK LT
EWK TT
a.u.
$p_T(Z_2)$ [GeV]


QCD qq
QCD gg
a.u.
$p_T(Z_2)$ [GeV]


EWK LL
EWK LT
EWK TT
a.u.
η(Z1)


QCD qq
QCD gg
a.u.
η(Z1)


EWK LL
EWK LT
EWK TT
a.u.
η(Z2)


QCD qq
QCD gg
a.u.
η(Z2)

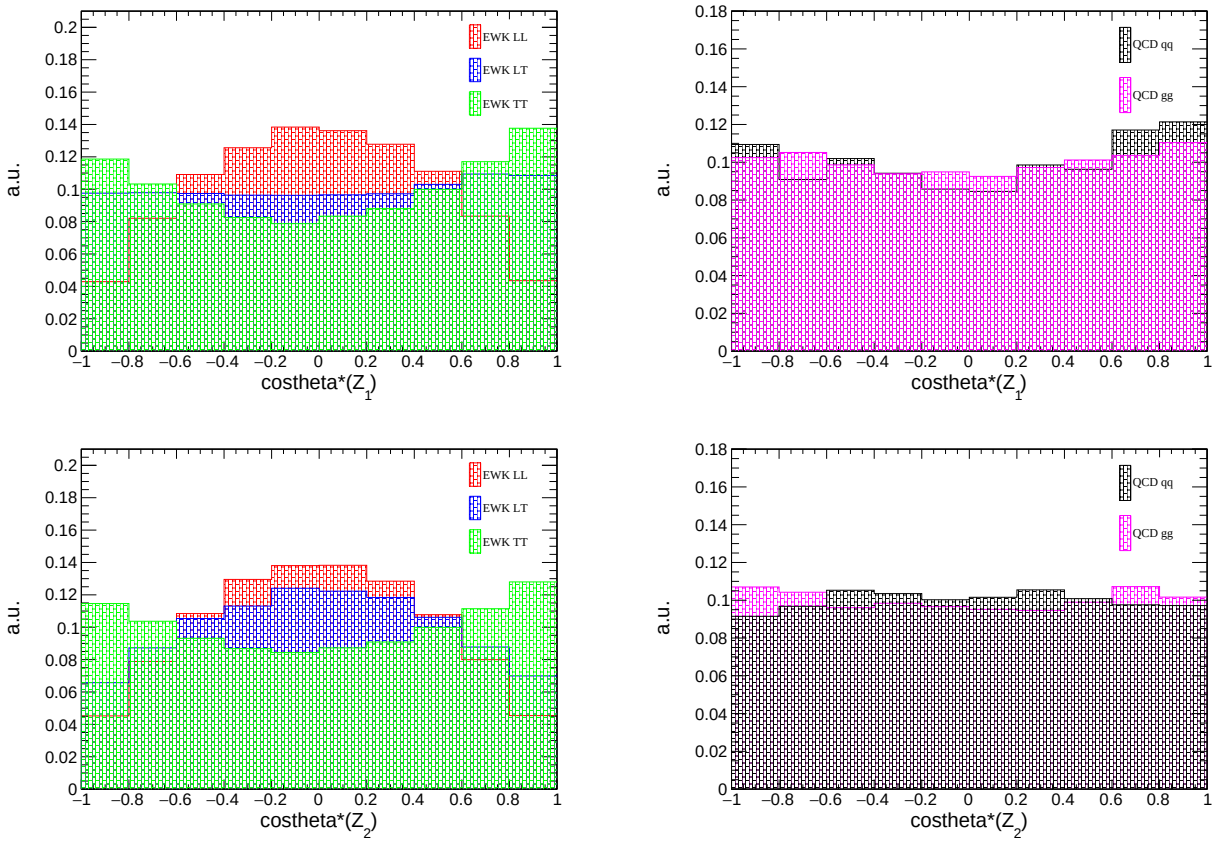
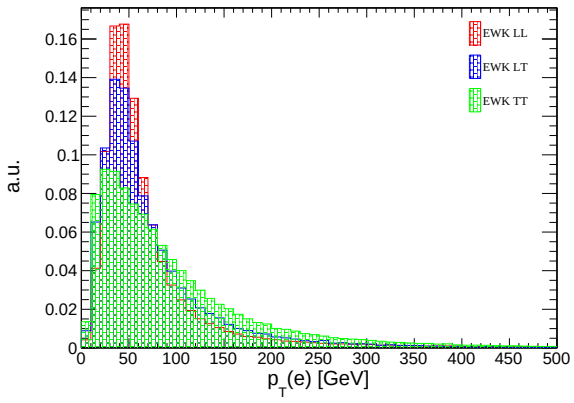
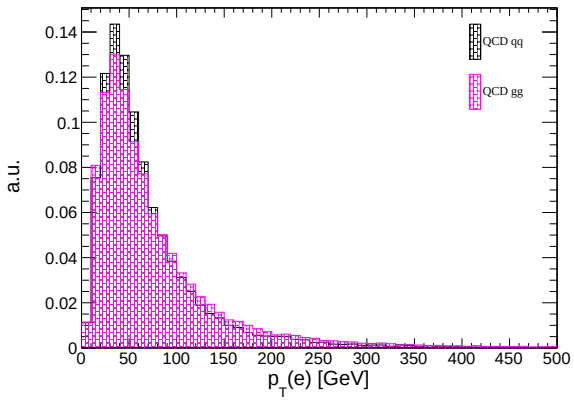
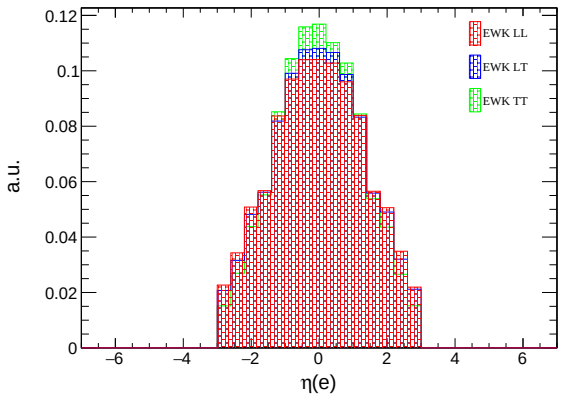
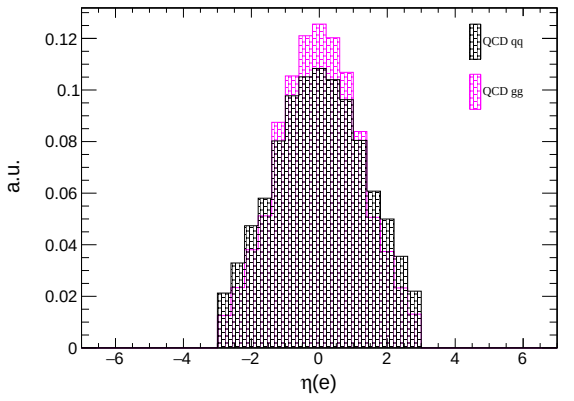


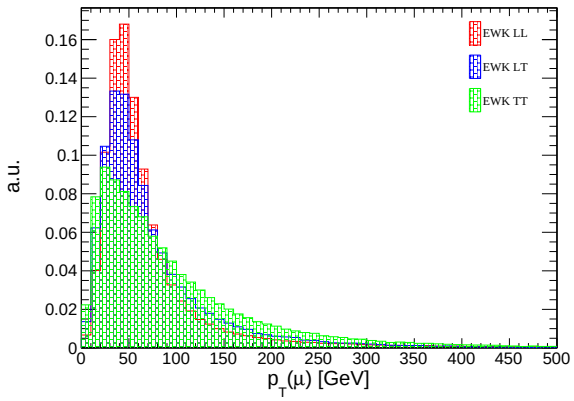
Figure B.1: Kinematics of VBS (left) and QCD (right) processes at 14 TeV after the baseline selection.

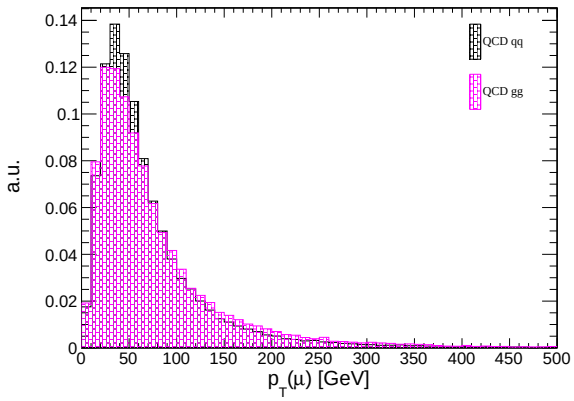

EWK LL
EWK LT
EWK TT
a.u.
p_T(e) [GeV]

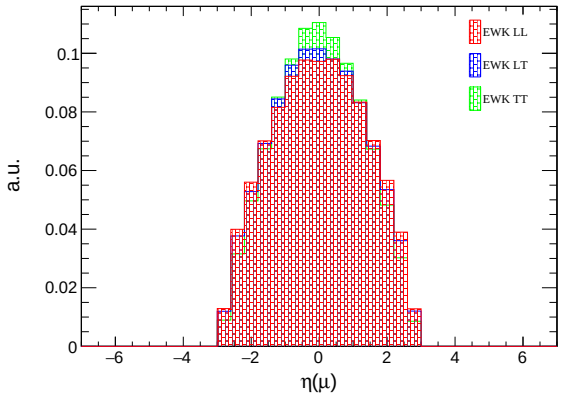

QCD qq
QCD gg
a.u.
p_T(e) [GeV]

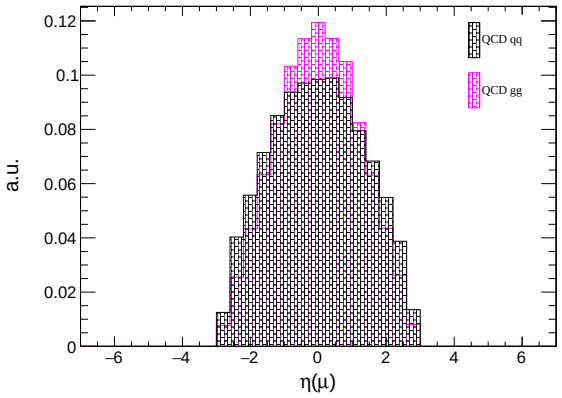

EWK LL
EWK LT
EWK TT
a.u.
η(e)

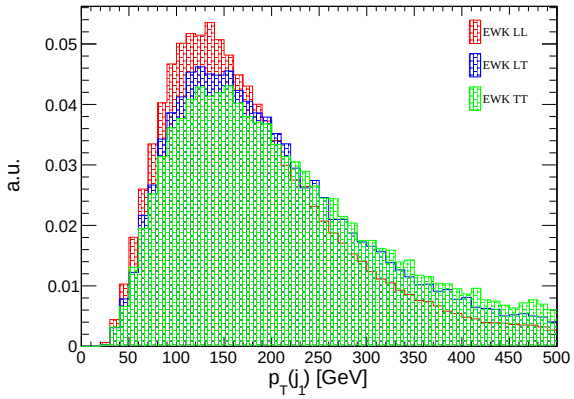

QCD qq
QCD gg
a.u.
η(e)

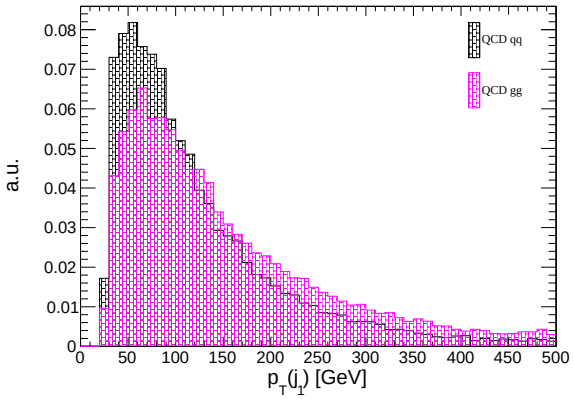

EWK LL
EWK LT
EWK TT
a.u.
p_T(μ) [GeV]

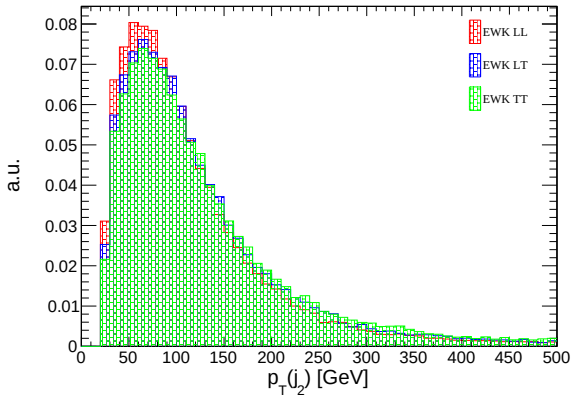

QCD qq
QCD gg
a.u.
p_T(μ) [GeV]

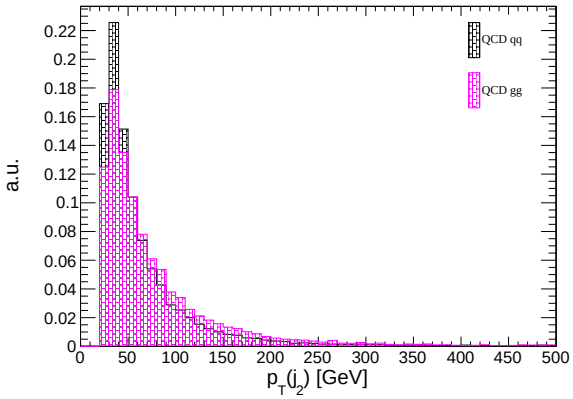

EWK LL
EWK LT
EWK TT
a.u.
η(μ)

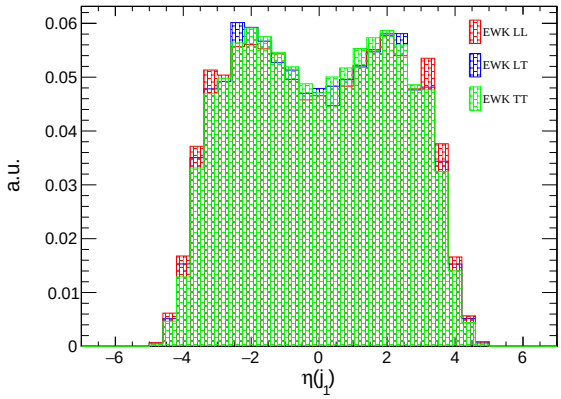

QCD qq
QCD gg
a.u.
η(μ)

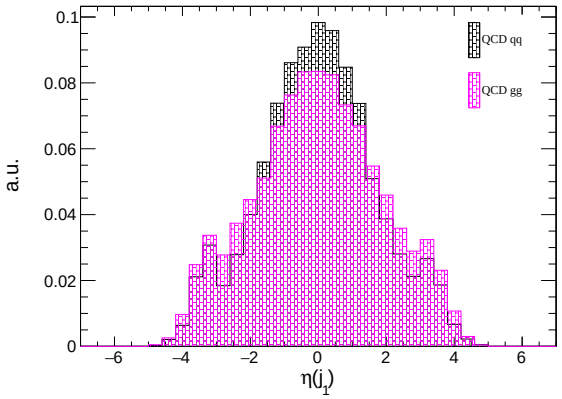

EWK LL
EWK LT
EWK TT
a.u.
p_T(j_1) [GeV]

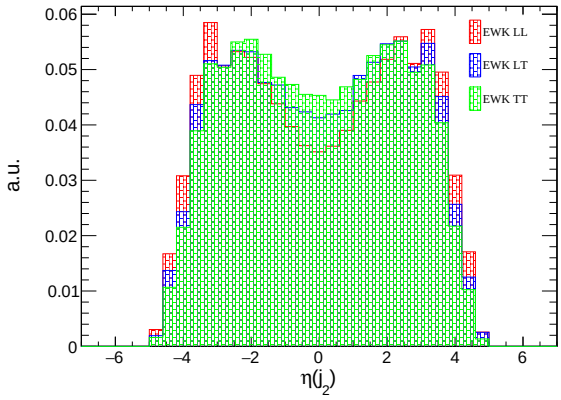

QCD qq
QCD gg
a.u.
p_T(j_1) [GeV]

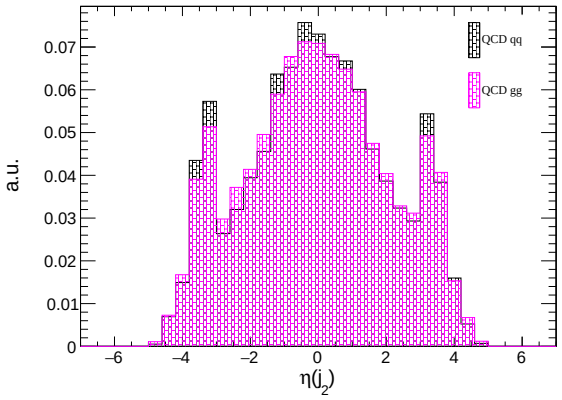

EWK LL
EWK LT
EWK TT
a.u.
p_T(j_2) [GeV]

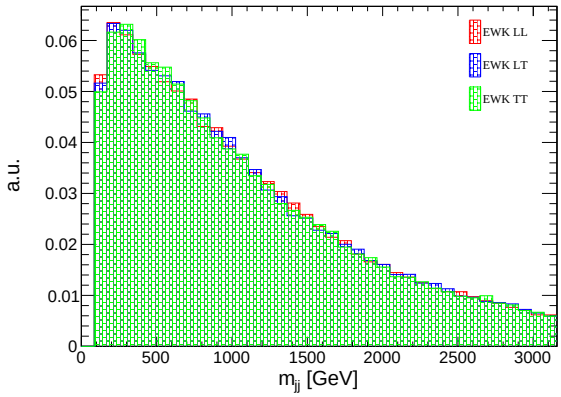

QCD qq
QCD gg
a.u.
p_T(j_2) [GeV]

