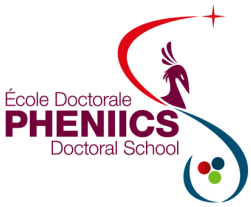
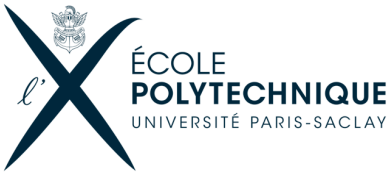
Study of the EWK double Z production in the four leptons final state withthe CMS experiment at the LHC

Philipp Pigard

NNT : 2017 SACLX039





THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS-SACLAY PREPARÉE À L'ÉCOLE POLYTECHNIQUE

École doctorale n°576
Particules, Hadrons, Énergie et Noyau, Instrumentation, Image, Cosmos et Simulation (PHENIICS)
Spécialité de doctorat : Physique des particules

par

PHILIPP PIGARD

Electron studies and search for vector boson scattering in events with four leptons and two jets with the CMS detector at the LHC

Identification des électrons et mise en évidence de la diffusion de bosons massifs dans les évènements à quatre leptons et deux jets avec le détecteur CMS auprès du LHC

Thèse présentée et soutenue à l'École polytechnique le 12 juillet 2017
devant le jury composé de :

M. PHILIPPE BUSSON	LLR, Palaiseau	Président du jury
M. TIZIANO CAMPORESI	CERN, Meyrin	Rapporteur
M. MATTHIAS SCHOTT	Universität Mainz	Rapporteur
Mme KERSTIN BORRAS	DESY et RWTH Aachen	Examinateur
Mme SINEAD FARRINGTON	University of Warwick	Examinateur
M. DIETER ZEPPENFELD	Karlsruhe Institute of Technology	Examinateur
M. CLAUDE CHARLOT	LLR, Palaiseau	Directeur de thèse

Abstract

This thesis reports the first experimental investigation into vector boson scattering (VBS) in the ZZ channel, where both Z bosons are required to decay into electrons or muons and are accompanied by at least two hadronic jets ($ZZjj \to \ell\ell\ell'\ell' jj$, where $\ell, \ell' = e$ or μ). VBS is a key process for elucidating the physics of electroweak symmetry breaking and the role of the recently discovered Higgs boson. This study analyses $35.9\,\text{fb}^{-1}$ of proton–proton collisions collected with the CMS experiment at the CERN Large Hadron Collider at a center-of-mass energy of 13 TeV. A multivariate analysis (MVA) technique is exploited to separate the electroweak signal from the QCD irreducible background and to measure the signal strength μ, i.e., the ratio of the observed number of events to the standard model expectation. The observed signal strength is $\mu = 1.39^{+0.86}_{-0.65}$ which excludes the background-only hypothesis at 2.7 standard deviations (1.6 standard deviations expected). Limits on physics beyond the standard model are derived in terms of anomalous quartic gauge couplings in the effective field theory approach, providing the most stringent constraints to date on the couplings for the operators T8 and T9.

The $ZZjj$ VBS analysis requires an accurate modeling of the signal and irreducible background processes, going beyond the existing simulations. Extensive work on generating and comparing the theory predictions from several Monte-Carlo event generators is presented. The detailed understanding of the signal and background kinematics is used to develop and systematically optimize a boosted decision tree (BDT) classifier. A matrix element discriminant is also developed and its classification performance compared to the BDT, finding comparable performance and indicating that the BDT is adequate. The signal extraction technique via a template fit of all $ZZjj$ events also permits to constrain the normalization of the QCD background using the data.

Multilepton analyses like the search for VBS in the ZZ channel depend on the ability to efficiently reconstruct and identify the final state leptons. This work presents the optimizations of the multivariate electron identification algorithms used in the first data at 13 TeV in 2015. A study on extending the use of tracking information in the MVA resulted in the reduction of the non-prompt electron background by up to 50 %. Monitoring the changes to the reconstructed electron objects and continuous optimizations allowed to improve or maintain the performance of the electron MVA ID algorithms, despite the harsher pileup conditions in the 2016 data. The electron efficiency measurements performed for the 2016 multilepton analyses in CMS are also documented.

Résumé

Cette thèse présente la première étude expérimentale de la diffusion de bosons massifs (VBS) dans le canal ZZ au LHC.

La structure non-abélienne du groupe de jauge électrofaible implique des interactions entre bosons électrofaibles via des vertex triples et quartiques. Les amplitudes de pure jauge pour les polarisations longitudinales des bosons violent l'unitarité pour des énergies supérieures à 1 TeV. Dans le modèle standard (MS) équipé de son secteur scalaire minimal, l'unitarité est restaurée par l'interférence avec les amplitudes faisant intervenir le boson de Higgs. La découverte en 2012 d'un boson scalaire par les expériences CMS et ATLAS auprès du Grand collisionneur de hadrons (LHC) du CERN a marqué le début de l'étude expérimentale de la brisure de la symétrie électrofaible. À ce jour, tous les résultats expérimentaux concernant ce nouveau boson et ses couplages aux bosons électrofaibles sont compatibles avec les prédictions du MS.

Toute déviation des couplages entre le boson de Higgs et les bosons de jauge (HVV) par rapport au MS empêcherait la régularisation des amplitudes VBS. L'étude des processus VBS à haute énergie permet donc une mesure des couplages HVV, indépendamment des couplages du boson de Higgs aux fermions. Cette approche est donc complémentaire aux mesures directes des taux de production et désintégration du boson de Higgs. L'intérêt pour les polarisations longitudinales est également une conséquence de leur origine dans le secteur scalaire, où ces états de polarisation sont identifiés avec les bosons de Goldstone. VBS constitue donc un processus clef dans la compréhension de la physique de la brisure de la symétrie électrofaible.

De nombreuses théories de physique au-delà du modèle standard impliquent une modification des couplages entre bosons électrofaibles. Les manifestations à basse énergie d'une telle théorie peuvent être paramétrées dans le cadre d'une théorie des champs effective. La topologie VBS est particulièrement sensible à des contributions aux interactions quartiques, permettant de rechercher des couplages quartiques anormaux (aQGC).

L'étude de VBS présentée dans cette thèse cible le canal ZZ, dans le cas où les deux bosons Z se désintègrent en paires de muons ou d'électrons et sont produits en association avec deux jets hadroniques ($ZZjj \to \ell\ell\ell'\ell'jj$, avec $\ell, \ell' = e$ ou μ). Cet état final offre une signature expérimentale propre avec un bruit de fond instrumental faible. Toutes les particules de l'état final peuvent être reconstruites, l'énergie de la diffusion des bosons est connue, et les angles de désintégration des fermions permettent de distinguer les polarisations des bosons Z. Cependant, le canal ZZ comporte des défis particuliers. D'abord, la section efficace du canal ZZ est la plus basse parmi tous les canaux VBS, et le rapport d'embranchement de la désintégration des bosons Z en leptons est faible. Le nombre d'événements attendus pour le signal $\ell\ell\ell'\ell'jj$ est donc très limité, de l'ordre d'une dizaine d'événements dans les données 2016. Le premier défi est donc de reconstruire et sélectionner un maximum de ces événements. Le sec-

ond est la maîtrise du bruit de fond constitué par la production $ZZjj$ ne résultant pas uniquement de couplages électrofaibles. La section efficace de ce bruit de fond principal est supérieure de plus d'un ordre de grandeur à celle du signal, nécessitant une compréhension précise et une stratégie d'extraction du signal adaptée.

Les analyses multi-leptons telles que la recherche du processus VBS dans le canal $ZZjj$ reposent sur la capacité à reconstruire et identifier de façon efficace les leptons de l'état final. Grâce à leur signature propre dans le détecteur CMS, les muons sont reconstruits et sélectionnés avec une très haute efficacité. La sélection des électrons soulève en revanche des défis plus importants, et repose sur un classificateur multivarié à base d'arbres de décision boostés (BDT) qui a été développé dans CMS pour le Run I. Un effort important est consacré à la préparation et à l'optimisation de cet algorithme pour le Run II et l'analyse $ZZjj$.

L'identification des électrons est traditionnellement dominée par la calorimétrie, et le matériau du trajectographe en silicium dans CMS soulève un problème particulier dû à la perte d'énergie des électrons par rayonnement de photons de bremsstrahlung. L'introduction du Gaussian Sum Filter (GSF) dans la reconstruction des électrons dans CMS ouvre la voie à des critères d'identification basés sur les traces, et cette thèse présente une telle étude.

Une première source de bruit de fond est constituée des jets hadroniques, dans lesquels la trace d'un hadron chargé et l'agrégat électromagnétique d'un hadron neutre du jet peuvent être reconstruits par erreur comme un électron. Pour ce type de bruit de fond, le rayon de courbure de la trace est constant, en raison de l'absence de pertes radiatives pour les hadrons. La mesure de l'impulsion et les incertitudes fournis par l'algorithme GSF sont étudiées dans la simulation avec l'objectif de discriminer ce bruit de fond. Des observables nouvelles et puissantes sensibles aux changements de courbure locale tout au long de la trajectoire dans le trajectographe sont présentées. Une deuxième source de bruit de fond est la conversion de photons en paires d'électrons dans le matériau du trajectographe. Des observables sensibles à ce phénomène sont étudiées, et leur utilisation dans le BDT d'identification des électrons permet de réduire ce bruit de fond de 50 %.

Cet algorithme optimisé et les nouvelles observables d'identification des électrons ont été validés et mis en service pour les premières collisions proton-proton à 13 TeV en 2015, et par la suite adoptés officiellement dans CMS. Grâce au suivi des changements dans la reconstruction des électrons et à une optimisation continue des algorithmes, les performances d'identification des électrons ont été préservées dans la prise de données de 2016, malgré la présence d'un empilement plus sévère. Ce document présente également la mesure d'efficacité de la sélection des électrons dans les données enregistrées en 2016. Cette mesure permet d'appliquer à la simulation des facteurs correctifs, qui sont utilisés dans tous les analyses multi-leptons de CMS.

L'analyse VBS dans le canal $ZZjj$ requiert une modélisation précise du signal et du bruit de fond irréductible, ce qui nécessite des simulations dédiées. Un effort important est consacré à la génération et à la comparaison des prédictions théoriques de plusieurs générateurs d'événements Monte Carlo. La compréhension détaillée du signal et des bruits de fond est exploitée pour développer et optimiser de façon systématique un BDT. Un classificateur basé sur les éléments de matrices est également développé, et sa puissance est comparée à celle du BDT, montrant des performances similaires. Le signal est extrait par un ajustement des distributions attendues aux données, incluant tous les événements $ZZjj$. Cette méthode permet aussi de contraindre la normalisation du bruit de fond principal par les données.

L'algorithme d'identification des électrons et la modélisation précise du signal et du bruit de fond sont les deux éléments cruciaux dans la recherche du processus VBS dans le canal ZZ. Cette recherche exploite ici les collisions proton-proton enregistrés par CMS en 2016 à 13 TeV, qui correspondent à une luminosité intégrée de 35.9 fb^{-1}. Le discriminant multivarié est utilisé pour séparer le signal électrofaible du bruit de fond irréductible QCD et pour mesurer la force du signal μ, définie comme le quotient des taux d'événements observés et attendus. La force du signal observée est de $\mu = 1.39^{+0.72}_{-0.57}\,(\text{stat})\,^{+0.46}_{-0.31}\,(\text{syst}) = 1.39^{+0.86}_{-0.65}$, excluant l'hypothèse de l'absence de signal à hauteur de 2.7 écarts-types, pour 1.6 écarts-types attendus. La force du signal est interprétée comme une section efficace fiducielle $\sigma_{\text{EW}}(\text{pp} \to ZZjj \to \ell\ell\ell'\ell' jj) = 0.40^{+0.21}_{-0.16}\,(\text{stat})\,^{+0.13}_{-0.09}\,(\text{syst})\,\text{fb}$, qui est en accord avec la prédiction du MS de $0.29^{+0.02}_{-0.03}$ fb. Les événements $ZZjj$ observés sont également exploités pour placer des limites sur la physique au-delà du MS pour les couplages quartiques. Les limites présentées dans cette thèse sont les plus strictes à ce jour sur les couplages des opérateurs $\mathcal{O}_{T,0}$, $\mathcal{O}_{T,1}$ et $\mathcal{O}_{T,2}$, et des opérateurs neutres $\mathcal{O}_{T,8}$ et $\mathcal{O}_{T,9}$.

Acknowledgements

This thesis would have never been possible without the unwavering support and patient guidance of my supervisor Claude. His mentorship and expertise are the foundation of this research and the publication of the $ZZjj$ paper.

I want to express my sincere gratitude to the CMS group at Laboratoire Leprince-Ringuet, which welcomed and taught me everything that was needed to produce the results of this thesis. I thoroughly appreciated learning from Christophe, Florian, Giacomo, Jean-Baptiste, Olivier, Philippe, and Roberto. A special word of appreciation is due to my office mate Stéphanie: in addition to her tremendous technical expertise, she was a patient teacher of the French language. I will never forget her support during dark days of struggle with the french health care administration. Praise is due for the exceptional technical support at LLR, and for the supportive administration.

I also want to thank my fellow PhD students at LLR for making the past three years a joyful adventure: Floriana, Luca (C. and M.), Simon, Thomas, and Yiurii. Toni was the best gym buddy one could ever want and Alex made those stays at CERN so much more entertaining.

Contents

Chapter 1

Vector boson scattering and experimental status

The discovery of the Higgs boson provided the first experimental evidence for the breaking of the electroweak symmetry as predicted by the minimal scalar sector. In addition to explaining the origin of mass, the scalar sector also provides the longitudinal polarizations of the massive gauge bosons which absorb the Goldstone bosons of the broken electroweak symmetry. At large scattering energies, the interaction between the gauge bosons is dominated by the longitudinal polarizations, which are equivalent to the Goldstone bosons. The study of vector boson scattering (VBS) thus allows elucidating the breaking of electroweak symmetry via the interplay of the electroweak and scalar sectors. After a brief reminder of the standard model, this chapter summarizes the theory of VBS and provides a phenomenological account of the signal process. This chapter also presents the effective field theory framework which is used in this thesis work to report model-independent limits on physics beyond the standard model. The status of experimental searches for VBS and anomalous quartic couplings is summarized.

1.1 The standard model

The standard model (SM) is the theory of fundamental interactions at the subatomic scale. Its predictions have been confirmed by many experiments. The fundamental constituents of the theory, i.e., the particles which are assumed to have no further substructure, are the listed in Fig. 1.1. Matter is described by spin-1/2 fermions while the interactions are mediated by spin-1 bosons. The first three columns in Fig. 1.1 correspond to the three generations, whereby generation two and three are essentially heavier, unstable copies of the first. Each particle in Fig. 1.1 also has an anti-particle which has the opposite quantum numbers and is usually denoted with a bar above the particle symbol.

Only the particles in the first column are stable: the electron (e), the up- (u) and down-type quarks (d) and the electron-neutrino (ν_e). The first generation forms the everyday matter: protons are made of three quarks (uud) as are neutrons (udd). Atomic nuclei are bound states of neutrons and protons and atoms, in turn, are bound states of nuclei and electrons. While electrons can be separated from atoms and can be observed in unbound states, the same is not true for quarks. The strong interaction mediated by gluons (g) causes quarks to only exist in bound states called hadrons due to confine-

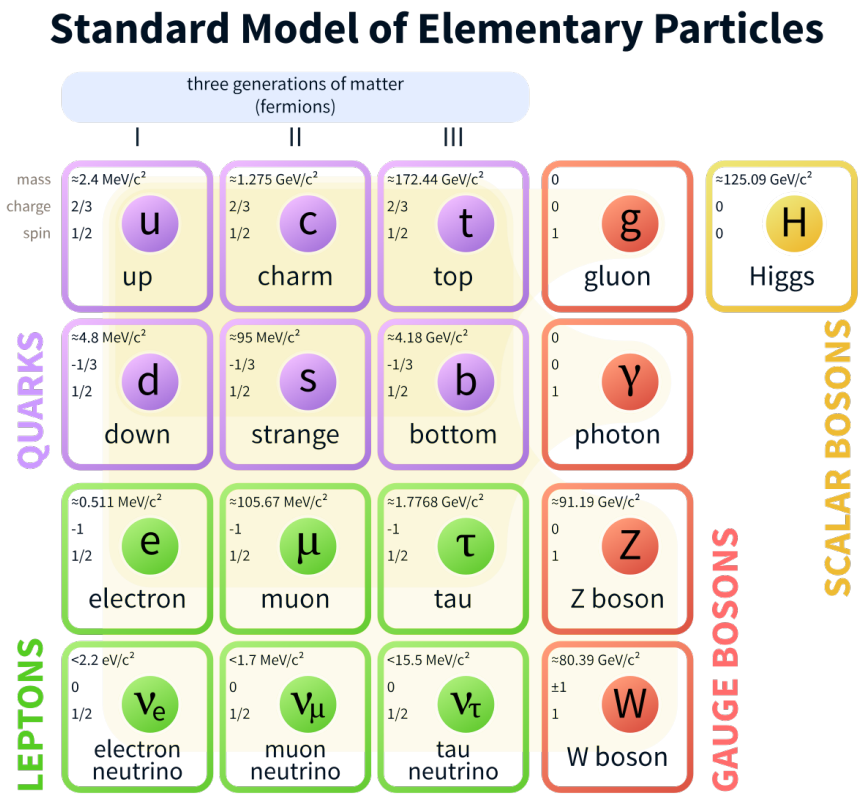


Figure 1.1: Particle content of the standard model [1].

ment. Mesons are bound states of a quark and anti-quark, with the neutral ($u\bar{u}$) and charged ($u\bar{d}$) pions being the most prominent examples. The charge of the strong interaction is associated with one of three colors (red, green, blue). In addition to the strong interaction, quarks also take part in the electroweak interaction which can change the flavor and generation of a quark via the charged weak boson.

Leptons on the other hand exclusively couple to the electroweak interaction. The electron is the lightest and only stable electrically charged lepton, the muon (μ) and tau leptons (τ) are heavier replicas. The uncharged neutrinos are assumed massless in the SM but the observation of neutrino oscillations implies that they must have small but nonzero mass.

The SM describes the interaction between the fermionic matter in the framework of a quantum field theory, where the interactions are derived by local gauge symmetries and particles are excitations of the quantum fields. It is the combination of quantum chromodynamics (QCD) and the electroweak theory [2–4]. The gauge group of the SM is $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$, where the indices stand for *color*, *left-handed*, and *hypercharge*.

1.1.1 Electromagnetism as a local gauge theory

The Lagrangian of a free fermionic field ψ is given by

$$\mathcal{L}_{\text{free}} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi \tag{1.1}$$

and exhibits a global $U(1)$ symmetry, i.e., $\mathcal{L}_{\text{free}}$ is invariant under a complex phase rotation by Λ:

$$\psi \to \psi' = e^{-i\Lambda}\psi \quad \Rightarrow \quad \mathcal{L}'_{\text{free}} = \mathcal{L}_{\text{free}}. \tag{1.2}$$

The transformation parameter Λ is global, i.e., constant and not a function on the space-time position x^{μ}. $\mathcal{L}_{\text{free}}$ is not invariant under local transformations, that is transformations $U(x) = \exp(-i\Lambda(x))$, as these would result in an additional term due to the non-vanishing derivative $\partial_{\mu}\Lambda(x)$. These derivative terms can, however, be absorbed by introducing a new field A^{μ} which transforms as $A'^{\mu} = A^{\mu} - g^{-1}\partial^{\mu}\Lambda(x)$. Replacing the partial derivative with the covariant derivative

$$D_{\mu} = \partial_{\mu} + igA_{\mu}(x) \tag{1.3}$$

and writing the kinematic term for the field $A_{\mu}(x)$ in terms of the field strength tensor

$$F_{\mu\nu} = \frac{i}{g}[D_{\mu}, D_{\nu}] = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu} \tag{1.4}$$

one obtains a Lagrangian that is invariant under arbitrary local transformations:

$$\mathcal{L}_{\text{QED}} = i\bar{\psi}\gamma^{\mu}D_{\mu}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \tag{1.5}$$

This is the Lagrangian of quantum electrodynamics (QED) where the field ψ corresponds to massless electrons and $g = e$. The interaction between the matter fields ψ and the photon field A was obtained by extending the global symmetry of the free Lagrangian $\mathcal{L}_{\text{free}}$ under transformations $U(1)$ to a local symmetry.

The free photon field A^{μ} satisfies the Maxwell equations, in particular, the transversality condition $\vec{\nabla} \cdot \vec{A} = 0$. The photon is also massless ($A^2 = 0$), which imposes another restriction on the four components of the four-vector field, leaving only two independent degrees of freedom. These two degrees of freedom correspond to the polarization states of the photon which for a photon of momentum $k^{\mu} = (E, 0, 0, E)^{\mu}$ are commonly parametrized by the polarization vectors

$$\epsilon^{\mu}_{\pm} = \frac{1}{\sqrt{2}}(0, 1, \pm i, 0)^{\mu}. \tag{1.6}$$

1.1.2 Quantum chromodynamics as a non-Abelian gauge theory

The three color states of the quarks (named red, green, blue) can be arranged in a triplet of quark fields $Q = (\psi_r, \psi_g, \psi_b)$ and the free quarks are described by the Lagrangian

$$\mathcal{L}^{\text{quarks}}_{\text{free}} = i\overline{Q}\gamma^{\mu}\partial_{\mu}Q \tag{1.7}$$

Similarly to Eq. (1.1), this Lagrangian is invariant under global transformations, though under the more complex symmetry group $SU(3)_C$. While transformations of $U(1)$ are

specified by a single rotation angle, transformations of $SU(3)$ require 8 such parameters Λ^a, $a = 1, ..., 8$ one for each of the 8 generators $\lambda^a$of the group. A generic $SU(3)$ transformation can then be written as $U = \exp(-i\Lambda^a\lambda^a/2)$.

Like for the derivation of the QED Lagrangian, one can now promote this global symmetry to a local one where the Λ^a parameters are functions of space-time. The covariant derivative then is

$$D_\mu = \partial_\mu + ig_s\frac{\lambda^a}{2}G^a_\mu, \tag{1.8}$$

where g_s is again a scalar coupling constant. The transformation of the gauge fields will be more complex due to the non-Abelian structure of $SU(3)$:

$$G'^a_\mu = G^a_\mu + f^{abc}\Lambda^b(x)G^c_\mu + \frac{1}{g_s}\partial_\mu\Lambda^a(x). \tag{1.9}$$

The f^{abc} are the structure constants of the group $SU(3)$ and obey the relationship $[\lambda^a, \lambda^b] = if^{abc}2\lambda^c$. The nonzero commutator between the generators also gives rise to an extra term in the field strength tensor associated with the gauge fields G^a_μ:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f^{abc}G^b_\mu G^c_\nu. \tag{1.10}$$

The full Lagrangian of the gauged theory of color-charged quarks then reads as

$$\mathcal{L}_{\mathrm{QCD}} = i\overline{Q}\gamma^\mu D_\mu Q - \frac{1}{4}G^{a,\mu\nu}G^a_{\mu\nu} \tag{1.11}$$

with a striking resemblance to the QED Lagrangian, where the photon field A has been replaced by 8 vector fields G_a that correspond to the gluons. A key difference to the QED case is apparent in the kinetic term of the gluon field, which features the commutator of the $SU(3)$ generators. This leads to terms that feature three and four gluon fields, corresponding to gauge field self-interactions. This feature is absent for theories of Abelian gauge groups like QED and is the origin of the hadronic confinement property of the strong interaction.

1.1.3 Unification of electromagnetism and the weak force

The electroweak theory provides a concise description of electromagnetic and weak phenomena by unifying these forces into one interaction. The underlying non-Abelian gauge group is $SU(2)_L \otimes U(1)_Y$. A particularity of the weak interaction is parity violation - the weak interaction only couples to the left-handed chiral component of the fermion fields ψ_L. It does not distinguish between charged and uncharged leptons, which leads one to introduce $SU(2)_L$ doublets of the left-handed leptons as $L = (\psi_L, \nu)$. The global symmetry transformation can then be written as

$$L' = \exp\Big(-i\alpha_i\frac{\sigma^i}{2} - i\beta\frac{Y}{2}\Big)\,L \tag{1.12}$$

$$\psi'_R = \exp\Big(-i\beta\frac{Y}{2}\Big)\,\psi_R, \tag{1.13}$$

where the $\sigma^i/2$ are the generators of $SU(2)_L$ and Y is the hypercharge operator, i.e., it returns the hypercharge of the field. The global symmetry is promoted to a local one by introducing the covariant derivatives

$$D_\mu = \partial_\mu + i g_w \frac{\sigma^i}{2} W_\mu^i + i g \frac{Y}{2} B_\mu. \tag{1.14}$$

The gauge fields W^i and B transform as expected

$$W_\mu'^i = W_\mu^i + \epsilon^{ijk}\alpha^j(x) W_\mu^k + \frac{1}{g_w}\partial_\mu \alpha^i(x) \tag{1.15}$$

$$B_\mu' = B_\mu + \frac{1}{g}\partial_\mu \beta(x) \tag{1.16}$$

and the ϵ^{ijk} are the structure constants of $SU(2)$. The electroweak Lagrangian can then be succinctly written as

$$\mathcal{L}_{\mathrm{EW}} = i\bar{L}\gamma^\mu D_\mu L + i\bar{\psi}_R \gamma^\mu D_\mu \psi_R - \frac{1}{4} W^{i,\mu\nu} W^i_{\mu\nu} - \frac{1}{4} B^{\mu\nu} B_{\mu\nu} \tag{1.17}$$

using the usual definitions of the field strength tensors

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \quad \text{and} \tag{1.18}$$

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g_w \epsilon^{ijk} W^j_\mu W^k_\nu. \tag{1.19}$$

The fields of the charged gauge bosons $W^\pm$ and the neutral Z boson are obtained via linear combinations of the gauge fields using the weak mixing angle θ_w:

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \mp i W^2_\mu) \tag{1.20}$$

$$Z_\mu = \cos\theta_w W^3_\mu - \sin\theta_w B_\mu, \tag{1.21}$$

and the photon A field

$$A_\mu = \sin\theta_w \, W^3_\mu + \cos\theta_w \, B_\mu. \tag{1.22}$$

An expansion of $\mathcal{L}_{\mathrm{EW}}$ reveals the triple (ZWW, γWW) and quartic ($ZZWW$, γZWW, $\gamma\gamma WW$, and $WWWW$) gauge boson self-interactions that are a consequence of the non-Abelian nature of the underlying gauge group.

1.1.4 Electroweak symmetry breaking and the minimal scalar sector

The electroweak Lagrangian in Eq. (1.17) does not include mass terms for the W or Z gauge bosons. This is a clear contradiction of the observation of gauge boson masses and the finite range of the weak interaction. Inserting mass terms such as $-m_W^2 W^\mu W_\mu$ however would explicitly break the gauge invariance of the Lagrangian. Gauge boson masses can be generated without explicitly breaking gauge invariance via the Brout–Englert–Higgs mechanism (BEH) [5–7].

The idea of the BEH mechanism is to create mass terms by introducing a scalar field with nonzero vacuum expectation value. The minimal implementation of this scalar sector consists of a $SU(2)_L$ doublet of a complex scalar field:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^1 + i\phi^2 \\ \phi^3 + i\phi^4 \end{pmatrix}. \tag{1.23}$$

The kinematic term and interaction of the scalar field with the gauge bosons is obtained from the covariant derivative of the electroweak theory (Eq. (1.14)):

$$\mathcal{L}_{\text{BEH}} = (D^{\mu}\Phi)^{\dagger}(D_{\mu}\Phi) + V(\Phi^{\dagger}\Phi). \tag{1.24}$$

Central to the BEH mechanism is the form of the self-interaction potential of the doublet field $V(\Phi^{\dagger}\Phi)$, which is parametrized as

$$V(\Phi^{\dagger}\Phi) = \mu^2\Phi^{\dagger}\Phi + \lambda(\Phi^{\dagger}\Phi)^2. \tag{1.25}$$

The parameter λ is required to be positive to bound the potential from below. The energy minimum of this potential is given by

$$\Phi^{\dagger}\Phi = -\frac{\mu^2}{2\lambda}. \tag{1.26}$$

For $\mu^2 < 0$ this corresponds to nonzero field values and the ground state of the theory acquires a nonzero expectation value v. Fluctuations around this ground state can be parametrized as

$$\Phi(x) = \frac{1}{\sqrt{2}} \exp\left(iw^i(x)\sigma^i\right) \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{1.27}$$

where $h(x)$ is the Higgs field, the σ^i are the generators of $SU(2)_L$, and the w^i are massless Goldstone bosons. Excitations of the Higgs field correspond to the Higgs boson particle, whose mass $m_{\text{H}} = \sqrt{2}|\mu|$ is the only free parameter of the BEH theory.

The Goldstone bosons w_i are a consequence of the Goldstone theorem, which states that the spontaneous breaking of a continuous symmetry leads to as many massless scalar fields as there are broken generators. The BEH mechanism spontaneously breaks the electroweak symmetry with 4 generators to the symmetry of electromagnetism of 1 generator $SU(2)_L \otimes U(1)_Y \to U(1)_{\text{em}}$.

After electroweak symmetry breaking, the BEH Lagrangian Eq. (1.24) reads:

$$\begin{aligned} \mathcal{L}_{\text{BEH}}^{\text{EWSB}} \quad = \quad & \frac{g_w^2 v^2}{4} W_{\mu}^{-} W^{+\mu} + \frac{g_w^2 v^2}{8\cos^2\theta_w} Z_{\mu} Z^{\mu} + \mu^2 h^2 \\ & + \frac{g_w^2 v}{2} h W_{\mu}^{-} W^{+\mu} + \frac{g_w^2 v}{4\cos^2\theta_w} h Z_{\mu} Z^{\mu} \\ & + \frac{g_w^2}{4} h^2 W_{\mu}^{-} W^{+\mu} + \frac{g_w^2}{8\cos^2\theta_w} h^2 Z_{\mu} Z^{\mu} \\ & + \frac{1}{2}\partial_{\mu} h \partial^{\mu} h + \frac{\mu^2}{v} h^3 + \frac{\mu^2}{4v^2} h^4. \end{aligned} \tag{1.28}$$

The first line features the mass terms for the gauge fields and the Higgs boson. The second and third lines respectively show the trilinear and quartic interactions between the gauge fields and the Higgs field.

The gauge fields in $\mathcal{L}_{\text{BEH}}^{\text{EWSB}}$ are not the same as in the pre-EWSB electroweak Lagrangian given in Eq. (1.20). After EWSB, the massive gauge fields absorb the Goldstone bosons, i.e., the fields w_i become part of the massive gauge bosons:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}\left(\left(W^1 - \frac{1}{gv}\partial_\mu w^1\right) \mp i\left(W_\mu^2 - \frac{1}{gv}\partial_\mu w^2\right)\right), \tag{1.29}$$

$$Z_\mu = \cos\theta_w\left(W_\mu^3 - \frac{1}{gv}\partial_\mu w^3\right) - \sin\theta_w B_\mu \tag{1.30}$$

while the photon field A remains unmodified and massless.

The first term in Eq. (1.28) provides the mass of the W boson and allows to connect the Higgs field vacuum expectation value v to the energy scale of the weak interaction, given by the Fermi constant G_F:

$$v = 2\frac{m_W}{g_w} = (\sqrt{2}G_F)^{-1/2} \approx 246\,\text{GeV}. \tag{1.31}$$

The most general renormalizable and gauge invariant Lagrangian for the complex scalar field ϕ and the fermion fields ψ includes interactions terms. After spontaneous symmetry breaking, these Yukawa interaction terms lead to interactions between the fermions and to mass terms for the fermion fields. The coupling of the Higgs field to the fermions is determined by the mass of the fermion, but the latter are free parameters of the theory.

Since its discovery by the ATLAS and CMS Collaborations in 2012 [8, 9], the properties of the Higgs boson have been studied in detail and the data are compatible with the minimal scalar sector of the standard model. In particular, the measured coupling strengths of the scalar to the gauge bosons and fermions are consistent with those predicted by the BEH theory [10], see Fig. 1.2.

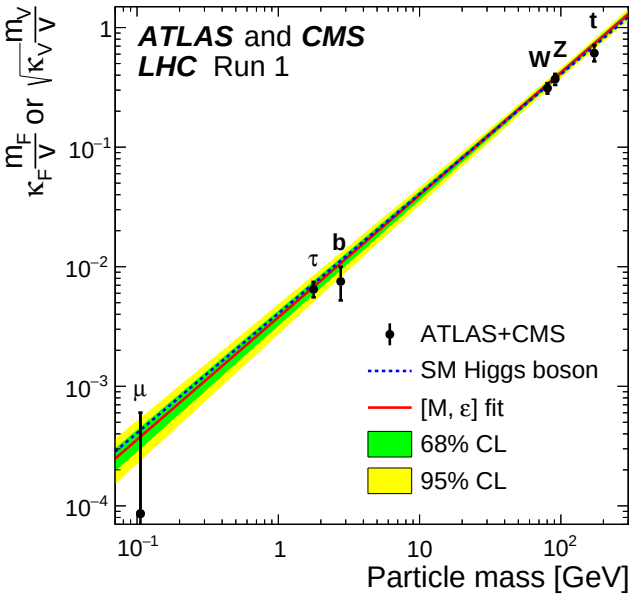


Figure 1.2: The measured coupling strengths of the Higgs boson with the gauge bosons and fermions, compared to the standard model expectation [10].

The Goldstone bosons fields that are merged with the massless gauge fields to form the massive gauge bosons in Eq. (1.29) constitute degrees of freedom, i.e., they are fields of the overall theory. However, the excitations of these fields do not give rise to

resonances that can be observed independently. Instead, these fields are a part of the massive gauge boson fields which are superpositions between the pure gauge fields $W^{\pm}/Z$ and the Goldstone bosons w_i. In fact the three pure gauge fields in Eq. (1.29) each have gained a new degree of freedom, the *longitudinal* polarization. For a massive vector boson of mass m and momentum $k^{\mu} = (E, 0, 0, k_z)^{\mu}$ the longitudinal polarization vector reads:

$$\epsilon_{\mathrm{L}}^{\mu} = \frac{1}{m}(k_z, 0, 0, E)^{\mu}. \tag{1.32}$$

1.2 Vector boson scattering

Gauge bosons of mass m have three degrees of freedom, two transverse and one longitudinal polarization:

$$\epsilon_{\mathrm{T}}^{\mu} = (0, 1, \pm i, 0)^{\mu} \tag{1.33}$$

$$\epsilon_{\mathrm{L}}^{\mu} = \frac{1}{m}(k_z, 0, 0, E)^{\mu}. \tag{1.34}$$

The two polarization vectors have a strikingly different high-energy behavior. The components of the transverse polarizations are constant, while the components of the longitudinal polarization scale as E/m. This means that the relative importance of the longitudinal polarizations will increase at high energies $E \gg m$, and eventually dominate over the transverse components that do not exhibit this energy dependence.

This difference between the transverse and longitudinal polarization in the high-energy limit is due to their respective origins. The transverse components correspond to the original gauge bosons that are already present in the unbroken electroweak theory of massless gauge bosons. In contrast, the longitudinal polarizations only arise for massive gauge bosons which absorbed the Goldstone bosons of EWSB in the Higgs sector. In the high-energy limit the two polarizations can be separated and the scattering of longitudinal vector bosons is equivalent to the scattering of Goldstone bosons, also known as the Goldstone boson equivalence theorem.

Using the Goldstone boson equivalence theorem, one can demonstrate that the scattering of longitudinal vector bosons exhibits a particular high-energy behavior, which is illustrated using the example of $W_{\mathrm{L}}^{+}W_{\mathrm{L}}^{-} \to W_{\mathrm{L}}^{+}W_{\mathrm{L}}^{-}$ scattering [11–13]. The dominant terms of the scattering amplitude read

$$\mathcal{A} \approx -i\frac{m_{\mathrm{H}}^2}{v^2}\left[2 + \frac{m_{\mathrm{H}}^2}{s - m_{\mathrm{H}}^2} + \frac{m_{\mathrm{H}}^2}{t - m_{\mathrm{H}}^2}\right], \tag{1.35}$$

where s and t denote the Mandelstam variables.

In the high-energy limit where $s, t \gg m_{\mathrm{H}}$ this amplitude becomes a constant and the cross section (σ) will decrease linearly with the scattering energy, $\sigma \sim 1/s$. It turns out that this finite cross section at large energies is the result of cancellations between the Higgs diagrams and the pure-Goldstone boson diagrams. In absence of a Higgs boson ($m_{\mathrm{H}} \to \infty$) or if the Higgs-to-gauge boson (HVV) couplings differ from their SM values, these cancellations are incomplete and the cross section diverges at sufficiently large scattering energies, eventually violating the unitarity at an energy of around 1.2 TeV [11, 12]. Prior to the discovery of the Higgs boson with a mass of $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, unitarity violation in the scattering of massive gauge bosons provided an important theory argument for yet-unobserved physics at the TeV-scale: either the cross section is regularized by a light Higgs boson or some other regularization mechanism has to exist.

With the discovery of the Higgs boson, the question of unitarity violation becomes less urgent, as the SM with the minimal scalar sector provides a UV-complete theory. However, this UV-completeness depends on the delicate cancellation between divergent scattering amplitudes and assumes the HVV couplings and thus permits to test the HVV coupling, complementing direct measurements of Higgs boson production and decay rates.

The massive gauge bosons of the SM are curious objects in the sense that they are the superposition of the pure electroweak gauge bosons and the Goldstone bosons of EWSB. Studying the longitudinal polarizations of the W- and Z-bosons thus permits to probe the mechanism of EWSB.

The scattering of vector bosons furthermore allows studying the non-Abelian structure of the electroweak sector by probing the quartic vertices. The electroweak Lagrangian leads to $WWZZ$, $WWZ\gamma$, $WW\gamma\gamma$, and $WWWW$ vertices whose couplings are fully specified by the gauge structure. There are no quartic vertices involving only the neutral gauge bosons, i.e., no $ZZZZ$ or $ZZ\gamma\gamma$ vertices. Physics beyond the SM could manifest itself in modifications to the quartic gauge couplings or introduce new vertices not present in the electroweak theory.

1.2.1 Phenomenology of vector boson scattering

At the LHC the scattering of vector bosons is initiated by the quarks in the initial state protons: one of the quarks in each proton (p) radiates off a vector boson which then interact. The top row of Fig. 1.3 shows some of the Feynman diagrams of VBS in the $pp \to ZZjj \to \ell\ell\ell'\ell' jj$ channel, where ℓ denotes electrons or muons and j denotes hadronic jets associated with the outgoing quarks. The first diagram in the top row features the quartic vertex, the center diagram illustrates the scattering via double trilinear couplings, and the right diagram the t-channel exchange of a Higgs boson. The exchange of a Higgs boson in the s-channel corresponds to its VBF production mode, a process sharing many similarities with the VBS signal discussed in this work. Including the decay of the vector bosons, the process is of order six in the weak coupling constant (α_{EW}^6).

Other diagrams leading to the same final state exist and are necessary to maintain gauge invariance, see center row of Fig. 1.3. These include diagrams where one (center left), or both vector bosons are radiate-off the outgoing quark lines. Diagrams where an off-shell boson splits into the two final state bosons are also possible (center right), though only the Higgs boson contributes in the ZZ channel due to the absence of an all-neutral triple gauge coupling in the SM.

Finally, the bottom row of Fig. 1.3 shows some pure-electroweak diagrams that are not relevant for the study of VBS and that can be suppressed by appropriate phase space selections. The first diagram of the bottom row illustrates one of many non-resonant diagrams where the final state leptons do not originate from an on-shell Z decay. These amplitudes are strongly suppressed when selecting on-shell Z bosons. The bottom right diagram shows triboson production where one of the gauge bosons decays hadronically. The outgoing quarks from such hadronic gauge boson decays will result in dijets with masses of around 100 GeV, meaning these contributions can be suppressed by imposing an m_{jj} threshold.

The Hallmark signs of VBS are the *tagging jets* which originate from the outgoing quarks in the first row diagrams of Fig. 1.3. It can be shown that the leading contribution to the squared scattering amplitude is proportional to

$$|\mathcal{A}|^2 \sim \frac{p_1 \cdot p_2 \;\; p_3 \cdot p_4}{(q_1^2 - M_Z^2)^2 (q_2^2 - M_Z^2)^2}, \tag{1.36}$$

where the $p_{1,2}$ refer to the momenta of the incoming quarks, $p_{3,4}$ are the momenta of the outgoing quarks, and the momenta of the intermediate gauge bosons are given

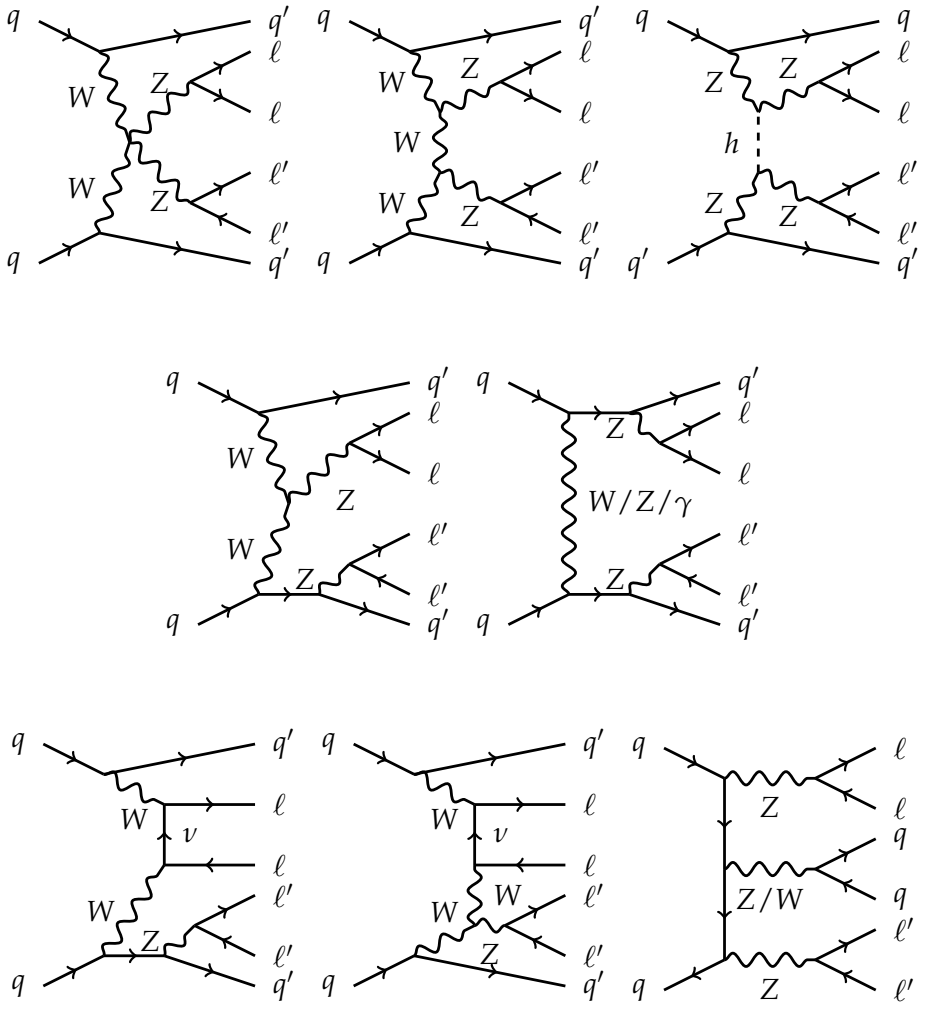


Figure 1.3: Feynman diagrams for the electroweak production of the $\ell\ell\ell'\ell' jj$ final state. The top row illustrates diagrams of the VBS signal. The scattering of massive gauge bosons as depicted in the first two diagrams of the top row is unitarized by the interference with diagrams that feature the Higgs boson (top right). The center row features diagrams that are required to ensure gauge invariance. The bottom row illustrates non-resonant production and triboson processes.

by $q_{1,(2)} = p_{1,(2)} - p_{3,(4)}$ [14]. For a fixed partonic scattering energy $\hat{s} = \sqrt{p_1 \cdot p_2}$, the expression in Eq. (1.36) is largest when the nominator is large or the denominator is small. The latter happens when the q_i are small. The square of q_1 can be written in terms of the scattering angle (ϑ), the energy of the incoming (E_1) and outgoing quarks (E_3) as well as the transverse momentum of the outgoing quark ($p_{\mathrm{T},3}$):

$$q_1^2 = -2p_1 \cdot p_2 = -2E_1E_3(1 - \cos\vartheta_1) = -\frac{2}{1+\cos\vartheta_1}\frac{E_1}{E_3}p_{\mathrm{T},3}^2 \tag{1.37}$$

The expression in Eq. (1.37) is small when the scattering angle is small ($\vartheta \to 0$), or when the transverse momentum of the outgoing quark is small. However, the quarks are recoiling against the vector bosons that they radiate off, and these gauge bosons need sufficient energy to create the on-shell Z boson of the final state. The transverse momentum of the outgoing jets will thus be of the order of the gauge boson mass $p_{\mathrm{T}}^j \approx m_Z$.

The VBS scattering amplitude of Eq. (1.36) is also large when the nominator $p_3 \cdot p_4 = m_{jj}$ is large. The dijet mass can be written as

$$m_{jj}^2 \approx 2p_{\mathrm{T}}^{j_1}\, p_{\mathrm{T}}^{j_2}\left[\cosh\left(\eta_{j_1} - \eta_{j_2}\right) - \cos\left(\phi_{j_1} - \phi_{j_2}\right)\right]. \tag{1.38}$$

At constant transverse momenta of the jets, this expression is largest for a large pseudorapidity gap between the jets, and when the jets are back-to-back ($\varphi_{j_1} - \varphi_{j_2} \to \pi$).

Another feature of the electroweak production of the $ZZjj$ final state is the kinematics of the vector boson with respect to the tagging jets. The jets are preferably at low scattering angles and the gauge bosons tend to be inside the rapidity gap between these jets. The concept can be formalized by measuring the pseudorapidity of a particle X with respect to the tagging jets [15]:

$$\eta_X^* = \eta_X - \frac{1}{2}\left(\eta_{j_1} + \eta_{j_2}\right), \tag{1.39}$$

and the centrality:

$$C_X = \frac{\eta_X^*}{|\eta_{j_1} - \eta_{j_2}|}. \tag{1.40}$$

Vector boson scattering processes feature central gauge bosons with $\eta_Z^* \approx 0$.

A final characteristic of VBS and VBF processes is the suppression of central hadronic activity[1]. Additional parton emissions in these processes are reduced and tend to be collinear with the tagging jets ($C_{\mathrm{jet}\,3} \approx 1/2$).

[1] This *color decoherence* can be understood with the following argument: non-collinear gluon emissions give rise to infrared divergencies which must cancel with the divergencies from the virtual gluon exchange to result in a finite cross section. However, the virtual gluon exchange is strongly suppressed because such a gluon changes the color structure of the diagram and the interference with the corresponding tree-level diagram vanishes. The suppression of the virtual gluon exchange thus means a suppression of non-collinear emissions.

1.2.2 Effective field theory

Physics beyond the SM such as anomalous quartic gauge couplings (aQGCs) can be described in the effective field theory (EFT) framework [16, 17]. The idea behind the EFT approach is to describe the low-energy effects of BSM physics at an energy scale Λ which cannot be probed directly by the experiment.

The contact interaction model of charged currents proposed by Fermi to describe beta decay of muons is an example of an EFT. Though not a UV-complete theory, it nonetheless provides a useful description of weak phenomena below the energy scale Λ set by the electroweak gauge boson masses. Fermi's theory effectively integrated out the gauge boson degrees of freedom of the electroweak theory. Similarly, EFTs can be used to formulate potential BSM physics in a consistent framework.

The EFT Lagrangian is given by an expansion in terms of BSM operators $\mathcal{O}_i^d$ of energy dimension d and corresponding Wilson coefficients or *coupling strengths* f_i :

$$\mathcal{L}_{\text{EFT}} = \mathcal{L}_{\text{SM}} + \sum_{d>4} \sum_i \frac{f_i \mathcal{O}_i^d}{\Lambda^{d-4}} \tag{1.41}$$

The operators are constructed from the SM fields and a common restriction is to demand invariance under the SM symmetries. The construction of the EFT Lagrangian follows the same logic of allowing all terms consistent with the symmetries of the SM, but lifts the restriction of renormalizability, i.e., operators of $d > 4$ are allowed.

Putting aside operators of odd dimensions which lead to lepton or baryon number violation, the lowest dimension operators in the EFT expansion are of order 6. These lead to anomalous triple gauge couplings (aTGCs), which can be probed in diboson production. The VBS analysis presented in this thesis work provides sensitivity to the aQGCs which are described by dimension 8 operators. Table 1.1 lists all aQGC operators in the linear Higgs-doublet representation [18], where the modified field-strength tensors are given by

$$\widehat{W}^{\mu\nu} = i g_w \frac{\sigma^j}{2} W^{j,\mu\nu} \tag{1.42}$$

$$\widehat{B}^{\mu\nu} = i g \frac{1}{2} B^{\mu\nu}. \tag{1.43}$$

Figure 1.4 presents the mapping of each operator to the quartic vertices that it modifies.

Anomalous couplings that only involve the electroweak fields are given by the tensor operators $\mathcal{O}_T$. The analysis of the $ZZjj$ channel presented in this thesis work is sensitive to the operators $\mathcal{O}_{T,0,1,2}$ as well as the neutral-current operators $\mathcal{O}_{T,8}$ and $\mathcal{O}_{T,9}$. The former can also be probed in final states involving charged gauge bosons.

The generic effect of the aQGC operators is to enhance production cross section for large boson scattering energies. The fully-leptonic final state of the $ZZjj$ analysis presented in this work permits to reconstruct the boson scattering energy, which is equal to the invariant mass of the four leptons.

Table 1.1: Effective field theory operators of dimension eight. Table taken from [14]. The limits on aQGCs presented in this work are based on the operator definitions given in [18], which are related to the definitions presented here and used in [19] by a rescaling.

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

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

Figure 1.4: Overview of the gauge boson vertices which are modified by a given aQGC operator. Table taken from [14].

1.3 Status of experimental searches for vector boson scattering

The experimental search for the scattering of massive gauge bosons is a recent scientific endeavor, owing to the small cross sections of these α_{EW}^{6} processes. The first results on the scattering of weak bosons were presented by the ATLAS and CMS Collaborations in the same-sign WW channel, using the 8 TeV datasets of around 19 fb^{-1} and reporting observed (expected) signal significances of 3.6 (2.8) and 1.9 (2.9) standard deviations respectively [20, 21]. Both results were released in the summer of 2014, shortly before the start of the PhD research project presented in this thesis.

The first observation of massive gauge boson scattering was made by the CMS Collaboration in the same-sign WW channel using 35.9 fb^{-1} of proton–proton collision data at 13 TeV [22]. The same-sign WW channel provides an excellent signal-to-background ratio because the charge configuration arises only in a limited number of QCD Feynman diagrams. The fully-leptonic final states ($e^{\pm}e^{\pm}jj$, $\mu^{\pm}\mu^{\pm}jj$, and $e^{\pm}\mu^{\pm}jj$) exploited by this analysis thus provides a clean signal with little backgrounds, particularly in the signal region defined by $m_{jj} > 500\,\mathrm{GeV}$, $|\Delta\eta_{jj}| > 2.5$, and $\max_{\ell} |C_{\ell}| < 0.75$, as can be seen in Fig. 1.5. The observed (expected) significance of the electroweak signal is 5.5 (5.7) standard deviations with a measured fiducial cross section that is compatible with the SM prediction. This observation of VBS was made public at the same time as the $ZZjj$ analysis that was developed during the thesis work presented here [23, 24]. These are the only channels of massive gauge boson scattering investigated to date.

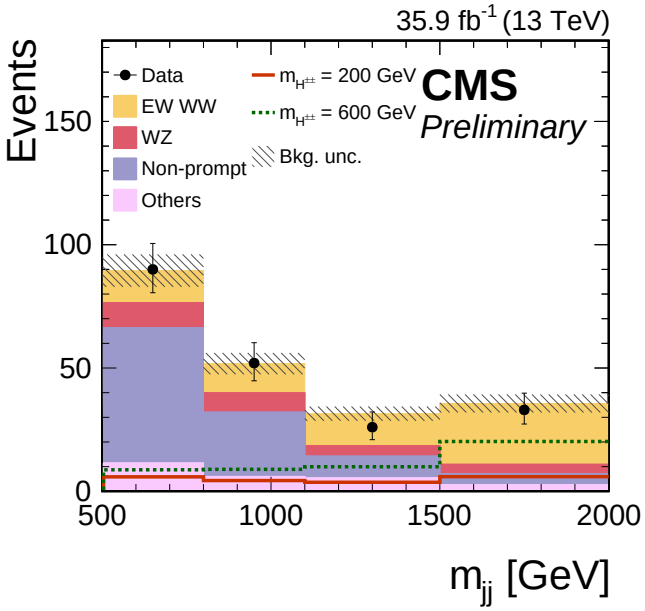


Figure 1.5: Tagging jet invariant mass distribution in the same-sign WW measurement, as reported in [22].

The ATLAS Collaboration reported limits on aQGCs and a fiducial cross section measurement of the electroweak production of the fully-leptonic final state in the WZ channel [25]. A search for aQGCs in the semi-leptonic final state of the WW/WZ+jj channel was also reported by the ATLAS Collaboration [26].

While not sensitive to the scattering of longitudinally polarized gauge bosons, the

study of final states involving photons allows to test the non-Abelian structure of the electroweak sector and to constrain anomalous couplings. These processes exhibit a similar tagging jet topology and reduced hadronic activity due to color decoherence characteristic for the scattering of massive gauge bosons. Using the 8 TeV data, the CMS Collaboration studied the electroweak production of $W\gamma jj$ and $Z\gamma jj$, reported observed (expected) signal significances of 2.7 (1.5) and 3.0 (2.1) standard deviations respectively [27, 28]. Based on the 8 TeV data, the ATLAS Collaboration reported an observed (expected) significance for the electroweak production of $Z\gamma jj$ of 2.0 (1.8) [29]. Both Collaborations reported limits on aQGCs in the studied channels.

The fusion of weak bosons also shares the tagging jet topology and reduced hadronic activity due to color decoherence phenomenology of the VBS process class. The electroweak production of a massive gauge boson in association with two jets thus permits to perform auxiliary measurements for the investigation of VBS. Figure 1.6 illustrates such measurements in the case of electroweak production of a Z boson in association with two jets at 13 TeV, presented by the CMS Collaboration [30]. The ATLAS and CMS Collaborations performed similar measurements of the electroweak production of single gauge bosons at 7 and 8 TeV [31–35].

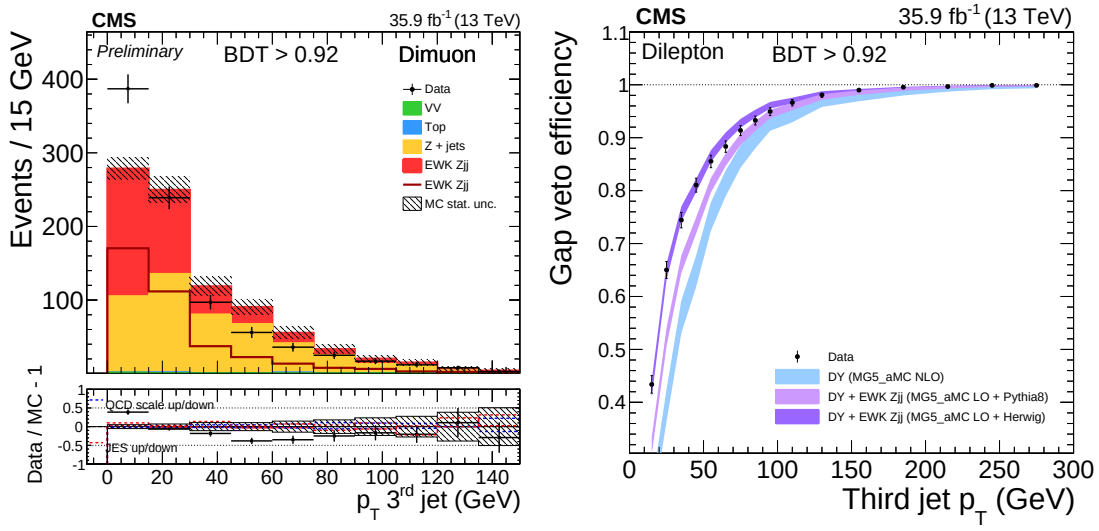


Figure 1.6: Distribution of the transverse momentum of the third soft jet in Zjj events (left panel) and efficiency of an event veto based on such jets as a function of the minimum transverse momentum of the third jet (right panel) [30]. Events without a third jet are added to the lowest p_{T} bin. The events considered in the distributions feature signal-like kinematics and are selected by a multivariate discriminant ('BDT > 0.92').

The results of the fiducial cross section measurements for the VBS and VBF processes carried out by the ATLAS and CMS Collaborations are summarized in Table 1.2. The current best limits on aQGCs, including the results presented in this thesis, are reported in the conclusions at the end of this document.

While involving photons and therefore not sensitive to EWSB, the study of exclusive or quasi exclusive production of W bosons ($\gamma\gamma \to W^+W^-$) via processes of the form $pp \to p^{(*)}W^+W^-p^{(*)} \to p^{(*)}\ell^+\ell^-p^{(*)}$, allow to study the non-Abelian structure of the electroweak interaction and permit to set limits on aQGCs. The CMS Collaboration performed such studies at 7 and 8 TeV [36–38] and the ATLAS Collaboration reported

results at 8 TeV [39]. Anomalous quartic couplings can also be constrained from diboson and triboson production [40–44].

Table 1.2: Fiducial cross section measurements of VBS and VBF processes at the LHC.

Channel	Measured fid. cross section [fb]	SM prediction [fb]	$\sqrt{s}$ [TeV]	$\int \mathcal{L} dt$ [fb^{-1}]	Collaboration	Ref.
$W^{\pm}W^{\pm}jj$	3.83 ± 0.66 (stat) ± 0.35 (syst)	4.25 ± 0.21	13	35.9	CMS	[22]
$W^{\pm}W^{\pm}jj$	$4.0^{+2.4}_{-2.0}$ (stat) $^{+1.1}_{-1.0}$ (syst)	5.8 ± 1.2	8	19.4	CMS	[21]
$W^{\pm}W^{\pm}jj$	1.3 ± 0.4 (stat) ± 0.2 (syst)	0.95 ± 0.06	8	20.3	ATLAS	[20]
$W^{\pm}Zjj$	$0.29^{+0.14}_{-0.12}$ (stat) $^{+0.09}_{-0.10}$ (syst)	0.13 ± 0.01	8	20.3	ATLAS	[25]
$W\gamma jj$	10.8 ± 4.1 (stat) ± 3.4 (syst) ± 0.3 (lumi)	6.1 ± 1.2	8	19.7	CMS	[27]
$Z\gamma jj$	1.1 ± 0.5 (stat) ± 0.4 (syst)	0.94 ± 0.09	8	20.3	ATLAS	[29]
$Z\gamma jj$	$1.86^{+0.90}_{-0.75}$ (stat) $^{+0.34}_{-0.26}$ (syst) ± 0.05 (lumi)	1.27 ± 0.12	8	19.7	CMS	[28]
Zjj	552 ± 19 (stat) ± 55 (syst)	543 ± 24	13	35.9	CMS	[30]
Zjj	54.7 ± 4.6 (stat) $^{+9.9}_{-10.5}$ (syst)	46.1 ± 1.0	8	20.3	ATLAS	[32]
Zjj	174 ± 15 (stat) ± 40 (syst)	208 ± 18	8	19.7	CMS	[33]
Zjj	154 ± 24 (stat) ± 53 (syst)	166	7	5	CMS	[31]
Wjj	159 ± 10 (stat) ± 17 (exp) ± 20 (th)	198 ± 12	8	20.2	ATLAS	[35]
Wjj	420 ± 40 (stat) ± 90 (exp) ± 10 (lumi)	500 ± 28	8	19.3	CMS	[34]
Wjj	144 ± 23 (stat) ± 23 (exp) ± 13 (th)	144 ± 11	7	4.7	ATLAS	[35]

Chapter 2

The CMS experiment at the CERN LHC

The search for the electroweak production of the $ZZjj$ final state presented in this work is carried out at the CERN Large Hadron Collider (LHC). The proton–proton collisions at a center-of-mass energy of 13 TeV studied in this analysis were recorded by the CMS experiment. This chapter provides a brief account of the LHC accelerator and introduction to the CMS detector.

2.1 The Large Hadron Collider

The Large Hadron Collider, the world's most powerful particle accelerator in terms of design collision energy (14 TeV) and instantaneous luminosity ($10 \times 10^{34}\,\mathrm{cm}^2\,\mathrm{s}^{-1}$), is housed in a 26.7 km circumference tunnel between the Swiss Alps and the French Jura mountains near Geneva. Its purpose is to provide the LHC experiments with a steady pace of high-energy particle collisions. The accelerator, often referred to as the *machine*, is part of a larger accelerator complex illustrated in Fig. 2.1, which exploits several of CERNs previous colliders: The proton beam is formed and accelerated to 50 MeV in the LINAC2, then injected into the Proton Synchrotron Booster (PSB) and Proton Synchrotron (PS), which increase the beam energy by three orders of magnitude to 26 GeV. The beam is then injected into the Super Proton Synchrotron (SPS), which brings the energy to 450 GeV before it is split and injected into the two opposing LHC beam pipes, where the maximal beam energy of 6.5 TeV is attained. Some 1232 superconducting dipole magnets with a magnetic field of 8.9 T are necessary to bend the protons into orbit, while quadru- and octupoles keep the beams focused. Once the beams are brought to nominal energy and declared *stable*, they are made to cross at the interaction points of the experiments, which record the outgoing particles resulting from the collisions.

Within the beam, protons are not spread out continuously, but grouped together in bunches, each bunch containing some $N = 1.1 \times 10^{11}$ protons.

The rate of collisions per unit time can be factorized into two parts: the cross section σ and the instantaneous luminosity $\mathcal{L}$. The former only depends on the kinematics of the interacting particles and the interaction strength; it is a fixed number and in particular independent of the characteristics of the colliding beams. In order to maximize the rate of events, one thus wants large instantaneous luminosities, which can be expressed in

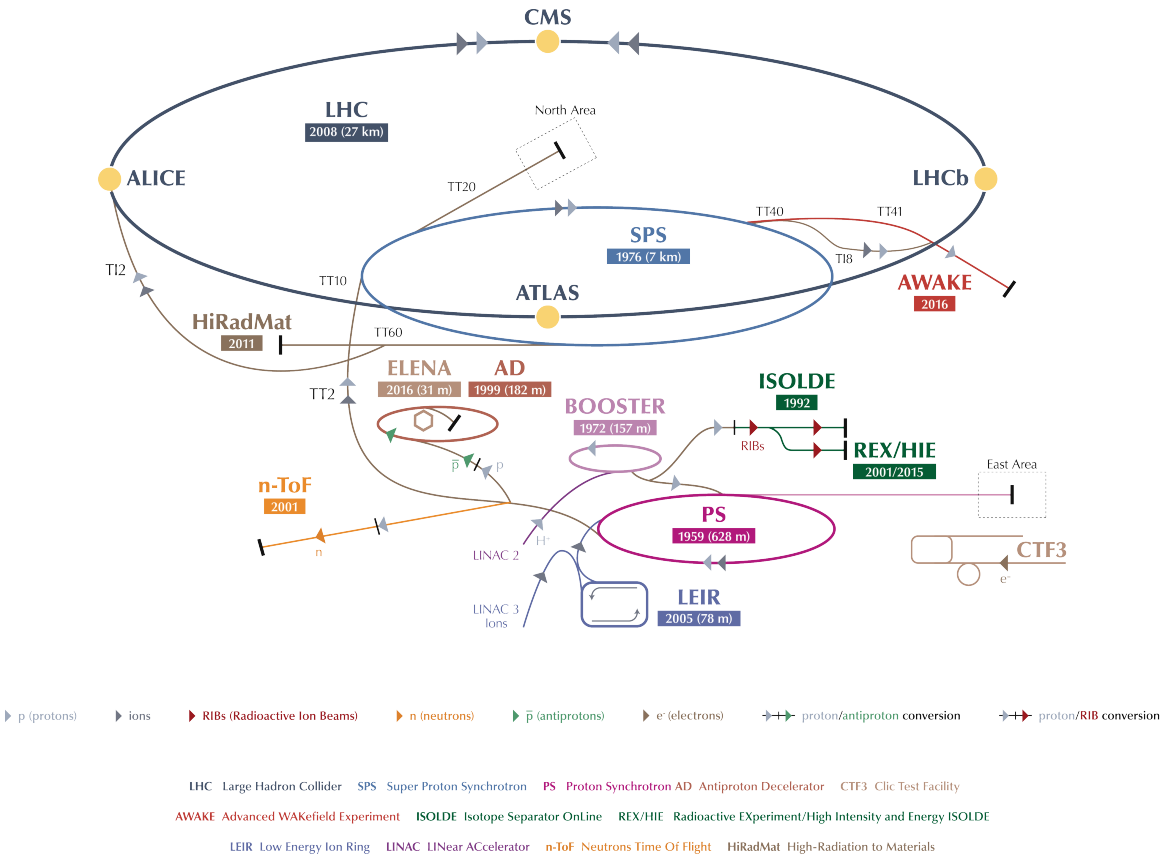


Figure 2.1: The LHC accelerator complex at CERN [45]. Before entering the LHC, protons are accelerated by the LINAC2, the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS), and finally the Super Proton Synchrotron (SPS).

terms of the beam parameters as (figures from the 2016 run)

$$\mathcal{L} = \gamma \frac{f n_b N^2}{4\pi \epsilon_n \beta^*} R \tag{2.1}$$

where $\gamma = E/m$ is the relativistic Lorentz factor of the protons, n_b is the number of bunches (2220), N is the aforementioned number of protons per bunch, f the revolution frequency, R is a reduction factor due to the beam crossing angle, $\beta^* = 40\,\text{cm}$ is the beam beta function at the collision point, and $\epsilon_n = 2.6\,\mu\text{m}$ is the normalized transverse beam emittance. Assuming the nominal beam and collision parameters, the design instantaneous luminosity of the LHC is $\mathcal{L} \approx 1 \times 10^{34}\,\text{cm}^2\,\text{s}^{-1}$. In 2016 the highest instantaneous luminosities yet of $1.4 \times 10^{34}\,\text{cm}^2\,\text{s}^{-1}$ were achieved, going beyond the nominal design specifications. The last pre-LHC hadron collider, the Tevatron, delivered peak instantaneous luminosities of $\mathcal{L} \approx 1 \times 10^{32}\,\text{cm}^2\,\text{s}^{-1}$, two orders of magnitude smaller.

The pursuit of large instantaneous luminosities, or large event rates, is driven by the need to accumulate sufficient data to study rare physics processes like Higgs boson production. However, the large instantaneous luminosity increases not just the rate at which interesting and potentially new physics processes take place, but also the rate of known processes like the QCD production of jets. The latter processes are well understood and not of primary interest for the physics program, but due to their large cross section they will occur at a high rate and overlap with the rare and interesting events. This overlap of collisions, called *pileup*, can obscure the signals and it is

one of the major experimental challenges for the LHC detectors and physics analyses. Figure 2.2 shows the event display of a typical LHC collision as recorded by the CMS detector in 2016. Each of the yellow curves corresponds to the track of a charged particle. During the 2016 data taking the average number of pileup interactions was about 20. This challenge of disentangling the interesting and rare physics was known since the conception of the LHC and the CMS detector is designed to cope with these conditions.