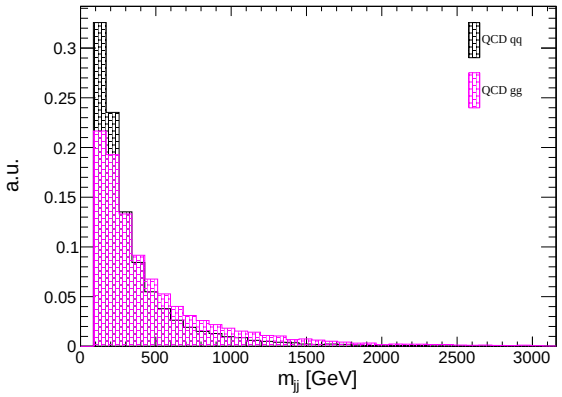

EWK LL
EWK LT
EWK TT
a.u.
η(j_1)

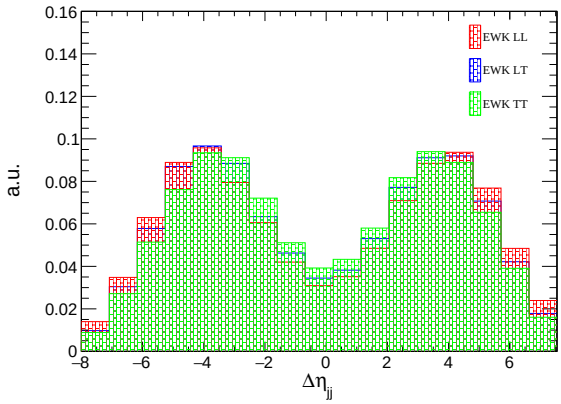

QCD qq
QCD gg
a.u.
η(j_1)

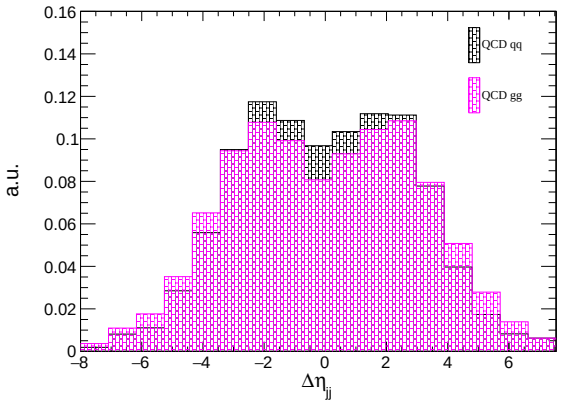

EWK LL
EWK LT
EWK TT
a.u.
$\eta(j_2)$

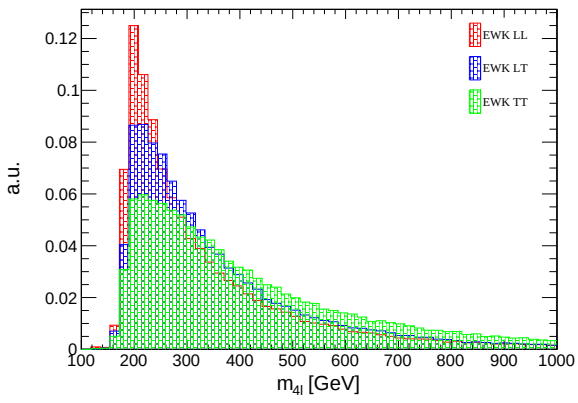

QCD qq
QCD gg
a.u.
$\eta(j_2)$

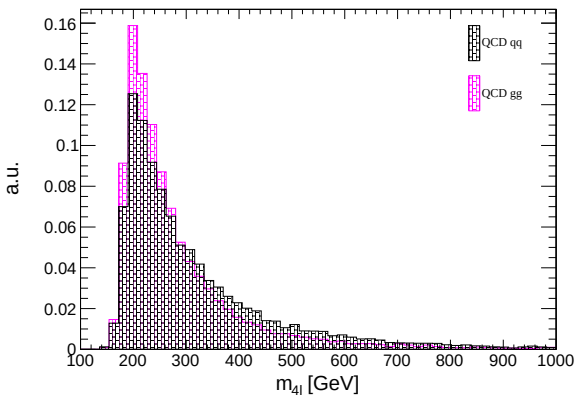

EWK LL
EWK LT
EWK TT
a.u.
m_{jj} [GeV]

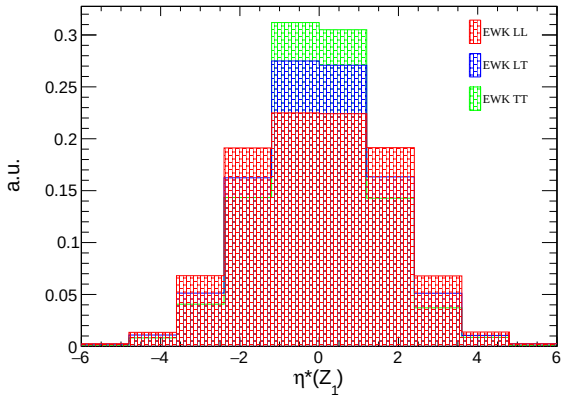

QCD qq
QCD gg
a.u.
m_{jj} [GeV]

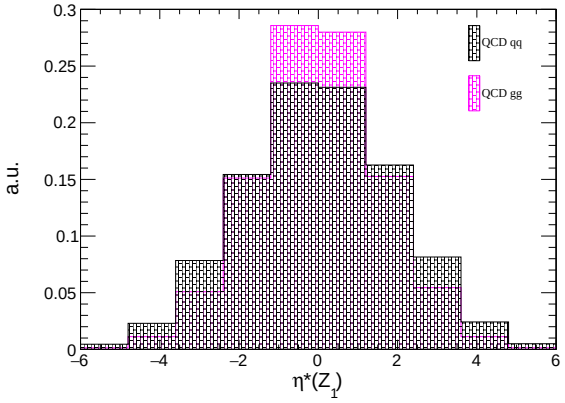

EWK LL
EWK LT
EWK TT
a.u.
$\Delta\eta_{jj}$

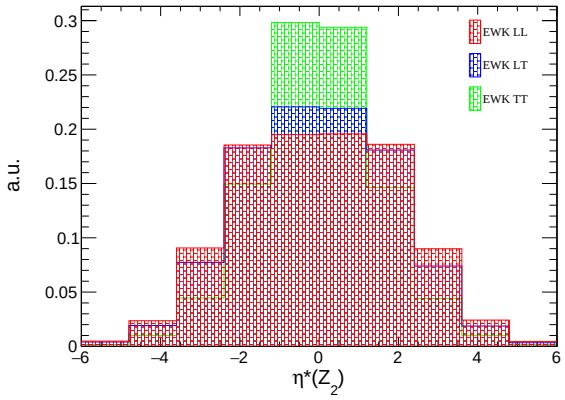

QCD qq
QCD gg
a.u.
$\Delta\eta_{jj}$

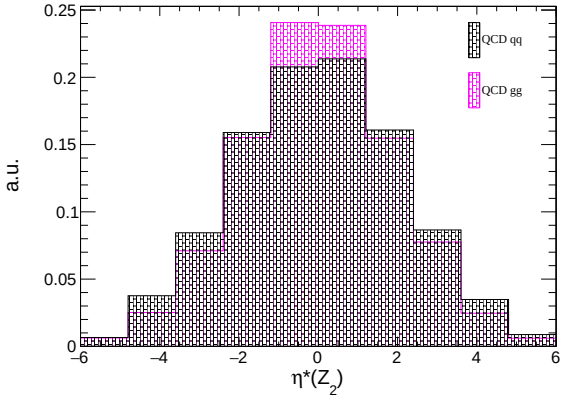

EWK LL
EWK LT
EWK TT
a.u.
m_{4l} [GeV]

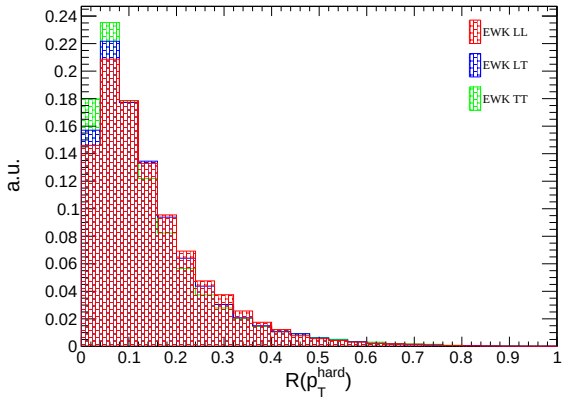

QCD qq
QCD gg
a.u.
m_{4l} [GeV]

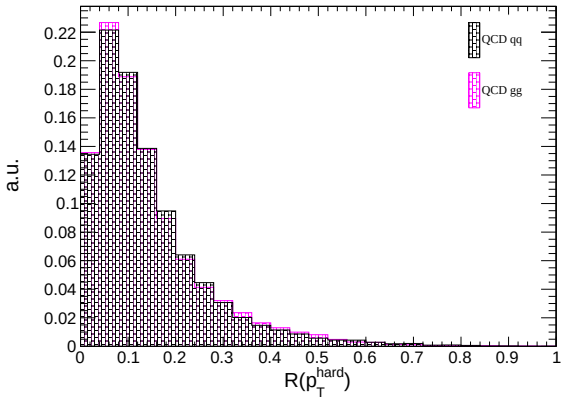

EWK LL
EWK LT
EWK TT
a.u.
η*(Z1)

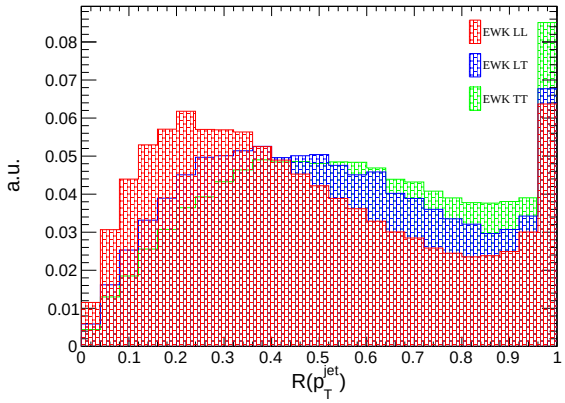

QCD qq
QCD gg
a.u.
η*(Z1)

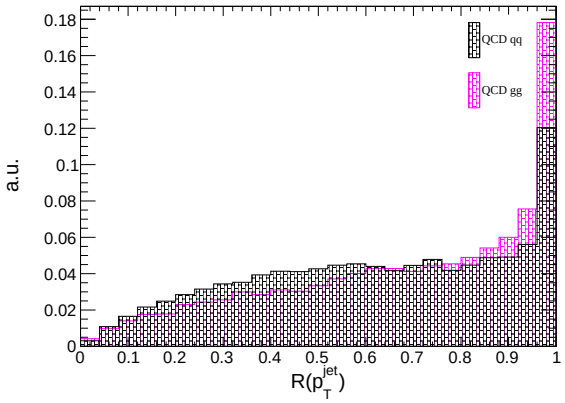

EWK LL
EWK LT
EWK TT
a.u.
η*(Z2)

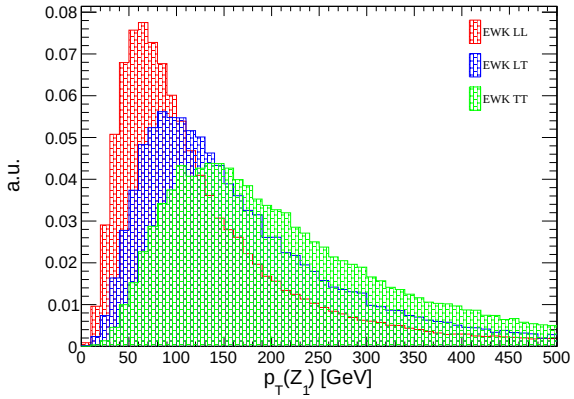

QCD qq
QCD gg
a.u.
η*(Z2)

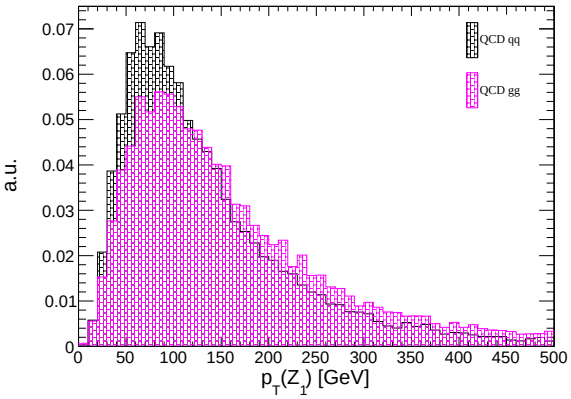

EWK LL
EWK LT
EWK TT
a.u.
R(pT hard)

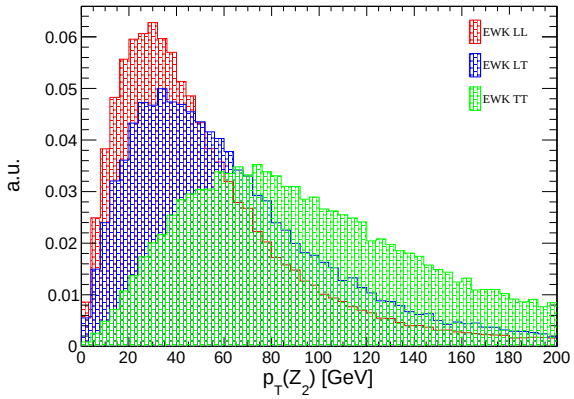

QCD qq
QCD gg
a.u.
R(pT hard)

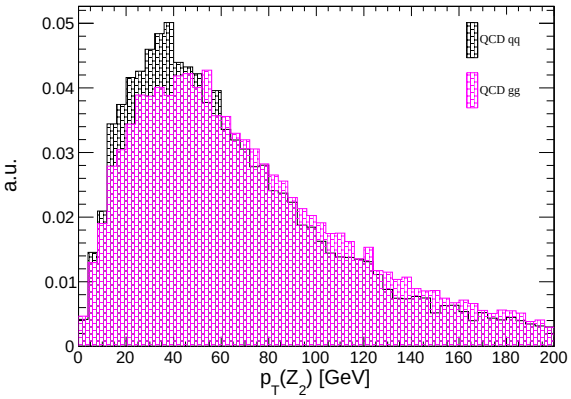

EWK LL
EWK LT
EWK TT
a.u.
R(pT jet)

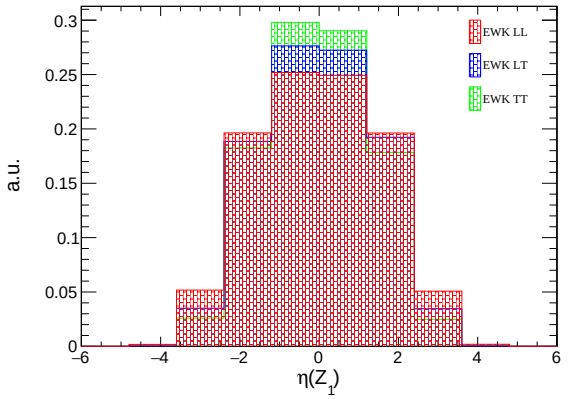

QCD qq
QCD gg
a.u.
R(pT jet)

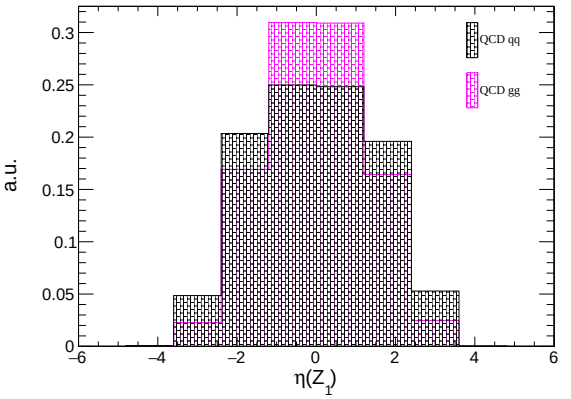

EWK LL
EWK LT
EWK TT
a.u.
$p_T(Z_1)$ [GeV]

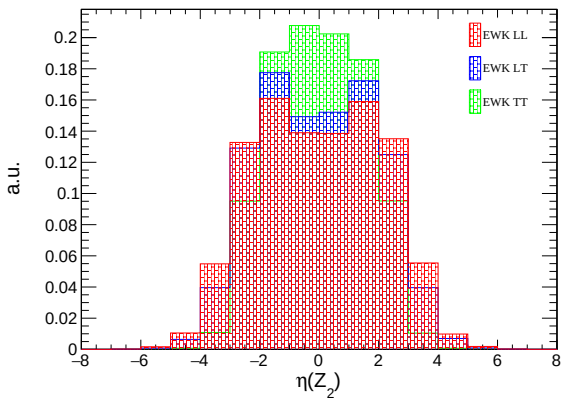

QCD qq
QCD gg
a.u.
$p_T(Z_1)$ [GeV]

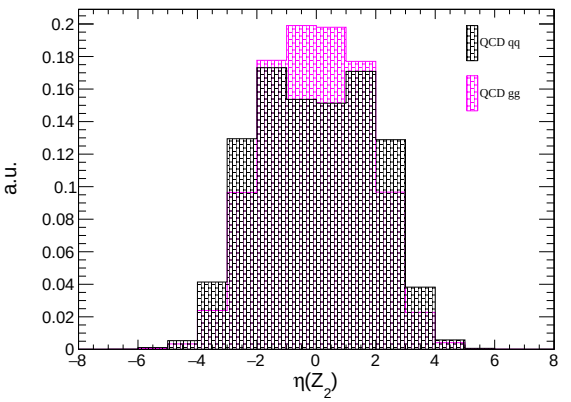

EWK LL
EWK LT
EWK TT
a.u.
$p_T(Z_2)$ [GeV]


QCD qq
QCD gg
a.u.
$p_T(Z_2)$ [GeV]


EWK LL
EWK LT
EWK TT
a.u.
η(Z1)


QCD qq
QCD gg
a.u.
η(Z1)


EWK LL
EWK LT
EWK TT
a.u.
η(Z2)


QCD qq
QCD gg
a.u.
η(Z2)

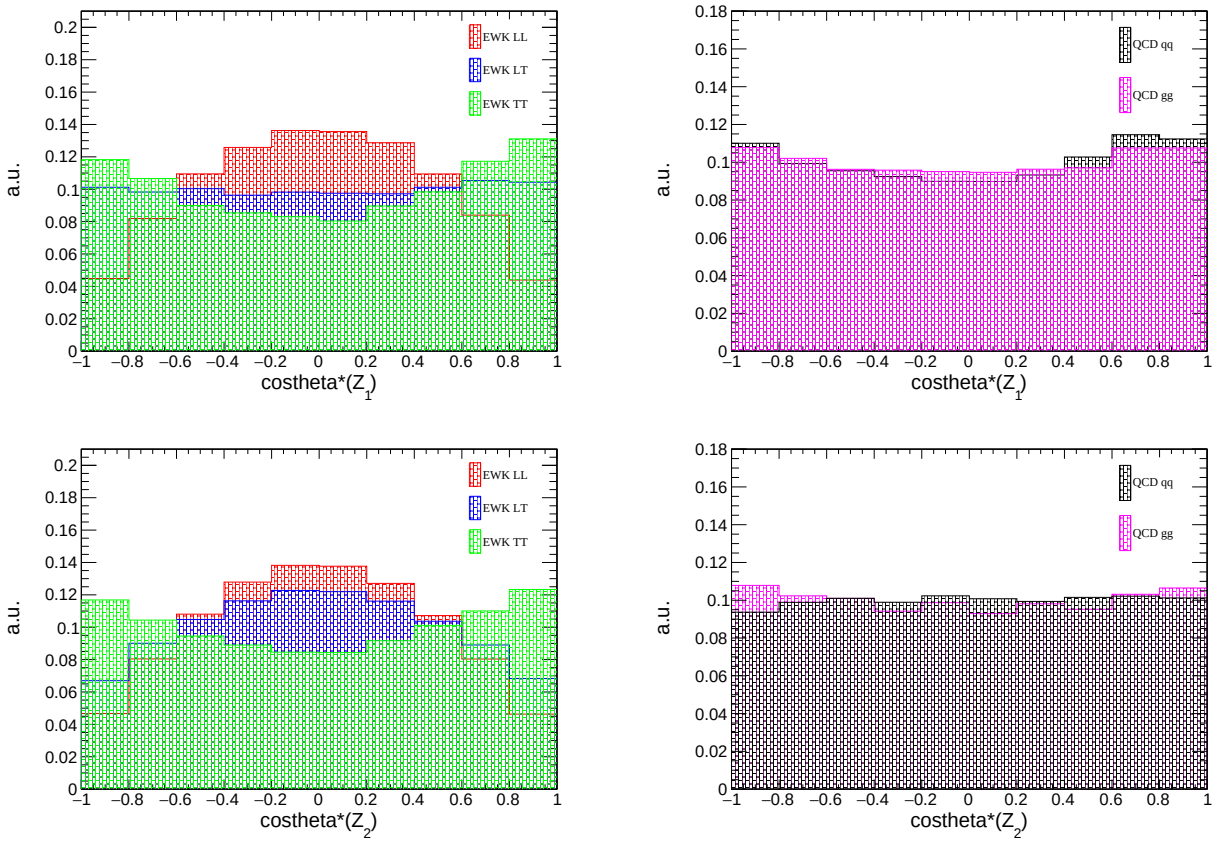
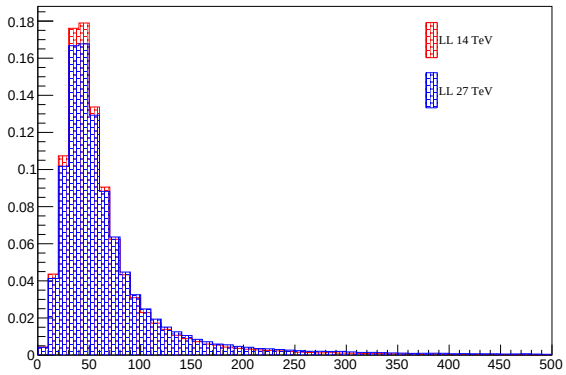
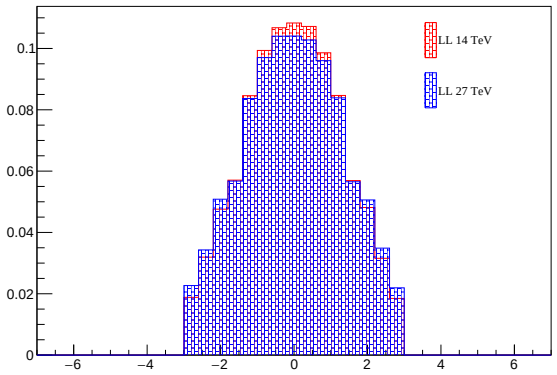


Figure B.2: Kinematics of VBS (left) and QCD (right) processes at 27 TeV after the baseline selection.

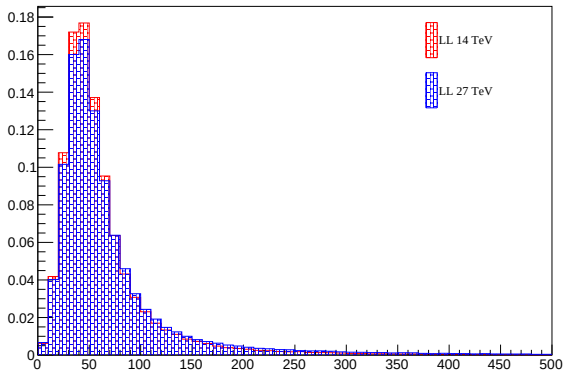
p_T(electrons) (PU200, LL)



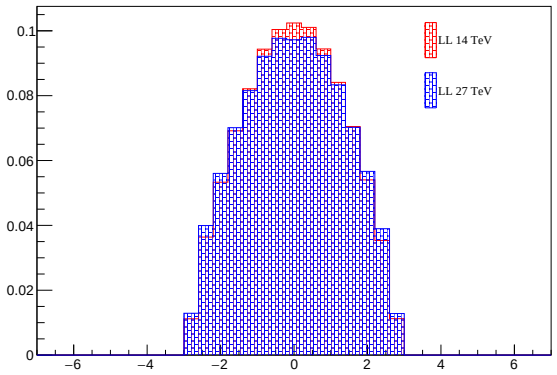
η(electrons) (PU200, LL)



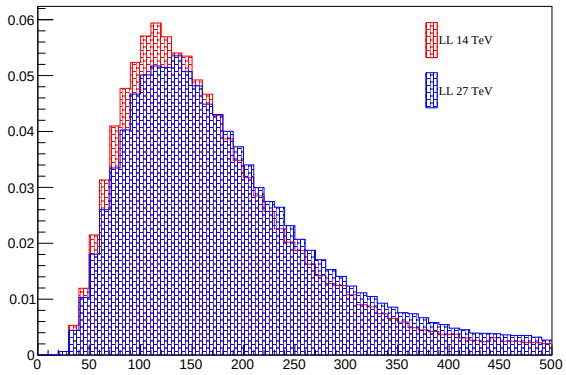
p_T(muons) (PU200, LL)



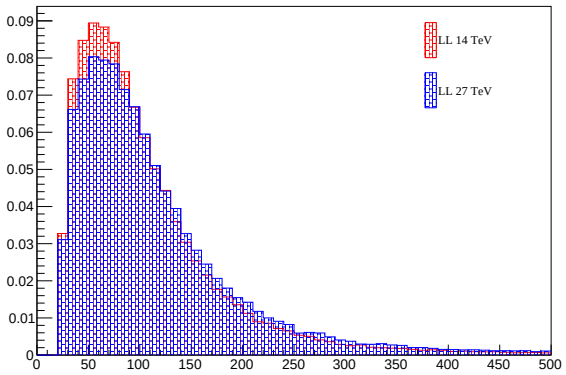
η(muons) (PU200, LL)



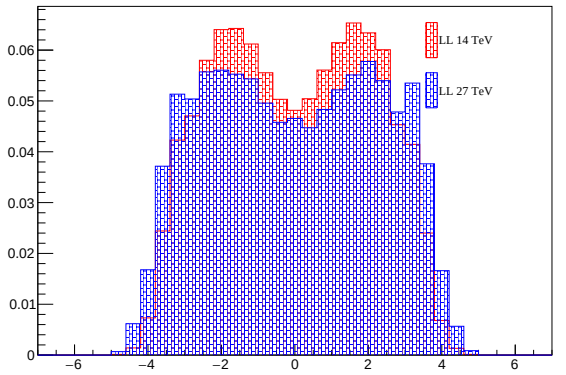
$p_T(j_1)$ (PU200, LL)



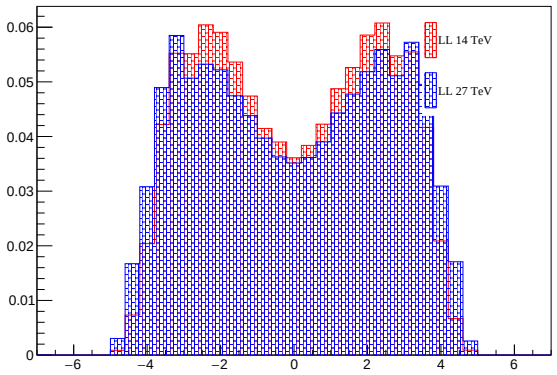
$p_T(j_2)$ (PU200, LL)



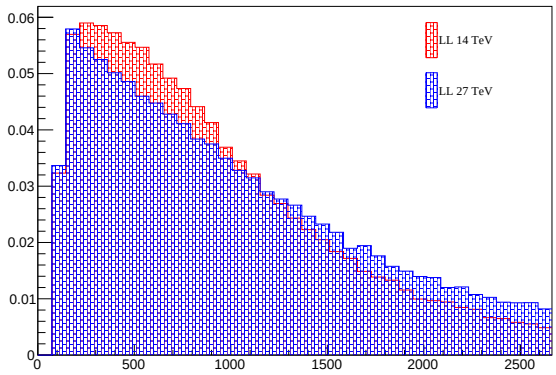
$\eta(j_1)$ (PU200, LL)



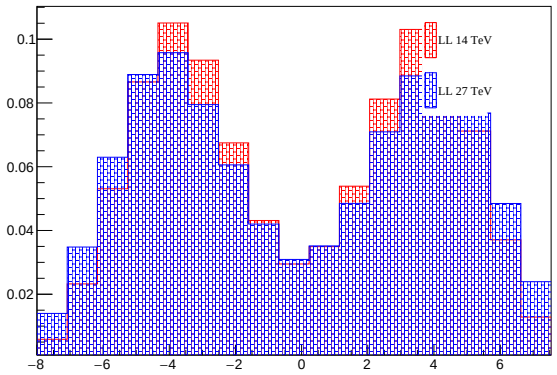
$\eta(j_2)$ (PU200, LL)



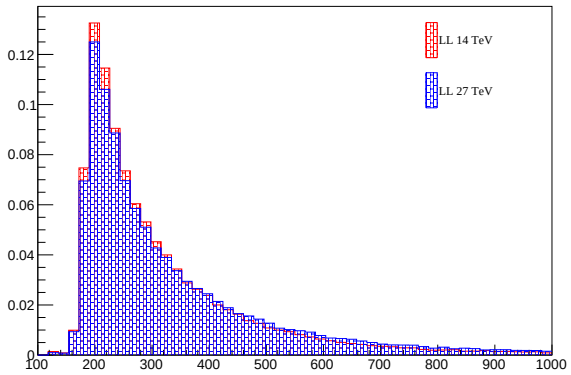
m_{jj} (PU200, LL)



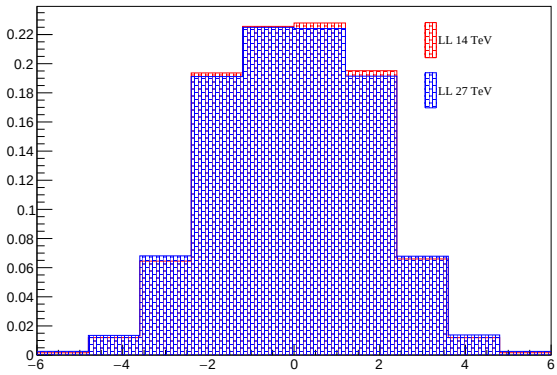
$deta_{jj}$ (PU200, LL)



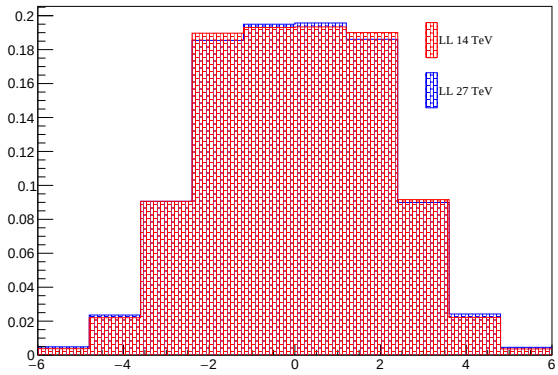
m_{4l} (PU200, LL)



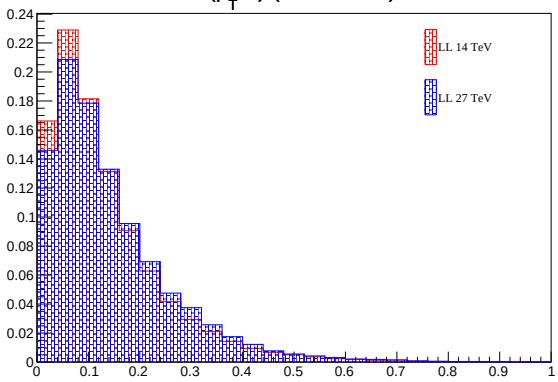
$\eta^*(Z_1)$ (PU200, LL)



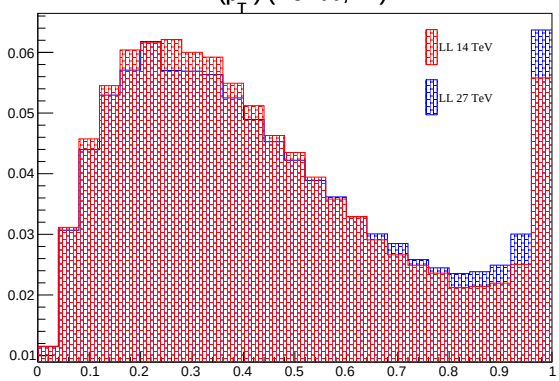
$\eta^*(Z_2)$ (PU200, LL)



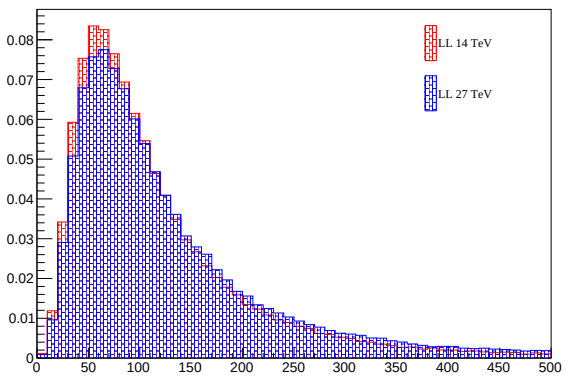
$R(p_T^{hard})$ (PU200, LL)



$R(p_T^{jet})$ (PU200, LL)



$pt(Z_1)$ (PU200, LL)