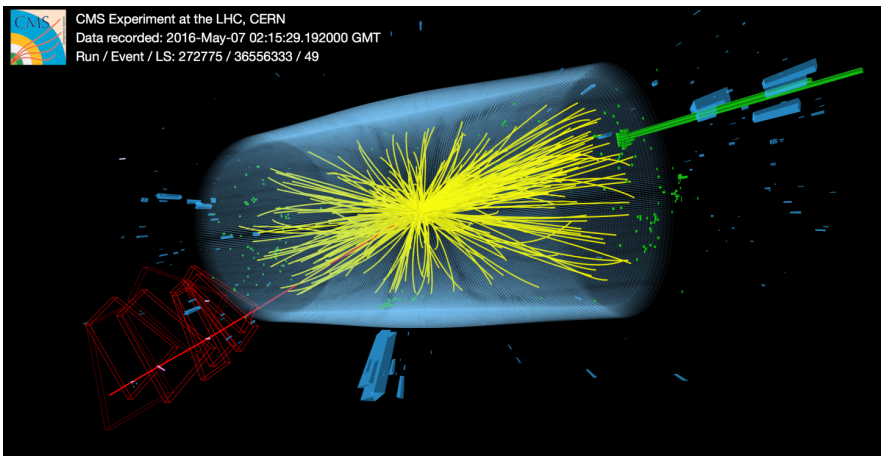


Figure 2.2: A typical proton-proton collision recorded by the CMS detector during the 2016 run [46]. The yellow curves correspond to the reconstructed tracks of charged particles, which are produced copiously in pileup interactions.

The search for the electroweak production of two Z bosons and two jets presented in this analysis is based on proton–proton collisions recorded in 2016 at a center of mass energy of 13 TeV.

2.2 The CMS experiment

2.2.1 Design philosophy and overview

The Compact Muon Solenoid (CMS) experiment is a multi-purpose detector, designed to achieve the LHC physics program which includes the search for the SM Higgs boson and the investigation of EWSB, the search for Supersymmetry and other physics beyond the Standard Model (BSM) at the TeV scale.

Since the mass of the Higgs boson is a free parameter in the SM, it had to be searched for in a large mass window from 100 GeV to about 1 TeV. Due to the strong dependence of the Higgs boson branching ratios on its mass, this meant that the detector has to be able to reconstruct and identify the physics objects of the final states best suited for a given mass. Among the key analyses that enabled the 2012 discovery of the Higgs boson at approximately 125 GeV was the $H \to ZZ^* \to 4\ell$ channel. With an excellent signal to background ratio but low signal rate, the main challenge in this analysis is the reconstruction of all four final state leptons as well as an excellent lepton momentum measurement. This means that the detector has to cover a large portion of the 4π solid angle, i.e., be *hermetic* and allow lepton reconstruction down to a few GeV. The $H \to \gamma\gamma$ analysis requires an excellent diphoton mass resolution to render the small Higgs peak visible on top of a large falling background. This means

a percent-level energy resolution for photons. Reconstruction of particle tracks and the ability to find primary and secondary vertices are crucial for searches involving decays of tau leptons or B mesons in $H \to b\bar{b}$ or $H \to \tau\tau$ analyses.

The requirements for a successful Higgs search program at low masses are supplemented by the needs posed by searches for physics beyond the SM. Some of these BSM models predict high-mass resonances that decay into energetic photons, leptons, or jets. In order to enable the searches involving these high-p_T objects, the calorimeters have to be able to reliably measure energies up to several TeV (a good *linearity*). The quality of the muon p_T measurement on the other hand is largely determined by the bending power or strength of the magnetic field, that is one needs a large magnetic field and long lever arm.

The large instantaneous luminosities of the LHC impose further constraints on the detector design. The bunch crossing rate of 40 MHz requires sensitive materials and readout electronics that allow association of a signal to a given bunch crossing and fast extraction of the signal in order to avoid dead time. Occupancy, particularly in the tracking detectors, has to be kept low in order to allow the reconstruction of individual particles. This is achieved with a high granularity which is crucial to disentangle the signal of the hard scattering interaction from pileup.

Finally, the sensitive material and the on-detector electronics need to be capable of operating in the high-radiation environment at a low failure rate. The latter is particularly relevant because the detector is inaccessible during the data-taking due to high radiation levels and even during the technical stops only some parts of certain sub-systems can be replaced at considerable effort.

Figure 2.3 shows a schematic overview of the CMS detector and its sub-systems. In the central or *barrel* part of the detector, the layers of the sub-systems are cylindrical around the beam axis, while they are perpendicular to the beam in the *endcaps*. The full detector name - Compact Muon Solenoid - indicates some of the specific choices made in its design in order to achieve the goals of the LHC physics program. The detector is small with a length of 22 m and a diameter of 15 m, compared to the ATLAS detector, which measures 46 m in length and 25 m in diameter. The defining feature is the superconducting solenoid which houses the all-silicon tracker and the compact electromagnetic and hadronic calorimeters. The muon system is the only subdetector situated outside the solenoid and embedded in the steel return yoke.

CMS coordinate system

A standardized coordinate system is adopted to describe the CMS detector and the reconstructed particles. Its origin is at the nominal interaction point at the center of the detector. The x axis points to the center of the LHC ring and the y axis points upwards. The z axis is tangential to the beam direction and points towards the Jura mountains, giving a right-handed coordinate system. Cylindrical coordinates are also commonly used, where the transverse plane is given by the x-y plane. The transverse momentum is the projection of any momentum $\vec{p}$ onto the x-y plane and often used to mean the magnitude of this projected vector:

$$p_T = \sqrt{p_x^2 + p_y^2} \tag{2.2}$$

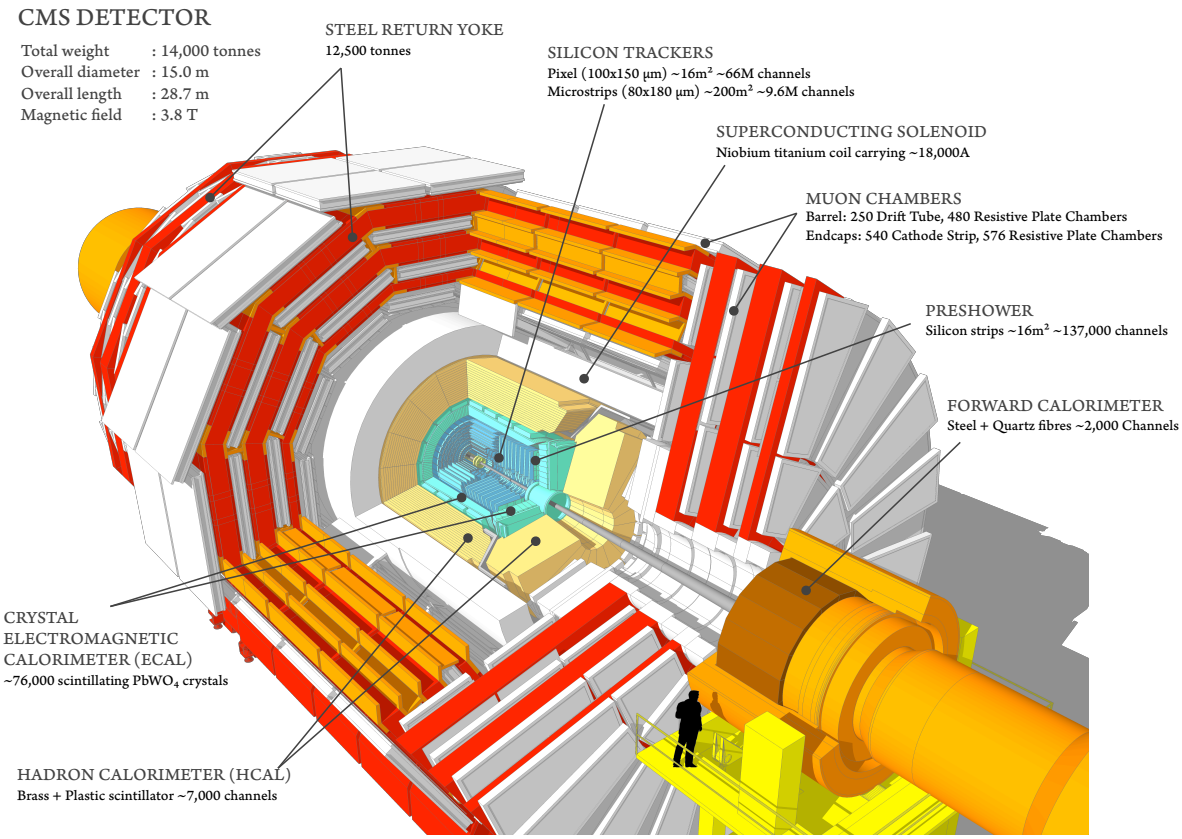


Figure 2.3: Schematic overview of the CMS detector and its key sub-systems, illustrating the split into the barrel and endcap sections. Updated figure from [47].

The rapidity of a particle is given by

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right). \tag{2.3}$$

and approaches the pseudorapidity in the limit where the mass is negligible compared to the energy, $m/E \ll 1$:

$$\eta = \frac{1}{2} \ln \left(\frac{p + p_z}{p - p_z} \right) = -\ln \tan \frac{\vartheta}{2}. \tag{2.4}$$

The angular distance between two particles with azimuthal angles φ_i and pseudorapidities η_i is commonly expressed as ΔR which is defined as:

$$\Delta R = \sqrt{(\eta_1 - \eta_2)^2 + (\varphi_1 - \varphi_2)^2} \tag{2.5}$$

Differences of pseudorapidities are invariant under longitudinal Lorentz boosts. Finally, one defines the transverse missing energy $E_{\mathrm{T}}^{\mathrm{miss}}$ as the negative momentum sum of all reconstructed particles projected onto the transverse plane.

2.2.2 Tracking system

The closest detector to the particle collisions, and therefore the first detector traversed by the particles coming out of the collisions is the tracking system, or *tracker*. The

objective of the tracker is to reconstruct the tracks, i.e., the trajectories of all charged particles resulting from a collision. By measuring the curvature of the track in the magnetic field, the momentum of the particle is measured. Considering the large number of interactions per bunch crossing of the LHC, a crucial task of the tracking system is to resolve the large number of pileup interactions and separate them from the interesting hard interaction. Failure to separate the particles of the hard interaction from the soft pileup vertices would degrade the quality of the physics results, in particular, the study of the Higgs boson. The innermost layers of the tracker allow identifying and resolving secondary vertices which are central to identifying hadronic jets from b quarks.

Tracking is achieved by having successive layers of sensitive elements that are capable of registering the passage of a charged particle via its ionization effect in the element. A schematic representation of the sensitive layers of the CMS tracking system is shown in Fig. 2.4.

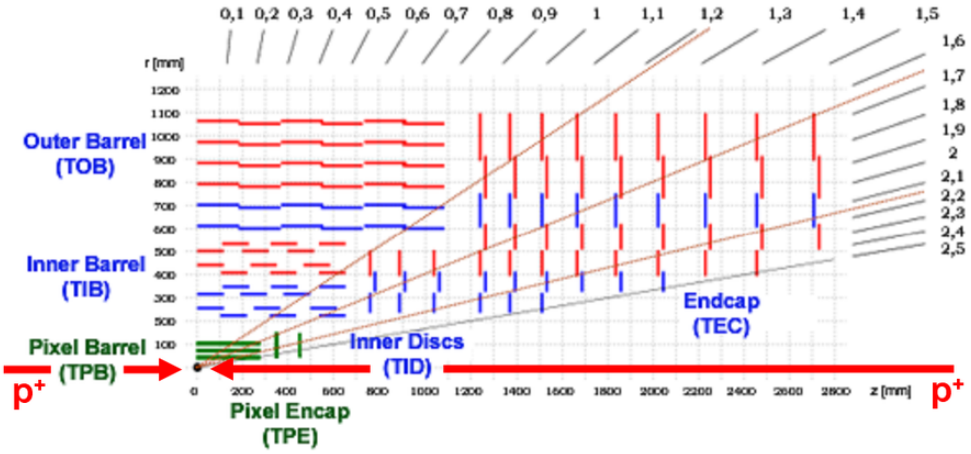


Figure 2.4: Geometry of the sensitive elements of the CMS tracker in the longitudinal plane [48]. Only a single quadrant of the full detector is shown.

The active elements of the CMS tracker are organized in layers and made of thin silicon sensors which are further split in the x-y plane into individual readout cells of rectangular shapes. Two different technologies are used: silicon pixel and silicon strip sensors. A charged particle crossing one of these cells will give rise to an electric signal which the local reconstruction turns into a *hit*. By having a precise knowledge of the spatial position of the silicon sensors, one can then infer the position of these hits which are used to reconstruct the particle trajectory.

The innermost layers of the CMS tracker are of the pixel detector, which is split into the barrel and two endcaps. As shown in Fig. 2.4, the Tracker Pixel Barrel (TPB) consists of three cylindrical layers that measure in z and are located at $r = 4.4, 7.3$ and $10.2\,\text{cm}$. The Tracker Pixel Endcap (TPE) is made of two vertical disks positioned at $|z| = 34.5$ and $46.5\,\text{cm}$, with the active elements located between $r = 6$ and $15\,\text{cm}$. In total, the pixel detector counts 65 million silicon pixels of $100\,\mu\text{m} \times 150\,\mu\text{m}$. The size of the pixels allows for an excellent resolution of $10\,\mu\text{m}$ in (x, y) and $20\,\mu\text{m}$ along z. The principal reason for the small pixel size, or high granularity, is the need to separate the hits from particles that are near-by and to resolve the z coordinate of particles coming from different vertices to suppress pileup. Another goal of the high granularity is the capability to identify secondary vertices in the decay of heavy mesons.

However, having such a large number of readout channels requires more on-chip electronics for the signal readout and high-voltage supply which in turn demands a larger cooling capacity. All this adds passive or *dead* material to the detector, which increases the chance of particle interactions within the detector and as a consequence has adverse effects on performance (see in particular Section 3.2.2 on electron reconstruction). All of these points also add to the cost of the detector. Equipping the entire tracking detector with pixels is not necessary as the increase in radius and the resulting growth of surface per solid angle will allow to use larger silicon cells: silicon strips.

Silicon strip detectors allow covering large surface areas by reducing the number of readout channels and the required electronics. This is achieved by increasing the length of a single silicon cell from a hundred µm to several cm. Here length refers to the coordinate that matters the least for the curvature and momentum measurement: the z coordinate in the barrel and the radial direction in the endcap disks.

The CMS strip detector is separated into the Tracker Inner Barrel (TIB), Tracker Outer Barrel (TOB), Tracker Inner Disk (TID), and the Tracker Endcap (TEC). It covers a total surface area of 65.6 m^2 with 4.6 million channels. The TIB consists of 4 layers located between $r = 30$ and 55 cm that extend up to $|z| = 65$ cm with strips of width (*pitch*) between 80 and 120 µm. Three disks on each side of the TIB complement the coverage at similar strip pitches. The outer barrel system consists of six layers up to $r = 116$ cm and $|z| = 110$ cm with pitches ranging from 120 to 180 µm. The nine TEC disks feature pitches similar to those found in the TOB and the last disk is located at $|z| = 264$ cm, providing a coverage up to $|\eta| = 2.5$.

An inherent disadvantage of the strip layout of the silicon sensor is the poor hit resolution along the strip. *Stereo* layers (highlighted in blue in Fig. 2.4) drastically increase this resolution by having two strip modules in the same layer with one being tilted by 12.6° with respect to the z axis, allowing to infer the z coordinate from the two strip sensors.

2.2.3 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) provides the energy measurement for electrons and photons, and crucially, allows triggering on these objects. The CMS ECAL is a homogenous scintillating crystal calorimeter, meaning all the energy of the electromagnetic shower is deposited in instrumented detector material. Similarly to the tracker, the ECAL is split into a barrel and two endcap sections, as shown in Fig. 2.5. The 75 848 crystals are arranged in a quasi-projective geometry, which tilts the crystals by 3° with respect to the nominal interaction point to avoid projective gaps between crystals. Figure 2.6 shows the layout of the ECAL in the longitudinal plane, highlighting the maximum coverage up to $|\eta| = 3$ and the barrel-endcap transition regions between $1.479 < |\eta| < 1.566$.

The compactness of the CMS ECAL is achieved with lead-tungstate ($PbWO_4$) scintillating crystals, which have a short radiation length of $X_0 = 0.89$ cm and small Molière radius of $R_M = 2.2$ cm. Another important crystal parameter is the short scintillation light emission time (80% within 25 ns). Operating a lead tungstate calorimeter at a hadron collider poses a challenge due to irradiation damage to the crystal structure, which reduces the optical transparency of the crystals. This loss of transparency due to radiation damage is monitored by illuminating the crystals with a calibrated laser pulse and recording the apparent drop in the energy response, as illustrated in Fig. 2.7.

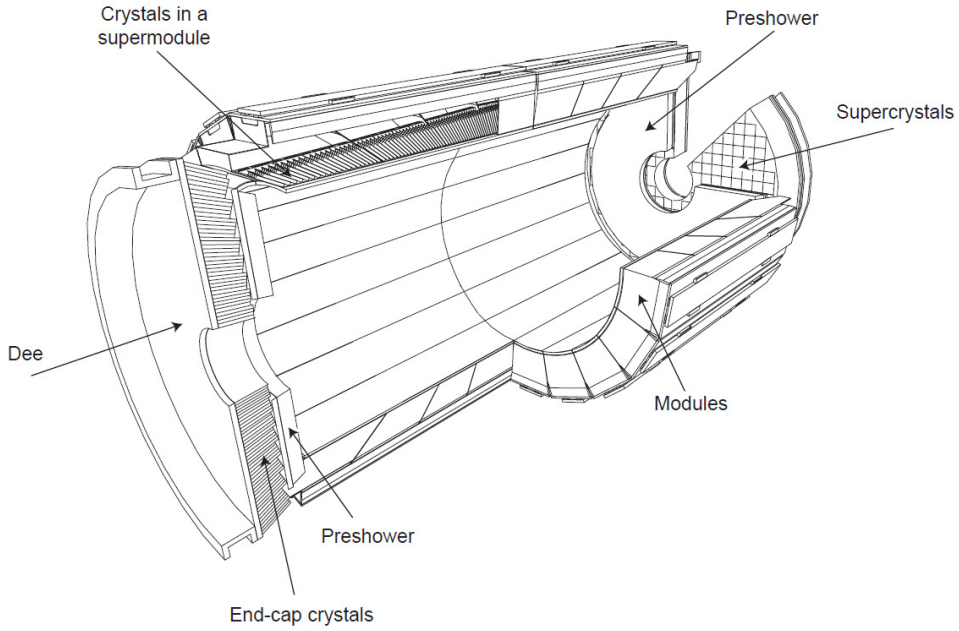


Figure 2.5: Schematic view of the CMS ECAL and the mechanical structure [49]. In the barrel, crystals are grouped into modules and super-modules. Each endcap consists of two half-disks or *dees*. The pre-shower detector covers most of the endcap surface.

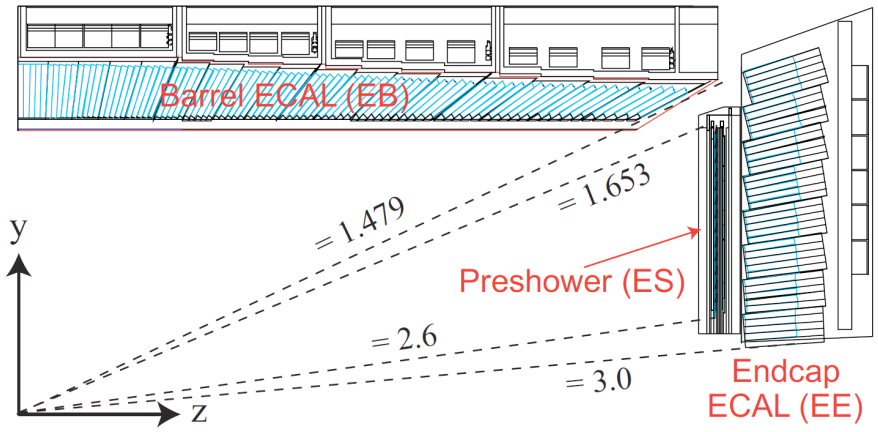


Figure 2.6: Longitudinal view of a ECAL quadrant [49]. The pseudorapidity coverage of the barrel, endcap, and preshower systems are indicated.

The ECAL is split into a barrel (EB) and two endcap (EE) systems. Barrel crystals cover 0.0174 in η and ϕ, corresponding to a front face cross section of 2.2 cm $\times$ 2.2 cm, and a length of 23 cm (25.8 X_0). In the barrel, crystals are grouped into modules, which in turn are mounted into 36 super-modules comprising 1 700 crystals each. Small gaps between super-modules are needed for mechanical support and result in small inefficiencies in the energy measurement in these regions. Light detection in the barrel is done via avalanche photo diodes (APD), while vacuum phototriodes (VPT) are used in the endcap due to higher radiation levels. The endcap crystals feature a larger front face cross section of 2.86 cm $\times$ 2.86 cm and are 22 cm long (24.7X_0). The endcap ECAL in the range $1.65 < |\eta| < 2.6$ is supplemented by the preshower (ES), a lead-silicon sampling structure, which helps to separate single prompt photons from collimated

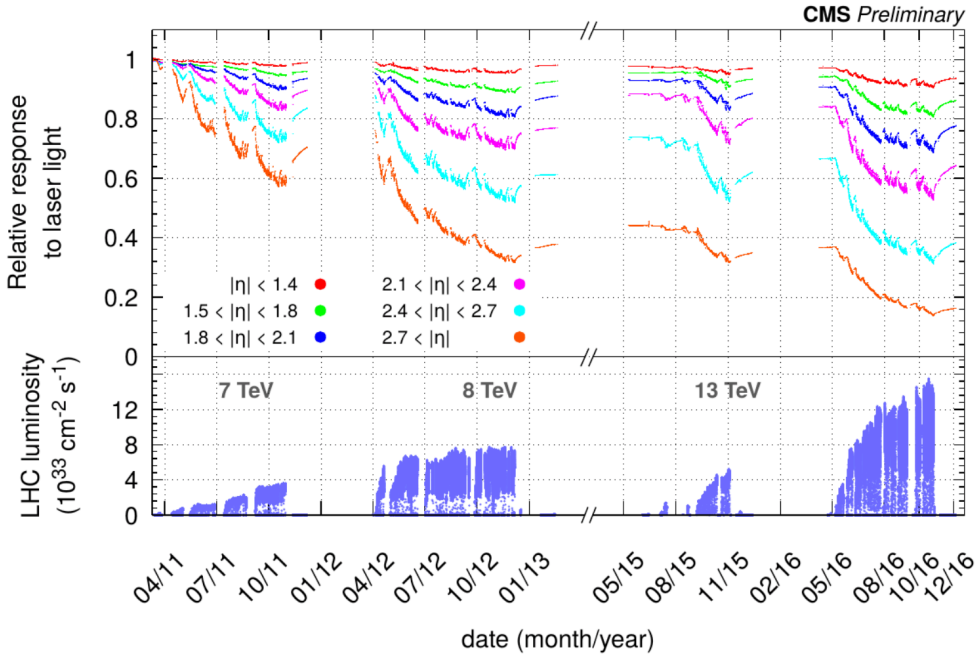


Figure 2.7: Time evolution of the ECAL response to the monitoring laser $R(t)$ [50]. The reduction in response during data-taking periods, most pronounced for large pseudorapidities, is caused by radiation damage to the ECAL crystals. The response is monitored and ultimately corrected via the laser monitoring system. Some recovery of the response during collision-free periods is also visible.

photon pairs from $\pi^0 \to \gamma\gamma$ decays. Approximately 6 % to 8 % of the energy in an electromagnetic shower is deposited in the ES. Thanks to its homogeneity, the ECAL achieves an excellent energy resolution of the order of a few percent for electrons at 15 GeV and about 1.7% at 45 GeV.

Electron test beam measurements on a 3×3 matrix of ECAL crystals show that the measured energy resolution is described by the usual parameterization [51]:

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E}}\right)^2 \oplus \left(\frac{12\%}{E}\right)^2 \oplus (0.3\%)^2\,, \tag{2.6}$$

where the energy E is given in GeV. Noise from the readout electronics, corresponding to the second term, contributes only at very low energies. At intermediate energies the first term contributes most, accounting in particular for shower-by-shower variations in the scintillation light yield (*stochastic term*). For electrons above 50 GeV the resolution is mostly determined by the constant term.

The quality and stability of the energy measurement from the ECAL is achieved by a series of calibrations and corrections. The reconstructed energy for an electromagnetic cluster is decomposed as

$$E_{e,\gamma} = F_{e,\gamma} \cdot \left[G \cdot \sum_i S_i(t) \cdot C_i \cdot A_i + E_{\mathrm{ES}} \right] \tag{2.7}$$

where the sum includes all crystals within the cluster and each term captures a different aspect of the calibration:

- **Pulse amplitude (A_i):** the energy deposited in a crystal is digitized by the analogue-to-digital converters (ADC). The pulse is sampled 10 times at 40 MHz and the three samples prior to the rise of the pulse are used to establish the pedestal.
- **Intercalibration coefficiencts (C_i):** the initial set of relative channel-by-channel calibrations was obtained from laboratory measurements, beam tests, and cosmic muons. Since the start of the LHC, the intercalibration coefficients are obtained from collision data by exploiting three methods, each performed for a ring of crystals at constant pseudorapidity. The φ-symmetry method is based on the expectation that the average transverse energy in minimum bias events is independent of the angle φ. The invariant mass of π and η meson decays into photons is also used. Finally, the comparison of the calorimeter energy and the tracker momentum for electrons from W and Z boson decays are exploited.
- **Response corrections ($S_i(t)$):** the changes $R_i(t)$ of the response due to irradiation as discussed previously and illustrated in Fig. 2.7 are tracked by the laser monitoring system. The relation between the response to the laser light and scintillation light is given by a power law with exponent α.
- **Preshower energy (E_{ES}):** the preshower energy is obtained as the weighted sum of the two responses of the two preshower planes to minimum ionizing particles and is ultimately calibrated to GeV.
- **Energy scale (G):** The absolute energy scale is established using $Z \to e^+e^-$ events in data. The invariant mass distribution for electrons in the barrel and endcap are fitted separately. The corrections are validated using final state radiation photons in $Z \to \mu^+\mu^-\gamma$ events as well as E/p comparisons for electrons from W and Z boson decays. The constant G corresponds to the ADC-to-GeV conversion factor for the APDs (VPTs) in the barrel (endcaps).
- **Energy containment corrections ($F_{e,\gamma}$):** superclusters (see Section 3.1.1 for details) are corrected for energy containment effects. These arise from geometric effects like energy losses in calorimeter gaps and energy losses due to upstream material. These semi-parametric corrections are derived from the simulation using a multivariate technique. The corrections are derived separately for electrons and photons to account for the different interactions of both particle species with the upstream material.

2.2.4 Hadronic calorimeter

The hadronic calorimeter (HCAL) serves to measure the energy of long-lived hadrons that traverse the tracker and ECAL. It provides the only energy measurement for neutral hadrons and complements the momentum measurement of the tracker for charged hadrons. The CMS HCAL calorimeter is a compact sampling calorimeter located within the solenoid. Brass is used for the mechanic structure and energy absorber in the barrel and endcaps. Plastic scintillating tiles are coupled to wavelength shifting fibers which transmit the signal to multi-channel hybrid photodiodes for readout. Figure 2.8 shows a longitudinal view of the HCAL geometry.

Coverage in the barrel ($|\eta| < 1.4$) is provided by the hadron barrel (HB), which features 2304 towers of $\Delta\eta \times \Delta\varphi = 0.087 \times 0.087$ that are read out as a whole. The hadron outer (HO) calorimeter is located outside the solenoid and covers $|\eta| < 1.26$. It increases the

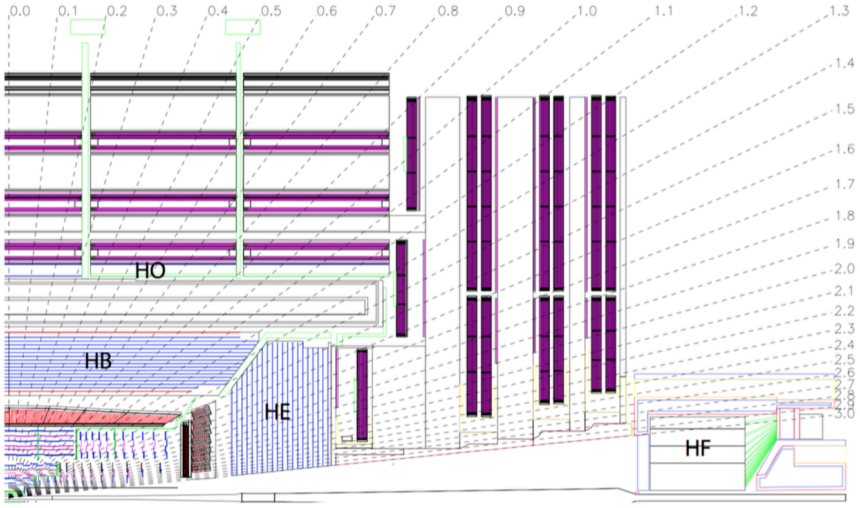


Figure 2.8: Longitudinal view of the geometry in a HCAL quadrant [52]. The location and pseudorapidity coverage of the barrel (HB), outer (HO), endcap (HE), and very forward (HF) hadron calorimeters are illustrated.

total number of interaction lengths to 10, decreasing the leakage of energetic hadronic jets into the muon system, which helps to reduce non-Gaussian tails in the energy resolution as well as lowering the odds of such jets being misidentified as muons. The endcaps are instrumented with the hadron endcap (HE), covering $1.3 < |\eta| < 3.0$ and extended coverage up to $|\eta| = 5.0$ is provided by the hadron forward (HF) system. Quartz fibers in the HF are used collect the energy of the showers developing in the iron absorber and to produce the signal by exploiting the Cherenkov effect.

2.2.5 Solenoid

The superconducting solenoid is crucial to the detectors' performance by providing magnetic field exploited in the track momentum and charge measurements. The magnetic field is strongest at 3.8 T in the central region of the detector and permeates the tracker, ECAL, and HCAL, see Fig. 2.9. The iron return yoke provides the field strength for the muon system and guarantees a low magnetic field in the CMS cavern. The solenoid is about 13 m long and the coils which carry the 20 kA current are located at a radius of about 3 m from the beam line.

2.2.6 Muon system

Muons are the only detectable particle species that traverses the entire CMS detector, resulting in a clean signal that is exploited in many physics analyses including the one presented in this thesis. The muon system also provides inputs to the trigger system, exploiting the low background rate and its excellent timing resolution. The muon system is the outermost subdetector of the CMS system with an instrumented surface area of about 25 000 m^2. Covering such large surfaces with detectors at a reasonable cost is typically done using gaseous detectors, three types of which are used in CMS. Figure 2.10 shows a longitudinal view of the CMS muon system and the three types of detectors.

In the barrel region, where the magnetic field in the return yoke is small and homo-

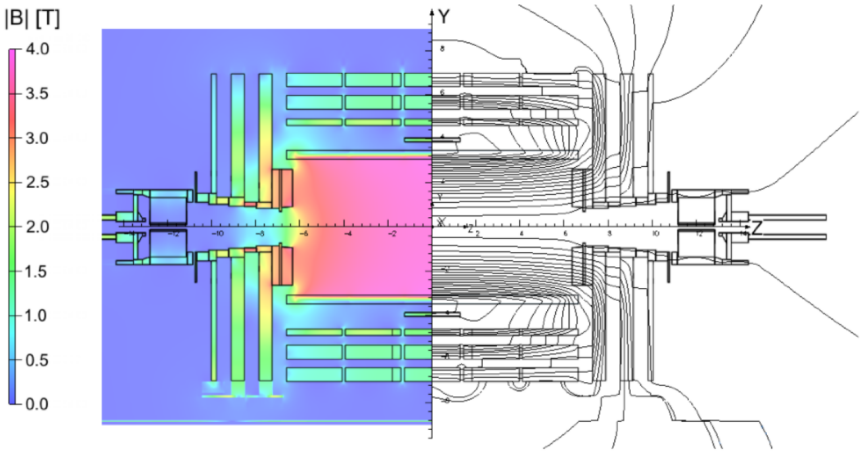


Figure 2.9: Map of the magnetic field strength in the CMS detector [53].

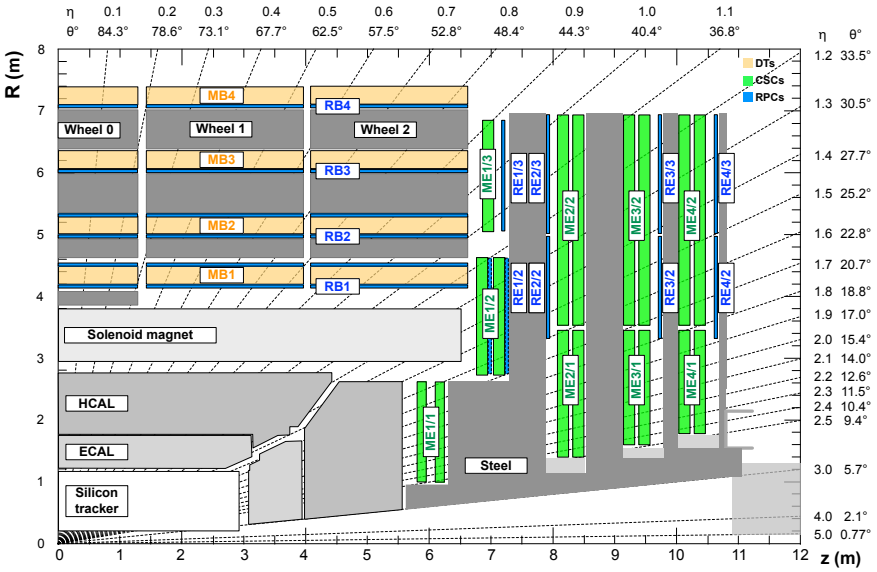


Figure 2.10: Longitudinal view of the muon system showing the position of the drift tube (DT) detectors in the barrel, the cathode strip chambers (CSC) detectors in the endcap and the resistive plate chambers (RPC) used for triggering [53].

geneous, the drift tube (DT) technology is chosen with coverage up to $|\eta| = 1.2$. Hits in the DT chambers are reconstructed based on the timing information of the electron avalanche caused by the crossing muon, with a resolution of about 100 µm in the r-φ plane. Three concentric muon stations are embedded in the iron return yoke, with a fourth station on the outside. The first station is positioned at a distance of 4.5 m from the center of the detector, the fourth station is located at 7.5 m. The barrel is furthermore split into 5 rings, each about 2.5 m long.

The endcaps are equipped with radiation hard cathode strip chambers (CSC), which cover the pseudorapidity range $0.9 < |\eta| < 2.4$. In CSCs the electron avalanche is collected by an anode wire, giving rise to an image charge on the cathode strips. The CSC chambers are trapezoidal in shape and grouped into 4 stations per wedge which compose six CSC layers. The hit resolutions range between 75 and 150 µm in the azimuthal direction and about 200 µm in the radial direction.

In addition to the DT and CSC, a third type of detector is used to guarantee the quality of the muon trigger decisions even at the highest LHC luminosities. Resistive plate chambers (RPC) provide a moderate spatial hit resolution, but the time resolution of about 1 ns allows unambiguous assignment of a muon to a bunch crossing. The RPCs are double gap chambers operated in avalanche mode to ensure reliable operation at high rates. Six RPC layers are installed in the barrel and three in each endcap.

2.2.7 Trigger system

The LHC bunch crossing rate of 40 MHz makes it impossible to store the detector readout for every collision. A dedicated readout system called *trigger* is used to filter out those events that are potentially interesting for physics analysis. The CMS trigger system consists of two systems that progressively read out the detector and decide to forward interesting events for further processing.

The first layer of the CMS trigger, the Level-1 or L1 trigger, is implemented on custom hardware and reduces the event rate from the bunch crossing rate of 40 MHz to a maximum of 100 kHz. It performs a limited readout of the calorimeters and the muon system with reduced granularity and has about 4 µs to decide whether an event is interesting for further analysis. In the first step the information of the calorimeters and muon system are treated independently. The *calorimeter trigger* reconstructs electromagnetic and hadronic clusters (the former are referred to as $e\gamma$ or *EG* candidates) while the *muon trigger* is responsible for reconstructing muon candidates. The two trigger flows are then combined for a more sophisticated analysis of the event. The Level-1 ultimately decides whether to pass the event on to the second trigger layer or whether to discard it. Only events that satisfy the requirements of at least one of the L1 *seeds* that form the L1 *trigger menu* are retained (in 2016 around 200 L1 seeds were used out of the 512 allowed by the L1 trigger logic). Each L1 seed specifies a list of requirements that need to be satisfied in order to *fire* the trigger, i.e., to pass the event to the next trigger level. The readout and electronics of the L1 trigger were significantly improved during the long shutdown and following the 2015 run, allowing for more sophisticated algorithms to be run, improving the position and energy resolution for jets and EG candidates in particular.

Events that pass the L1 trigger are passed on to the second stage of the CMS trigger system called High Level Trigger (HLT) which filters the events to achieve a maximum event rate of 1 kHz. Contrary to the L1, the HLT is a pure software trigger run on commercial computers. It exploits the full granularity of the entire CMS detector and runs a streamlined version of the offline event reconstruction algorithms[1] that reduces the event reconstruction time to about 150 ms or about 1/100 of the offline reconstruction.

In order to pass the HLT an event needs to satisfy the requirements of at least one of its *paths*, which are defined in the HLT menu, similar to the L1 trigger. Each trigger path targets a certain event topology, e.g., the presence of two prompt and isolated electrons or muons. The trigger path defines a sequence of modules which are run sequentially and either reconstruct a certain object or perform a selection based on these reconstructed objects (*filter*). The sequence is organized such that computationally expensive modules are run last in order to speed up the overall execution. If an event satisfies one of the trigger paths, i.e., it passes at least one *final filter*, it is marked for permanent storage and transferred to the CERN T0 in one or more *data streams*. Data

[1] It is common to refer to the HLT reconstruction algorithm as the *online* reconstruction, as opposed to the offline reconstruction which uses all of the advanced reconstruction algorithms.

streams gather similar trigger paths that are commonly used by the offline analyses, e.g., the DoubleEG data stream will contain all events that fired one of the dielectron trigger paths.

The HLT software was upgraded in order to cope with the larger event rates at higher luminosity and increased pileup of the run II. A trivial way to reduce the event rate would be to simply increase the p_{T} thresholds in the trigger paths, but this would severely degrade the physics program, particularly the $\mathrm{H}\to ZZ^{*}\to 4\ell$ analysis. Instead, the approach is to port some of the advanced algorithms of the offline reconstruction to the HLT, in particular the particle flow reconstruction and the associated particle identification and isolation algorithms [54] as well as the Gaussian Sum Filter track fitting for electrons (the electron tracking is described in Section 3.2.1).

2.3 LHC and CMS operations

The first LHC collisions recorded by CMS during this thesis project happened in April 2015, marking the end of the long shutdown I and start of the run II. The 2015 data taking allowed to commission the LHC and the detectors at the increased center-of-mass energy of 13 TeV. As part of this thesis work, this early data also enabled the commissioning of the improved electron identification algorithm. In total, some $4\,\mathrm{fb}^{-1}$ of collisions were delivered to the experiments in 2015, but part of the data recorded by CMS was at a reduced or no magnetic field due to problems with the cryogenics system of the solenoid magnet.

The commissioning of the LHC in 2015 enabled an exceptionally productive 2016 data taking, marked by beyond-design instantaneous luminosities of $1.4\times10^{34}\,\mathrm{cm}^{-1}\,\mathrm{s}^{-1}$. High instantaneous luminosities and the excellent operational availability of the LHC accelerator complex allowed to deliver a total of over $40\,\mathrm{fb}^{-1}$ to the experiments in 2016, see Fig. 2.11. The yellow histogram in Fig. 2.12 shows the amount of data that was recorded by the detector, and after requiring that all sub-systems were functional and the data of high quality, the total amount of data exploitable for physics analyses is $35.9\,\mathrm{fb}^{-1}$ [55]. The search for vector boson scattering presented in this thesis is based on this dataset.

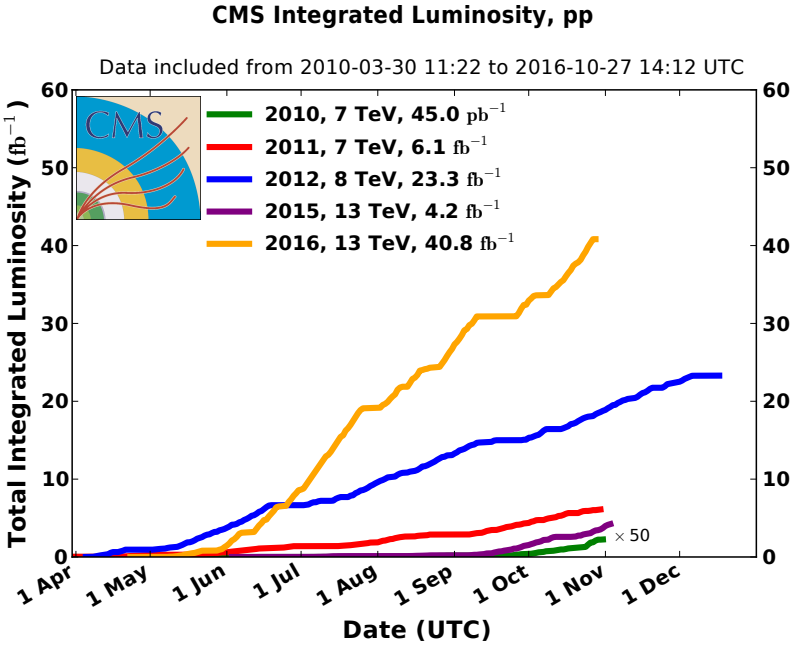


Figure 2.11: Cumulative integrated luminosities per day as delivered to CMS in the 2016 and previous proton–proton (pp) data-taking periods [56].

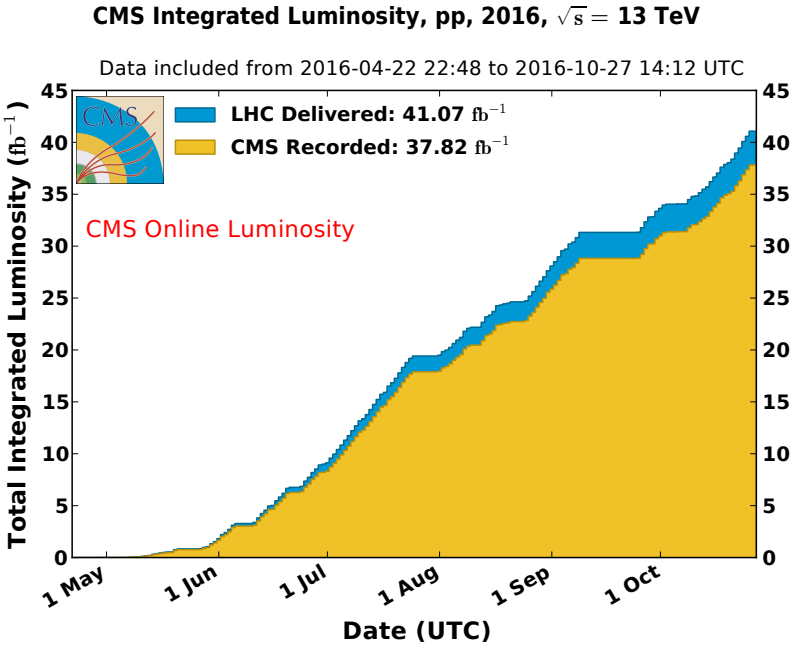


Figure 2.12: Cumulative integrated luminosities per day as delivered to CMS [56]. The total integrated luminosity for physics analysis of the 2016 dataset is 35.9 fb^{-1} [55].

Chapter 3

Physics object reconstruction and selection

The electroweak production of $ZZjj$ in the fully leptonic final state is an extremely rare process and the 2016 dataset is expected to contain only about ten such events. The ability to efficiently reconstruct and identify the leptons in the complex final states produced by the LHC at high luminosities is thus crucial to the success of this analysis. Advanced reconstruction and selection algorithms that combine the information from all subdetectors are used to select the physics objects on which the $ZZjj$ analysis is based. The following discussion of the particle-flow reconstruction algorithms starts with the calorimeter clustering and the reconstruction of charged particle tracks. The reconstruction of the objects most relevant in the $ZZjj$ analysis – leptons and jets – is then presented in detail. For each object the CMS-wide reconstruction algorithm is described first, followed by a discussion of the object selection specific to the $ZZjj$ analysis.

3.1 Event reconstruction and the particle-flow algorithm

With several million readout channels the CMS detector is able to collect a great wealth of information from each collision. In order to make sense of these readouts, e.g., an energy reading from an ECAL crystal[1], and analyze them for interesting physics processes, the signals coming from the subdetectors need to aggregated into abstract objects, called *physics objects*, that can serve as a starting point for further analysis. Examples include the grouping, or clustering, of several energy deposits in adjacent ECAL crystals into an energy cluster or the association of multiple tracker hits to form a particle track. These physics objects can then be used to construct even more involved physics objects, for example, tracks and ECAL clusters are used to reconstruct electrons.

Apart from practical aspects like runtime and computing resource requirements, an important metric of any reconstruction is its efficiency and the purity of the resulting collections. The general purpose of the reconstruction is to provide a common collection of particle candidates with maximal efficiency at an affordable fake rate. As an example, the electron reconstruction has a signal efficiency of about 95 % but will

[1] The energy measurement in a single crystal is the result of the *local reconstruction*. The analog amplitudes from the photo-detectors are sampled at 40 MHz and digitized at 12 bit. An advanced fitting algorithm then determines the signal amplitude, also reducing the bias of out-of-time pileup.

result in a large contamination of fake electrons from hadronic jets or non-prompt electrons from photon conversions. It is thus necessary to apply further selections on the reconstructed objects to increase the purity of the selection. Common selections include an identification (ID) selection based on more refined observables (in the case of electrons this could be the shape of the electromagnetic cluster or the quality of the track-cluster matching), and isolation (ISO), that is the absence of other energetic particles close to the candidate. This multi-step approach enables a common definition of particle candidates at the reconstruction level, followed by further selections to reduce the fake rates as appropriate for the physics analysis in question.

The traditional approach to reconstructing high-level objects such as electrons or jets is to focus on the subdetectors most relevant to the object at hand. Electrons and photons are largely reconstructed using ECAL information, hadronic jets are based on the calorimeter clusters, and tracking is mostly done to distinguish electrons and photons, to identify displaced vertices for the identification of b-hadron decays.

While simple and effective for objects like prompt electrons, this approach has drawbacks which we illustrate for the case of jet energy measurement and missing transverse momentum. Jets and their energy have traditionally been determined solely based on the constituent calorimetric clusters. The energy in a typical hadronic jet will be split into about 65 % charged hadrons, 25 % photons from π or η meson decays, and 10 % from long-lived neutral hadrons. The energy resolution for photons is a few percent, but the resolution of the hadronic energy measurement in the HCAL will be several tens of percent, thus dominating the uncertainty of the jet energy measurement.

However, the bulk of the jet energy is carried by charged hadrons, which opens up the possibility to replace or combine the calorimetric measurement with the much more precise track momentum measurements. This allows the jet energy resolution to be improved by a factor two or more. Such an improvement of the jet energy resolution naturally helps to improve the resolution on the missing transverse momentum $E_{\mathrm{T}}^{\mathrm{miss}}$. Another potential problem for the $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement in the traditional approach is double counting of energy deposits: an electron from a b-hadron decay might be simultaneously reconstructed as an electron and as a jet. In summary, the traditional detector-centric approach to particle reconstruction does not take advantage of all the information available from the subdetectors and the output of the reconstruction algorithms is not guaranteed to form a list of mutually exclusive and collectively exhaustive particle candidates.

In contrast, the *particle-flow* (PF) algorithm aims to optimize the physics object reconstruction by combining the available information from all subdetectors and by exploiting redundant measurements. The output of the particle-flow algorithm is a list of all stable final state particles: muons, electrons, photons as well as charged and neutral hadrons.

3.1.1 Clustering

The information on energy deposits in the calorimeters is a crucial ingredient to the PF reconstruction, allowing to detect uncharged particles, supplementing the track-based measurements for charged hadrons, and enabling an efficient reconstruction of photons and electrons. The energy measurements of individual calorimeter *cells*, e.g. ECAL crystals, need to be aggregated or *clustered* for further processing. In this context, the objective of the clustering algorithm is to

- detect and measure the energy and direction of stable neutral particles, notably photons and neutral hadrons,
- enable the separation of the energy deposits from neutral and charged particles,
- enable the reconstruction of electrons and the associated Bremsstrahlung photons,
- supplement the momentum measurement for charged hadrons.

The clustering algorithm developed for the particle-flow reconstruction is performed independently in the ECAL and HCAL, and separately for the barrel and endcap detectors[2]. The parameters of the clustering algorithm are optimized for each subdetector, and their values are reported in Table 3.1. The algorithm starts by constructing *cluster seeds*, which are local energy maxima in a calorimeter cell with respect to its neighboring cells. Minimum energy thresholds on the cells and seeds suppress detector noise. An additional E_T threshold is applied in the ECAL endcap to cope with noise levels that increase with detector η. *Topological clusters* are grown from seeds by aggregating cells that share at least one corner with an already clustered cell.

Table 3.1: Parameters of the particle-flow clustering algorithm.

	ECAL		HCAL		preshower
	EB	EE	HB	HE	
Cell E threshold [MeV]	80	300	800	800	0.06
Number of closest cells to seed	8	8	4	4	8
Seed E threshold [MeV]	230	600	800	1100	0.12
Seed E_T threshold [MeV]	0	150	0	0	0
Gaussian width σ [cm]	1.5	1.5	10.0	10.0	0.2

Topological clusters provide only a crude representation of the calorimetric information, particularly if the energy deposits of several particles are merged. A finer picture is obtained by assuming that the topological cluster is the result of N energy deposits, where N is the number of seeds in the cluster. The shapes of the energy deposits are assumed to be Gaussian and the energy amplitudes (A_i) and the positions of the peaks ($\vec{\mu}_i$) are to be inferred from the M cells constituting the cluster. The widths of the Gaussian energy deposits (σ) are fixed to the pre-defined values given in Table 3.1. This model allows for the possibility that the energy in a cell is due to several particles, meaning it accommodates overlap. The fitting of the model is performed via an iterative expectation-maximization algorithm. At the start of each iteration, the expected energy fraction $f_{i\alpha}$ of cell α in the total energy of Gaussian i is calculated as:

$$f_{i\alpha} = \frac{A_i e^{-(\vec{c_\alpha}-\vec{\mu_i})/2\sigma^2}}{\sum_{k=1}^{N} A_k e^{-(\vec{c_\alpha}-\vec{\mu_k})/2\sigma^2}}, \tag{3.1}$$

where $\vec{c_\alpha}$ denotes the position of cell α. A maximum-likelihood fit is then performed to estimate the model parameters:

$$A_i = \sum_{\alpha}^{M} f_{i\alpha} E_\alpha, \tag{3.2}$$

[2]No clustering is done for the HF. Instead, the energy deposits in the large HF cells are directly transformed into clusters.

$$\vec{\mu}_i = \sum_{\alpha}^{M} f_{i\alpha} E_\alpha \vec{c_\alpha}. \tag{3.3}$$

The initial values for their Gaussian parameters are taken from the cells and the algorithm is repeated until convergence. The stability of the fit is increased by attributing the energy of the seeds to the closest Gaussian component. The fitted parameters of the model are then used to define *PF clusters*.

These PF clusters are then calibrated, in particular, to compensate the bias in the energy arising from the finite cell thresholds during topological clustering. This calibration is done for the ECAL and HCAL separately, whereby hadronic clusters are also corrected for the energy lost in the dead material between the ECAL and HCAL.

One of the goals of the clustering step is to aggregate all energy deposits of a particle. In order to collect the energy of electrons and converted photons, which can exhibit a large spread in φ due to bremsstrahlung, an additional clustering step is needed. The resulting clusters are referred to as *superclusters* (SCs). In run I, superclustering was done in a fixed $\Delta\eta$-$\Delta\varphi$-rectangular region around a seed crystal. The rectangular region has to be sufficiently large to capture bremsstrahlung far from the primary electron. A large clustering region will, however, pose problems in the presence of energy deposits from pileup interactions close to the electromagnetic cluster, biasing the SC energy and cluster shapes. In view of the increased pileup in run II the superclustering was improved to avoid the use of a large $\Delta\varphi$ region for high-E_T deposits, and also to accommodate the separation in η for very low-E_T clusters.

3.1.2 Tracking

As outlined in the previous section, the efficient reconstruction of charged particle tracks plays a crucial role in the particle-flow algorithm. Conceptually, tracking can be split into two consecutive tasks: pattern recognition, i.e., the association of multiple hits into a track, and parameter estimation, which is the measurement of track observables like transverse momentum. The challenge for the pattern recognition step is the sheer multiplicity of tracks in any given collision and the resulting combinatorics for the hits, particularly for low momentum tracks. The driving principle of track reconstruction is thus to reduce the number of potential hit associations by reconstructing prompt and high-p_T tracks first, removing the hits associated with these tracks, and then continuing to reconstruct more challenging tracks from the reduced set of tracker hits. A total of ten such iterations are run, with progressive iterations targeting tracks not reconstructed by the previous step by relaxing the quality parameters. In each iteration a potential track is constructed from a seed, i.e., a triplet or pair of hits that point towards the interaction region, by a Kalman Filter (KF).

3.1.3 The particle-flow link algorithm

Reconstructed tracks and clusters are the input to the particle-flow *link* algorithm, which proceeds to build the final list of stable particles. The inputs of the link algorithm are referred to as *elements*, and the goal of the link algorithm is to identify elements that are likely to originate from the same particle and should thus be grouped together. The quality of the links is quantified by a suitable metric, e.g., the spatial distance between a track and a cluster. Elements are ultimately grouped if the link

is of sufficient quality. Linked elements are grouped into *blocks*, which allows to parallelize the event reconstruction. The particle-flow reconstruction then proceeds to reconstruct *PF candidates*, by running the following sequence on the PF blocks:

1. Muon candidates are reconstructed based on the criteria outlined in Section 3.3.1. Any PF elements used to build PF muons are removed from the block.

2. Electron and photon candidates are reconstructed following the algorithm of Section 3.2.2. This includes sophisticated techniques to identify bremsstrahlung photons and conversions. In order to be labeled a PF electron, the candidates are required to pass the PF electron ID. The later is not identical to the electron ID selection used in the $ZZjj$ analyses but exploits very similar observables to reduce the fake contamination. The PF elements used to built PF electrons and photons are again removed from the block.

3. A track cleaning is performed to reduce the number of fake tracks, particularly at high p_{T}. Tracks with fit uncertainties larger than the expected calorimetric energy resolution for hadrons are removed. This cleaning step only affects 0.2 % of tracks in multijet events, 90 % of which are actual fake tracks from random hit associations.

4. The redundancy of the track and calorimeter measurements are furthermore used to identify muons within jets and fake tracks, both of which can cause the sum of the track momenta to be much smaller than the sum of cluster energies. Muons are selected from the global muon collection with relaxed quality requirements and their tracks removed from the block. If the reduced track momentum is still larger than the sum of cluster energies, fake tracks are selected and discarded from the block by ordering all tracks according to their p_{T} uncertainty $\sigma_{p_{\mathrm{T}}}$ and removing those with $\sigma_{p_{\mathrm{T}}} > 1\,\mathrm{GeV}$ until the p_{T}-sum of the remaining tracks would be smaller than the energy sum. This cleaning procedure only affects 0.3 per mil of tracks in multijet events.

5. *Charged hadron* candidates are created for each of the remaining tracks in the block, with their momenta set equal to the track momenta. In cases where the sum of track momenta is compatible with the sum of cluster energies within the measured uncertainties, the hadron momenta are redefined to the result of a global fit of the tracks and clusters.

6. If the sum of the cluster energies is larger than the sum of track momenta, the excess is used to create PF photons and neutral hadrons. In cases where the excess is smaller than or equal to the total ECAL energy, the excess is interpreted as a PF photon. In the remaining cases, the ECAL energy is interpreted as a PF photon and the remaining excess as a PF neutral hadron.

7. Finally, clusters not linked to tracks are used to create PF photons and neutral hadrons. Within the tracker acceptance ($|\eta| < 2.5$), all ECAL/HCAL clusters give rise to photons/neutral hadrons. Outside the tracker acceptance, all ECAL clusters linked to HCAL clusters are interpreted as (neutral) hadrons, while those not linked to HCAL clusters spawn PF photons.

The output of the particle-flow link algorithm is a list of mutually exclusive PF candidates, which are then used for further processing like jet reconstruction, sophisticated particle-flow isolation, or the calculation of event-level quantities like missing transverse energy.

3.2 Electrons

Electrons are important for many analyses, in particular for multilepton analyses like $ZZjj$. Their efficient reconstruction together with an excellent energy resolution is crucial, given that 3/4 of the final states in $ZZ \to 4\ell$ involve at least two electrons. Energy loss in the CMS tracker via bremsstrahlung and backgrounds from hadronic jets make electron measurements experimentally challenging. This section describes the CMS-wide reconstruction of electrons, followed by the electron selection specific to the $ZZjj$ analysis. The electron selection used in the CMS multilepton analyses, and, in particular, the electron identification, is the result of extensive optimization efforts made during this thesis work. These studies are documented in detail in Chapter 4.

3.2.1 Tracking for electrons

A particular challenge for the tracking and for achieving a global event description is the large material budget of the CMS tracker and its impact on electrons. Figure 3.1 shows the number of radiation lengths upstream of the ECAL, which is about $0.3X_0$ around $|\eta| \approx 0$ and peaks at almost $2X_0$ near $|\eta| \approx 1.3$. Traversing this material causes electrons to lose a substantial fraction of their energy via *bremsstrahlung*. The energy loss due to bremsstrahlung is stochastic, i.e., it cannot be predicted on a per-electron basis. Electrons near $|\eta| \approx 0$ will on average lose about 33 % of their energy, while for $|\eta| \approx 1.3$ the average loss is 86 %. Because bremsstrahlung photons are unaffected by the magnetic field which bends the electron trajectory along φ, the electron energy in the calorimeter will be spread out along this direction. The energy loss furthermore reduces the radius of the electron helix, posing a challenge to the iterative track pattern recognition and track parameter fitting.

Two illustrative cases of poor track reconstruction due to bremsstrahlung losses are a single-large emission and the case of several small emissions. A single large emission will cause the electron track to exhibit a kink and the pattern recognition can fail to collect the post-emission hits, resulting in a short track with few hits. Several small emissions, corresponding to a gradual change of track curvature, will not impact the hit collection as much as the quality of the final fit, resulting in a poor track momentum measurement and large χ^2.

A further challenge arises from bremsstrahlung photons that undergo electron-positron pair production, leading to a complex shower pattern of potentially very short tracks and missing energy. These effects complicate the track reconstruction for electrons, and a dedicated tracking algorithm has been developed to improve the efficiency of electron track finding and the accuracy of the parameter estimation, the *Gaussian Sum Filter* (GSF). GSF tracking enhances the usefulness of the tracking information to the electron reconstruction at large, but its algorithmic complexity makes it CPU intensive. In order to keep the per-event processing time manageable, the GSF tracking is only performed on tracks that are likely to originate from electrons, i.e., one first defines electron seeds which are then treated by the GSF. Two complementary seeding algorithms are used to construct electron seeds:

- **ECAL-driven seeding:** Starting from suitable electromagnetic clusters, an attempt is made to find tracker hits from which to build the electron track seeds (*outside-in seeding*). Only energetic superclusters ($E_T > 4\,\text{GeV}$) are used in this procedure, in order to limit the CPU time spent on fakes as this procedure is best suited for high-p_T electrons. The position and energy of the

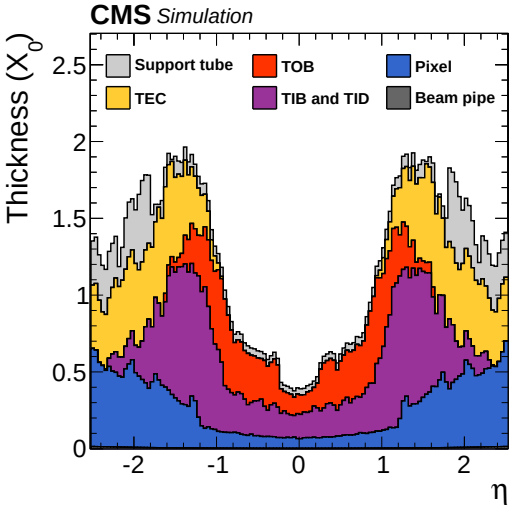


Figure 3.1: Material budget in front of the ECAL in units of radiation lengths X_0 as a function of the detector η [57]. The total material thickness of the silicon pixel tracking detector, the silicon strip detectors in the barrel (TIB and TOB) and endcap (TID and TEC), as well as the tracker support tube, are shown separately.

selected superclusters are used to construct two trajectories, corresponding to the positive and negative charge hypotheses, which are then propagated from the ECAL surface to the innermost layers of the tracker. Tracker seeds are selected if they are compatible with either trajectory and electron seeds are formed if pairs or triplets of hits are matched. The size of the geometric matching window in $\Delta\varphi$ and $\Delta z/r$ between the extrapolated trajectory and a hit as well as the minimum number of matched hits required to form a seed depends on the tracker subdetector. Again, these parameters are optimized to compromise between the efficiency for true electrons and the rate of fake electrons.

- **Tracker-driven seeding:** Starting from the tracks found with the standard iterative tracking, an attempt to match a track to a PF cluster (*inside-out seeding*) is made. All tracks with $p_T > 2\,\text{GeV}$ are used to search for matching clusters, though some pre-identification selections are made in order to reduce the fake rate. Tracks of sufficient quality, e.g., those for electrons that emit little bremsstrahlung, are propagated to the ECAL surface[3] and are matched to the closest ECAL cluster. The track-cluster pair[4] is used to define an electron seed only if the ratio of the cluster energy to the track momentum is close to unity and if the extrapolation of the track to the ECAL surface and the cluster position are within a $\Delta\varphi$ and $\Delta\eta$ window. Tracks of poor quality are more challenging to match to the correct clusters and are treated separately. Specifically, tracks that have a sufficient number of hits but a large χ^2, i.e., those coming from electrons with small successive energy losses, are refit using a light version of the GSF fit. The light fit uses a reduced number of components in the energy loss modeling in order to speed up the execution. The final decision to consider the track seed of a refitted track as an electron seed is based on a *pre-identification* boosted de-

[3]The track position is corrected for the position bias introduced by the nonnegligible depth of electromagnetic showers.

[4]Technically, the electron seed links the seed of a track and not the track itself.

cision tree. The latter exploits the track quality parameters of the KF and light GSF fit, together with the $\Delta\varphi$ and $\Delta\eta$ between the cluster and refitted track.

Because the ECAL-driven seeding performs well on isolated energetic electrons, it has been the traditional seeding algorithm. The tracker-driven seeding mainly serves to recover efficiency for low-p_{T} electrons and electrons in jets. The former are a particular challenge for the ECAL-driven seeding because the spatial spread of the bremsstrahlung energy losses increases at low momenta, which can cause a fraction of the electron energy to be excluded from the supercluster. An underestimate of the energy will, in turn, result in a poor hit matching efficiency. The output of both seeding algorithms are ultimately merged and the seeds submitted to the electron track finding and GSF fitting.

The dedicated electron track finding uses the electron seeds as input and attempts to collect those hits that are lost in the regular tracking due to the large curvature change following bremsstrahlung emissions. Similar to the regular Kalman filter, the current trajectory state is propagated to the next tracker layer and the predicted hit position is calculated. The increase in hit collection efficiency is achieved by loosening the compatibility requirements between the predicted and found hits and by explicitly modeling the probability of energy losses due to bremsstrahlung. If multiple hits are found to be compatible, several trajectory candidates are developed for each found hit, with a maximum of five candidates per tracker layer. Up to one expected-but-missing hit is allowed per trajectory candidate, but a high χ^2 penalty is applied in these cases to suppress cases of bremsstrahlung photon conversion close to the primary track. As a result of these modifications to the pattern recognition, the number of hits per track are significantly increased, allowing the subsequent parameter estimate to extract a greater wealth of information about the electron and its interaction with the tracker material. This parameter estimate is the core of the GSF algorithm.

The central idea of the GSF fit is to extend the Gaussian error modeling in the regular Kalman filter with a sum of Gaussians that approximates the expected distribution of energy losses between tracker layers.

The probability density of the remaining fractional energy z of an electron after having traversed a thin layer of material of radiation length $t = x/X_0$ is given by a formula first developed by Bethe and Heitler [58]:

$$f(z) = \frac{[-\ln z]^{c-1}}{\Gamma(c)}, \tag{3.4}$$

where c is a rescaling of the material thickness $c = t/\ln 2$. Figure 3.2 shows this highly non-Gaussian distribution for several values of the material thickness t. The values of the material thickness chosen in the figure correspond to thin layers, for example, a single tracker sensor.

The insufficiency of approximating this function with a single Gaussian $G(x, \mu, \sigma)$ is immediately apparent. However, the approximation of the Bethe-Heitler function can be drastically improved by using not just one, but multiple Gaussians:

$$f_A(z) = \sum_{i=1}^{M} w_i \, G(z, \mu_i, \sigma_i), \tag{3.5}$$

where w_i are appropriately chosen weights and M is the number of Gaussians used

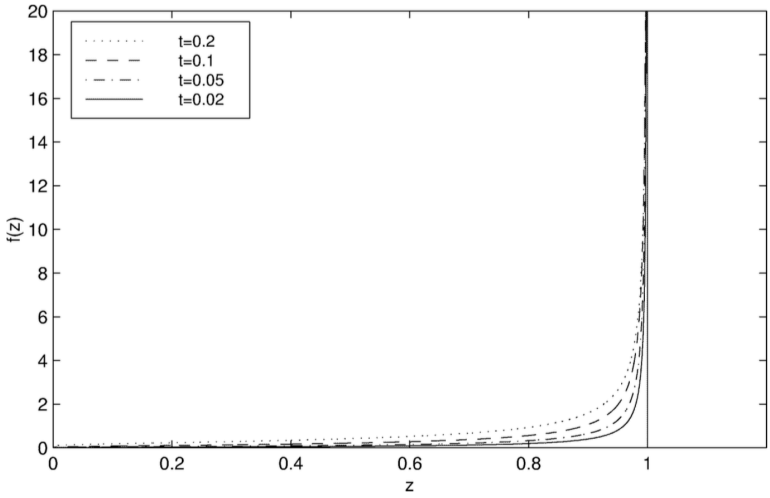


Figure 3.2: Probability density distribution of the remaining energy fraction z as predicted by the Bethe-Heitler model [59]. Upon traversing a material of thickness t, an electron of initial energy E_i will exit the material with a final energy E_f, for an energy fraction $z = E_f / E_i$. Several scenarios for the material thickness t are shown. The bulk of the distribution is concentrated near $z \approx 1$, indicating that most electrons only loose a very small fraction of their energy. Large energy losses are however possible, corresponding to the nonzero probability near $z \approx 0$. The total energy loss of electrons in the CMS tracker is the result of traversing many material layers of the tracker and its support structure, resulting in large cumulative energy losses.

in the approximation. The free parameters of the model are the means and widths of the M Gaussians, as well as the $M-1$ weights. These $3M-1$ parameters are tuned to reduce the difference between the approximation and the Bethe-Heitler model. Studies based on simulation show that the best results are obtained with a metric based on the cumulative distribution functions (CDF) of the exact PDF $F(z)$ and the Gaussian approximation $F_A(z)$

$$D_{\text{CDF}} = \int_{-\infty}^{+\infty} |F(z) - F_A(z)| \, \mathrm{d}z. \tag{3.6}$$

The optimal model parameters for a given material thickness t are then determined by the minimum of D_{CDF}. The number of Gaussians M is a hyper-parameter of the overall model, with larger values of M yielding a finer approximation. The CMS implementation uses $M = 6$, which compromises between the accuracy and computational complexity. Finally, by running the optimization of the model parameters for a grid of values in t and fitting each model parameter with a fifth-degree polynomial, a continuous model in t is obtained. When running the GSF algorithm for a specified value of t, the calculation of the approximation is thus reduced to the straightforward calculation of $3M-1$ polynomials[5].