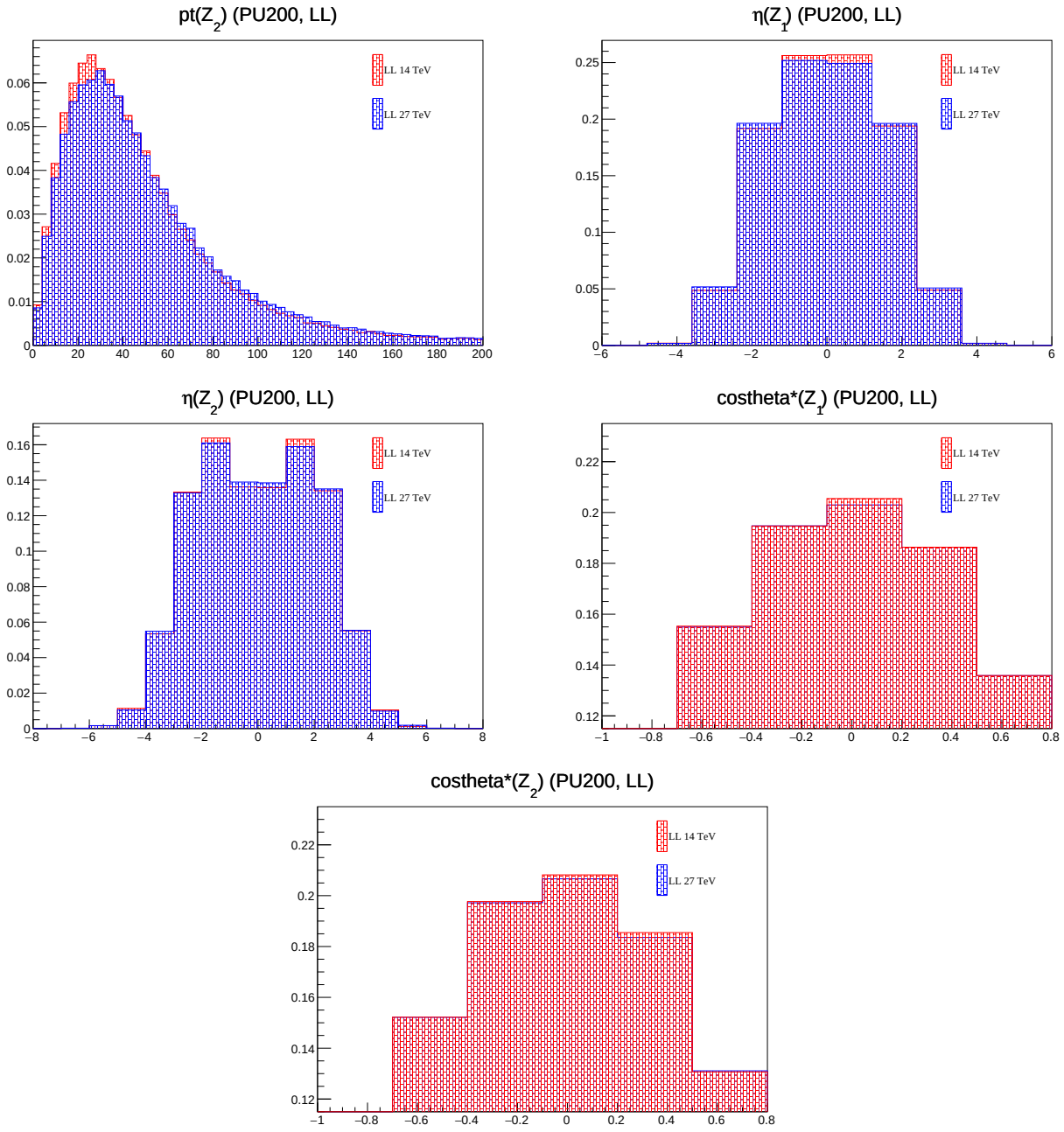
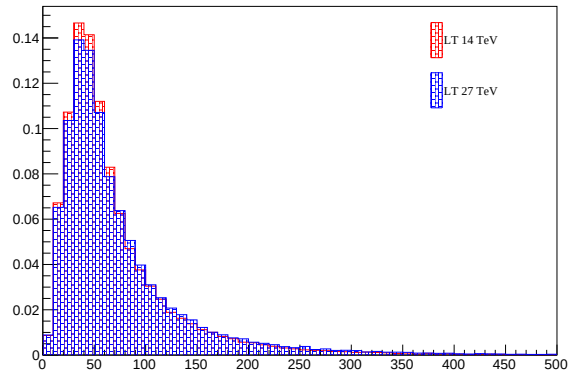
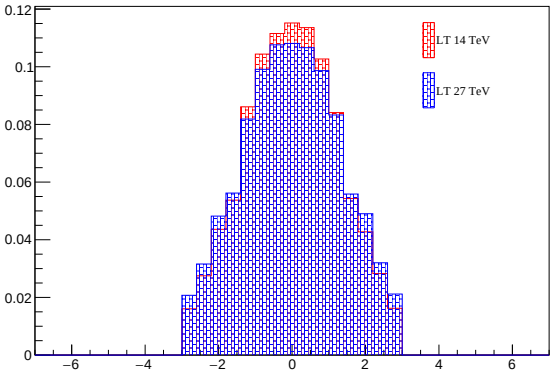


Figure B.3: Kinematics of the LL process at 14 and 27 TeV after the baseline selection. [will be updated]

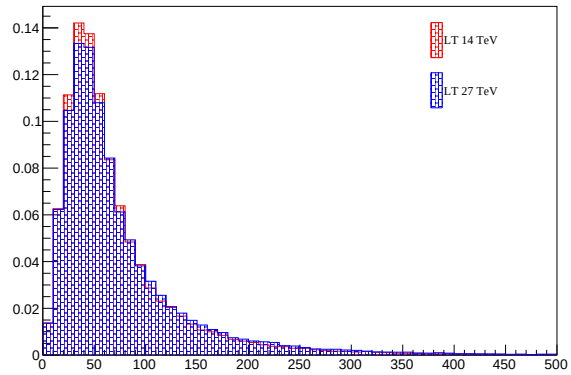
p_T(electrons) (PU200, LT)



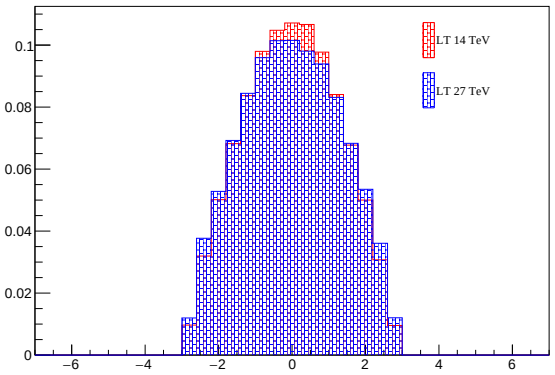
η(electrons) (PU200, LT)



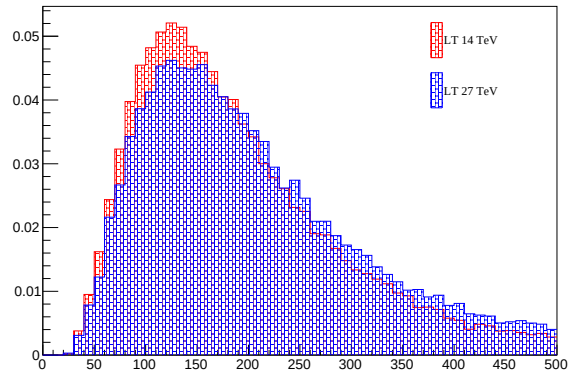
p_T(muons) (PU200, LT)



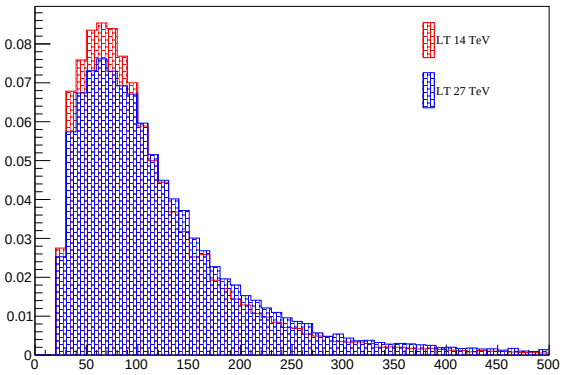
η(muons) (PU200, LT)



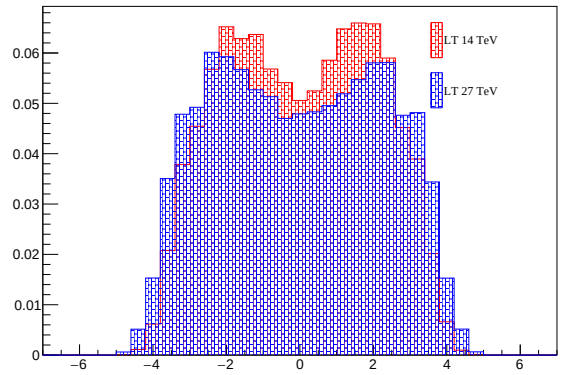
$p_T(j_1)$ (PU200, LT)



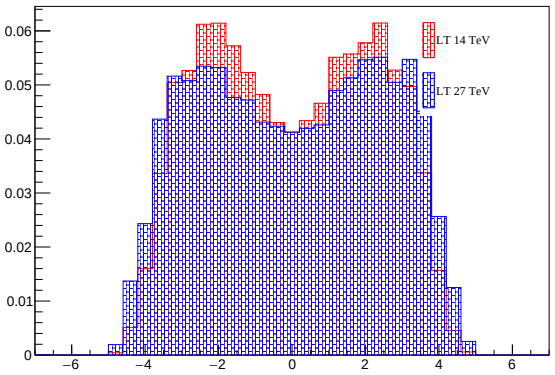
$p_T(j_2)$ (PU200, LT)



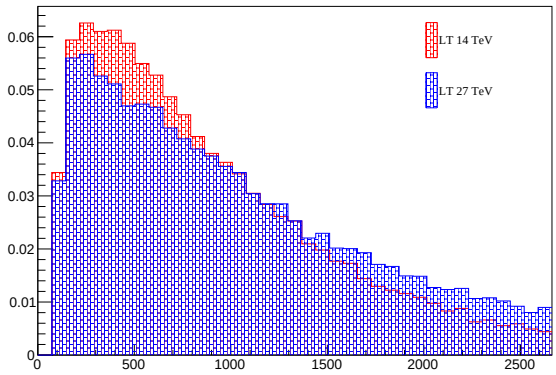
$\eta(j_1)$ (PU200, LT)



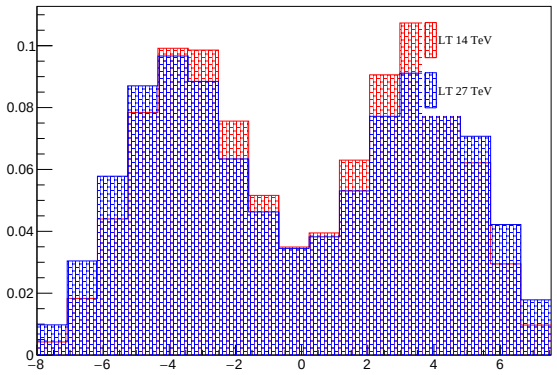
$\eta(j_2)$ (PU200, LT)



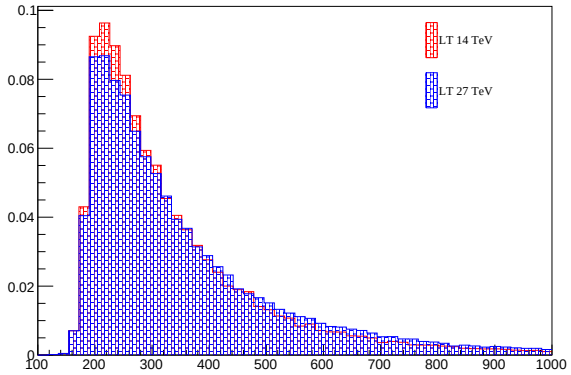
m_{jj} (PU200, LT)



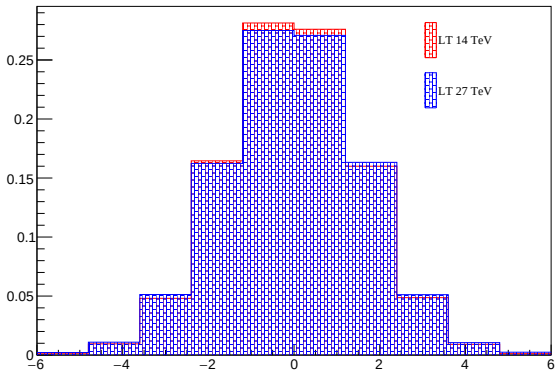
$deta_{jj}$ (PU200, LT)



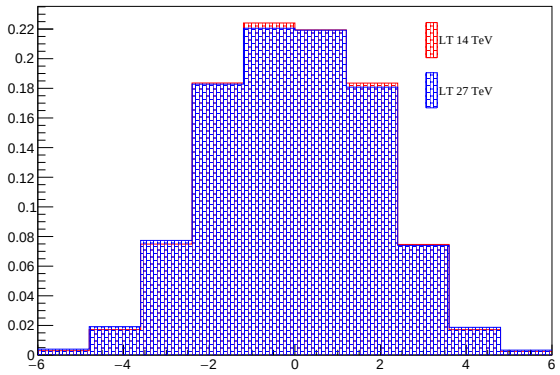
m_{4l} (PU200, LT)



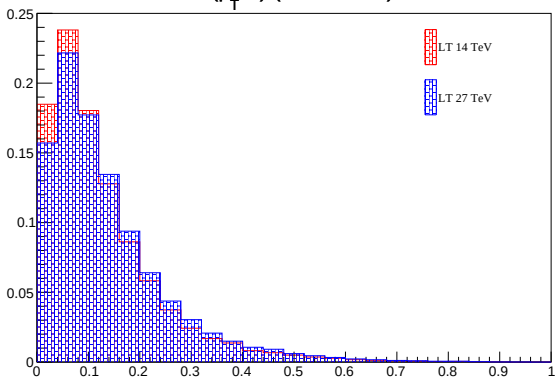
$\eta^*(Z_1)$ (PU200, LT)



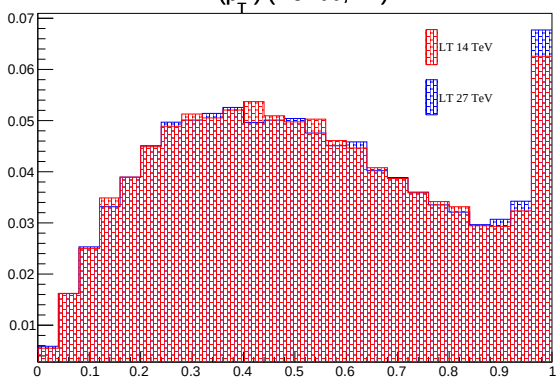
$\eta^*(Z_2)$ (PU200, LT)



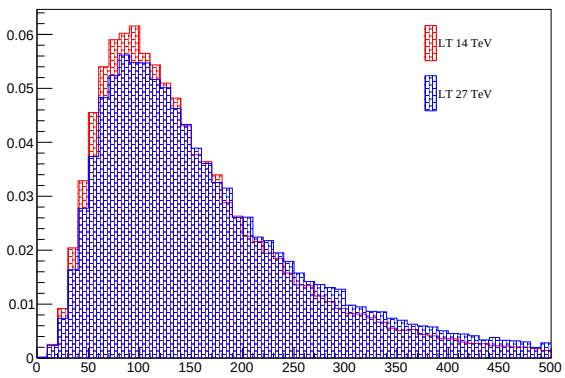
$R(p_T^{hard})$ (PU200, LT)



$R(p_T^{jet})$ (PU200, LT)



$pt(Z_1)$ (PU200, LT)



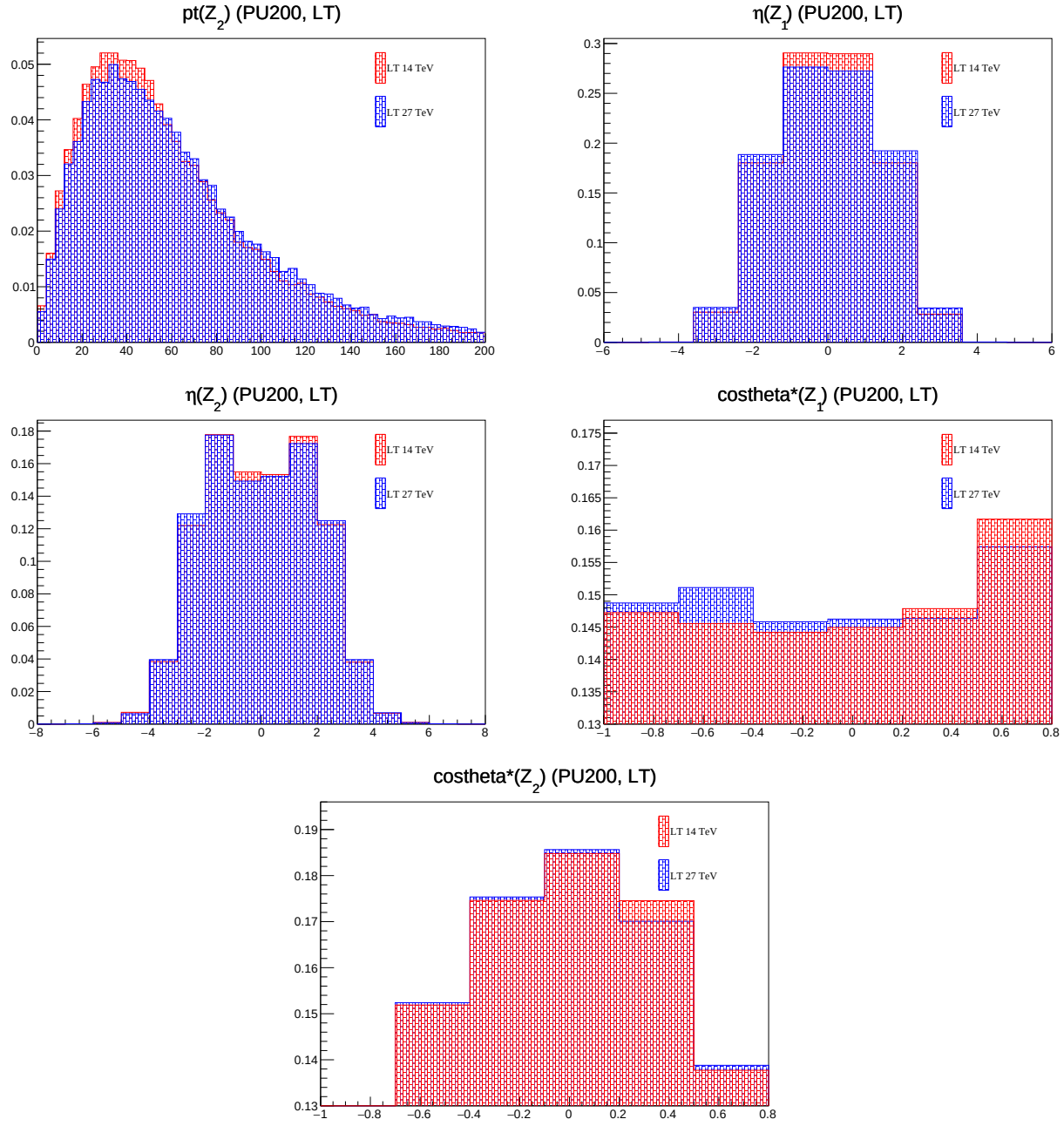
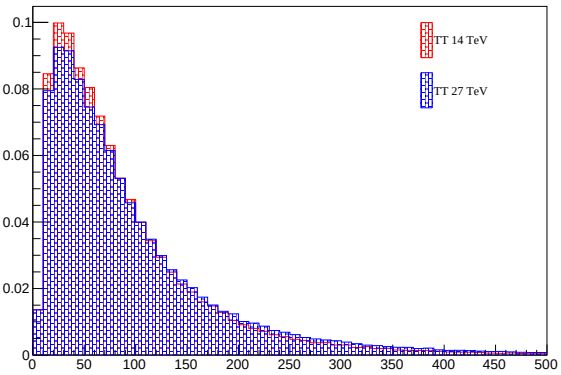
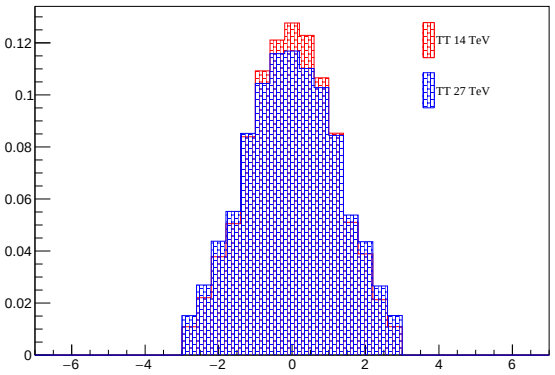


Figure B.4: Kinematics of the the LT process at 14 and 27 TeV after the baseline selection. [will be updated]

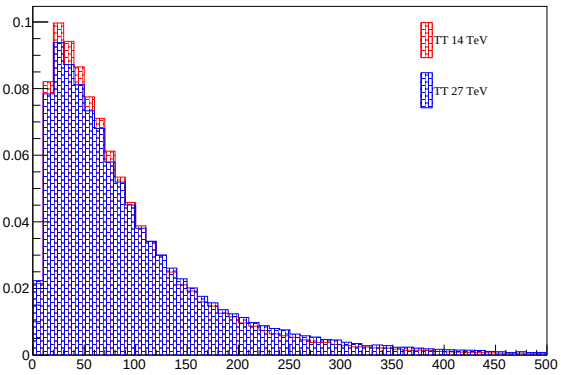
p_T(electrons) (PU200, TT)



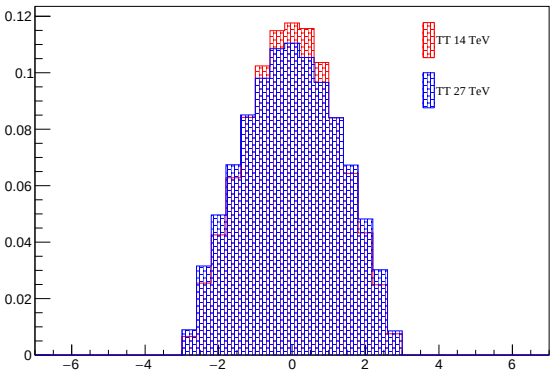
η(electrons) (PU200, TT)



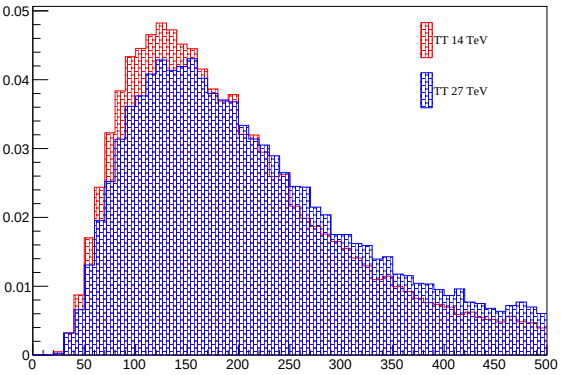
p_T(muons) (PU200, TT)



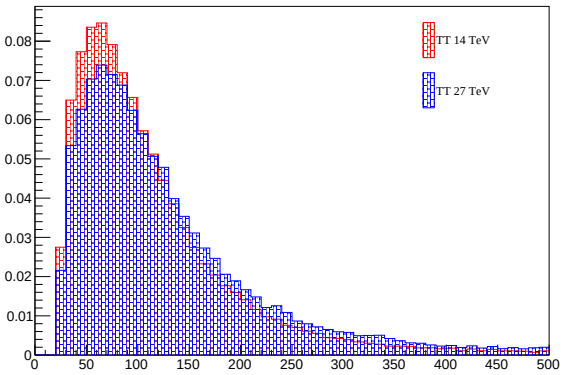
η(muons) (PU200, TT)



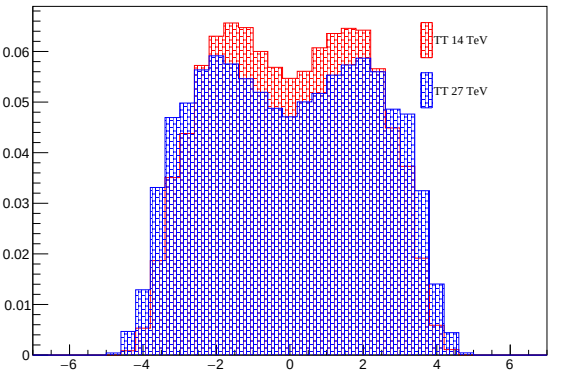
$p_T(j_1)$ (PU200, TT)



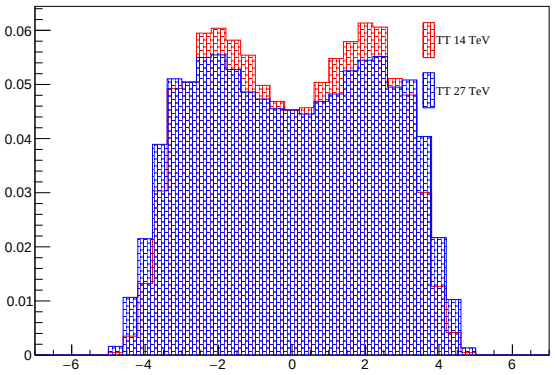
$p_T(j_2)$ (PU200, TT)



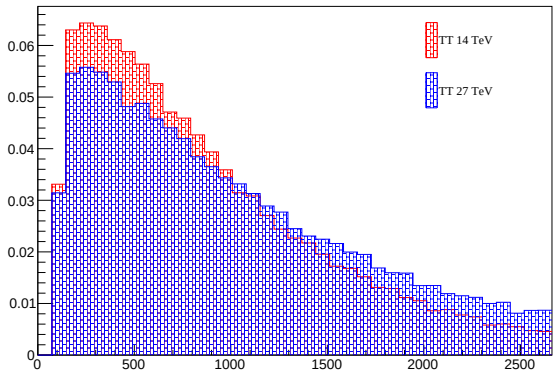
$\eta(j_1)$ (PU200, TT)



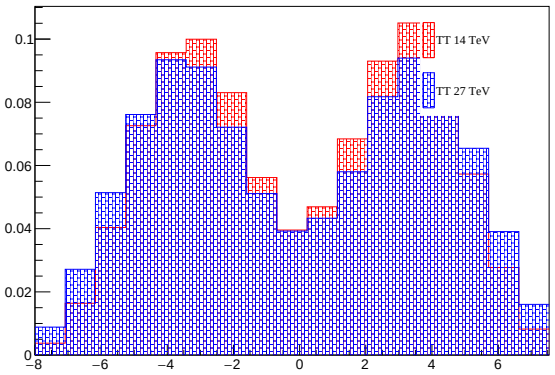
$\eta(j_2)$ (PU200, TT)



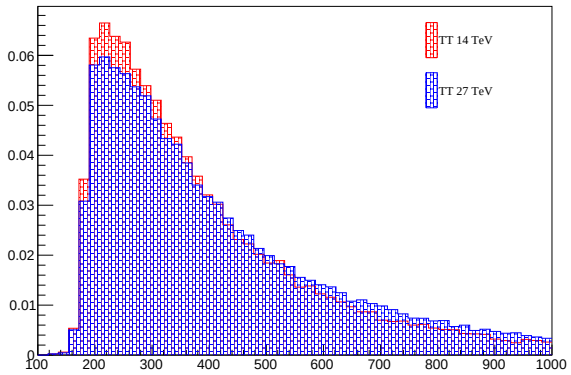
m_{jj} (PU200, TT)



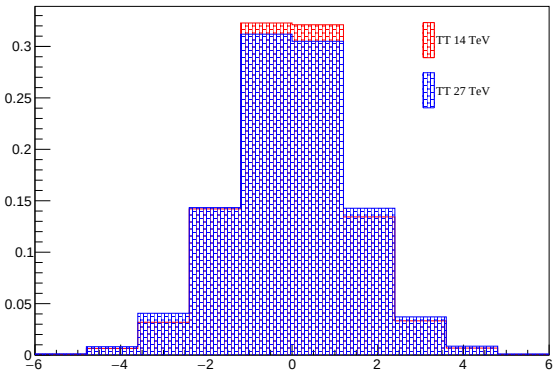
$deta_{jj}$ (PU200, TT)



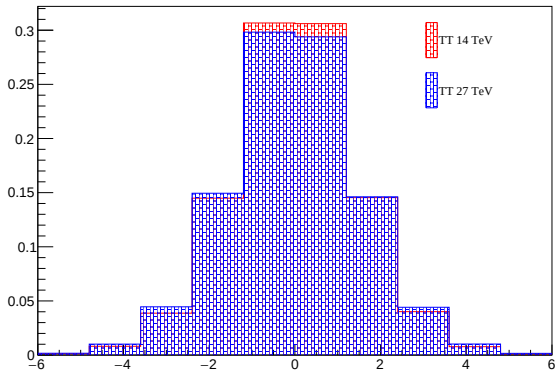
m_{4l} (PU200, TT)



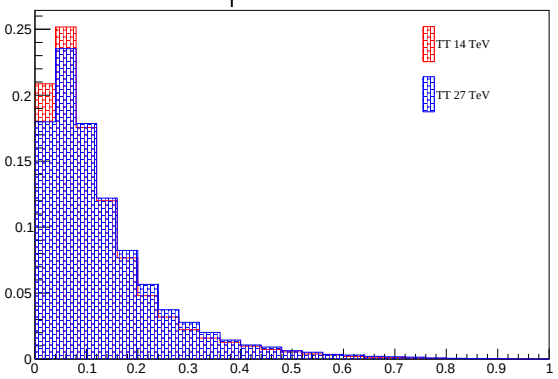
$\eta^*(Z_1)$ (PU200, TT)



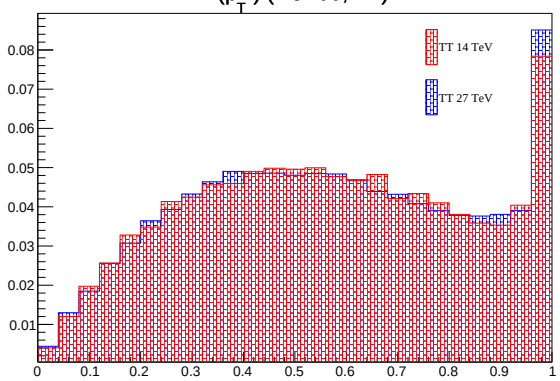
$\eta^*(Z_2)$ (PU200, TT)



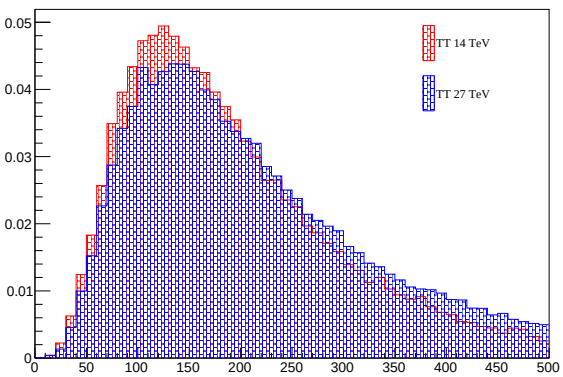
$R(p_T^{hard})$ (PU200, TT)



$R(p_T^{jet})$ (PU200, TT)



$pt(Z_1)$ (PU200, TT)



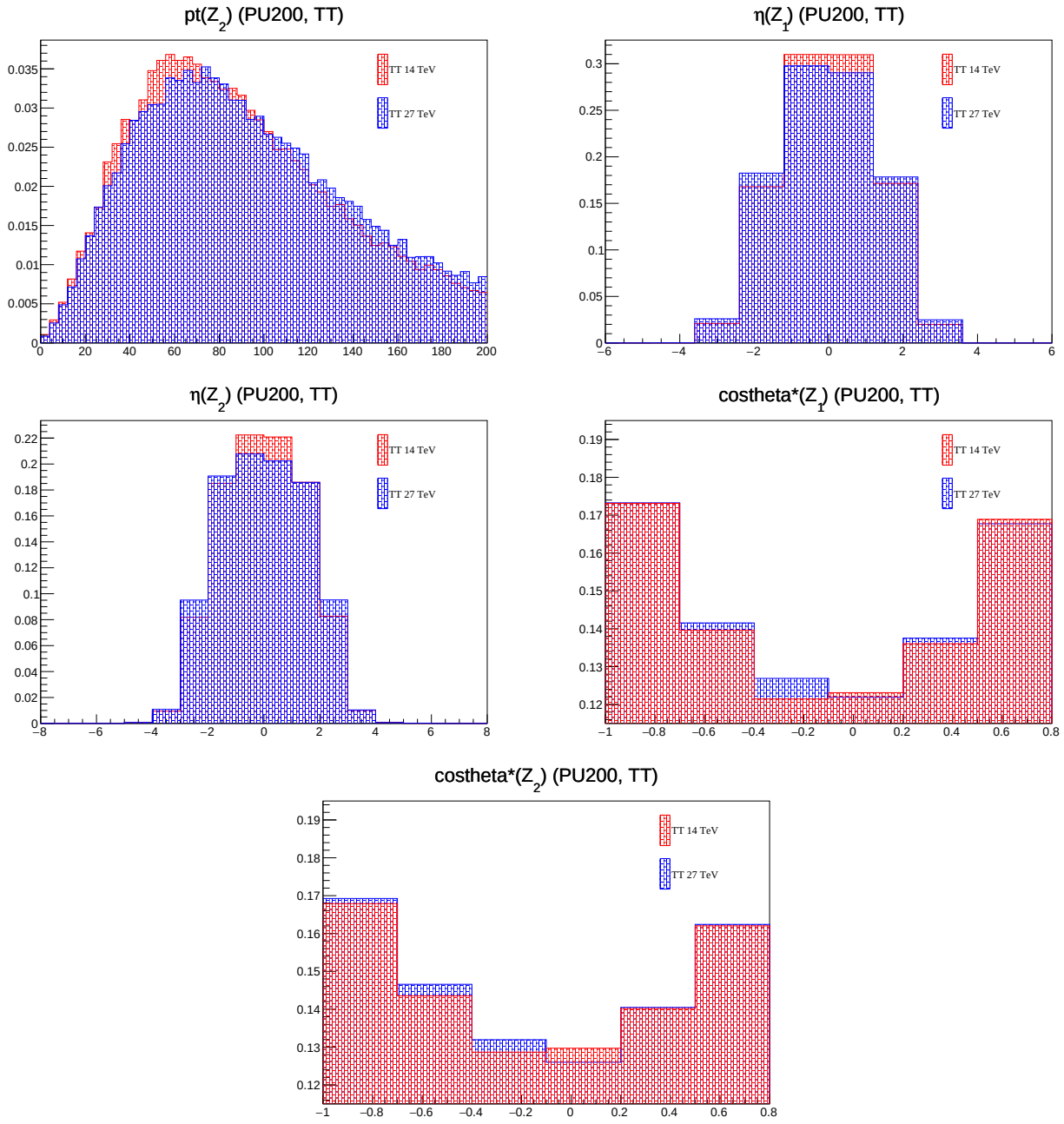
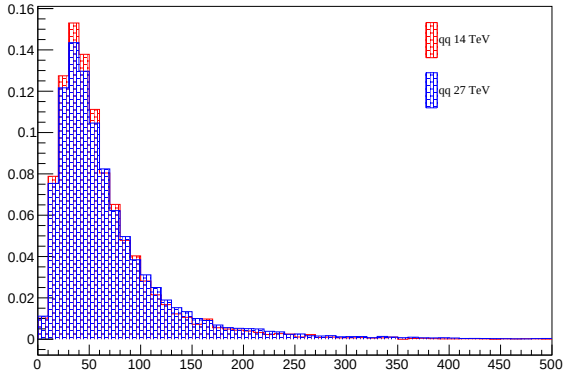
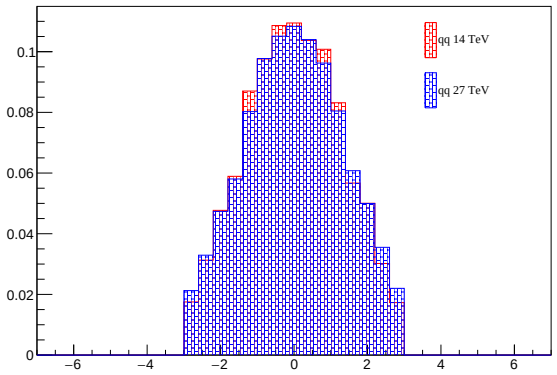