The GSF algorithm then uses each of the M Gaussians in an independent Kalman filter update. Starting with one state vector, the result of the first GSF update are M state vectors, called components, together with their weights w_i. Without any reduction in the number of components, the next update would feature M^2 components and

[5]The thickness of the material t is an input to the GSF fit and is based on the detector description implemented in the CMS simulation. A simplified material model is built by projecting the material between two successive tracker layers onto one of the layers, accounting for the angle of incidence.

so on. This exponential growth in the number of components is limited by merging components after each update and adjusting the component weights accordingly. The strategy adopted in the CMS implementation iteratively merges the components $f(z)$ and $g(z)$ with lowest *Kullback-Leibler* distance, given by

$$D_{\mathrm{KL}} = \int_{-\infty}^{+\infty} f(z) \ln \left[f(z)/g(z) \right] \mathrm{d}z. \tag{3.7}$$

The moments of the merged component are set equal to the moments of the sum of the individual components and the new weight is the sum of the individual weights. This merging is repeated until the maximum number of components $M_{\mathrm{max}} = 12$ is reached.

Aside from the presence of several components, the GSF then proceeds just like the regular Kalman filter, including the backward filtering and the final smoothing. The final output of the GSF is a sum of Gaussian state vectors corresponding to a multimodal PDF of the estimated track parameters. In practice the information included in the PDF is usually reduced to a single number since working with the full PDFs in the electron reconstruction is cumbersome. Common choices include the mean or mode of the distribution, illustrated in Fig. 3.3. Though the average provides an unbiased estimate of the momentum, the mode is centered around the true value and features a better resolution and is thus chosen as the statistic for the nominal track parameters.

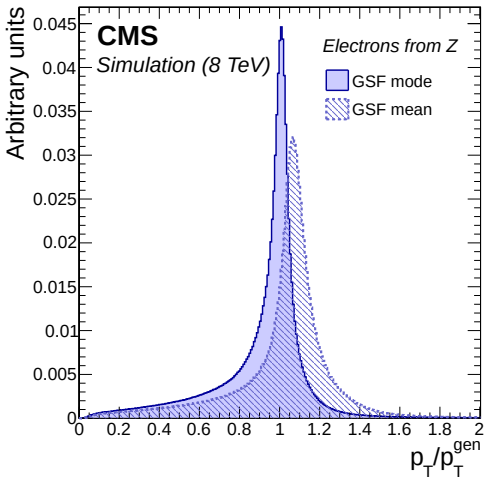


Figure 3.3: Ratio $p_{\mathrm{T}}/p_{\mathrm{T}}^{\mathrm{GEN}}$ for the mean and mode statistics of the final GSF track parameter PDF for electrons from Z boson decays [57].

By explicitly modeling the possibility of potentially large energy losses along the electron trajectory, the GSF provides an improved estimate of the track parameters, notably the transverse momentum compared to the regular Kalman filter. The GSF furthermore provides observables on the change of the track parameters along the trajectory, which are exploited in the electron identification algorithm.

3.2.2 Electron reconstruction

The starting point of the electron reconstruction are the electron track seeds described in Section 3.2.1 and the superclusters detailed in Section 3.1.1. Although the superclustering algorithm is designed to collect the energy deposits from bremsstrahlung it

can fail to do so, mainly when the primary and bremsstrahlung clusters are far apart spatially or in case of converted bremsstrahlung photons. Aside from recovering the missed energy, an important reason to collect these emissions for multilepton analyses is to reduce their impact on the electron isolation.

An attempt is thus made to collect these deposits, targeting the unconverted and converted bremsstrahlung photons respectively. For the former case, a tangent to the GSF track is built at each tracker layer, and any ECAL cluster compatible with the tangent is PF-linked to the supercluster. Figure 3.4 illustrates this idea of the GSF track tangents and the complex bremsstrahlung patterns of electrons. The hypothetical photon emissions given by the GSF track tangents are also used to identify the cases where the bremsstrahlung photon undergoes conversion in the tracker material.

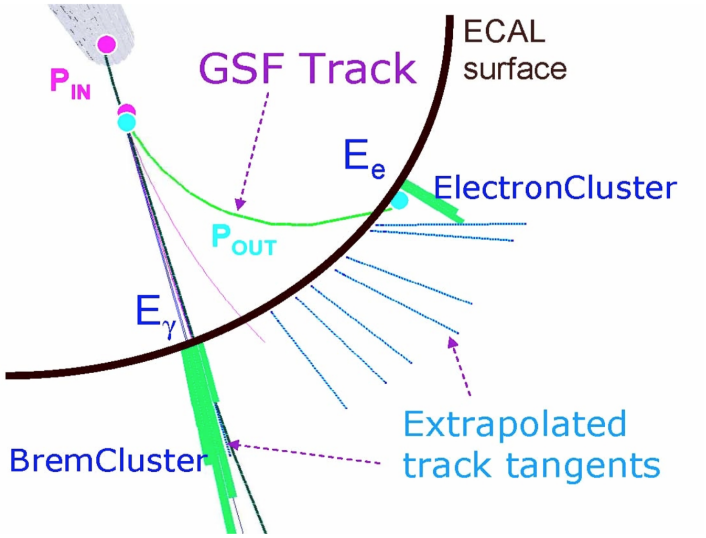


Figure 3.4: Illustration of an electron undergoing bremsstrahlung emission and the components of the electron reconstruction implemented in the particle-flow algorithm [60]. The initial electron (magenta line) emits a bremsstrahlung photon (gray line), giving rise to two distinctive electromagnetic clusters in the calorimeter (green bars). The GSF tracking accommodates the change of curvature of the electron track (green line) and allows to measure the incoming (p_{in}) and outgoing momenta (p_{out}). Finally, the cluster of the bremsstrahlung photon is linked to the electron cluster via the GSF track tangents, giving the refined supercluster as described in the text.

Two algorithms attempt to identify photon conversions, targeting the cases where one or both tracks of the conversion electrons are reconstructed by the iterative tracking algorithm. A dedicated conversion finder attempts to find track pairs from conversion vertices. If the direction of the converted photon as given by the sum of the two conversion tracks is compatible with one of the GSF tangents, the tracks are linked to the GSF track. The conversion tracks are then used to search for ECAL clusters, which are linked to the supercluster if the ratio of the cluster energy and conversion track momentum is compatible with the electron hypothesis.

Reconstructing the tracks of both conversion electrons is challenging, because the conversion can happen late in the tracker and thus yield few hits, or because the conversion is asymmetric with one low-p_{T} electron. These cases are targeted by a single-leg conversion identification algorithm. For each GSF tangent, the closest KF track in ΔR is identified. Tracks passing a preselection on $\Delta\eta$ and $\Delta\varphi$ are then submitted to a multivariate discriminant to suppress backgrounds from spurious associations aris-

ing mostly from charged pions in proximity to the primary electron. In addition to the spatial compatibility of the track and the GSF tangent, the conversion BDT also exploits the radius of the innermost hit on the KF track, the transverse impact parameter with respect to the primary vertex, as well as the E/p of the KF track and the ECAL clusters linked to it. Selected KF tracks and their associated clusters are again linked to the GSF track and the supercluster respectively. A *refined super cluster* is ultimately defined based on the merger of the supercluster and the ECAL clusters linked to it via the bremsstrahlung recovery algorithms.

Electron energy calibration and final momentum estimate

Without a final calibration, the electron energy scale in data would exhibit a residual shift because of imperfect corrections of the transparency loss of the ECAL crystals due to irradiation as well as of other effects. This is corrected by monitoring the measured mass of the Z boson and shifting the electron energy scale such that the corrected mass is equal to the mass in the simulation. It should be stressed that the electron energy scale is not corrected to be equal to some experimental value of the Z boson mass, e.g., the world best average determined by the Particle Data Group, but it is to match the reconstructed peak position in the simulation. A mass measurement of say the Higgs boson, will take its absolute mass scale from the input parameters to the simulation, which will in general be larger than the reconstructed mass by up to 100 MeV. In practice, the final energy calibration is with $Z \to ee$ is derived for several consecutive data runs and parameterized in p_{T}, $|\eta|$, and R_9[6].

The GSF track and the calibrated supercluster each provide a measurement of the electron momentum. At momenta below 15 GeV or in the ECAL gap regions, the momentum resolution achieved with the GSF track is better than the supercluster energy resolution, while at larger transverse momenta the supercluster energy is more accurate. A regression BDT combines both estimates and its output is used as the final electron momentum.

In addition to calibrating the electron energy scale in data, a smearing of the energy in the simulation is performed. The goal is to improve the data-simulation agreement for the energy resolution. Without such a correction, the resolution in the simulation would be too optimistic. The smearing is done by scaling the nominal energy by a random number sampled from a Gaussian distribution whose width is parameterized in terms of p_{T} and $|\eta|$. Effectively, this corresponds to convoluting the mass spectrum with a Gaussian.

Electron charge estimate

The final step in the electron reconstruction is the estimate of its charge. The GSF track naturally provides a charge estimate, however, it can lead to large charge misidentification rates when the track includes hits from converted bremsstrahlung photon. This happens most frequently for electrons at high $|\eta|$ where the material budget is large in the innermost part of the tracker. Two alternative estimates are thus considered. The first is based on the charge of the KF track associated with the GSF track, if any. The second is based on the sign of the difference in φ between the vector joining the beam spot to the supercluster position and the vector joining the beam spot and the first

[6]The R_9 observable is sensitive to the amount of bremsstrahlung. It is defined as the ratio of the energy in a 3×3 crystal matrix centered around the seed crystal and the supercluster energy.

hit of the electron GSF track. The nominal charge assigned to an electron candidate is then given by the majority of the three estimates. Requiring all three estimates to agree allows to further reduce the charge mis-identification rate, if needed for physics analysis.

3.2.3 Electron selection

The goal of the electron selection is to reduce the rate of electron candidates from background processes, while balancing the efficiency for signal electrons. In the case of the $ZZjj$ analysis, which benefits from the large background suppression provided by the $ZZ \to 4\ell$ selection, this means a rather loose electron selection with signal efficiencies of 90 % to 95 % and background rates of 10 % to 20 %. Other considerations in the design of the electron selection, or object selections in general, include:

- dependence of the signal and background efficiencies on the candidate kinematics: processes that contribute to the reducible background in multilepton analyses feature mostly low-p_{T} electrons,
- stability of the signal and background efficiencies with respect to pileup,
- monitor and understand the quality of the data-simulation agreement for the observables exploited in the selection,
- the ability to perform cross-checks on background modeling, to understand the composition of the background at various levels of the selection, and to validate the data-simulation agreement with sufficient statistics.

The last point is the main reason that the electron selection in multilepton analyses is split into three subselections: impact parameter requirement, identification, and isolation.

Impact parameter selection

The impact parameter selection aims to reduce backgrounds that result in electron candidates that do not originate from the hard interaction, but from subsequent decays. The most important example include B meson decays that arise in $\mathrm{t\bar{t}}$ production and photon conversions, since the tracks of these electron candidates will generally not point to the primary vertex. Algorithmically, one determines the *impact parameter* $\mathrm{IP_{3D}}$ between the candidate and the primary vertex, which is defined as the minimal Euclidean distance between the two. Using the impact parameter, loose vertex requirements are imposed:

$$|d_{xy}| < 0.5\,\mathrm{cm} \text{ and } |d_z| < 1\,\mathrm{cm}, \tag{3.8}$$

where d_z denotes the longitudinal component and d_{xy} refers to the distance in the transverse plane. A more refined observable can be constructed by also considering the tracking uncertainty on the impact parameter $\sigma_{\mathrm{IP_{3D}}}$, and one requires:

$$\mathrm{SIP_{3D}} = \frac{|\mathrm{IP_{3D}}|}{\sigma_{\mathrm{IP_{3D}}}} < 4. \tag{3.9}$$

Identification

The electron identification (ID) aims to reduce backgrounds arising from hadronic jets and photon conversions. Hadronic jets can mimic the electron signature via accidental association during reconstruction: the reconstructed track from a charged hadron like a $\pi^{\pm}$ can be in close vicinity to an electromagnetic cluster of $\pi^0 \to \gamma\gamma$ decays. This is, in fact, the dominant source of electron backgrounds for most analyses, including the multilepton analyses.

Because electrons are constructed from an electromagnetic cluster and a track, one usually categorizes the observables used to separate prompt electrons from the backgrounds into three classes:

- observables based on the shape of the electromagnetic cluster, e.g., the width of the cluster along the η direction,
- observables based on tracking information, e.g., the momentum loss due to bremsstrahlung $f_{brem} = 1 - p_{out}/p_{in}$ where p_{in} and p_{out} are the track momenta at the vertex and the ECAL surface respectively,
- the quality of the matching between the supercluster and the track, e.g., the ratio of the supercluster energy over the track momentum.

The electron identification is based on a boosted decision tree (BDT) that combines 20 variables from the above categories. The BDT returns a real number (*score*) for each electron candidate, where large positive values are signal-like. Figure 3.5 shows the background versus signal efficiencies of the BDT and the chosen working points obtained by only selecting electron with BDT scores above a given threshold. Details on the development and optimization of the BDT are given in Section 4.2.1.

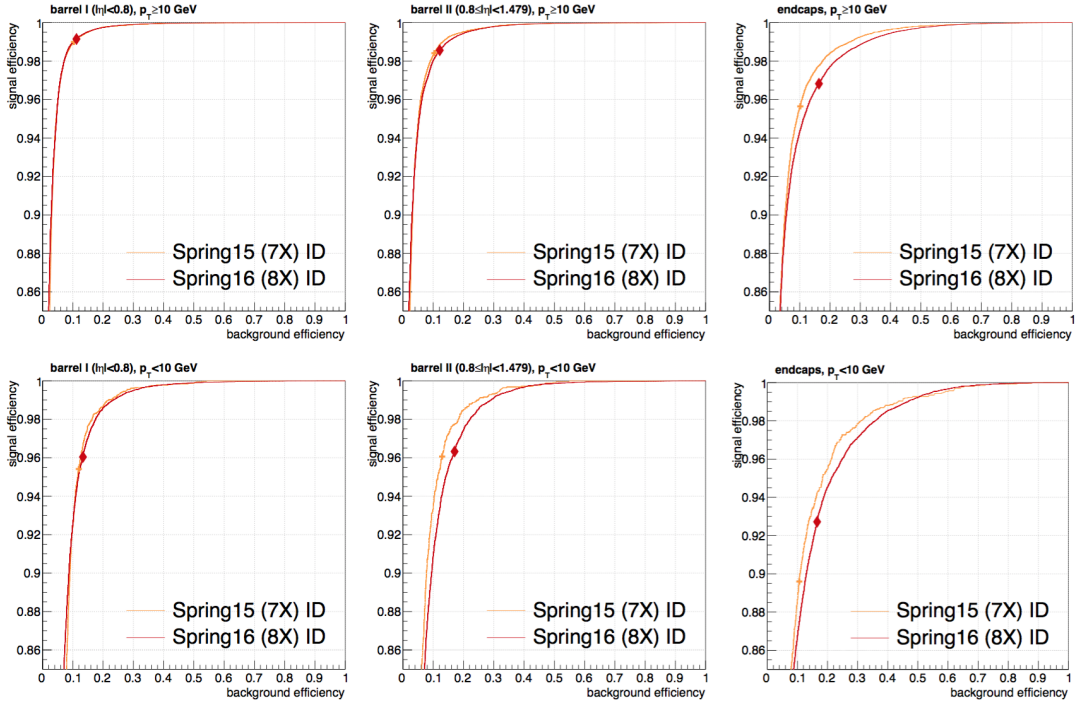


Figure 3.5: Performance comparison of the electron BDT developed for the $H \to ZZ^* \to 4\ell$ analysis on 2015 data and the retraining for the 2016 conditions. The respective working points are indicated by the markers.

Electron isolation

Electron isolation is a powerful tool to reduce backgrounds coming from hadronic jets. The basic notion is that a prompt electron will not be surrounded by other particles coming from the same hard interaction, i.e., it will be *isolated*. In practice one sums the p_{T} of all other reconstructed PF candidates around the electron. This is done separately for charged hadrons, neutral hadrons and photons. The lack of a reliable vertexing for photons and neutral hadrons makes the latter two contributions susceptible to the amount of pileup in the event and a correction based on the hadronic activity in the event is used to reduce this pileup dependence. Denoting this correction with $p_{\mathrm{T}}^{\mathrm{PU}}$ one can write the per-electron isolation as

$$\mathcal{I} = \sum_{\substack{\text{charged}\\ \text{hadrons}}} p_{\mathrm{T}} + \max\left[0, \sum_{\substack{\text{neutral}\\ \text{hadrons}}} p_{\mathrm{T}} + \sum_{\text{photons}} p_{\mathrm{T}} - p_{\mathrm{T}}^{\mathrm{PU}}\right], \tag{3.10}$$

where all candidates within $\Delta R < 0.3$ of the electron are considered in the sums. Photons selected by the FSR algorithm described in Section 3.5 are ignored in the sum[7]. The pileup correction for electrons is based on the effective area technique:

$$p_{\mathrm{T}}^{\mathrm{PU}} = \rho \times A_{\mathrm{eff}} \tag{3.11}$$

where ρ is the mean energy density in the event and the effective area A_{eff} is determined in five bins of $|\eta|$ and defined as the ratio between the slope of the average isolation and that of ρ as a function of the number of reconstructed vertices.

Two problems arise with the *absolute* isolation defined in Eq. (3.10). Firstly, it does not consider the momentum of the electron p_{T}^{e} for which the isolation is calculated. Secondly, a cut on absolute isolation is done in units of GeV and it will thus depend explicitly on the quality of the energy scale measurement. Both issues are avoided in the *relative* isolation:

$$\mathcal{I}_{\mathrm{rel.}} = \mathcal{I} / p_{\mathrm{T}}^{e}. \tag{3.12}$$

The $\Delta R = 0.3$ parameter, also referred to as the *cone size*, as well as the final isolation cut value, were optimized for the $\mathrm{H}\to ZZ^{*}\to 4\ell$ analysis. Electrons with a relative isolation below 0.35 are considered to pass the isolation requirement.

Electrons which satisfy the impact parameter requirements, pass the identification and isolation are used to select Z boson candidates in the $ZZjj$ analysis.

3.2.4 Electron efficiency measurements

Selection efficiency measurement and tag-and-probe technique

The previous sections on the reconstruction and selection of electrons showed that the final objects used in a physics analysis are the result of hundreds of selections on

[7] The explicit recovery of final state radiation is specific to multilepton analyses and not part of the general lepton reconstruction. No veto of FSR photon candidates is performed in the general PF isolation calculation.

detector measurements. For example, the ID selection exploits observables like the geometric distance between the reconstructed track and the position of the energy cluster. The efficiency of such a selection will depend on the quality of the alignment between the tracker and calorimeter.

Consequentially, the efficiency of the selection in data might not be equal to that predicted by the simulation. Although one monitors and improves the simulation to match what is observed in data, it is not feasible to guarantee perfect agreement between the simulation and the data for the many observables that are used in the reconstruction or enter as inputs to the identification BDT. Also, the distributions in data will, in general, depend on the run conditions like the amount of pileup or time-depended effects like shifts in energy scale. This motivates the measurement of the electron efficiency in data and to use that measurement to correct the simulation.

Measuring the efficiency of a selection requires having a pure set of electrons that are unbiased with respect to that selection, i.e., one has to ensure that the selection used to obtain the set of electrons is uncorrelated to the cut whose efficiency one wishes to measure. Such a set can be obtained by the tag-and-probe (T&P) technique, which selects the decay products of resonances like the Z boson to ensure high purity. The T&P method is used for all efficiency measurements in this analysis, be it the trigger efficiency, the electron reconstruction, or the muon selection efficiency. We illustrate its application for the measurement of the electron selection efficiency $\epsilon_{\text{sel.}|\text{reco.}}$, that is we measure the efficiency of reconstructed electrons to pass the electron selection outlined in section Section 3.2.3: impact parameters, isolation, and multivariate ID. This measurement and the resulting corrections to the simulation are used in the $ZZjj$ analysis and in all other CMS multilepton analyses based on the 2016 data.

The starting point for measuring the selection efficiency is a set Z boson decays in data. These are selected by requiring the presence of two electrons of opposite charge with an invariant mass in the range $60 < m_{e^+e^-} < 120\,\text{GeV}$. This selection will be enriched in true $Z \to e^+e^-$ decays, but multijet and $t\bar{t}$ events will also pass the selection. Such background contributions can be suppressed by imposing stringent quality requirements on one of the electrons - this is the *tag* electron. The tag electron cannot be used anymore for the efficiency measurement, but the remaining electron, called *probe*, is still unbiased and can be used for the measurement. To reduce the low-p_T QCD background, the tag has to satisfy $p_\text{T} > 30\,\text{GeV}$. Electrons in the EB-EE transition regions ($1.4442 < |\eta| < 1.566$) are rejected because of the large background rates in these parts of the detector. The tag electron also has to pass the tight working point of the cut-based electron ID.

A crucial aspect ignored in the above description is the trigger. After all, for an event to be recorded and to be available for offline analysis it needs to have passed a trigger path. There are no trigger paths that only demand the presence of one or more reconstructed electrons without further quality requirements because of the large fake rates from QCD multijet events. All electron triggers thus impose selections on the electron quality beyond the reconstruction. This would bias the efficiency measurement because the offline selection will rely on the same observables. The solution is to use only single electron triggers and to exclude the electrons that pass the trigger from the efficiency measurement, i.e., those that are *matched* to the trigger. This is equivalent to requiring that the tag electron be matched to the electron that passed the single electron trigger. Because of the high background rates in the very forward region of the detector, the single electron trigger is restricted to $|\eta_\text{SC}| < 2.1$ and the offline tag selection imposes the same cut to remove pathological electron candidates.

For probes with $p_T \gtrsim 40\,\text{GeV}$, the tag selection reduces the background contamination to a low level, as illustrated in the top row of Fig. 3.6. At lower momenta the background can be considerable and not subtracting it would severely bias the efficiency measurement as shown in the bottom row of Fig. 3.6. In practice the signal and background yields are estimated by fitting the sum of a signal and a background shape to the data. The background shape is taken as the sum of an error function and an exponential function. Besides the empirical agreement with the data, this shape is motivated by the exponentially falling p_T spectrum of jets, which constitute the bulk of the background. The signal shape is taken from the Drell-Yan simulation, where the template histogram is furthermore convoluted with a normal distribution to capture differences in the energy resolution between the data and the simulation. This fit is performed independently on the data distributions for electrons passing and failing the electron selection[8], and the estimated number of passing N_p and failing electrons N_f is determined by taking the integral of the post-fit signal shapes. The final efficiency $\epsilon_{\text{sel.}}$ is then given by

$$\epsilon_{\text{sel.}} = \frac{N_p}{N_p + N_f}. \tag{3.13}$$

Because the selection efficiency depends on the kinematics of the electron, this measurement is performed in bins of transverse momentum and supercluster pseudorapidity $\epsilon_{\text{sel.}} = \epsilon_{\text{sel.}}(p_T, \eta_{SC})$. Furthermore, the efficiencies for electrons in the EB-EE transition regions, in the gaps between supermodules in the barrel, and in the dee gaps of the endcaps are treated separately. A particularity of these *gap* electrons is the much larger fraction of poor energy measurements. This is most evident in the EB-EE transition region, where a large part of the energy of the electromagnetic shower can fall into noninstrumented material. This loss of energy is challenging to compensate in the calibration, as it depends on the azimuthal angle of incidence and the starting point of the electromagnetic shower, both of which are measured with limited accuracy. As a result of this energy loss, the invariant mass spectrum of gap electrons exhibits a second bump left of the Z mass peak and the position of this peak depends on the probe p_T. Figure 3.7 shows the p_T dependence of the measured selection efficiency for non-gap (left) and gap electrons (right).

A final particularity of the electron selection used in the multilepton analyses is the FSR recovery, which is also used in the efficiency measurement presented here. This means that any FSR photon matched to an electron is implicit in its kinematics, the tag-probe invariant mass, and FSR photons are excluded in the isolation sums.

The accuracy of the selection efficiency measurement relies on the modeling of the signal and background contributions, which introduces systematic uncertainties into the measurement. The following variations of fit model are considered to estimate these uncertainties:

- **Uncertainty in the accuracy of the signal model:** Variation of the signal shape from a simulation-based template to an analytic shape (Crystal Ball),
- **Uncertainty in the background modeling:** Variation of the background model to an exponential function,
- **Uncertainty in the background coming from the tag selection:** Tightening of the tag selection to $p_T > 35\,\text{GeV}$ and tight MVA-based ID,

[8] A simultaneous fit was also explored but the poor convergence and fit instabilities favor the independent fit.

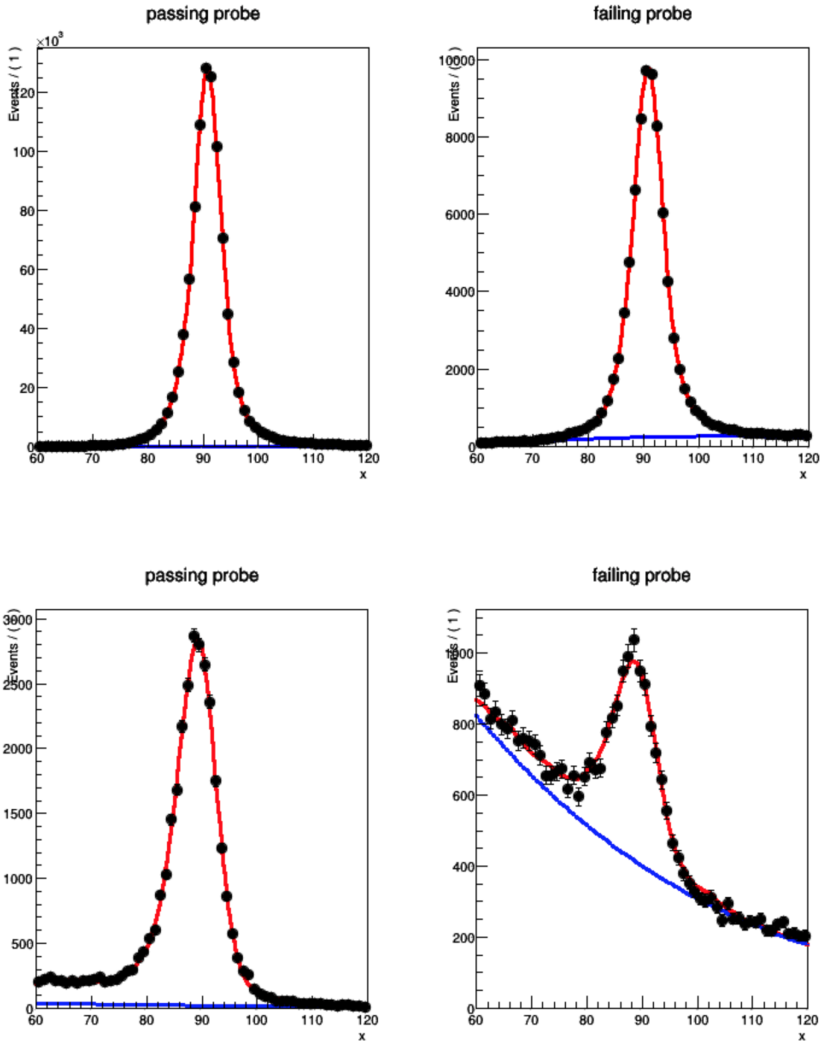


Figure 3.6: Example tag-and-probe invariant mass distributions for probe electrons passing the selection (left panels) and those failing the selection (right panel) in the 2016 dataset. The top panel shows the distribution for probe electrons with $0 < \eta_{SC} < 0.8$ and $40 < p_T < 50\,\text{GeV}$ while the bottom row shows those with $\eta_{SC} > 2.0$ and $15 < p_T < 20\,\text{GeV}$. In each plot, the blue line shows the fitted background model and the red line corresponds to the sum of the signal and background distribution.

- **Uncertainty in the overall event description:** Using an NLO Drell-Yan simulation for the signal templates.

The total uncertainty for the measurement of the efficiency is the quadratic sum of these systematic uncertainties and the statistical uncertainties returned from the fit. With the exception of very forward electrons ($|\eta_{SC}| > 2.0$) with $p_T > 100\,\text{GeV}$, the measurement is limited by the systematic uncertainties.

The selection efficiencies measured in data are used to correct the simulation by weighing each electron selected in the simulation by the ratio of the measured efficiency in data and the efficiency in the simulation, the *scale factor*:

$$\text{SF}(p_T, \eta_{SC}) = \frac{\epsilon_{\text{data}}}{\epsilon_{\text{MC}}}. \tag{3.14}$$

Figure 3.8 shows these scale factors, again separately for non-gap and gap electrons as

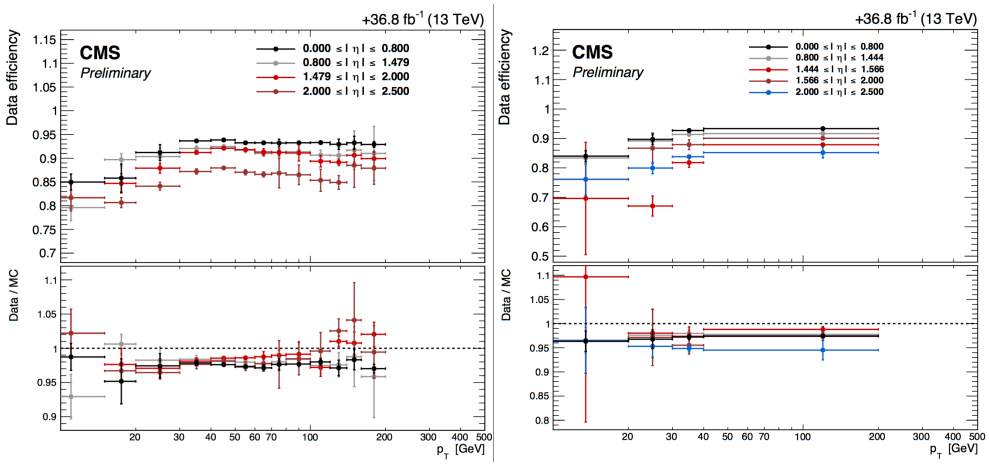


Figure 3.7: Electron selection efficiencies in the 2016 dataset measured with the tag-and-probe technique described in the text, for non-gap electrons (left) and gap electrons (right).

well as the total uncertainties as measured for the 2016 dataset. The scale factors for electrons with transverse momenta above 20 GeV are between 0.97 and 0.98, and exhibit an increase for the endcaps where the scale factor exceeds unity for $p_{\mathrm{T}} > 120\,\mathrm{GeV}$. The uncertainty on the scale factors amounts to a few percent below 30 GeV and is reduced to 0.3 % for electrons with 50 GeV in the central part of the detector.

Reconstruction efficiency measurement

The tag-and-probe method is also used to measure the electron reconstruction efficiency in data. In the 2015 dataset this measurement was only performed to validate the accuracy of the simulation, with no further correction of the predicted efficiencies. In the 2016 dataset the tracking efficiency was reduced due to a lower hit reconstruction efficiency in the silicon strip detector (called *HIP* effect). This necessitated the measurement of the reconstruction efficiency in data to correct the simulation which does not model the inefficiency adequately.

The dominant inefficiency in the reconstruction of electrons is the ability to reconstruct the track. The commissioning in run I showed that the efficiency of the electron cluster reconstruction is very close to 100 %. The electron reconstruction efficiency is thus taken to be the GSF tracking efficiency[9]. The latter is measured in data by using superclusters as the probe/denominator and determining the fraction of those clusters that are used in reconstructed electrons. This association between clusters and reconstructed electrons is thus done directly on the cluster objects and not based on a geometric matching. Aside from these technical details, the measurement proceeds just like the one for the selection efficiency, including the evaluation of the systematic uncertainties.

The efficiencies are used to calculate data-to-simulation scale factors $\epsilon_{\mathrm{reco.}}$, which are shown in Fig. 3.9 as a function of η_{SC}. The error bars show the systematic uncertainties which also cover the minor p_{T}-dependence.

[9] Very loose selections on H/E and track-cluster matching variables are part of the electron seed filtering to reduce the background rate, but have negligible impact on the signal.

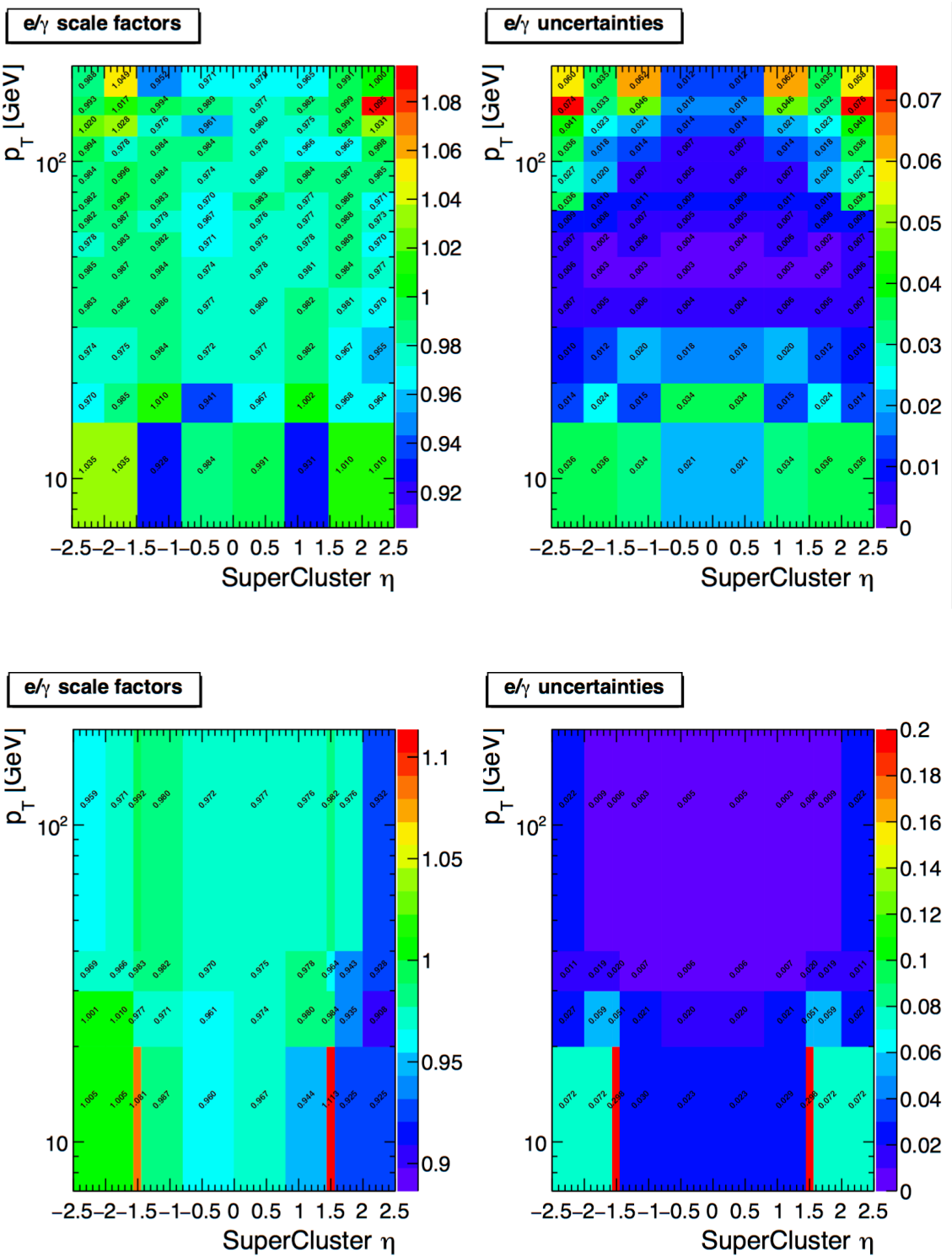


Figure 3.8: Electron selection efficiencies in the 2016 dataset measured using the tag-and-probe technique (top row), for non-gap electrons (left) and gap electrons (right). The bottom row shows the corresponding total uncertainties.

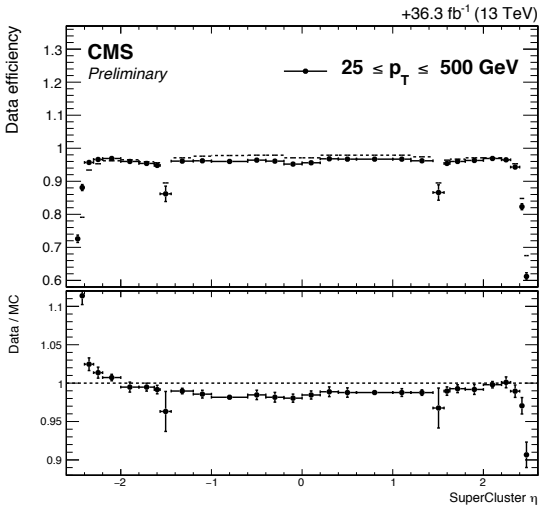


Figure 3.9: Electron reconstruction efficiency versus η_{SC} and data-to-simulation scale factors for the 2016 dataset [61]. The error bars report the sum of the statistical and systematic uncertainties. An additional 1 % uncertainty is to be added for electrons with $p_{\mathrm{T}} < 20\,\mathrm{GeV}$ and $p_{\mathrm{T}} > 100\,\mathrm{GeV}$.

3.3 Muons

Muons are crucial for many analyses in CMS, be it because they are part of the final state as is the case in the $ZZjj$ analysis or because they can help to identify heavy meson decays. This section first describes the standardized muon reconstruction as implemented in the particle-flow algorithm, followed by the muon selection specific to the $ZZjj$ analysis.

3.3.1 Muon reconstruction and identification

Reconstructing muons relies on the fact that muons are the only detectable particle species to fully traverse the CMS detector, notably the HCAL, the solenoid, and the iron return yoke[10]. The final list of reconstructed muons is the result of merging three collections:

- **Standalone muons** are reconstructed using only the muon spectrometer: hits in the each DT and CSC chamber are first used to construct segments or *track stubs*, which serve as seeds for a track reconstruction, which also exploits hits from the RPC. The output of the fit are *standalone muon tracks*.
- **Global muons** are those for which a tracker and standalone muon track can be geometrically matched. In this case, the hits of the two tracks are re-fit using a Kalman filter.
- **Tracker muons** recover efficiency for low-p_{T} muons with $p_{\mathrm{T}}^{\mu} \lesssim 5\,\mathrm{GeV}$, which do not always fully traverse the iron return yoke and muon spectrometer. Tracker tracks are extrapolated to the muon system, and matched to muon stubs from the DT and CSC chambers. If at least one such stub is matched to the track, the track is considered a tracker muon.

The combined reconstruction efficiency for global and tracker muons for muons with $p_{\mathrm{T}} \gtrsim 4\,\mathrm{GeV}$ is above 99 %. Standalone and global muon candidates that share a tracker track are merged to avoid double counting. Owing to the high reconstruction efficiency of this merged collection, standalone muons are not used in the $ZZjj$ analysis, also considering their reduced momentum resolution and the much higher background rate.

The large lever arm of the muon spectrometer compared to the tracker helps to improve the track parameter measurements, particularly for transverse momenta above several hundred GeV. However, the passage of the muon through the iron return yoke disturbs its original trajectory because of multiple Coulomb scattering and muon bremsstrahlung[11]. The later will give rise to electromagnetic showers in the muon stations, which in turn reduces the accuracy and precision of the hit localization. These effects limit the usefulness of the hits in the muon stations for the muon track parameter measurement in the global track fit compared to the tracker-only fit. As a consequence, the momentum of the PF muon is based on the tracker-only fit, unless:

- the p_{T} of the tracker-only and global fits are above 200 GeV, and
- the charge-to-momentum ratios q/p between the global and tracker-only fits agree to within $2\sigma_{q/p}$,

[10] The background from hadron punch- and sail-through rarely traverses the full muon system.

[11] The energy loss via bremsstrahlung for a particle of mass m scales as m^{-4} to m^{-6}, depending on the relative angle between the acceleration and velocity of the particle. Bremsstrahlung for muons is thus suppressed by a factor of at least $(m_e/m_\mu)^4 = 10^{-10}$.

in which case the parameters of the global fit are used. As will be shown in Section 5.3, the median transverse momentum of the leading muon in the $ZZjj$ analysis is around 100 GeV, for which the transverse momentum resolution is less than 2 % in the barrel, and better than 6 % in the endcap as illustrated in Fig. 3.10.

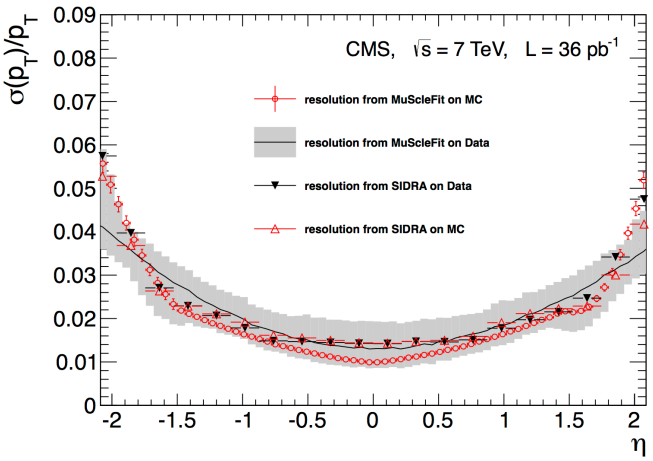


Figure 3.10: Resolution of the transverse momentum of muons in simulation and data for muons from Z boson decays. Two methods to measure the muon momentum resolution are compared [62].

3.3.2 Muon Selection

The muon selection for the $ZZjj$ analysis follows the same three-step approach used for electrons: impact parameter, identification, and isolation requirements.

Impact parameter selection

The impact parameter selection is identical to that used for electrons, detailed is Section 3.2.3. The cuts are

$$|d_{xy}| < 0.5\,\text{cm and } |d_z| < 1\,\text{cm and } |\text{SIP}_{3\text{D}}| < 4, \tag{3.15}$$

Identification

The muon identification in this analysis is identical to the identification used by the particle-flow muon algorithm, which proceeds in three stages referred to as *isolated*, *tight*, and *loose*. Because of the low level of ambiguity, the algorithms first selects isolated muons, that is muons which have a relative isolation of 0.1 where tracker tracks and calorimeter hits in a cone of $\Delta R = 0.3$ are summed. The second stage targets non-isolated muons by requiring a minimum number of hits in the muon track and compatibility of the muon segment and calorimeter deposit with templates derived from simulation. The final stage recovers efficiency by relaxing the number of hits on the muon track and the template matching is replaced by a compatibility requirement of the muon track to hits in the muon stations. In the $ZZjj$ analysis, muons are considered to pass the ID if they are selected by any of the three stages of the particle-flow algorithm.

Isolation

The implementation of the muon isolation selection is almost identical to the one presented for electrons, the only difference being the pileup subtraction. For muons the $p_{\mathrm{T}}^{\mathrm{PU}}$ term in Eq. (3.10) is determined via the $\Delta\beta$ *method*. This correction is based on the assumption that the energy from neutral particles inside the isolation cone is proportional, on average, to the energy from charged hadrons:

$$p_{\mathrm{T}}^{\mathrm{PU}} = 0.5 \times \sum_{\substack{\text{charged} \\ \text{PU hadrons}}} p_{\mathrm{T}}, \tag{3.16}$$

where the sum considers all charged hadrons from pileup, i.e., all PF hadrons which are associated with a vertex that is not the primary vertex. The factor 0.5 is the constant of proportionality and accounts for the extrapolation from charged hadrons to neutral particles.

Muons are considered isolated if their relative isolation in a 0.3 cone is less than 0.35.

An additional *ghost-cleaning* step is performed to deal with situations where a single muon is incorrectly reconstructed as two or more muons:

- tracker muons that are not global muons are required to be arbitrated,
- if two muons share 50 % or more of their segments, then the muon with lower quality is removed.

Muons that satisfy the impact parameter requirements, pass the identification and isolation, as well as the ghost-cleaning, are used to select the Z boson candidates in the $ZZjj$ analysis.

Muon momentum calibration

The momentum scale for muons is derived from data, from Z boson and for the low-p_{T} regime $J/\psi \to \mu^+\mu^-$ meson decays. The resulting calibrations are used in a refit of the Kalman filter track, whereby the dominant sources of corrections stem from inhomogeneities in the magnetic field, misalignment of the detector, and the limited accuracy in the modeling of the material budget.

3.3.3 Muon efficiency measurements

Muon efficiencies are measured in data with the Tag and Probe method, similar to the electron case. This relies on selecting $Z \to \mu^+\mu^-$ and $\mathrm{J}/\psi \to \mu^+\mu^-$ decays and determining the scale factors in bins of the probe muon p_{T} and η. The measurement proceeds along the four muon selections used in this analysis:

- **Tracking:** Standalone muon tracks are used to probe the efficiency to reconstruct a muon track with the inner tracker. This measurement uses $Z \to \mu^+\mu^-$ decays and the resulting scale factors are given as a function of η and given separately for muons below (above) 10 GeV, with scale factors of about 0.98 (0.99).
- **Reconstruction and identification:** Given a reconstructed inner track, the efficiency of reconstructing and identifying a probe as a loose PF muon is measured. The measurement exploits Z boson decays, and to increase the

statistics at low-p_T, $J/\psi \to \mu^+\mu^-$ decays. The scale factors are close to unity, with notable deviations for $p_T < 10\,\text{GeV}$ and for $|\eta| < 0.2$.

- **Impact parameters:** For muons that are reconstructed and pass the identification requirements, the efficiency of passing the SIP_{3D}, d_{xy}, and d_z cuts is measured using Z boson decays. The associated scale factors are above unity and exhibit a rise to 1.02 for muons with $p_T < 20\,\text{GeV}$.
- **Isolation:** This measurement relies on Z boson decays and excludes any matched FSR photons from the sum, like it is done for electrons. The resulting scale factors are compatible with unity, exhibiting only minor deviations for $p_T < 10\,\text{GeV}$.

The four scale factor measurements outlined above are multiplied to get an overall selection scale factor. The result is shown in Fig. 3.11 (left), and the associated systematic uncertainties are shown on the right-hand side of the same figure.

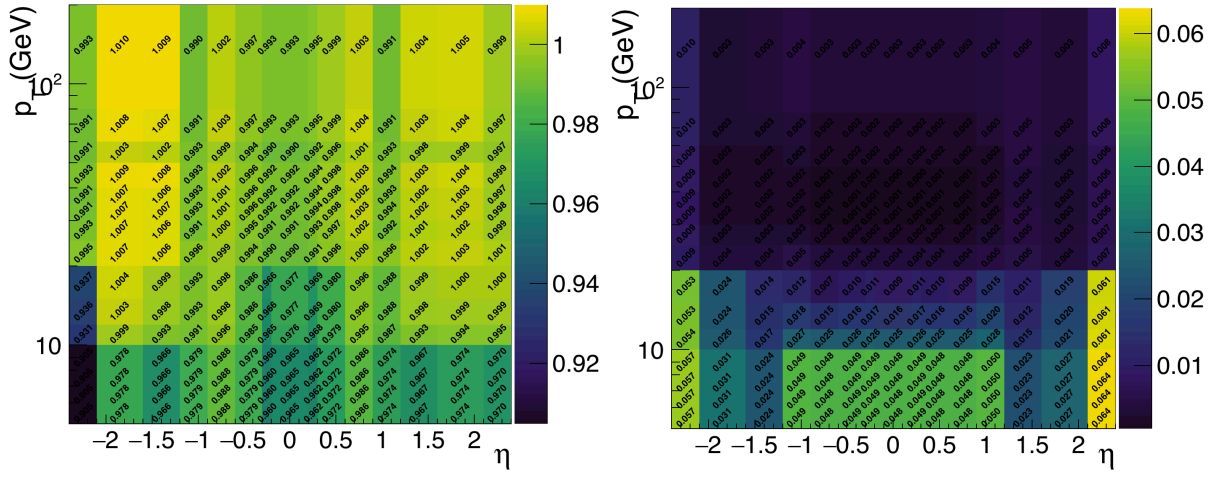


Figure 3.11: Data-to-simulation scale factors for muon reconstruction and selection (left) and the systematic uncertainties (right) for the 2016 dataset [63].

3.4 Jets

3.4.1 Jet reconstruction

The goal of the jet reconstruction is to provide a mean to detect and measure the kinematics of final state partons. Quarks and gluons themselves are not directly accessible with the detector, but produce a spray of collimated hadrons that are the result of the fragmentation and hadronization of color-charged final state partons. These *hadronic jets* or simply jets are reconstructed by means of a clustering algorithm. The most common jet reconstruction algorithm at the LHC is the anti-k_T algorithm, which is inspired by a reversal of the fragmentation process. It is infrared and collinear safe, meaning that the result of the algorithm is unaltered under additional soft gluon emissions or parton splitting. Jet reconstruction in CMS is based on the FASTJET [64] package and uses PF candidates as inputs to the clustering. Several jet collections corresponding to different choice of the cone sizes are available. In this study anti-k_T jets with a cone size of 0.4 and exploiting the *charged hadron subtraction* (CHS) technique are used. The

goal of CHS is to reduce the pileup dependence by removing all charged hadron PF candidates associated with pileup vertices before the jet clustering. Pileup vertices are defined as all reconstructed vertices, except the primary vertex. It should be noted that not all charged hadron tracks are associated with a vertex, because of the quality requirements in the vertex fitting.

3.4.2 Jet selection

In this thesis work, jets are required to be within $|\eta| < 4.7$ and have a transverse momentum above 30 GeV. To reduce the background coming from calorimeter noise, a loose PF jet ID is applied. It exploits jet observables related to the number of neutral and charged PF constituents, their respective energy fractions, and the fraction of electromagnetic energy for each of the two PF hadron classes.

Since reconstructed electrons and photons are also clustered into jets, an additional *jet cleaning* has to be performed. Specifically, all jets are required to be separated from the selected leptons and their FSR photons: $\Delta R(\text{jet, lepton or photon}) > 0.4$.

3.4.3 Jet energy calibration

The momentum of a reconstructed jet, as determined by summing the momenta of its constituents, is a proxy for the parton momentum and accurate to within 10 %. This uncertainty is reduced by applying the *jet energy corrections* (JEC). The CMS JEC are designed as a sequence of corrections, each targeting a specific effect, and implemented as a scaling of the jet momentum based on event- and jet-level observables. The first step removes the dependence of the jet energies on pileup and detector noise via a mean energy density method. The next step attempts to make the jet energy response uniform in η and p_{T}, with corrections derived from simulation and residual corrections obtained from dijet and γ + jet measurements. Figure 3.12 illustrates the jet response after the jet energy scale corrections for three recoil data samples and the final JEC derived by a global fit.

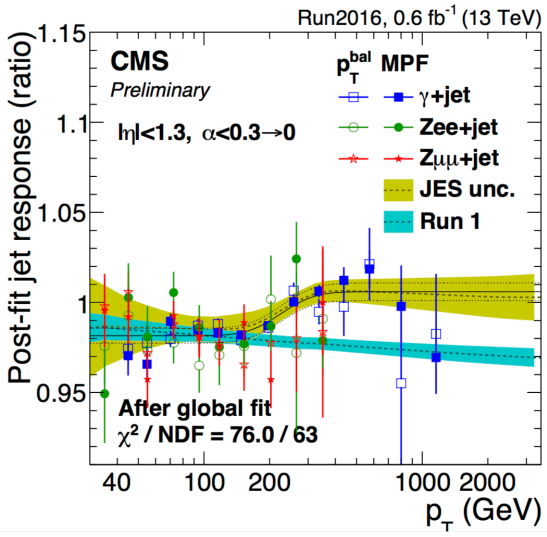


Figure 3.12: Jet energy response for three jet recoil data samples (γ + jet, $Z[\to ee]$ + jet, and $Z[\to \mu\mu,$ + jet) and two recoil methods (p_{T} balance and missing transverse momentum projection fraction (MPF)) and the final JEC derived from a global fit of all samples (solid black line) [65]. The total JES uncertainty is given by the yellow band, while the statistical uncertainty is given by the dashed curves.

3.5 Photons

The leptons exploited in the $ZZjj$ analysis are electromagnetically charged and as such, can emit energetic photons in a phenomenon called *final state radiation* (FSR). These photons tend to be collinear with the lepton and can carry away a significant fraction of the lepton's momentum. The FSR photons will reduce the signal efficiency of the isolation selection and degrade the momentum and mass resolution if they are not identified.

The starting point for the FSR algorithm are PF photons, with additional kinematic selections $|\eta^{\gamma}| < 2.4$ and $p_{\mathrm{T}}^{\gamma} > 2\,\mathrm{GeV}$. The FSR candidates are furthermore required to be isolated:

$$\mathcal{I}_{\mathrm{rel.}}^{\gamma} = \frac{1}{p_{\mathrm{T}}^{\gamma}} \left[\sum_{\text{photons}} p_{\mathrm{T}} + \sum_{\substack{\text{neutral}\\ \text{hadrons}}} p_{\mathrm{T}} + \sum_{\substack{\text{charged}\\ \text{hadrons}}} p_{\mathrm{T}} \right] \tag{3.17}$$

where the cone size is 0.3 and $\mathcal{I}_{\mathrm{rel.}}^{\gamma} < 1.8$. Photons that are linked to the supercluster of any electron that passes the impact parameters selection are not considered to avoid double counting.

Finally, the FSR candidates are required to satisfy kinematic selection $\Delta R(\gamma, \ell) < 0.5$ and $\Delta R(\gamma, \ell)/E_{\mathrm{T},\gamma}^2 < 0.012$ to further suppress backgrounds. In the rare case where several photons are matched to the same lepton, only the FSR candidate with lowest $\Delta R(\gamma, \ell)/E_{\mathrm{T},\gamma}^2$ is kept. As mentioned in the sections on the electron and muon selections, FSR photons are excluded from all lepton isolation sums. The FSR algorithm affects about 4 % of all events, with significantly lower rates for final states that feature electrons, for which the majority of FSR photons are already included in the refined supercluster.

Chapter 4

Electron studies

This chapter describes in detail the development and continued optimization of the multivariate electron identification algorithm (MVA ID) used for the 13 TeV data in CMS. These optimizations are crucial for multilepton analyses like $ZZjj$ where lepton selection efficiencies enter the event selection efficiencies to the fourth power. Compared to muons, electrons suffer from intrinsically higher background rates and larger selection performance improvements are to be expected. The chapter starts with an investigation into the use of tracking observables in the identification of electrons, resulting in a list of novel observables to discriminate electron backgrounds. Starting from the 8 TeV MVA ID, these novel observables are then evaluated in the context of the multivariate ID. A reduction of the electron fake rate by up to 50 % is achieved by introducing observables sensitive to the photon conversion background. These improvements are also implemented in the general purpose MVA ID, which is used by non-multilepton analyses that in general select only electrons which fired the trigger. A triggering selection which mimics the HLT selection as part of the general purpose ID is derived. The MVA ID used in the $ZZjj$ analysis and all multilepton studies based on the 2016 data is discussed in detail. Finally, the efficiency of the MVA ID as a function of the electron transverse momentum and pseudorapidity are studied.

4.1 Tracking observables for electron identification

This section presents a study on the use of observables derived from the electron tracking in the electron identification algorithm. The search for such novel variables is motivated by the fact that electron reconstruction and identification has traditionally relied heavily on calorimetric information, suggesting a potential for improvements coming from the excellent track reconstruction provided by the CMS tracker and the GSF fit. The latter furthermore provides a wealth of information that is usually only exploited in the form of a few simple statistics.

The two dominant sources of background to prompt electrons are the accidental overlap of a charged hadron track and an electromagnetic cluster from $\pi \to \gamma\gamma$ decays and the conversion of photons in the tracker material. The potential of exploiting tracking information to reduce both types of backgrounds are investigated.

The starting point of this study is the observation that the energy loss from bremsstrahlung for charged hadrons is much suppressed compared to electrons. Any tracking observable that correlates strongly with the true amount of energy losses due to

bremsstrahlung thus has the potential to discriminate against the background. This idea has been exploited in the run I MVA ID, which exploited the $f_{brem} = 1 - p_{out}/p_{in}$ observable, calculated on the GSF track momentum at the ECAL (p_{out}) and at the vertex (p_{in}). However, the accuracy of the GSF measurement, in particular on the outgoing momentum, has not been studied extensively. Understanding the correlation between f_{brem} and the true energy loss from bremsstrahlung is thus the first step in gauging the potential of novel tracking observables in the ID algorithm.

The regular MC datasets are insufficient for this study as they do not store the details of the track fitting, like individual tracker hits, to reduce the event size. For similar reasons, they also do not store the interaction records from the Geant simulation which are crucial to understanding the actual bremsstrahlung behavior of the electrons. The first step is thus to generate the extended simulation samples for the electron signal and the charged pion background. To isolate the relevant physics, these simulated events feature a single particle that is being fired at a fixed transverse momentum of 35 GeV and random pseudorapidity into the detector (a *particle gun*). These events are then reconstructed just like in the standard simulation, albeit with a small modification to the electron seeding parameters. The default electron reconstruction exploits very loose cuts on H/E, $\Delta\varphi$, and $\Delta\eta$ to reject background and to reduce the number of time-consuming GSF fits. These selections that combine calorimetric measurements with tracking information are relaxed in order to increase the electron reconstruction efficiency in the background sample, without introducing a bias for the pure track-related observables.

The Geant interaction record allows to trace the passage of the simulated particle through the tracker material. For any interaction of the electron with the tracker material it includes the position of that interaction and the outgoing particles. For electron bremsstrahlung it contains the position and outgoing kinematics of the $e \to e + \gamma$ emission. Figure 4.1 shows the position of all bremsstrahlung vertices of the simulated electron sample, projected onto the longitudinal plane of the detector. The visible structures clearly resemble the active material of the silicon tracker, but also reveal regions of large material density due to supporting structures.

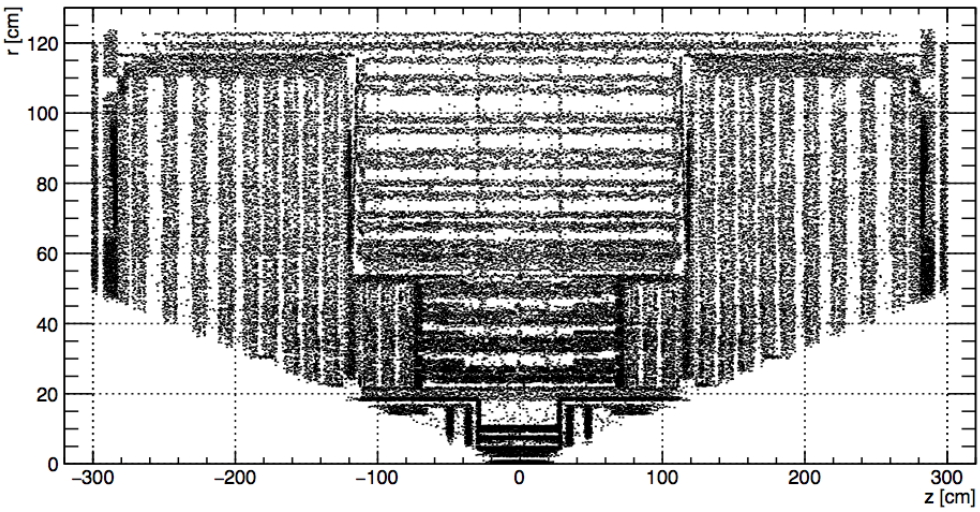


Figure 4.1: Projection of all electron bremsstrahlung vertices onto the longitudinal plane, indicating the regions of lower and higher material density in the CMS tracker. The underlying simulation uses electrons with $p_T = 35$ GeV and a uniform η distribution.

4.1.1 Study of momentum loss measurements

The Geant interaction record allows to access the relative momentum loss in the simulation f_{brem}^{Geant}, that is to access the true momentum of the electron after traversing and loosing energy in the tracker material. This quantity is of central interest because it encapsulates the difference between electron signal ($f_{brem}^{Geant} > 0$) and charged hadron backgrounds ($f_{brem}^{Geant} \approx 0$). A perfect reconstruction of this quantity in data, i.e. a perfect correlation between f_{brem}^{Geant} and f_{brem}, would provide a very powerful discriminator against charged hadron backgrounds.

However, this correlation is not perfect as shown in Fig. 4.2. A reliable reconstruction of the true momentum losses is only achieved for a fraction of electrons, with an appreciable underestimation particularly in the endcap. Figure 4.3 provides further details on the accuracy of the f_{brem} reconstruction as a function of the detector η and the character of the energy loss. The panels in the top row of Fig. 4.3 show the reconstructed f_{brem} for two sub-populations of electrons with $0.6 < f_{brem}^{Geant} < 0.7$ (left) and $0.9 < f_{brem}^{Geant}$ (right), separately electrons in the central and forward regions. For electrons with $0.6 < f_{brem}^{Geant} < 0.7$ in the central region a reasonable f_{brem} measurement is achieved, but a significant left tail is visible. In contrast, the f_{brem} measurement such electrons in the forward region does not feature any peak around $f_{brem} \approx 0.6$, but the reconstructed f_{brem} is distributed almost evenly between 0.2 and 0.6. For electrons that loose $> 90\,\%$ of their energy (top right panel) there is almost no difference between central and forward regions regarding the quality of the f_{brem} measurement, which features a peak at the expected value but large tails are again present. The bottom panels of Fig. 4.3 differentiates the electron sub-populations by the nature of the momentum loss: 'big brem' electrons are identified as those cases where at least 70% of the total radiative loss occurs in a single emission and 'no brem' are the corresponding complement. The distributions for the 'big brem' are bi-modal, where one of the peaks corresponds to a fair measurement of f_{brem}.

This seems to indicate that the momentum measurement by the GSF fit is reliable in certain cases, likely those where the single large emission occurs in the middle of the reconstructed tracks, leaving a sufficient number of tracker hits before and after the emission to determine p_T^{in} and p_T^{out}. The f_{brem} reconstruction for electrons that lose energy via multiple smaller emissions is less reliable. The demonstrated poor representation of the true momentum loss by the reconstructed f_{brem} can be interpreted as a sign that the physics of energy losses through bremsstrahlung is not yet optimally exploited with f_{brem}.

In those cases where the reconstructed f_{brem} is a poor estimator of the actual energy losses, one might assume that this poor measurement is reflected in the GSF fit uncertainties on the incoming and outgoing momenta. This assumption is not supported by the simulation, as shown in Fig. 4.4: neither the uncertainty on the incoming momentum (left panel) nor on the outgoing momentum (middle panel) correlate with the true energy loss. The same holds for the correlation between both uncertainties, as illustrated in the right panel of Fig. 4.4 which shows the significance of the f_{brem} measurement. The significance is defined as the ratio of f_{brem} and its uncertainty, the latter being the squared sum of the incoming and outgoing momentum uncertainties assuming no correlation between the two.

An interesting observation is the presence of sub-populations around 10 % and 20 % to 30 % in the uncertainty on the outgoing momentum. These can be traced to the η-dependance of the uncertainty as shown in the left panel in Fig. 4.5, which is rem-

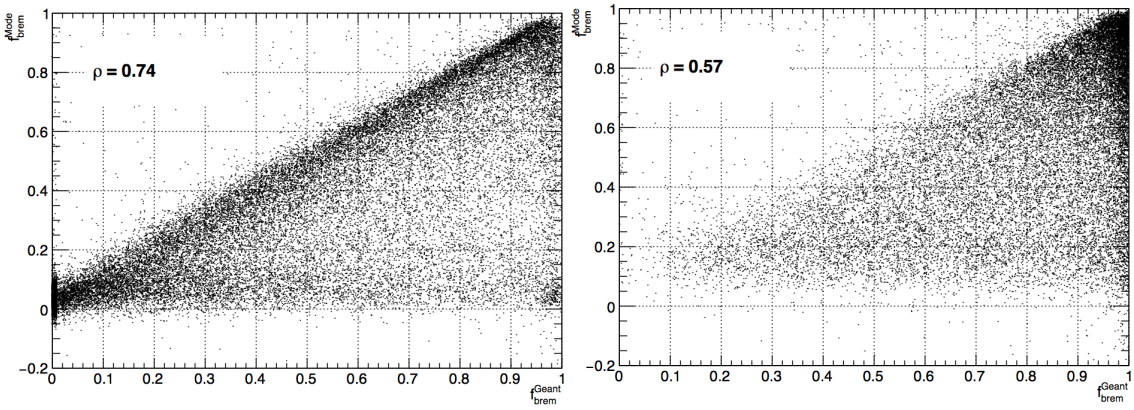


Figure 4.2: Distribution of the true momentum loss of electrons traversing the tracker $f_{\text{brem}}^{\text{Geant}}$ and the reconstructed $f_{\text{brem}}^{\text{Mode}}$ for electrons in the central region ($|\eta| < 0.8$, left) and the forward region ($|\eta| > 0.8$, right). The correlation coefficient ρ is superimposed and the superscript 'Mode' in $f_{\text{brem}}^{\text{Mode}}$ indicates that the momenta are based on the mode of the GSF states. The underlying simulation uses electrons with $p_{\text{T}} = 35\,\text{GeV}$ and a uniform η distribution.

iniscent of the material budget distribution of the CMS tracker. A dependance on the amount of traversed material is of course expected, but one also expect to find a fraction of tracks whose outgoing momenta are well-measured, even in regions of large material budget. The GSF fit uncertainty appears to be dominated by the material modeling provided as an input to the fitting procedure, not by the actual uncertainty of a particular track. The GSF fit uncertainty of the transverse momentum is the quadratic sum of two contributions:

$$\sum_{i}^{N} w_i \sigma_i^p, \tag{4.1}$$

$$\frac{1}{p} \sum_{i}^{N} \sum_{j}^{N} w_i w_j (p_i - p_j)^2. \tag{4.2}$$

The sums are over the components of the GSF fit and w_i, p_i, σ_i^p are the associated weights, momenta, and uncertainties on the momenta.

The covariance term is illustrated in the center panel of Fig. 4.5 and the term of the differences between the GSF states is shown in the right panel. Both exhibit a shape reminiscent of the material budget distribution, but the covariance distribution shows less dispersion. The covariances in a single Kalman filter state are the result of the combination of the expected and measured uncertainty from the per-hit resolutions. The lack of dispersion and the resemblance of the overall distribution of the covariance term seems to indicate that it is not the uncertainties of the measured hit positions that dominate the covariance matrix, but the expected uncertainty, which in turn is given by the material budget modeling.

The term capturing the differences between the GSF components results in a larger spread around the mean and appears less driven by the details of the material budget. This is expected as the components of the GSF fit correspond to the different hypotheses of energy losses. A sizable contribution in the sum requires that there are at least two components with non-negligible weight and different momentum estimates,

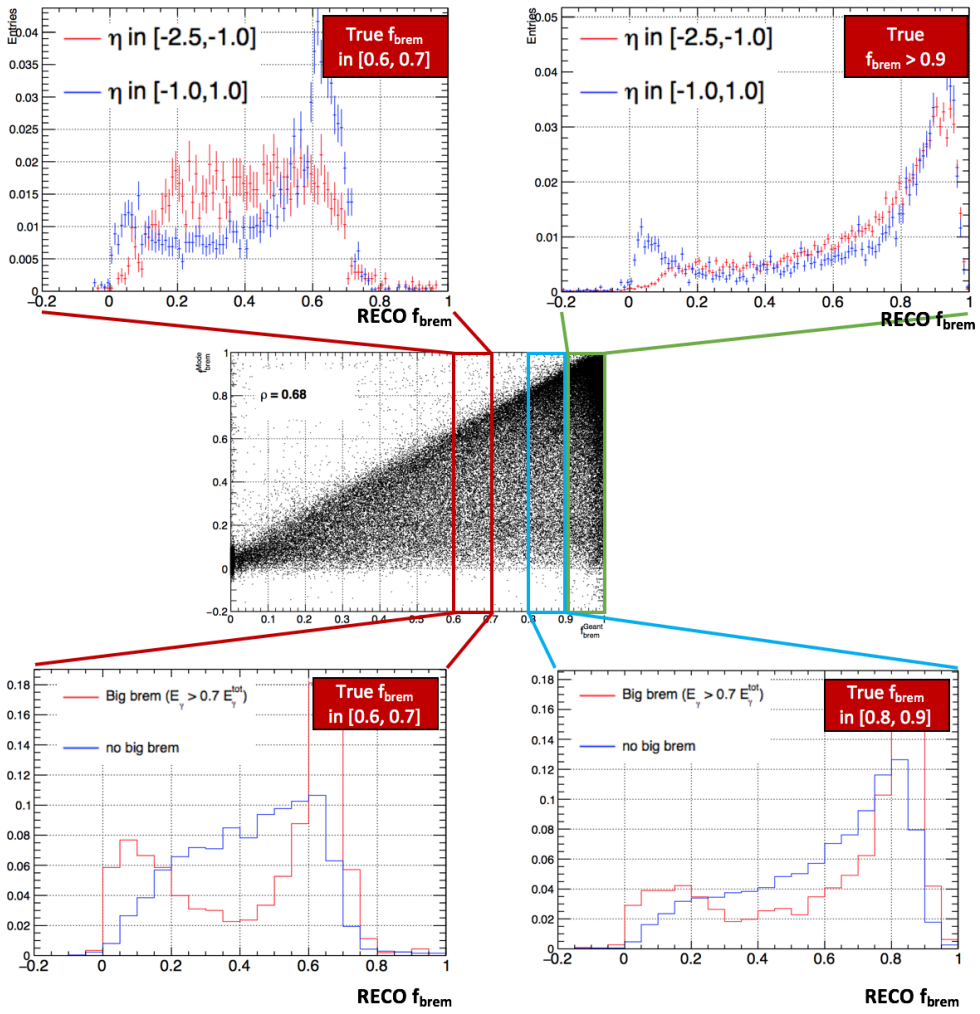


Figure 4.3: Distributions of the reconstructed f_{brem} for two selections of electrons with true momentum losses in the specified ranges. The distribution merges the central and forward regions separated in Fig. 4.2. The top row compares the quality of the f_{brem} reconstruction in the central and forward detector regions while the bottom row illustrates the impact of the difference between a single large and multiple small emissions.