Figure B.5: Kinematics of the TT process at 14 and 27 TeV after the baseline selection. [will be updated]

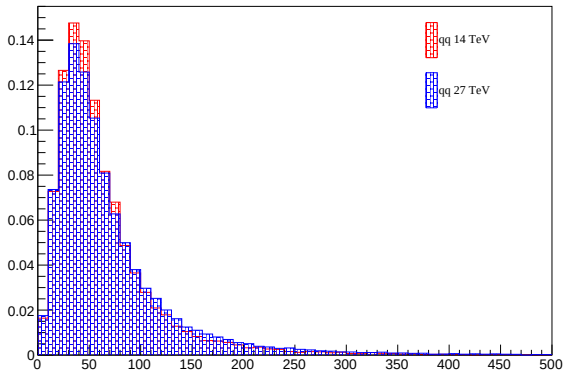
p_T(electrons) (PU200, qq)



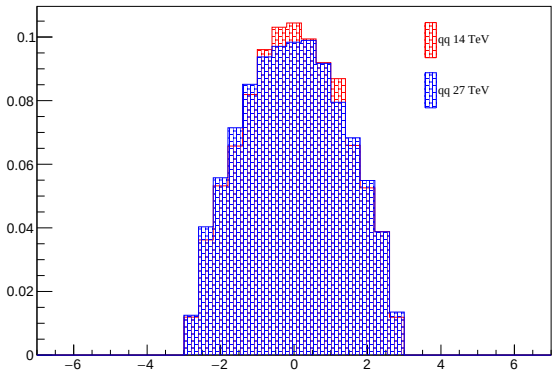
η(electrons) (PU200, qq)



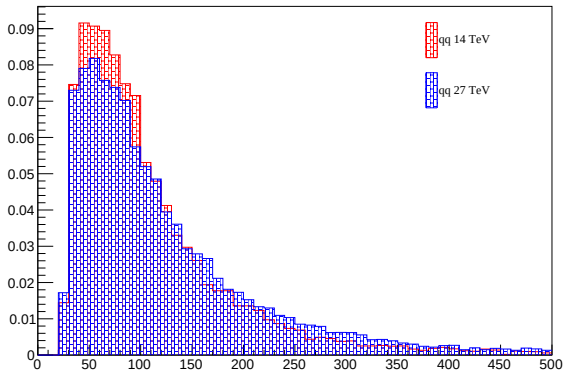
p_T(muons) (PU200, qq)



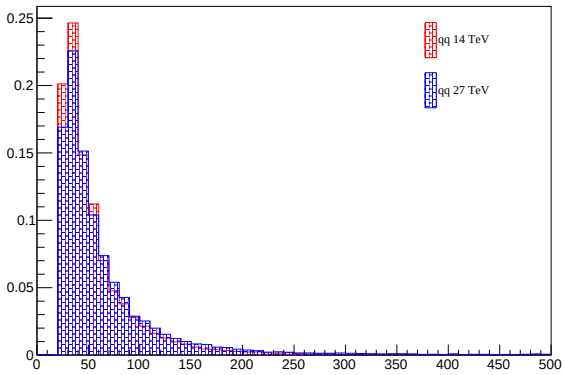
η(muons) (PU200, qq)



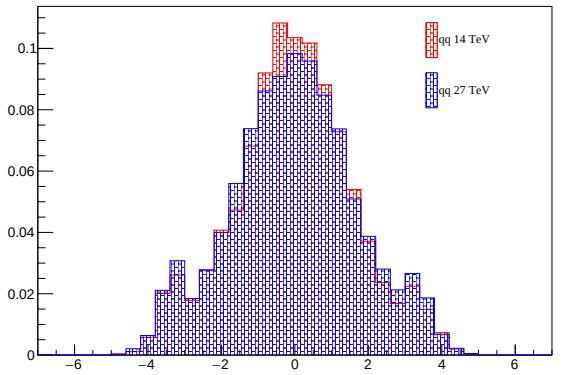
$p_T(j_1)$ (PU200, qq)



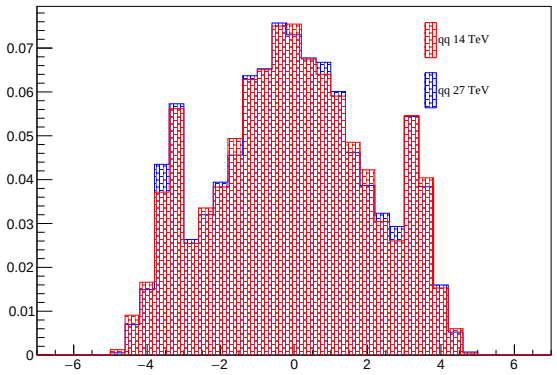
$p_T(j_2)$ (PU200, qq)



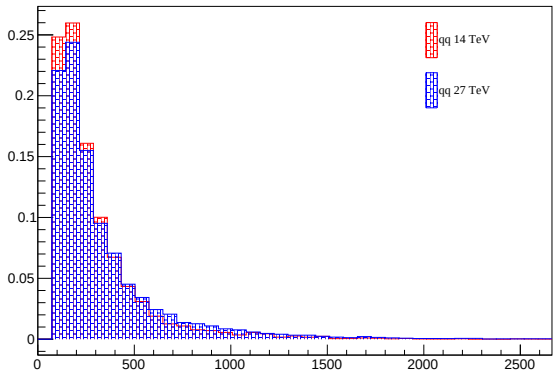
η(j_1) (PU200, qq)



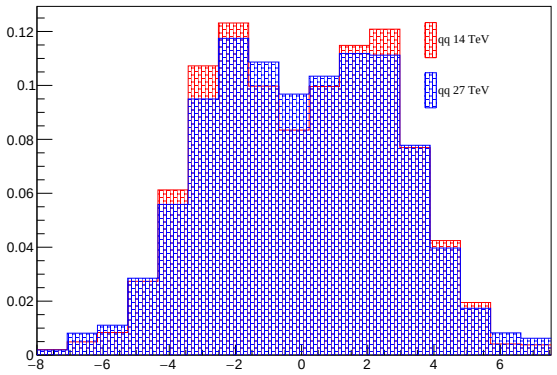
η(j_2) (PU200, qq)



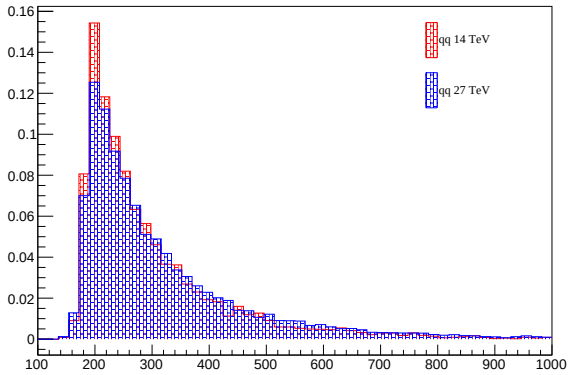
m_{jj} (PU200, qq)



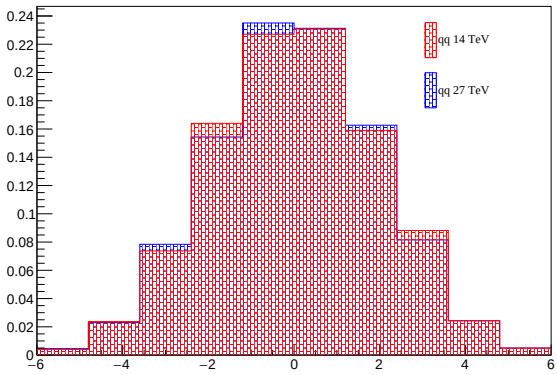
$deta_{jj}$ (PU200, qq)



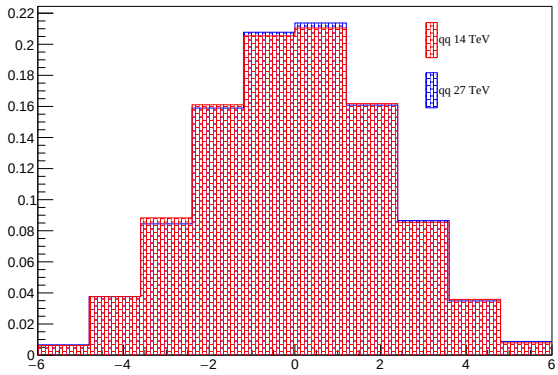
m_{4l} (PU200, qq)



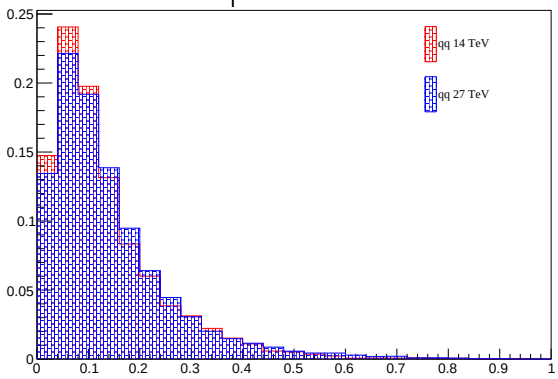
$\eta^*(Z_1)$ (PU200, qq)



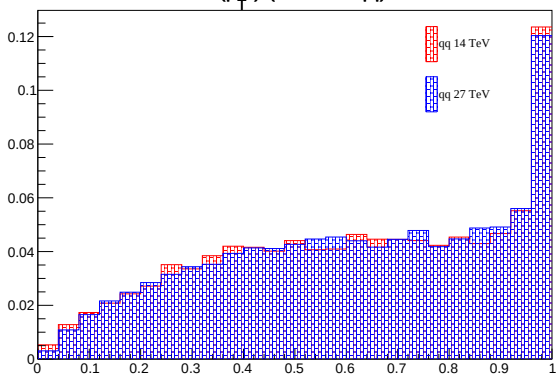
$\eta^*(Z_2)$ (PU200, qq)



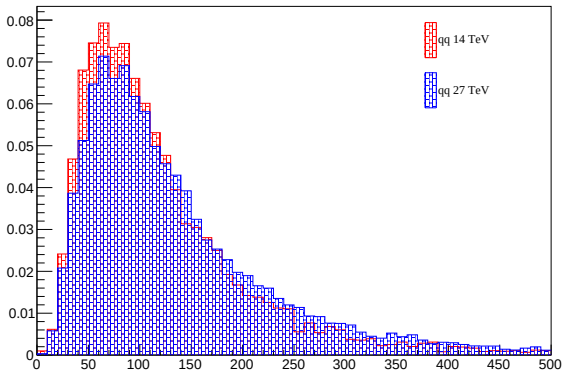
$R(p_T^{hard})$ (PU200, qq)



$R(p_T^{jet})$ (PU200, qq)



$pt(Z_1)$ (PU200, qq)



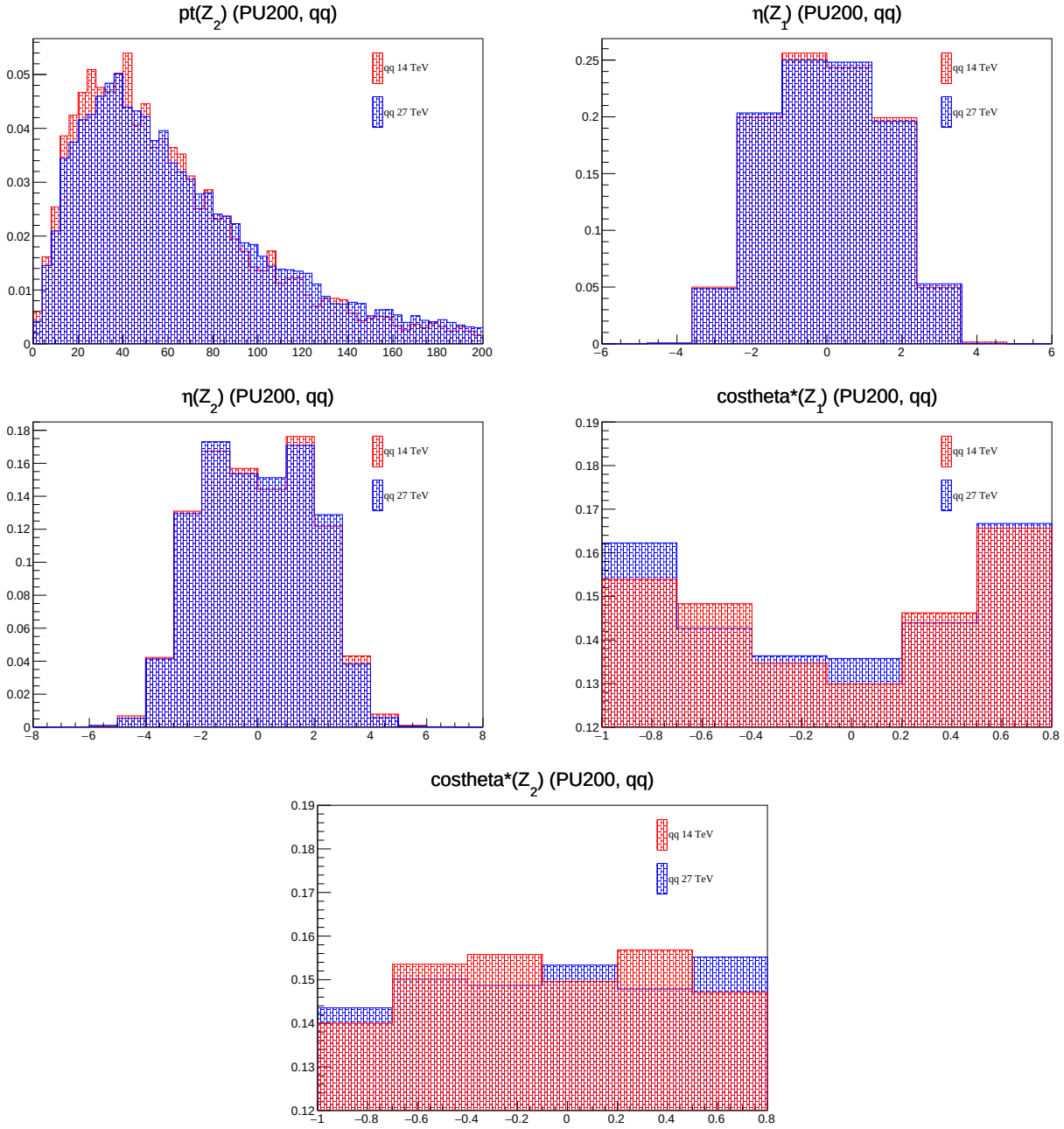
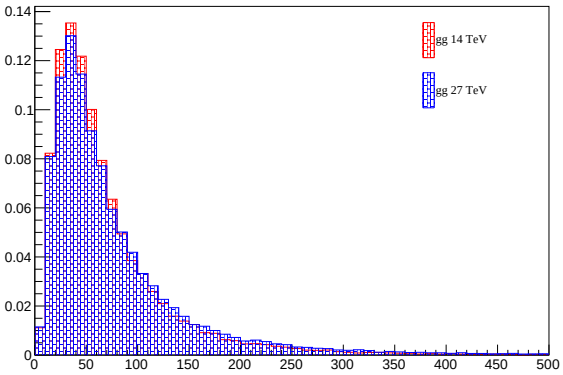
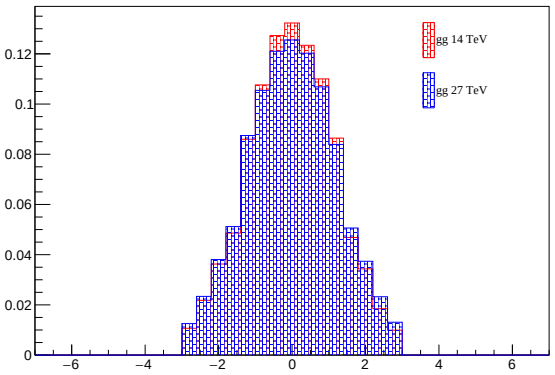


Figure B.6: Kinematics of the qq process at 14 and 27 TeV after the baseline selection. [will be updated]

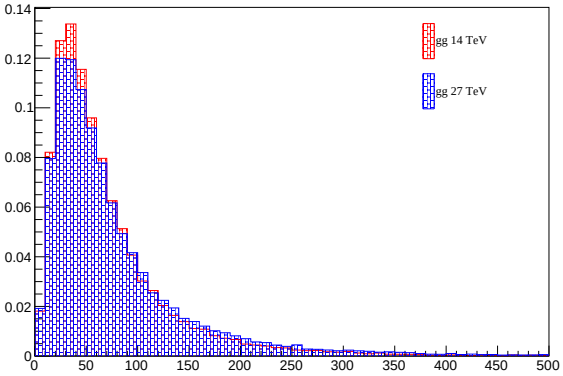
p_T(electrons) (PU200, gg)