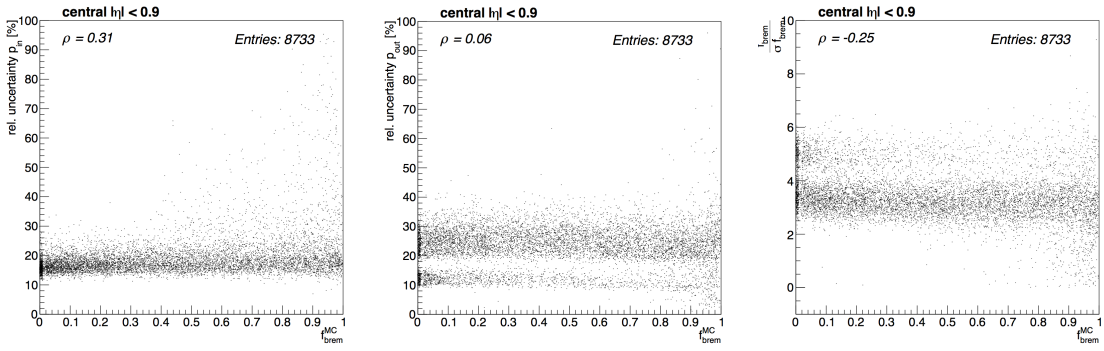


Figure 4.4: Uncertainties of the inner (left) and outer (center) electron momentum as returned from the GSF fit as well as the f_{brem} significance (right) as functions of the true energy loss. The uncertainties are largely uncorrelated to the energy losses. The underlying simulation features electrons of $p_T = 35\,\text{GeV}$ and a uniform η distribution.

a situation which occurs when the hit pattern does not admit a proper momentum measurement, e.g., following a change of curvature due to bremsstrahlung emission. However, the second term (Eq. (4.2)) also provides a poor estimate of the actual per-track uncertainty and is still dominated by the expected energy losses as can be seen in Fig. 4.6 which shows the uncertainty for electrons that loose at most 10 % of their energy (left) and for misreconstructed electron candidates with tracks from charged pions. Both both cases are expected to be measured well without any dependence on the expected bremsstrahlung losses in the GSF fit. Instead, one observes the same material structures, indicating that the GSF components corresponding to different energy loss hypotheses are not sufficiently suppressed.

One is of course not limited to the above statistics to determine the uncertainty and a variety of statistics on the GSF components was investigated, including the uncertainties of individual components and strategies to groom components of low weight. No promising candidate to reliably estimate the uncertainty on p_{out} was identified.

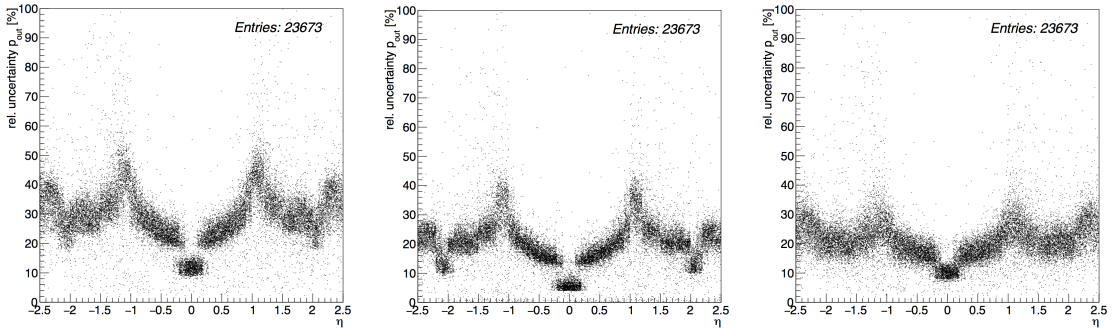


Figure 4.5: Pseudorapidity distributions of the GSF fit uncertainty on the outgoing momentum (left). The center and right panel show the two contributions to the total uncertainty given by Eq. (4.1) and Eq. (4.2) respectively. The distributions are obtained from a simulation of $p_{\text{T}} = 35\,\text{GeV}$ electrons with a uniform η distribution.

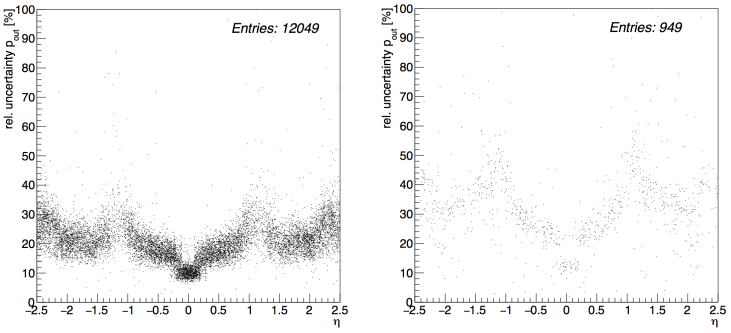


Figure 4.6: Pseudorapidity distributions of the GSF fit uncertainty on the outgoing momentum for electrons that loose a maximum of 10 % of their energy due to bremsstrahlung (left) and for misreconstructed electrons from pion tracks (right). The distributions are obtained from a simulation of $p_{\text{T}} = 35\,\text{GeV}$ electrons and pions, with a uniform η distribution.

These observations prompted an investigation into the accuracy of the incoming and

outgoing momenta, in order to assure that the material modeling inherent in the GSF fit does not bias the p_{out} measurement. Figure 4.7 shows the ratio of the reconstructed to the true incoming (left) and outgoing momenta (center) for tracks where the momentum on the relevant hits does not change. At least three such hits at constant curvature are needed for a reasonable momentum measurement, as could be expected.

The outgoing momentum is indeed very well measured for tracks that do not feature bremsstrahlung emissions and there seems to be no bias on p_{out} coming from the GSF fit and the assumed energy losses. This absence of a bias is supported by the observation that the reconstructed f_{brem} for tracks which feature at least 5 hits at constant momentum at the beginning and at the end is much better correlated with the true energy losses, as shown in the right panel of Fig. 4.7.

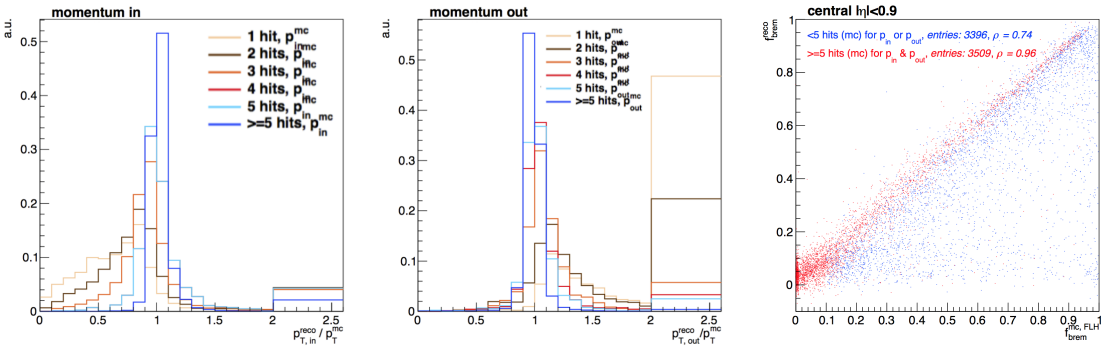


Figure 4.7: Ratio between the reconstructed and true transverse momentum of electrons at the first tracker hit (left) and at the last tracker hit (center). Tracks for which the true momentum is constant to within 10 % for a given number of hits are displayed separately. The distribution in the right panel highlights the improved correlation between the true and reconstructed energy losses for tracks that feature at least 5 hits at constant curvature (red points). The f_{brem} measurement of tracks that do not satisfy this selection exhibit significant tails (blue points). The distributions are obtained from a simulation of $p_{\text{T}} = 35\,\text{GeV}$ electrons with a uniform η distribution.

Conclusions on outgoing momentum measurement study

The reconstruction of the energy loss along the electron track is found to exhibit appreciable tails with respect to the true energy loss. The uncertainties returned by the GSF fit, as well as its subcomponents do not provide a proxy of the quality of the f_{brem} measurement and are not correlated to the true energy losses. The uncertainties largely reflect the expected spread of electron momenta due to expected bremsstrahlung losses, not the actual losses. The uncertainty on the outgoing momentum p_{out} in particular is dominated by the material modeling. However, the measurement of p_{out} is not biased by the material modeling, provided that the momentum can actually be measured from a sufficient number of tracker hits at a constant momentum. For such tracks the p_{out} measurement accurately reflects the true outgoing momentum, but the uncertainty estimates do not reflect the quality of the measurement. No promising observables for electron identification from the standard GSF fit output are found. In particular, no observables permitting to estimate the quality of the p_{out} measurement could be identified. The next idea was to resort to simpler track measurements that do not require the successful measurement of the outgoing momentum.

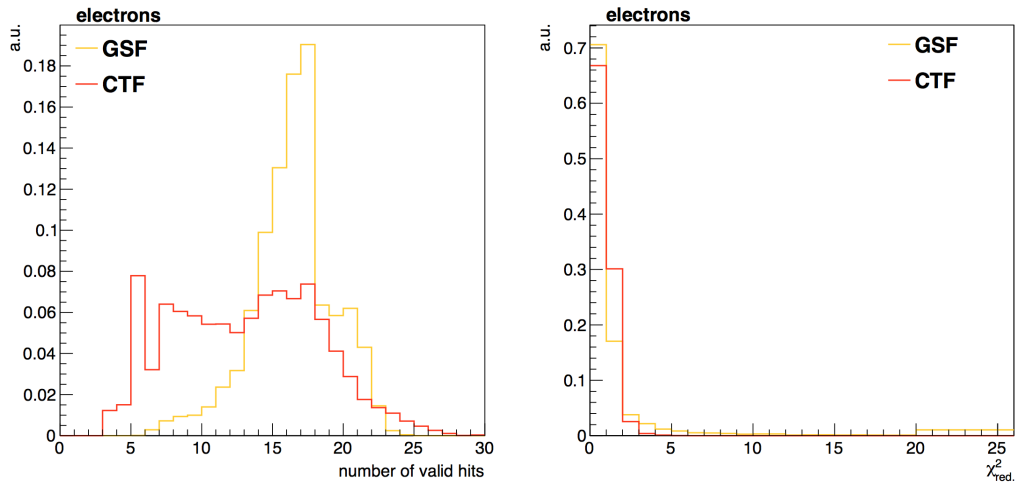


Figure 4.8: Distribution of the number of hits (left) and the reduced χ^2 (right) for the Kalman filter track (CTF) and the GSF track for true electrons. The distributions are obtained from a simulation of $p_T = 35\,\text{GeV}$ electrons with a uniform η distribution.

4.1.2 Extracting novel tracking observables for electron identification

The problem of electron identification based solely on the tracking measurements can be framed as an hypothesis test, where the reconstructed track is compared to the electron and charged hadron track models. In a way this test is already encapsulated in the reduced χ^2 observables of the Kalman and GSF track fit, χ^2_{KF} and χ^2_{GSF}. However, the two are based on different track candidates, i.e., the actual hit pattern used in the Kalman fit is different to that used in the GSF fit due to the relaxed pattern recognition employed in the latter.

Figure 4.8 (left) shows the number of hits for both track collections, illustrating that the KF tracks tend to be much shorter, illustrating the improvement provided by the dedicated track finding for electrons. While χ^2_{GSF} can thus be seen as testing the compatibly of a given track with the electron hypothesis, the same is not true for χ^2_{KF}. In fact the regular track building exploits the χ^2 variable to stop the hit collection. As a consequence, both χ^2 values are of order one for electrons, as shown in the right panel of Fig. 4.8.

The idea is thus to refit the hits of the GSF track with a regular Kalman filter that models only the small energy losses expected for charged hadrons. Figure 4.9 (left) presents the χ^2_{refit} distributions, that is the reduced χ^2 for the signal and background, clearly showing the increase in χ^2 for the electron signal when all hits are used in the Kalman filter fit. The right panel of Fig. 4.9 shows the considerable improvement to the ROC curve for a simple rectangular cut on the regular Kalman fit and the refit χ^2.

One shortcoming of the χ^2 observables is that they only take into account the absolute value of the discrepancy between the measured hit positions and those predicted by the respective track model. Considering the transverse plane, one can define a signum of a hit and a fitted track based on whether the hit is located 'inside' or 'outside' the circle obtained by extending the track at constant curvature. Figure 4.10 (left) illustrates this notion and the distributions of the signed per-hit deviations $\Delta\varphi$ of the GSF track for the signal and two background samples are shown in the center and right

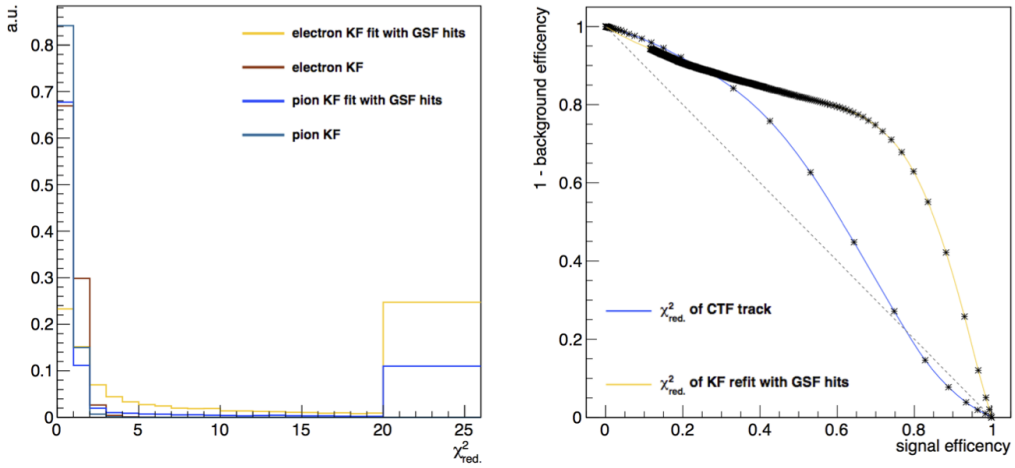


Figure 4.9: Distribution of the reduced χ^2 of the Kalman filter tracks (CTF) and the refit of the GSF hits with a Kalman filter for the signal and charged pion background (left). The ROC curve illustrates the improvement in separation power for the refitted Kalman filter track. The distributions are obtained from a simulation of $p_T = 35\,\text{GeV}$ electrons with a uniform η distribution.

panel. As expected, the background distributions are not centered at zero, but shifted towards negative values, that is the predicted hits are systematically 'outside' of the hypothesized track circle. In other words the electron energy loss model implemented in the GSF assumes that the next hit will likely be the result of a particle of lower momentum, in effect under-predicting the actual momentum for charged hadrons. A clear disadvantage of the per-hit deviations $\Delta\varphi$ is the fact that they are measured in radians and will thus depend on the specific tracking subdetector and the momentum of the track they are associated with.

Instead, a new observable that combines the notion of a hit signum and the per-hit increase of χ^2 is constructed. During the inside-out pass of the Kalman fit, each hit leads to an increase of the overall χ^2, which is referred to as the per-hit χ^2. A new observable $\sum_{\chi^2}$ sums these signed χ^2 and is then normalize to the total number of hits. Figure 4.11 (left) shows the distribution of this novel observable for the GSF track. Similarly to the per-hit deviations, the background distribution is systematically shifted towards negative values, allowing to separate the signal and background as illustrated in the right panel of Fig. 4.11. Finally, the refit of the hits of the GSF track with a regular Kalman filter and the notion of a per-hit deviation signum are combined into $\sum_{\chi^2}^{\text{refit}}$. Figure 4.12 shows the three novel track observables finally considered in this study: the χ^2_{refit} of the refit of the GSF track hits with a regular Kalman filter (left), the normalized sum of the signed per-hit χ^2 of the GSF track $\sum_{\chi^2}^{GSF}$ (center), and the normalized sum of the signed per-hit χ^2 of the refitted track $\sum_{\chi^2}^{\text{refit}}$ (right).

By themselves these new tracking observables, notably $\sum_{\chi^2}^{\text{refit}}$, provide a fair separation power to discriminate between the signal and the pion background. The next question is whether they add any information to the variables already exploited in the multivariate electron ID.

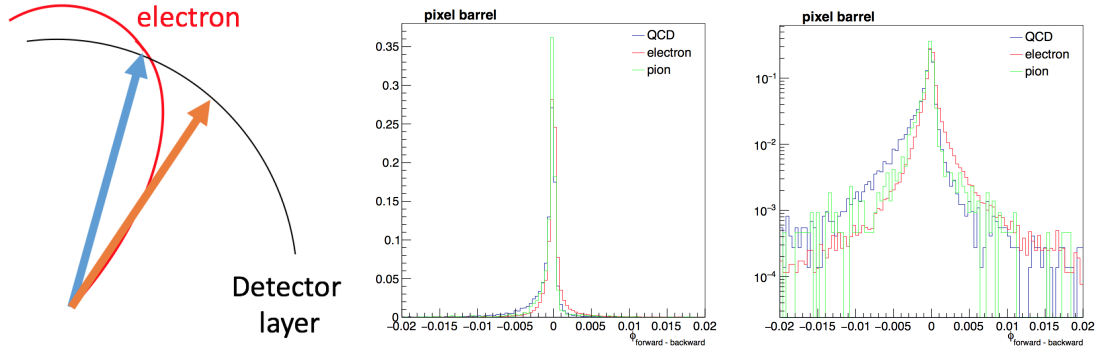


Figure 4.10: Illustration of the per-hit signum definition (left). Distribution of the per-hit deviations of electron tracks in the barrel pixel tracker (center and right). The underlying simulations use electrons and pions of $p_T = 35\,\text{GeV}$ and with a uniform η distribution as well as a QCD multijet sample.

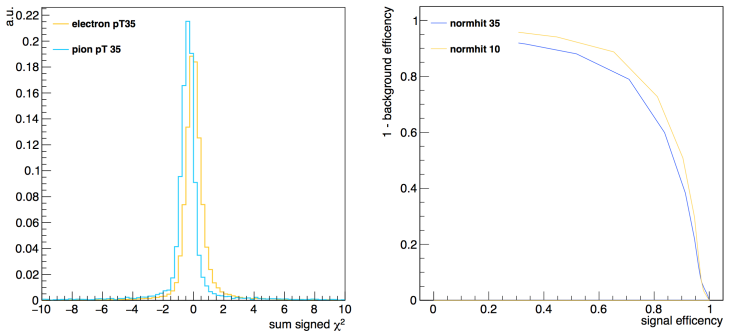


Figure 4.11: Distribution of the normalized sum of signed χ^2 for 35 GeV electrons and charged pions in the simulation (left) and the associated ROC curves for 10 and 35 GeV electrons and pions

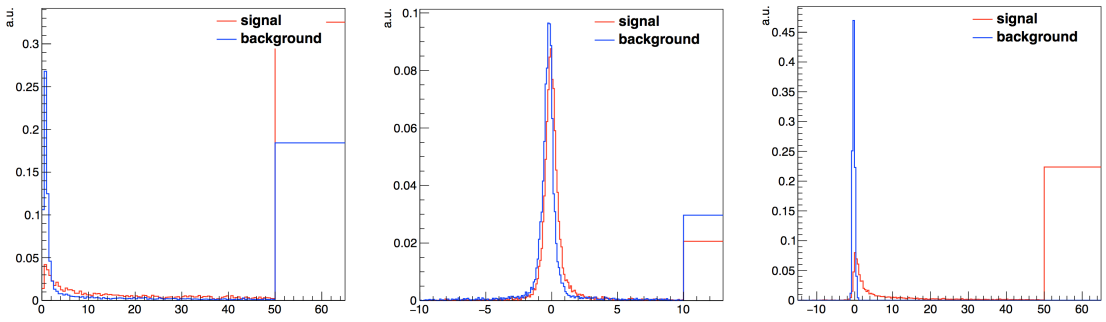


Figure 4.12: Signal and background distributions for χ^2_{refit} (left), $\sum_{\chi^2}^{\text{GSF}}$ (center), and $\sum_{\chi^2}^{\text{refit}}$ (right) described in the text. The underlying simulation is based on Z boson decays but only electrons with $9 < p_T < 12\,\text{GeV}$ are shown here.

4.2 Optimization of the 13 TeV multivariate electron ID

4.2.1 Introduction to the multivariate electron ID

The most straightforward way to exploit observables sensitive to the difference between signal and background is to consider one such variable and to make a single selection or *cut*, for example to require that f_{brem} be larger than a certain threshold. While one can extend the notion of rectangular cuts to several variables and optimize the determination of the cut values, this approach is still limited to a single selection per variable and will thus be suboptimal. This can be illustrated using the f_{brem} example: the material budget causing the bremsstrahlung exhibits a strong η dependence and the efficiency of a cut on f_{brem} will thus also have an η dependance.

From a machine learning perspective, the task of discriminating between two processes or classes – electron signal and background in this case – is considered a supervised classification problem. Given a set of N observables, the goal is to find the optimal separation boundary between the two classes in the N-dimensional space. The electron ID uses the gradient boosted decision tree (BDT) algorithm implemented in the TMVA framework to achieve this separation. The BDT is effectively a function that maps the N-dimensional space spanned by the input variables to a real number, $\mathrm{BDT} : \mathbb{R}^N \rightarrow \mathbb{R}$, and one defines a signal selection by cutting on its output.

In machine learning parlance, the BDT algorithm has to be *trained* on a sample of signal and background electrons. These samples are obtained by selecting reconstructed electrons in a Drell-Yan plus jets simulation, where signal electrons are geometrically matched to electrons from the Z boson decay. This matching exploits the MC truth record, and the first step consists in finding the generated electron that is closest in ΔR to the reconstructed electron candidate, if any. The exact classification of reconstructed electrons based on the MC truth record is then as follows[1]:

1. **Unmatched**: electron candidates that are not matched to a MC truth electron within $\Delta R < 0.1$
2. **Non-prompt**: electron candidates that are matched to a true electron, whose ancestor has a $|\mathrm{ID}_{\mathrm{PDG}}| > 50$ and is short-lived
3. **Tau decay**: electron candidates that are matched to a true electron, whose ancestor was a τ lepton
4. **Prompt**: all remaining

Prompt electrons are taken as the signal and the background is composed of the unmatched and non-prompt electron candidates. With the signal and background defined, the next step is to identify which observables are to be included in the training. Table 4.1 lists all observables used for the 2016 MVA ID.

The distributions of most of the observables in Table 4.1, e.g., f_{brem}, vary depending on the detector η or transverse momentum of the signal and background. These variations are driven by the change in material budget and differences related to the detector itself. With sufficient statistics, on could expect the BDT to learn the different

[1] The selection of signal and background electrons had to be updated with respect to run I. The upgrade from Pythia6 to Pythia8 changed the MC status codes and the legacy selection led to background contaminations in the signal sample.

Table 4.1: Overview of input variables to the identification classifier. Variables introduced as part of this thesis work are given in **bold font**.

Observable type	Observable	Definition
cluster shape	$\sigma_{i\eta i\eta}$	Standard deviation of the energy distribution in the cluster along the η direction. The η coordinate is given by crystal index instead of the actual detector η to avoid biases in clusters across ECAL gaps. The cluster used is not the mustache SC, but the 5x5 cluster used in run-I because it provides more discrimination power.
	$\sigma_{i\varphi i\varphi}$	Same as $\sigma_{i\eta i\eta}$, but along the φ direction
	$\Delta\eta_{\mathrm{SC}}$	Width of the supercluster along η
	$\Delta\varphi_{\mathrm{SC}}$	Width of the supercluster along φ
	H/E	Ratio of the HCAL energy in a cone of $\Delta R = 0.15$ centered at the SC position to the SC energy
	$(E_{5\times5} - E_{5\times1})/E_{5\times5}$	Circularity. The energy sums $E_{i\times j}$ of the i crystals in φ and j crystals in η centered on the seed crystal
	$R_9 = E_{3\times3}/E_{\mathrm{SC}}$	Ratio of the energy in a 3×3 (9 crystal) cluster around the seed over the SC energy
	$E_{\mathrm{PS}}/E_{\mathrm{raw}}$	For endcap training bins only: energy fraction in pre-shower over the raw SC energy
tracking	$\mathrm{f}_{\mathrm{brem}} = 1 - p_{\mathrm{out}}/p_{\mathrm{in}}$	Fractional momentum loss as measured by the GSF fit. The momenta p_{in} and p_{out} are extrapolations of the GSF track to the vertex and ECAL respectively.
	N_{KF}	Number of hits of the Kalman Filter track of the iterative combinatorial track finder, if any.
	N_{GSF}	**Number of hits of the GSF track**
	χ^2_{KF}	Reduced/normalized χ^2 of the KF track, if any
	χ^2_{GSF}	Reduced χ^2 of the GSF track fit
	$N_{\mathrm{miss.\ hits}}$	**Number of expected but missing inner hits**
	$P_{\mathrm{conv.}}$	**Probability transform of the conversion vertex fit χ^2, if any**
track-cluster matching	$E_{\mathrm{SC}}/p_{\mathrm{in}}$	Ratio of the SC energy and the track momentum at the innermost hit
	$E_{\mathrm{ele}}/p_{\mathrm{out}}$	Ratio of the energy of the cluster closest to the electron track and the track momentum at the outermost hit
	$1/E_{\mathrm{tot}} - 1/p_{\mathrm{in}}$	Energy-momentum agreement
	$\Delta\eta_{\mathrm{in}} = \lvert\eta_{\mathrm{SC}} - \eta_{\mathrm{in}}\rvert$	Distance between the energy-weighted center of the SC and the expected shower position as extrapolated from the GSF trajectory state at the vertex
	$\Delta\varphi_{\mathrm{in}} = \varphi_{\mathrm{SC}} - \varphi_{\mathrm{in}}$	Same as $\Delta\eta_{in}$ but along φ
	$\Delta\eta_{\mathrm{seed}} = \lvert\eta_{\mathrm{seed}} - \eta_{\mathrm{out}}\rvert$	Distance between the pseudorapidity of the seed cluster and the expected shower position as extrapolated from the GSF trajectory state of the outermost hit

$\mathrm{f}_{\mathrm{brem}}$ distributions for signal and background as a function of the electron η. However, this cannot be guaranteed and the time needed to train the BDT scales linearly with the sample size. One can use the understanding of the observables to define several categories in which to train the BDT. This can help to boost performance and reduces the training time. A split at $|\eta| = 1.479$ is motivated by the differences of the ECAL detector in the endcaps and barrel. Another split at $|\eta| = 0.8$ in the barrel is introduced to separate the region of low and high material budget. In order to assure an optimal treatment of low-p_{T} electrons, those with $5 < p_{\mathrm{T}} < 10\,\mathrm{GeV}$ are treated separately, making a total of six independent BDTs.

4.2.2 Optimization of the multivariate ID

The starting point for the run II MVA ID optimization is a retraining of the BDTs on the 13 TeV simulation using the same setup of six BDTs as previously used at 8 TeV. The input variables and hyper-parameters are also unchanged with respect to the run I MVA ID. The signal and background are selected from the $Z \to e^+e^-$ + jets simulation, following the algorithm presented in Section 4.2.1 and without any reweighing of the kinematics. This retraining serves as the baseline to evaluate the performance improvements coming from the new tracking observables.

Adding the new tracking observables detailed in Section 4.1.2 does indeed improve the background rejection by up to 20 %, as illustrated in the left plot in Fig. 4.13. However, identical improvements are obtained when adding the number of hits of the GSF track N_{GSF} to the training, as shown in the right panel of Fig. 4.13. Adding the new tracking observables on top of N_{GSF} does not improve the performance of the BDT, that is the information included in the new tracking observables is redundant with the

information coming from the other observables and their correlations. It was checked that the same holds for specific sub-populations of electrons, in particular those with superclusters that are composed of a single cluster and with $f_{brem} < 0.5$ ('golden' electrons).

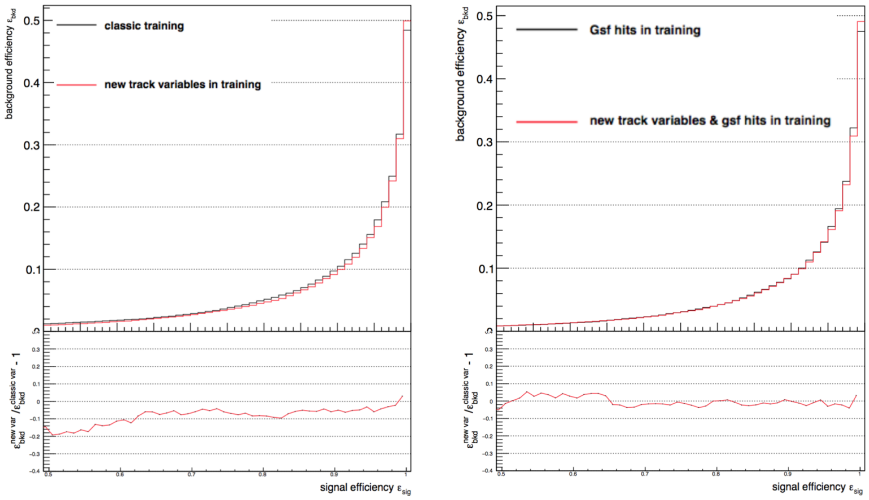


Figure 4.13: ROC curve comparisons for electrons with $p_T > 10\,\text{GeV}$ in the endcap training bin for variations of the BDT training. The left panel compares the BDT trained using the run I ('classic') set of input variables to a BDT training that includes the new tracking variables and N_{GSF}. The right panel shows the performance difference between the BDT trained on the new tracking variables plus N_{GSF} and a training that only adds N_{GSF}. Similar results are obtained in the 5 training bins not shown here.

The novel tracking observables explored so far targeted the electron background from hadronic jets where the track of the electron candidate originates from a charged hadron. Of similar importance is the background of non-prompt electrons from photon conversions in the tracker material, in particular in the forward region where the material budget is large and the electron track seeding requirements are relaxed to recover reconstruction efficiency. Tracking provides powerful observables to reduce this background from converted photons.

For some run I analyses, a dedicated conversion rejection selection was applied on-top of the electron ID. A central observable in this conversion rejection is the number of expected but *missing inner hits*, that is the tracker hits that should have been recorded because the extrapolated track goes through active and working tracker modules, but weren't. Barring the case of early photon conversions in the beam pipe or first layer of the pixel detector, electrons from conversions are very likely to feature at least one such missing inner hit. Photon conversions can also be directly identified if both tracks of the conversion electrons are reconstructed. A new conversion vertex finder was developed for the run II and the χ^2 of the this vertex fit is added in the BDT training[2]. Figure 4.14 shows the improvements achieved from including these conversion rejection observables in the BDT: the background rates for a fixed signal efficiency decrease by 20 % to 40 %, with larger reductions in the endcaps.

[2]The χ^2 of the vertex fit is converted into a probability, based on the number of degrees of freedom of the fit and assuming a chi-squared distribution. This transformation is beneficial for the BDT training, as it restricts the variable to lie between zero and one.