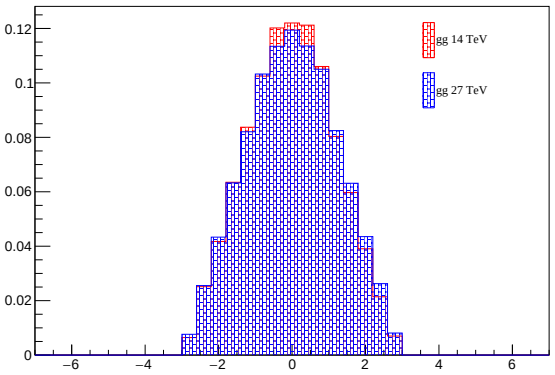
η(electrons) (PU200, gg)



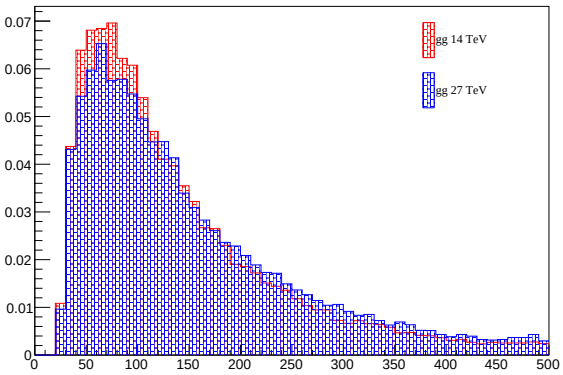
p_T(muons) (PU200, gg)



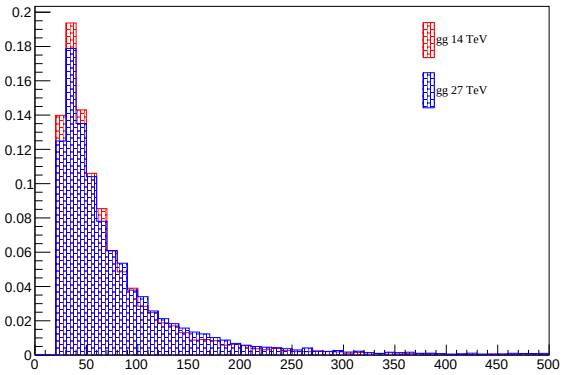
η(muons) (PU200, gg)



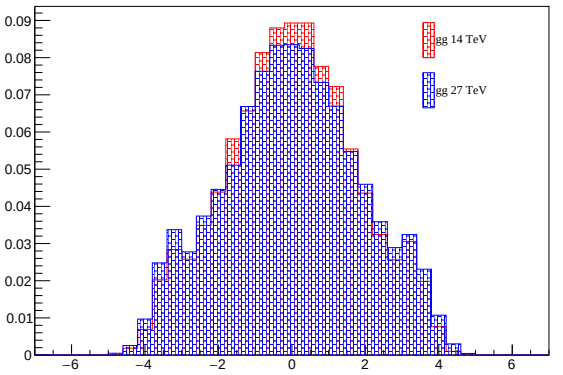
$p_T(j_1)$ (PU200, gg)



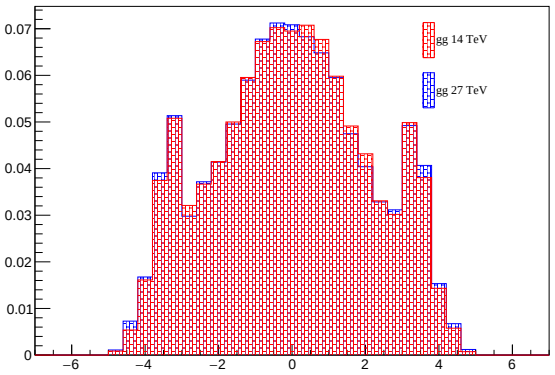
$p_T(j_2)$ (PU200, gg)



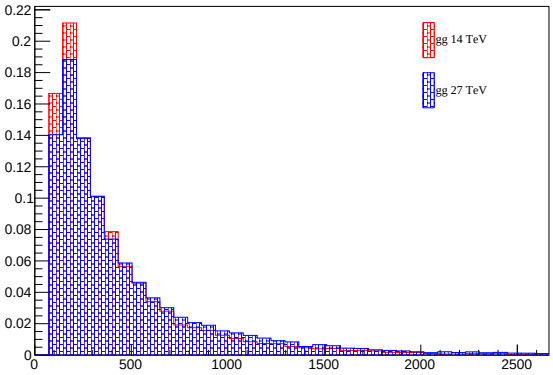
η(j_1) (PU200, gg)



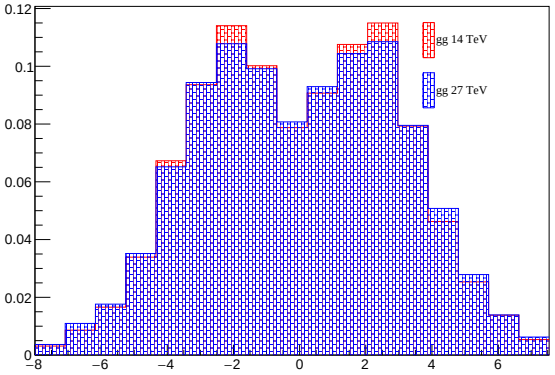
η(j_2) (PU200, gg)



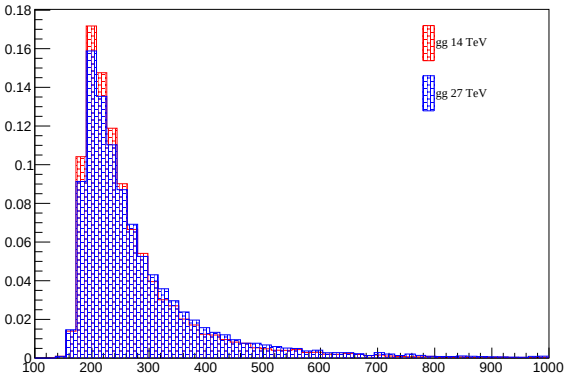
m_{jj} (PU200, gg)



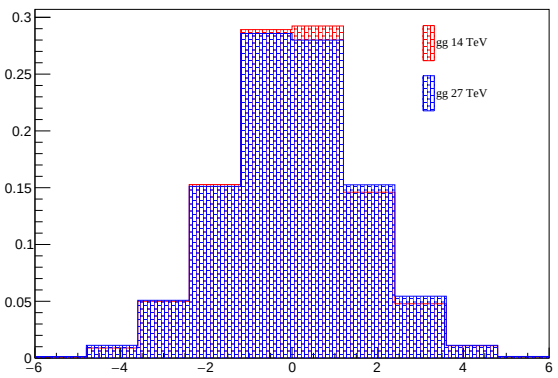
$deta_{jj}$ (PU200, gg)



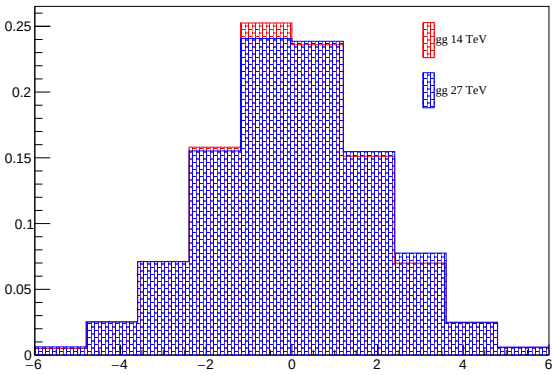
m_{4l} (PU200, gg)



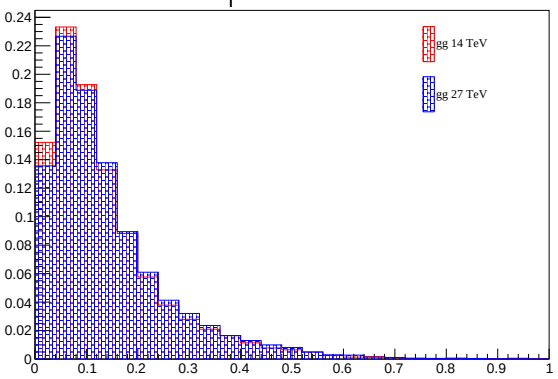
$\eta^*(Z_1)$ (PU200, gg)



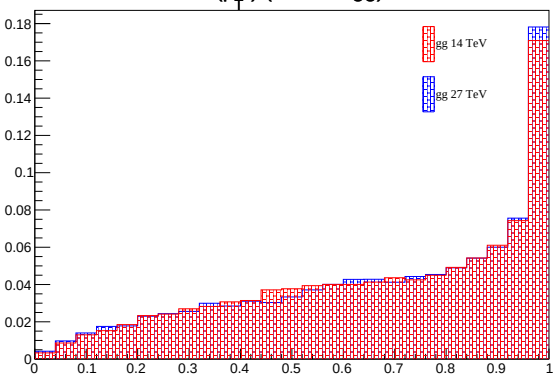
$\eta^*(Z_2)$ (PU200, gg)



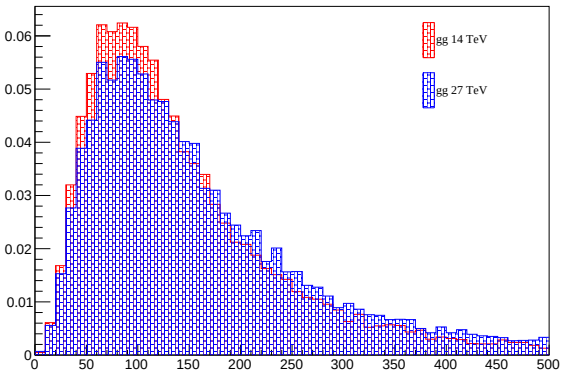
$R(p_T^{hard})$ (PU200, gg)



$R(p_T^{jet})$ (PU200, gg)



$pt(Z_1)$ (PU200, gg)



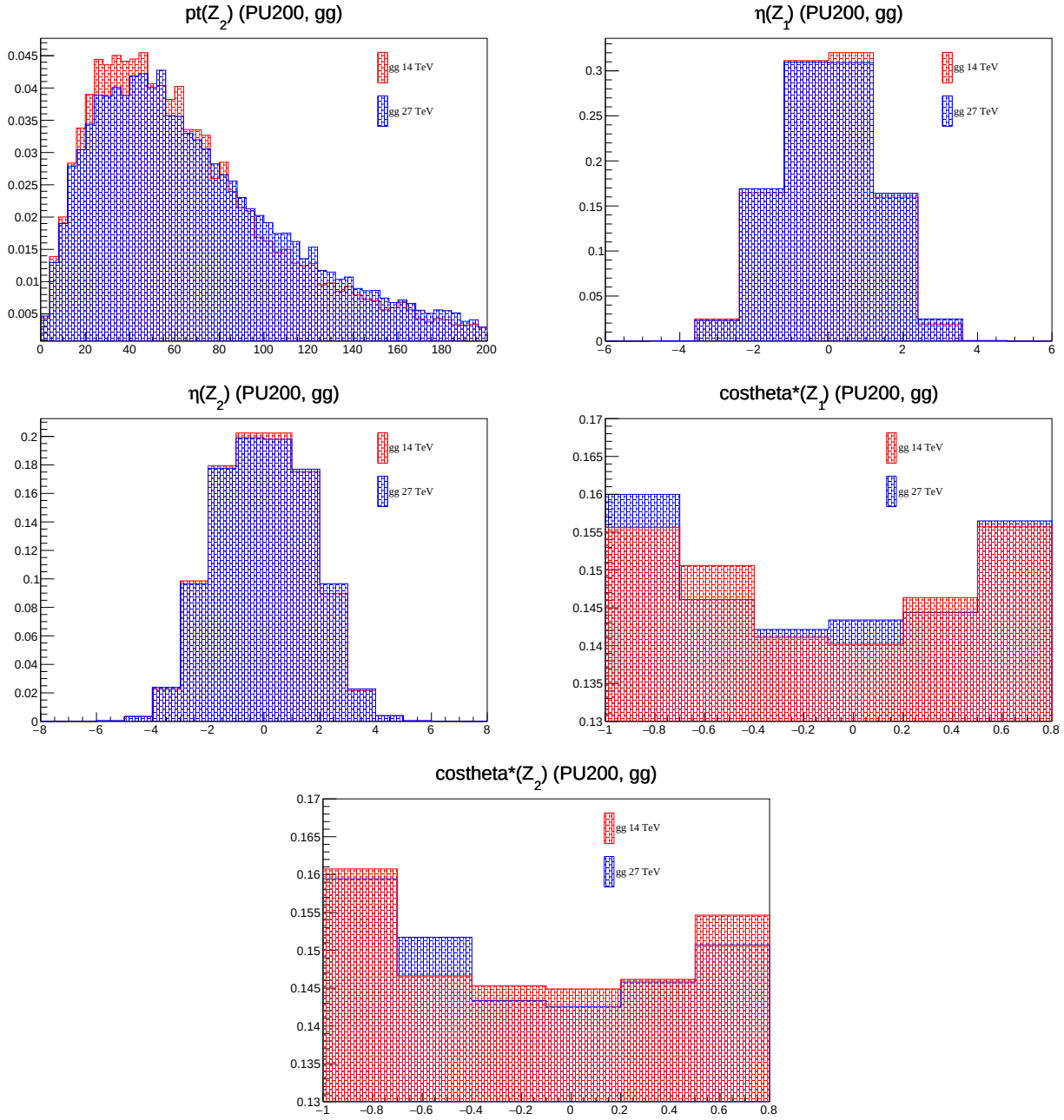


Figure B.7: Kinematics of the gg process at 14 and 27 TeV after the baseline selection. [will be updated]

Bibliography

[1] J. M. Cornwall, D. N. Levin, and G. Tiktopoulos, "Derivation of gauge invariance from high-energy unitarity bounds on the s matrix," *Phys. Rev. D*, vol. 10, pp. 1145–1167, Aug 1974.

[2] B. W. Lee, C. Quigg, and H. B. Thacker, "Weak interactions at very high energies: The role of the higgs-boson mass," *Phys. Rev. D*, vol. 16, pp. 1519–1531, Sep 1977.

[3] M. S. Chanowitz and M. K. Gaillard, "The tev physics of strongly interacting w's and z's," *Nuclear Physics B*, vol. 261, pp. 379–431, 1985.

[4] J. Brehmer, "Polarised WW Scattering at the LHC," Master's thesis, U. Heidelberg, ITP, 2014.

[5] B. W. Lee, C. Quigg, and H. B. Thacker, "Strength of weak interactions at very high energies and the higgs boson mass," *Phys. Rev. Lett.*, vol. 38, pp. 883–885, Apr 1977.

[6] A. Ballestrero, B. Biedermann, *et al.*, "Precise predictions for same-sign w-boson scattering at the LHC," *The European Physical Journal C*, vol. 78, aug 2018.

[7] D. Rainwater, R. Szalapski, and D. Zeppenfeld, "Probing color-singlet exchange in z + 2-jet events at the LHC," vol. 54, pp. 6680–6689, dec 1996.

[8] R. Gomez-Ambrosio, "Study of vbf/vbs in the lhc at 13 tev, the eft approach," 2016.

[9] C. Degrande, N. Greiner, W. Kilian, O. Mattelaer, H. Mebane, T. Stelzer, S. Willenbrock, and C. Zhang, "Effective field theory: A modern approach to anomalous couplings," *Annals of Physics*, vol. 335, pp. 21–32, aug 2013.

[10] G. Perez, M. Sekulla, and D. Zeppenfeld, "Anomalous quartic gauge couplings and unitarization for the vector boson scattering process $pp \to w^+w^+jjx \to l^+\nu_l l^+\nu_l jjx$," *The European Physical Journal C*, vol. 78, sep 2018.

[11] A. Dedes, P. Kozów, and M. Szleper, "Standard model eft effects in vector-boson scattering at the lhc," *Phys. Rev. D*, vol. 104, p. 013003, Jul 2021.

[12] S. Weinberg, "Baryon- and lepton-nonconserving processes," *Phys. Rev. Lett.*, vol. 43, pp. 1566–1570, Nov 1979.

[13] M. Rauch, "Vector-boson fusion and vector-boson scattering," 2016.

[14] O. J. P. É boli, M. C. Gonzalez-Garcia, and J. K. Mizukoshi, "$p\,p \to j\,j\,e+/-\,\mu+/-\,\nu\,\nu$ and $j\,j\,e+/-\,\mu+/-\,\nu\,\nu$ at $\mathcal{O}(\alpha_{em}^6)$ and $\mathcal{O}(\alpha_{em}^4\,\alpha_s^2)$ for the study of the Quartic Electroweak Gauge Boson Vertex at LHC," *Physical Review D*, vol. 74, oct 2006.

[15] M. Rauch, "Vector-boson fusion and vector-boson scattering," 2016.

[16] The CMS Collaboration, "Study of vector boson scattering and search for new physics in events with two same-sign leptons and two jets," *Physical Review Letters*, vol. 114, feb 2015.

[17] The ATLAS Collaboration, "Evidence for Electroweak Production of $W^{\pm}W^{pm}jj$ in pp Collisions at $\sqrt{s} = 8\ TeV$ with the ATLAS Detector," *Physical Review Letters*, vol. 113, oct 2014.

[18] The ATLAS Collaboration, "Measurement of $W^{\pm}Z$ production cross sections in pp collisions at $\sqrt{s} = 8\ TeV$ with the ATLAS detector and limits on anomalous gauge boson self-couplings," *Physical Review D*, vol. 93, may 2016.

[19] The ATLAS Collaboration, "Measurement of $w^{\pm}w^{\pm}$ vector-boson scattering and limits on anomalous quartic gauge couplings with the ATLAS detector," *Physical Review D*, vol. 96, jul 2017.

[20] The CMS Colaboration, "Measurement of electroweak-induced production of $w\gamma$ with two jets in pp collisions at $\sqrt{s} = 8\ tev$ and constraints on anomalous quartic gauge couplings," *Journal of High Energy Physics*, vol. 2017, jun 2017.

[21] The CMS Collaboration, "Measurement of the cross section for electroweak production of z gamma in association with two jets and constraints on anomalous quartic gauge couplings in proton-proton collisions at $\sqrt{s} = 8\ tev$," *Physics Letters B*, vol. 770, pp. 380–402, jul 2017.

[22] The ATLAS Collaboration, "Studies of $Z\gamma$ production in association with a high-mass dijet system in pp collisions at $\sqrt{s} = 8\ TeV$ with the ATLAS detector," *Journal of High Energy Physics*, vol. 2017, jul 2017.

[23] The CMS Collaboration, "Observation of electroweak production of same-sign W boson pairs in the two jet and two same-sign lepton final state in proton-proton collisions at $\sqrt{s} = 13\ TeV$," *Physical Review Letters*, vol. 120, feb 2018.

[24] The CMS Collaboration, "Measurement of vector boson scattering and constraints on anomalous quartic couplings from events with four leptons and two jets in proton-proton collisions at $\sqrt{s} = 13\ TeV$," *Physics Letters B*, vol. 774, pp. 682–705, nov 2017.

[25] The ATLAS Collaboration, "Observation of electroweak $W^{\pm}Z$ boson pair production in association with two jets in pp collisions at $\sqrt{s} = 13\ TeV$ with the ATLAS detector," *Physics Letters B*, vol. 793, pp. 469–492, jun 2019.

[26] The CMS Collaboration, "Measurement of electroweak WZ boson production and search for new physics in $WZ+$ two jets events in pp collisions at $\sqrt{s} = 13\ TeV$," *Physics Letters B*, vol. 795, pp. 281–307, aug 2019.

[27] The ATLAS Collaboration, "Observation 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," *Physical Review Letters*, vol. 123, oct 2019.

[28] The CMS Collaboration, "Search for anomalous electroweak production of vector boson pairs in association with two jets in proton-proton collisions at 13 TeV," *Physics Letters B*, vol. 798, p. 134985, nov 2019.

[29] ATLAS Collaboration, "Observation of electroweak production of two jets and a z-boson pair with the atlas detector at the lhc," 2020.

[30] CMS Collaboration, "Measurements of production cross sections of polarized same-sign W boson pairs in association with two jets in proton-proton collisions at $\sqrt{13}\ TeV$," 2020.

[31] "Measurement of the electroweak production of $Z\gamma$ and two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on dimension 8 operators," tech. rep., CERN, Geneva, 2021.

[32] T. Schörner-Sadenius, "The large hadron collider—background and history," *The Large Hadron Collider: Harvest of Run 1*, pp. 1–26, 05 2015.

[33] "LHC Design Report. 3. The LHC injector chain," 12 2004.

[34] E. Mobs, "The CERN accelerator complex. Complexe des accélérateurs du CERN," 2016. General Photo.

[35] P. Baudrenghien, L. Arnaudon, T. Bohl, O. Brunner, A. Butterworth, P. Maesen, J. E. Muller, G. Ravida, E. Shaposhnikova, and H. Timko, "Status and commissioning plans for LHC Run 2. The RF system.," in *5th Evian workshop on LHC beam operation*, (Geneva), pp. 99–104, CERN, 2014.

[36] L. R. Evans and P. Bryant, "LHC Machine," *JINST*, vol. 3, p. S08001. 164 p, 2008. This report is an abridged version of the LHC Design Report (CERN-2004-003).

[37] L. Evans and P. Bryant, "LHC machine," *Journal of Instrumentation*, vol. 3, pp. S08001–S08001, aug 2008.

[38] O. Aberle, I. Béjar Alonso, *et al.*, *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. CERN Yellow Reports: Monographs, Geneva: CERN, 2020.

[39] Q. Ingram, "Energy resolution of the barrel of the CMS electromagnetic calorimeter," *Journal of Instrumentation*, vol. 2, pp. P04004–P04004, apr 2007.

[40] The CMS Collaboration, "CMS physics: Technical design report volume 1: Detector performance," *CMS Technical Design Report CERN-LHCC-2006-001*, 2006.

[41] The CMS Collaboration, "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC," *Journal of Instrumentation*, vol. 16, p. P05014, may 2021.

[42] D. Valsecchi, "Deep learning techniques for energy clustering in the cms ecal," 2022.

[43] "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV," *Journal of Instrumentation*, vol. 10, pp. P06005–P06005, jun 2015.

[44] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, "Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC," *Journal of Physics G: Nuclear and Particle Physics*, vol. 31, pp. N9–N20, jul 2005.

[45] H. A. Bethe and L. C. Maximon, "Theory of bremsstrahlung and pair production. i. differential cross section," *Phys. Rev.*, vol. 93, pp. 768–784, Feb 1954.

[46] S. Baffioni, C. Charlot, *et al.*, "Electron reconstruction in CMS," *CMS-NOTE-2006-040*, Feb 2006.

[47] M. Oreglia, "A study of the reactions $\psi' \to \gamma\gamma\psi$," *SLAC Report SLAC-R-236*, 1980.

[48] The CMS Collaboration, "Measurement of the properties of a higgs boson in the four-lepton final state," *Physical Review D*, vol. 89, may 2014.

[49] The CMS Collaboration, "Measurements of properties of the higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV," *Journal of High Energy Physics*, vol. 2017, nov 2017.

[50] "Measurements of properties of the Higgs boson in the four-lepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV," tech. rep., CERN, Geneva, 2019.

[51] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas," *Physics Letters B*, vol. 659, pp. 119–126, jan 2008.

[52] M. Cacciari, G. P. Salam, and G. Soyez, "The catchment area of jets," *Journal of High Energy Physics*, vol. 2008, pp. 005–005, apr 2008.

[53] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual," *The European Physical Journal C*, vol. 72, mar 2012.

[54] Donato, Silvio, "Cms trigger performance," *EPJ Web Conf.*, vol. 182, p. 02037, 2018.

[55] The CMS Collaboration, "Performance of the CMS level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV," *Journal of Instrumentation*, vol. 15, pp. P10017–P10017, oct 2020.

[56] A. Hocker *et al.*, "TMVA - Toolkit for Multivariate Data Analysis," 3 2007.

[57] "Xgboost documentation." `https://xgboost.readthedocs.io/en/latest/index.html`.

[58] "Electron and Photon performance in CMS with the full 2017 data sample and additional 2016 highlights for the CALOR 2018 Conference," May 2018.

[59] J. Alwall, R. Frederix, *et al.*, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *Journal of High Energy Physics*, vol. 2014, jul 2014.

[60] The CMS Collaboration, "Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV," *Eur. Phys. J. C 81 (2021) 488*, 2019.

[61] P. Nason, "A new method for combining NLO QCD with shower monte carlo algorithms," *Journal of High Energy Physics*, vol. 2004, pp. 040–040, nov 2004.

[62] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method," *Journal of High Energy Physics*, vol. 2007, pp. 070–070, nov 2007.

[63] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower monte carlo programs: the POWHEG BOX," *Journal of High Energy Physics*, vol. 2010, jun 2010.

[64] P. Pigard, *Study of the EWK double Z production in the four leptons final state withthe CMS experiment at the LHC*. Theses, Université Paris-Saclay, July 2017.

[65] R. Frederix and I. Tsinikos, "On improving NLO merging for $t\bar{t}w$ production," *Journal of High Energy Physics*, vol. 2021, nov 2021.

[66] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO," *Journal of High Energy Physics*, vol. 2012, dec 2012.

[67] J. Alwall, R. Frederix, *et al.*, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," vol. 2014, no. 7, p. 79.

[68] R. Covarelli, Y. An, R. Bellan, M. Bonanomi, C. Charlot, E. Fontanesi, D. Giljanovic, H. He, D. Lelas, C. Li, Q. Li, G. Ortona, A. Savin, and T. Sculac, "Search for vector-boson scattering in the 4ljj final state with full Run2 data," *CMS AN 2019/172*.

[69] J. Alwall, S. Höche, *et al.*, "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions," *The European Physical Journal C*, vol. 53, pp. 473–500, dec 2007.

[70] V. Hirschi and P. Pigard, "Discussions on gluon-loop induced zz+2jets." `https://answers.launchpad.net/mg5amcnlo/+question/402723`, Oct. 2016.

[71] J. M. Campbell and R. Ellis, "MCFM for the tevatron and the LHC," *Nuclear Physics B - Proceedings Supplements*, vol. 205-206, pp. 10–15, aug 2010.

[72] The CMS Collaboration, "Event generator tunes obtained from underlying event and multiparton scattering measurements," *The European Physical Journal C*, vol. 76, mar 2016.

[73] The CMS collaboration, "Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements," *The European Physical Journal C*, vol. 80, jan 2020.

[74] R. D. Ball, V. Bertone, *et al.*, "Parton distributions for the LHC run II," *Journal of High Energy Physics*, vol. 2015, apr 2015.

[75] M. Grazzini, S. Kallweit, and D. Rathlev, "ZZ production at the LHC: Fiducial cross sections and distributions in NNLO QCD," *Physics Letters B*, vol. 750, pp. 407–410, nov 2015.

[76] S. Gieseke, T. Kasprzik, and J. H. Kühn, "Vector-boson pair production and electroweak corrections in herwig++," 2014.

[77] F. Caola, K. Melnikov, *et al.*, "QCD corrections to zz production in gluon fusion at the LHC," *Physical Review D*, vol. 92, nov 2015.

[78] F. Caola, M. Dowling, *et al.*, "QCD corrections to vector boson pair production in gluon fusion including interference effects with off-shell higgs at the LHC," *Journal of High Energy Physics*, vol. 2016, jul 2016.

[79] G. Petrucciani, A. Rizzi, and C. Vuosalo, "Mini-AOD: A new analysis data format for CMS," *Journal of Physics: Conference Series*, vol. 664, p. 072052, dec 2015.

[80] The CMS Collaboration, A. M. Sirunyan, and A. Tumasyan, "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s}$ =13 TeV," vol. 2017, no. 11, p. 47.

[81] A. Sirunyan, A. Tumasyan, *et al.*, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$ =13 TeV," vol. 13, no. 06, pp. P06015–P06015.

[82] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual," vol. 72, no. 3, p. 1896.

[83] The CMS Collaboration, "Identification of b-quark jets with the CMS experiment," *Journal of Instrumentation*, vol. 8, pp. P04013–P04013, apr 2013.

[84] D. Rainwater, R. Szalapski, and D. Zeppenfeld, "Probing color-singlet exchange in $Z+2$-jet events at the CERN LHC," vol. 54, no. 11, pp. 6680–6689.

[85] The CMS Collaboration, "Evidence for electroweak production of four charged leptons and two jets in proton-proton collisions at $\sqrt{s} = 13\ tev$," *Physics Letters B*, vol. 812, p. 135992, jan 2021.

[86] Y. Gao, A. V. Gritsan, Z. Guo, K. Melnikov, M. Schulze, and N. V. Tran, "Spin determination of single-produced resonances at hadron colliders," *Physical Review D*, vol. 81, apr 2010.

[87] S. Bolognesi, Y. Gao, A. V. Gritsan, K. Melnikov, M. Schulze, N. V. Tran, and A. Whitbeck, "Spin and parity of a single-produced resonance at the LHC," *Physical Review D*, vol. 86, nov 2012.

[88] I. Anderson, S. Bolognesi, F. Caola, Y. Gao, A. V. Gritsan, C. B. Martin, K. Melnikov, M. Schulze, N. V. Tran, A. Whitbeck, and Y. Zhou, "Constraining anomalous HVV interactions at proton and lepton colliders," *Physical Review D*, vol. 89, feb 2014.

[89] R. Brun and F. Rademakers, "ROOT: An object oriented data analysis framework," *Nucl. Instrum. Meth. A*, vol. 389, pp. 81–86, 1997.

[90] P. Speckmayer, A. Höcker, J. Stelzer, and H. Voss, "The toolkit for multivariate data analysis, TMVA 4," *Journal of Physics: Conference Series*, vol. 219, p. 032057, apr 2010.

[91] K. Arnold, M. Bähr, *et al.*, "Vbfnlo: A parton level monte carlo for processes with electroweak bosons," *Computer Physics Communications*, vol. 180, pp. 1661–1670, sep 2009.

[92] E. da Silva Almeida, O. J. P. Eboli, *et al.*, "Unitarity constraints on anomalous quartic couplings," *Physical Review D*, vol. 101, jun 2020.

[93] T. Plehn, "Lhc phenomenology for physics hunters," 2008.

[94] The CMS Collaboration, "CMS luminosity measurement for the 2016/2017/2018 data-taking period at $\sqrt{s} = 13$ TeV,"

[95] C. Ochando, T. Sculac, M. Xiao, *et al.*, "Measurements of properties of the higgs boson in the four-lepton final state at $\sqrt{s} = 13$ TeV with full Run II data," *CMS AN 2019/139*.

[96] B. Jäger, A. Karlberg, and G. Zanderighi, "Electroweak ZZjj production in the standard model and beyond in the POWHEG-BOX v2," *Journal of High Energy Physics*, vol. 2014, mar 2014.

[97] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower monte carlo programs: the POWHEG BOX," *Journal of High Energy Physics*, vol. 2010, jun 2010.

[98] A. Denner, R. Franken, M. Pellen, and T. Schmidt, "NLO QCD and EW corrections to vector-boson scattering into ZZ at the LHC," *Journal of High Energy Physics*, vol. 2020, nov 2020.

[99] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C*, vol. 71, feb 2011.

[100] The CMS Collaboration, "Measurement of the electroweak production of $z\gamma$ and two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on anomalous quartic gauge couplings," *Physical Review D*, vol. 104, oct 2021.

[101] C. Charlot, D. Lelas, D. Giljanovic, and A. Savin, "Vector boson scattering prospective studies for the high-luminosity lhc upgrade in the zz fully leptonic decay channel," *CMS AN 2018/072*.

[102] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," vol. 2014, no. 7, p. 79.

[103] S. Catani and M. H. Seymour, "A General Algorithm for Calculating Jet Cross Sections in NLO QCD," vol. 485, no. 1-2, pp. 291–419.

[104] J. M. Campbell, R. K. Ellis, and W. T. Giele, "A Multi-Threaded Version of MCFM.".

[105] G. Peter Lepage, "A new algorithm for adaptive multidimensional integration," vol. 27, no. 2, pp. 192–203.

[106] J. M. Campbell and R. K. Ellis, "An update on vector boson pair production at hadron colliders," vol. 60, no. 11, p. 113006.

[107] J. M. Campbell, R. K. Ellis, and C. Williams, "Vector boson pair production at the LHC," vol. 2011, no. 7, p. 18.

[108] J. Alwall, A. Ballestrero, P. Bartalini, S. Belov, E. Boos, A. Buckley, J. M. Butterworth, L. Dudko, S. Frixione, L. Garren, S. Gieseke, A. Gusev, I. Hinchliffe, J. Huston, B. Kersevan, F. Krauss, N. Lavesson, L. Lönnblad, E. Maina, F. Maltoni, M. L. Mangano, F. Moortgat, S. Mrenna, C. G. Papadopoulos, R. Pittau, P. Richardson, M. H. Seymour, A. Sherstnev, T. Sjöstrand, P. Skands, S. R. Slabospitsky, Z. Wcas, B. R. Webber, M. Worek, and D. Zeppenfeld, "A standard format for Les Houches Event Files.".

[109] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, "An Introduction to PYTHIA 8.2," vol. 191, pp. 159–177.

[110] S. Ovyn, X. Rouby, and V. Lemaitre, "Delphes, a framework for fast simulation of a generic collider experiment.".

[111] J. de Faverau, C. Delaere, P. Demin, A. Giammanco, V. Lemaitre, A. Martens, and M. Sekvaggi, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment," vol. 02, p. 057.

[112] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-kt jet clustering algorithm," p. 15.

[113] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. Nadolsky, and W. K. Tung, "New Generation of Parton Distributions with Uncertainties from Global QCD Analysis," vol. 2002, no. 07, pp. 012–012.

[114] J. Mousa, A. Tumasyan, W. Adam, F. Ambrogi, T. Bergauer, M. Dragicevic, J. Erö, A. Valle, M. Flechl, R. Frühwirth, M. Jeitler, N. Krammer, I. Krätschmer, D. Liko, T. Madlener, I. Mikulec, N. Rad, J. Schieck, S. Xie, and L. Finco, "Pileup mitigation at CMS in 13 TeV data," vol. 15, pp. P09018–P09018.

[115] T. Sjostrand, "A Model for Initial State Parton Showers," vol. 157, pp. 321–325.

[116] S. De, V. Rentala, and W. Shepherd, "Measuring the polarization of boosted, hadronic W bosons with jet substructure observables," 2020.

ECOLE DOCTORALE

Titre: Thesis title in French

Mots clés: CMS, LHC, VBS, ZZ

Résumé: Abstract in French

Title: Study of vector boson scattering in events with four leptons and two jets with CMS detector at the LHC

Keywords: CMS, HGCAL, LHC

Abstract: Abstract in English.

Institut Polytechnique de Paris
91120 Palaiseau, France