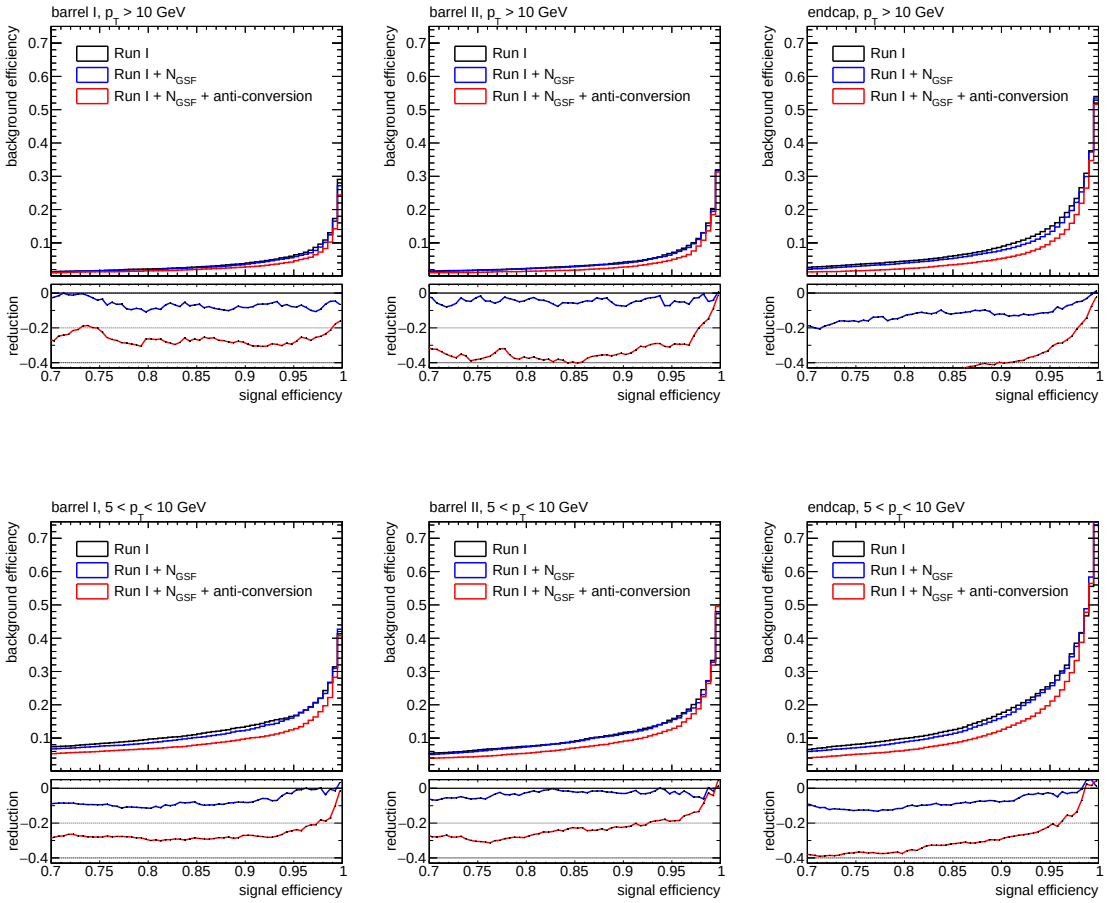


Figure 4.14: ROC curves for all training bins of the 2015 electron BDT for various input variables. The baseline BDT (black curve) uses the observables exploited in the run I ID. The blue curve shows the performance of the BDT trained with the number of hits of the GSF track (N_{GSF}) added as an input. The red curve shows the final BDT trained with the run I observables, N_{GSF}, as well as the conversion vertex fit probability, and the number of missing inner hits. The inclusion of the tracking variables reduces the background efficiency by 20 % to 40 %, depending on the working point.

Aside from adding new variables in the BDT training it was tested whether the performance can be improved further by introducing more training categories in order to allow the BDT to specialize on the specifics of the selected electron population. Two such ways of splitting the existing training bins were evaluated. The first is to introduce a split in the endcap at $|\eta| = 2$, similar to the split in the barrel, to separate regions of lower and higher material budget. This did not yield any significant improvements. Another potential split of training categories is motivated by the turn-on curves of the BDT presented in the left panel of Fig. 4.15, which show that the identification of electrons with $p_{\text{T}} < 20\,\text{GeV}$ is particularly challenging. However, no improvement in background rejection is observed when training the BDT exclusively on this particular p_{T} regime as can be seen from the ROC curves in the center and right panel of Fig. 4.15.

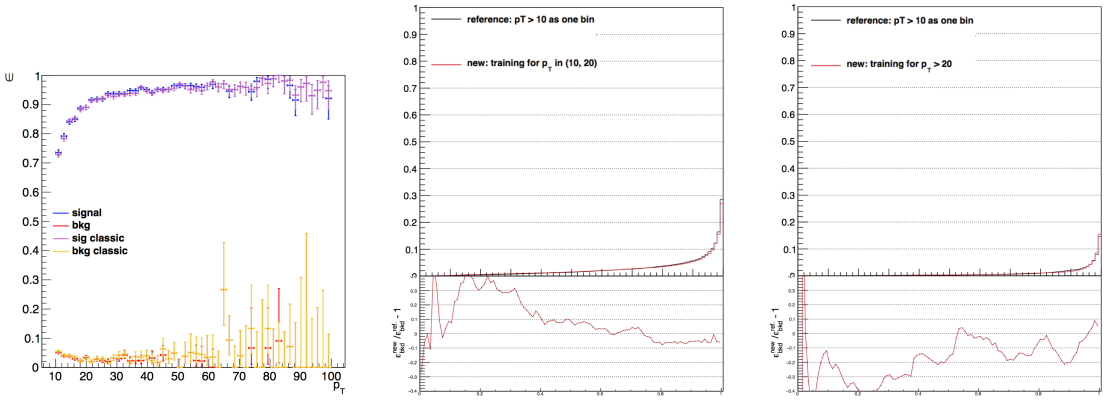


Figure 4.15: Signal and background efficiencies as functions of the electron p_T (left). Training the BDT specifically on electrons with $10 < p_T < 20\,\text{GeV}$ (center) does not improve performance, neither does it for electrons with $p_T > 20\,\text{GeV}$ (right).

4.3 Electron ID for the ZZjj analysis

The BDT training algorithm has several tunable parameters that can be adjusted to increase the performance. Table 4.2 shows these *hyper-parameters* and their values for the 2016 set of electron BDTs, which are the result of performing a grid search. One key consideration in choosing these hyper-parameters is their impact on the *overtraining* of the classifier. Overtraining refers to effect that an MVA can select discriminatory features in the training data set that are not really present in the distribution from which these data were sampled but the result of the finite statistics and the sparsity arising from the high dimensionality of the input vector. The MVA could for example select a small volume around one signal electron in the 20-dimensional space spanned by the input observables. While such a selection can be void of any background electrons in the training set, it is highly unlikely that it will generalize to the true distribution or actual data. In order to detect overtraining and to get a realistic estimate of the classifier's performance, one employs a set of signal and background samples that have not been used during the training, the *test set*. The test set will allow an unbiased estimate of the performance and the gap to the performance on the training set is the overtraining.

Table 4.2: Hyper-parameters used in the 2016 electron BDT. Parameter names are those used in TMVA.

Parameter	default value	optimal value
NTrees	800	2000
Shrinkage	1	0.1
MaxDepth	3	6
PruneStrength	0	5

The left panel of Fig. 4.16 shows the output of the BDT on the training and testing samples for true and fake electrons for the high-p_T training bin in the endcap. Comparing the BDT scores is a common way to detect overtraining, but cumulative effects

are hard to asses and it fails to quantify the effect. This is achieved by comparing the ROC curves of the training and testing set as shown in the right panel of Fig. 4.16. The background efficiency in the training set is about 20 % lower than in the test set, i.e., the BDT is slightly overtrained. Overtraining is an indication that the BDT is not optimal in the sense that it is fitting statistical fluctuations of the training sample instead of actual features - an ideal classifier would exhibit no overtraining. In practice, some overtraining is acceptable and might also be due to non-tunable parameters in the training algorithm or the algorithm itself. The grid search performed to identify the optimal hyper-parameters listed in Table 4.2 indicated that a reduction of overtraining was also accompanied by a reduction of performance on the test set.

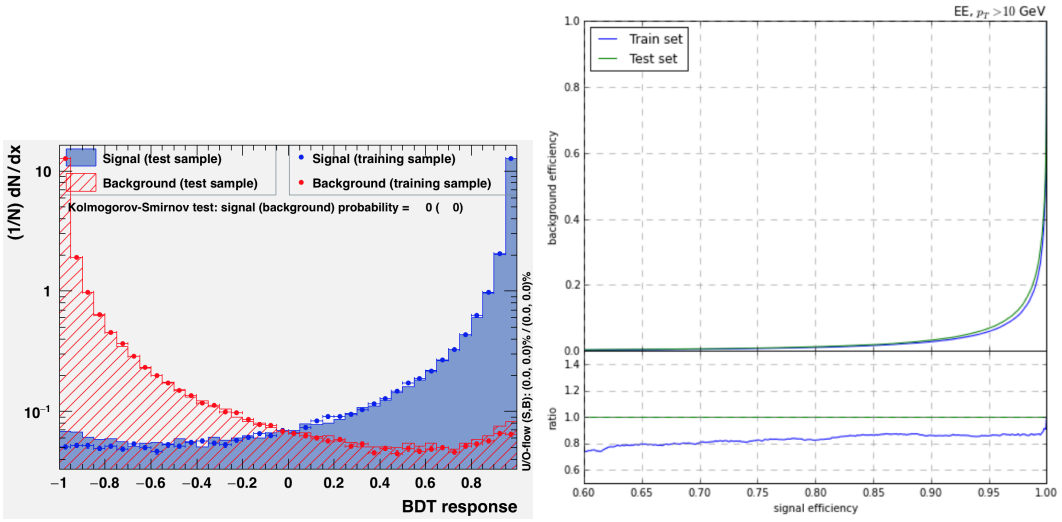


Figure 4.16: BDT output for the training and testing sample for true and fake electrons (left) and associated ROC curves (right) for the high-p_T endcap training bins.

Aside from performance considerations, care has to be taken not to introduce an overtraining bias into a physics analysis. If one where to use the same Drell-Yan plus jets simulation used to populate the BDT training sample to obtain the $Z \to ee$ selection efficiency, one would obtain a biased estimate. Practically, this is not a concern because only a small fraction of the full sample is actually used in the training, that is the overtraining effect only occurs for a small subset of all simulated events. In the final analysis one uses the data-driven electron efficiency corrections described in the Section 3.2.4, removing the bias completely.

The final step in using the BDT to identify electrons is to define a *working point*, i.e., to determine a cut value on the BDT score. Table 4.3 lists these minimal BDT score values for the six training categories. They were determined by considering the total event selection efficiency for $H \to ZZ^{(*)} \to 4e$ decays and requiring that the overall event selection efficiency be comparable to that achieved at 8 TeV. Figure 4.17 shows the ROC curves in the six training categories and the working points. Electrons with BDT scores above the respective cut value are considered to pass the MVA ID.

Table 4.3: Minimum BDT score required for passing the electron identification.

minimum BDT score	$\|\eta\| < 0.8$	$0.8 < \|\eta\| < 1.479$	$\|\eta\| > 1.479$
$5 < p_{\mathrm{T}} < 10\,\mathrm{GeV}$	-0.211	-0.396	-0.215
$p_{\mathrm{T}} > 10\,\mathrm{GeV}$	-0.870	-0.838	-0.763

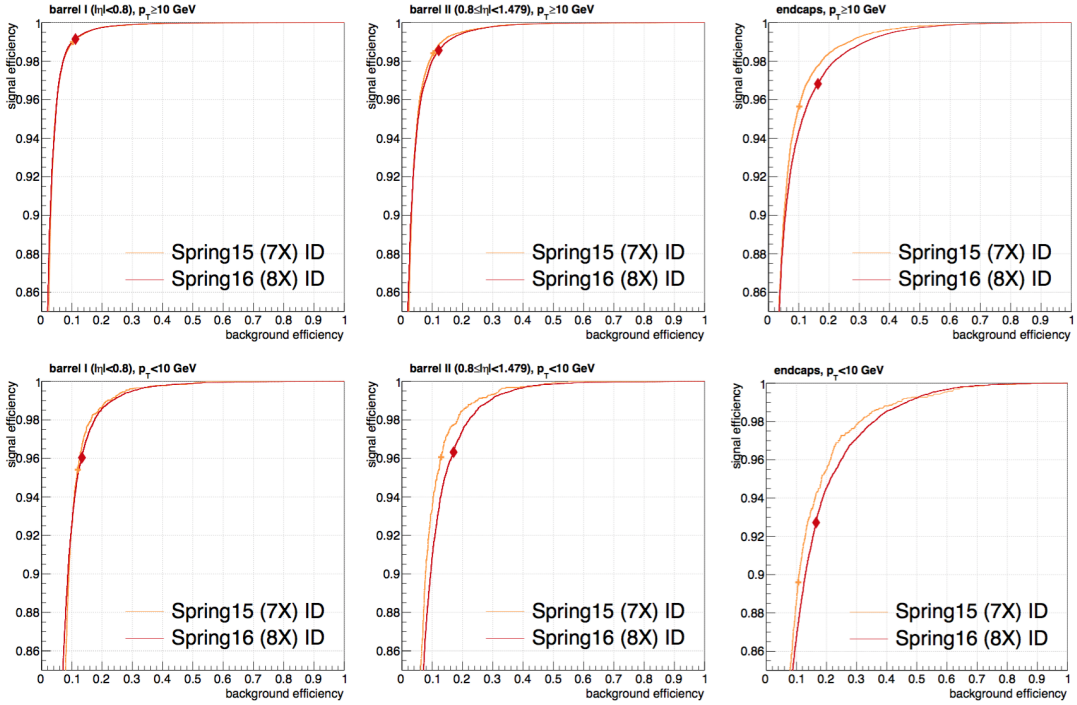


Figure 4.17: Performance comparison of the MVA trained for the 2015 analysis and the retraining for 2016 conditions. The respective working points are indicated by the markers.

4.4 Trigger preselection and the 2015 general purpose MVA ID

In most CMS analyses, electrons that are used in the offline analysis are required to have fired a trigger. The information on whether any given electron has fired a trigger is not readily available in the offline data format and the simulation does not always include the trigger menu deployed in data. Instead, the selection imposed by the trigger is re-applied during the offline analysis, to guarantee that the electrons used in the event selection are indeed the electrons that fired the trigger. This is not straightforward, as the online quantities used in the HLT are not available once the event has been reconstructed and instead one has to use their offline counterparts. The simplifications made in the HLT reconstruction cause the online and offline variables to differ and imposing the same online cut values on the offline variables will not necessarily result in the same electrons being selected. The goal is thus to find an offline trigger selection that is sufficiently close to the online selection. This selection will be then applied before the electron identification and is therefore called the *trigger preselection*.

In order to carry out the study of the correlations between online and offline variables, the HLT had to be run on-top of the simulation samples used to train the MVA ID. The HLT object information includes the four vector and enables a geometric matching

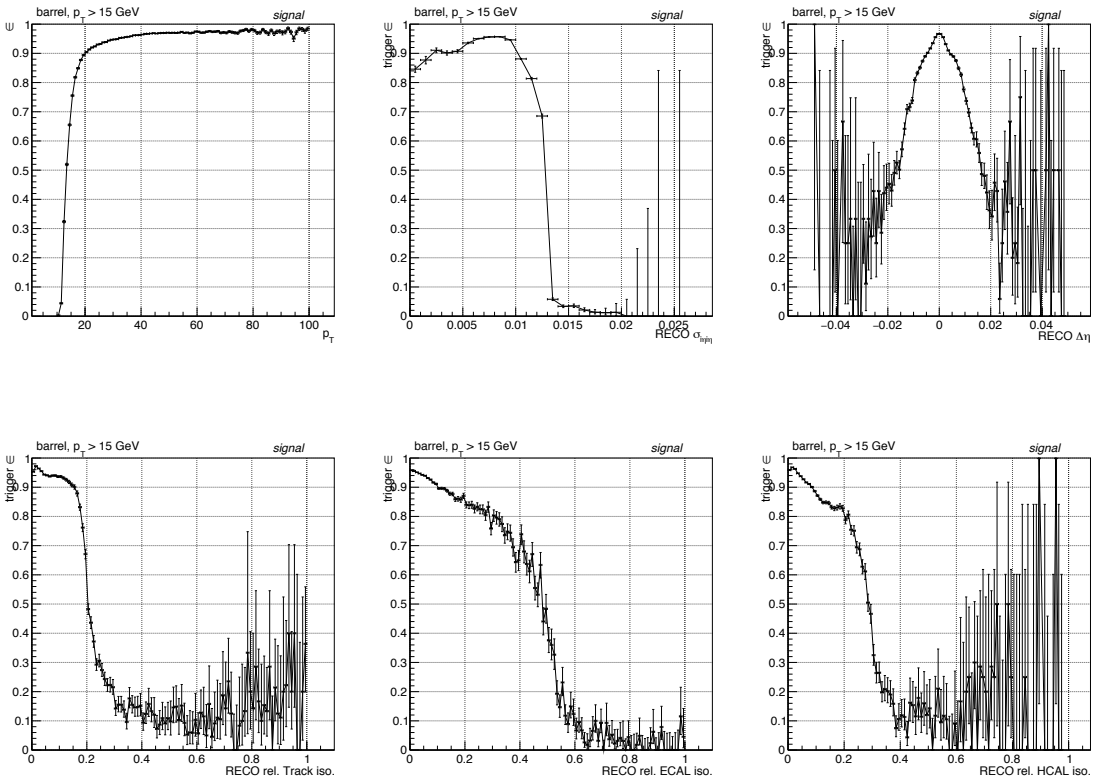


Figure 4.18: Efficiencies for true electrons in the $Z \to ee$ simulation to pass the dielectron trigger as a function of the reconstructed p_{T} (top left), the offline $\sigma_{i\eta i\eta}$ (top center), the track-cluster distance in η (top right), the relative track (bottom left), ECAL (bottom center), and HCAL (bottom right) isolations.

with the reconstructed electron allows to study the turn-on of a trigger filter as a function of the offline variable. The dielectron HLT trigger path used in this study, which requires that the electron candidates be isolated and that its electromagnetic cluster be signal-like. Figure 4.18 shows the turn-on curves for signal electrons for some of the variables used in the HLT menu for the 25 ns data in 2015. Based on these plots the triggering preselection cut values are determined as the offline value for which about 80 % of electrons from Z boson decays pass the trigger. The complete list of triggering preselection cut values is given in Table 4.4 and the overall selection efficiencies and impurities are summarized in Table 4.5. A similar study was carried out for the HLT menu used during the 50 ns data-taking period in 2015.

The application of the trigger preselection will result in a sample of electrons that is more signal-like, illustrated by the $\mathcal{O}(50\%)$ background efficiencies in Table 4.5. In order to provide an optimal separation power on this biased sample of true and fake electrons, the electron BDT is retrained on reconstructed electron candidates that pass the trigger preselection. This electron MVA ID is referred to as the *general purpose ID* as it covers the majority of use cases. Figure 4.19 shows the ROC curves of the general purpose ID, once trained using the run I set of observables ('classic') and once trained including the conversion rejection observables outlined in the previous chapter. Appreciable reductions in the background rates of around 50 % are achieved.

Table 4.4: Cut values of the 2015 triggering preselection for the 25 ns HLT menu. All values refer to the maximum value in order to pass the triggering selection, except for the minimum p_T requirement.

observable	barrel	endcap
p_T [GeV]	15	15
$\sigma_{i\eta i\eta}$	0.012	0.033
H/E	0.09	0.09
ISO_{ECAL}/p_T	0.37	0.45
ISO_{ECAL}/p_T	0.25	0.28
ISO_{track}/p_T	0.18	0.18
$\lvert\Delta\eta\rvert$	0.0095	no cut
$\lvert\Delta\varphi\rvert$	0.065	no cut

Table 4.5: Efficiencies of the triggering preselection for the 25 ns HLT menu used in 2015, evaluated from Drell-Yan simulation. The impurities, that is the fraction of electrons that did not pass the dielectron trigger but pass the trigger preselection, are also given. In each entry, the first number refers to electrons in the barrel while the number in parentheses refers to electrons in the endcap.

	Signal	Background
Efficiency	96% (97%)	54% (64%)
Impurity	4.8% (7.8%)	36% (59%)

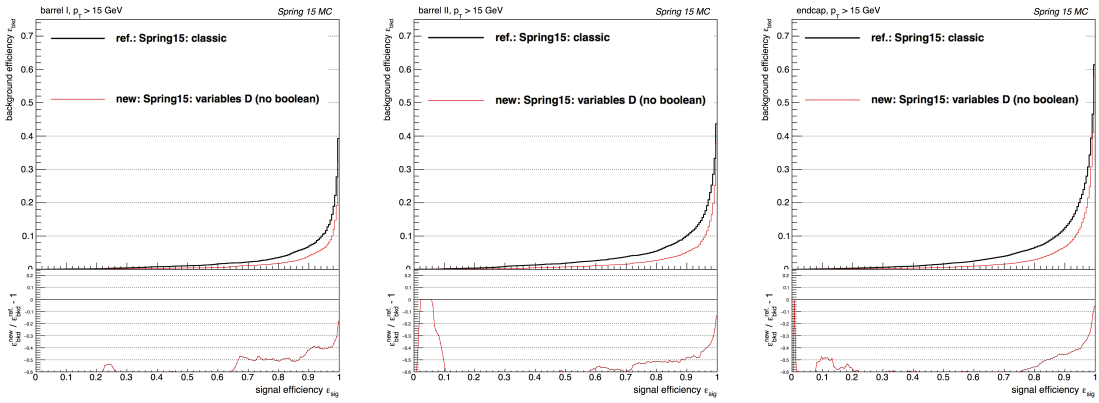


Figure 4.19: ROC curves of the 2015 general purpose ID (red curves) and a training of the BDT that does not include conversion rejection variables (black curves). Significant improvements from the inclusion of photon conversion rejection variables are obtained.

4.5 The 2016 general purpose ID and selection uniformity study

The general purpose MVA ID was retrained for the 2016 conditions, just like the MVA ID used in the multilepton analyses. One objective of the retraining was to reduce the kinematic dependence of the selection efficiency. A straightforward retraining of the

BDT on the simulation with the 2016 conditions proved to significantly worsen the efficiency turn-on of the ID, as can be seen by comparing the red and green curves in Fig. 4.20. The working points of the three BDTs correspond to a signal efficiency of 90 %. The signal and background selections are unchanged with respect to the previous BDTs: the signal is the standard selection of true electrons from Z boson decays and the background is from hadronic jets in the Drell-Yan simulation.

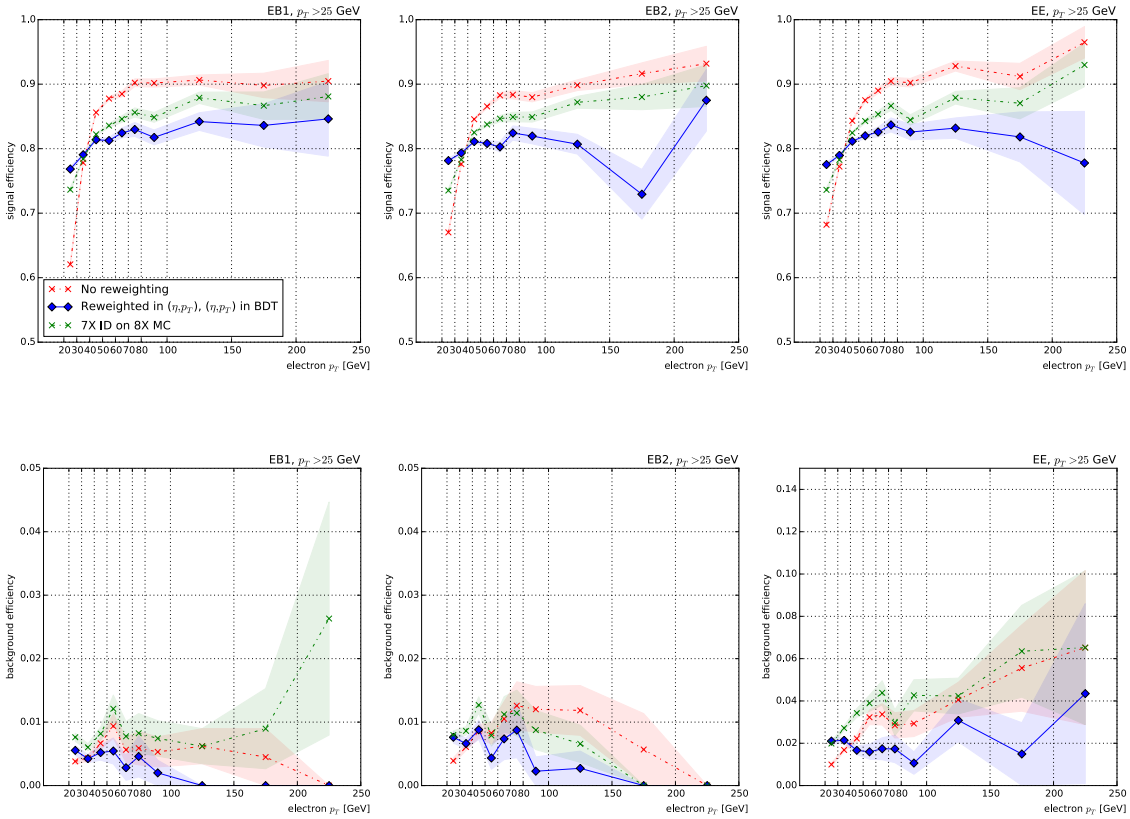


Figure 4.20: Efficiency of the electron BDT as a function of the electron p_T for a fixed working point corresponding to an integrated signal efficiency of 90 %. The top row shows the signal efficiency, the corresponding background efficiencies are shown in the bottom row. The red curve shows the BDT trained on the simulation of the 2016 conditions without any reweighing of the sample kinematics. The blue curve shows a BDT trained on a signal sample that was reweighted to reproduce the background distribution in p_T and η and in addition includes both observables in the training. Finally, the red curve shows the BDT trained on the 2015 conditions. All BDTs are evaluated on unreweighted samples of the 2016 conditions.

A common approach to reduce such a kinematic dependence is to reweight the training samples in that observable and to include the observable as an input in the training. A first attempt performing a reweighing such that both the signal and background sample feature a uniform p_T and η dependence resulted in a drastically reduced separation power. The p_T spectrum of the fake electrons is decreasing exponentially, following the spectrum of the hadronic jets. Reweighting to a uniform p_T spectrum thus reduces the statistical weight of the low-p_T background and increases the relative importance of the high-p_T tails. It is hypothesized that the loss in performance for the BDT trained on these reweighted samples is caused by the poor effective statistics for fake electrons with $p_T \gtrsim 30\,\text{GeV}$.

To avoid such reduction of the statistical power of the training samples, the reweighting was instead only performed on the signal sample, which features much higher statistics than the background sample. The signal is reweighted to reproduce the (p_T, η) distribution of the background and the BDT is trained with both p_T and η added as an input. The performance of this BDT is shown as the blue curve in Fig. 4.20. No reweighting of the test sample kinematics is performed, i.e., the BDT is applied on the p_T and η distributions of the Drell-Yan sample. A clear reduction of the kinematic dependence is observed and the reweighted BDT is adopted for the general purpose electron MVA ID for the 2016 data.

An alternative approach to achieving a reduced kinematics dependence is to replace the single working point with a series of working points, each adapted to produce a desired efficiency. This allows to obtain arbitrary turn-on curves on either the signal or the background, the latter case is illustrated in Fig. 4.21 for a p_T-dependent working point that results in a uniform 1 % background efficiency. The three curves in Fig. 4.21 correspond to the three BDT trainings discussed earlier. With these p_T-dependent working points, it can be seen that the reweighted (blue curve) and the unreweighted (red curve) provide almost identical separation power for a electron in a given p_T range. This means that the principal effect of the reweighting is to stabilize the BDT score under changes of the kinematics. The reweighing effectively provides a p_T-dependent transformation the BDT score which leaves the ROC curves invariant for a sufficiently narrow p_T range of the test samples.

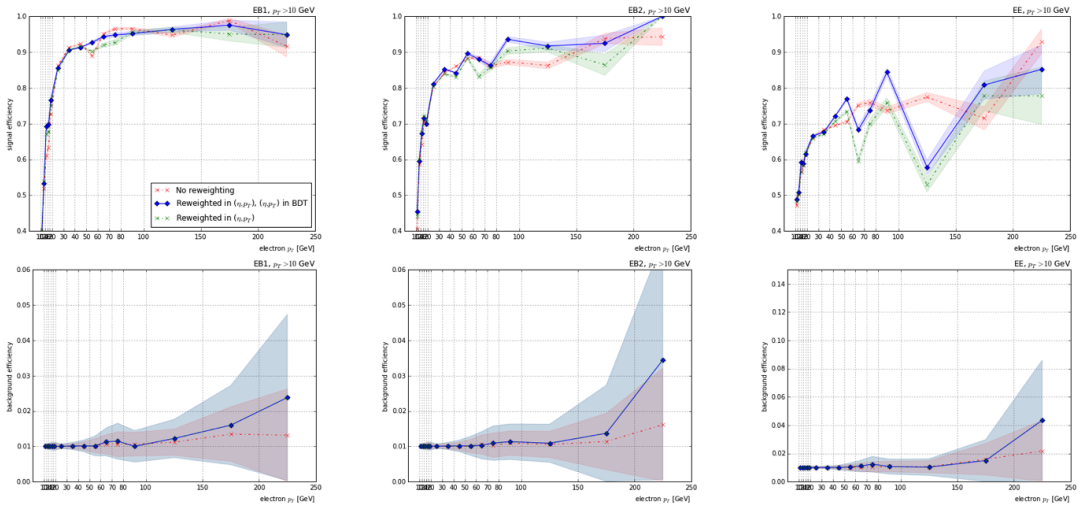


Figure 4.21: Efficiency of the electron BDT as a function of the electron p_T for a dynamic working point corresponding to a constant background efficiency of 1 %. The top row shows the signal efficiency, the corresponding background efficiencies are shown in the bottom row. The red curve shows the BDT trained on the simulation of the 2016 conditions without any reweighing of the sample kinematics. The blue curve shows a BDT trained on a signal sample that was reweighted to reproduce the background distribution in p_T and η and in addition includes both observables in the training. Finally, the red curve shows the BDT trained on the 2015 conditions. All BDTs are evaluated on unreweighted samples of the 2016 conditions.

The choice to shape the background efficiencies in the above example was mostly intended to illustrate the method, but choosing a uniform background efficiency over a uniform signal efficiency could be beneficial. The copious $Z \to ee$ decays allow to measure the electron selection efficiency in data with high accuracy and as a function of the electron kinematics. The determination of the fake rates on the other hand has a much lower statistics and reducing their potentially poorly-measured kinematic

dependence could reduce systematic uncertainties on the background estimate.

The above proposal to achieve a desired kinematic dependence of a classifier is admittedly trivial and more advanced approaches have been proposed. The approach presented in [66] introduces an additional loss term to the boosting algorithm and thus reduces the relative weight of decision trees that result in large kinematic dependencies. The method presented in [67] uses adversarial neural networks, a setup where the classifier network incurs a loss if the observer network is able to infer the value of an external observable based solely on the output of the classifier.

Chapter 5

Signal and background modeling and kinematics

Having a precise understanding of the signal and background yields and kinematics are mandatory in the search for the VBS process. This chapter describes the technical work carried out to produce these predictions, many of which did not exist prior to this work. The reliability of the new simulation samples is assessed by comparing the predictions delivered by different Monte Carlo event generators. The simulated samples are then used to establish the final state object kinematics for the signal and irreducible QCD background. Care is taken to understand the aspects that impact the event selection and the acceptance.

5.1 Monte Carlo simulations

5.1.1 Signal simulation and phase-space optimizations

The signal for this analysis is the purely electroweak production of two jets and two leptonically decaying Z bosons. At leading order this process involves six electroweak vertices, i.e., the signal is of order α^6 in the perturbative expansion. Figure 5.1 illustrates some of the Feynman diagrams that lead to such a final state.

The hard process of the signal is simulated with the MADGRAPH5_AMC@NLO [68] Monte Carlo event generator (henceforth abbreviated as MG5_AMC). The prediction obtained by explicitly reducing the number of allowed QCD vertices to zero[1]:

```
generate p p > z z j j QCD=0, z > l+ l-
```

This leading-order (LO) generation includes diagrams featuring the standard model Higgs boson ($m_{\mathrm{H}} = 126\,\mathrm{GeV}$) produced in vector boson fusion as well as the interference with the non-Higgs diagrams. Only diagrams that feature on-shell Z bosons are included in the signal, i.e., nonresonant diagrams are excluded. The use of this *decay chain syntax* reduces the $2 \to 6$ scattering to a $2 \to 4$ process, drastically reducing the

[1] The MG5_AMC package uses an internal weighting scheme to decide which diagrams to generate. A vertex of the strong interaction is assigned a weight of 1 while an electroweak vertex is assigned weight 2. The code then determines the minimal sum of weights needed for a given process and only generates the Feynman diagrams corresponding to the lowest weight of the couplings, effectively assuming that diagrams with higher weight are negligible.

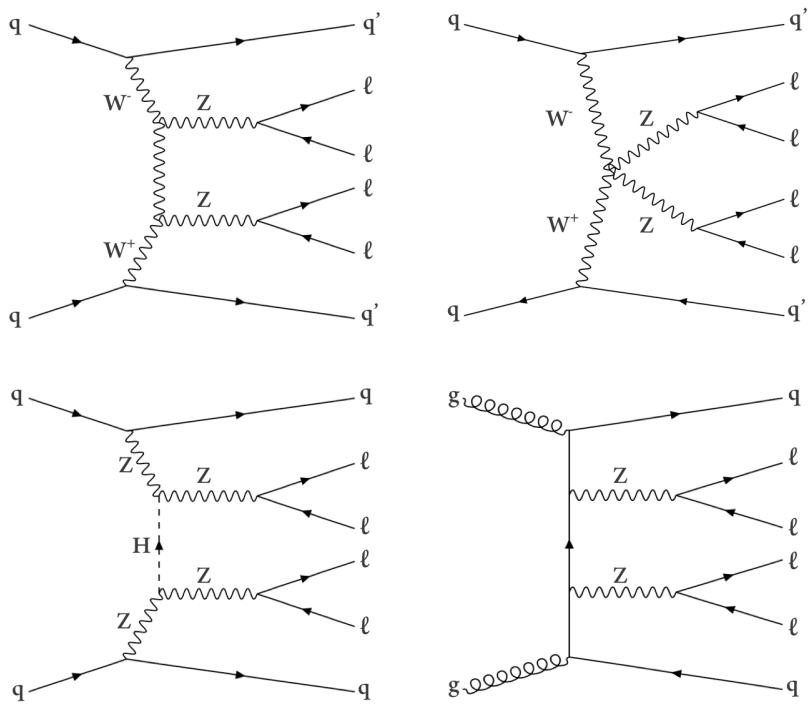


Figure 5.1: Representative Feynman diagrams for the electroweak- (top row and bottom left) and QCD-induced production (bottom right) of the $ZZjj \to \ell\ell\ell'\ell' jj$ ($\ell, \ell' = e$ or μ) final state. The scattering of massive gauge bosons as depicted in the top row is unitarized by the interference with diagrams that feature the Higgs boson (bottom left).

number of diagrams by a factor 100 and resulting in a considerable speed-up for the phase-space integration[2].

The $ZZjj$ analysis requires both Z bosons to be on-shell which suppresses the contribution of non-resonant amplitudes to below one percent and thus negligible compared to the uncertainty on the theory prediction. For the bulk of nonresonant events at least one of the four final state leptons is either too soft or is outside the detector acceptance. Generating these non-resonant contributions that are irrelevant to the analysis, would require generating about an order of magnitude more events for the same statistical precision in the analysis.

Finally, the decay chain syntax includes the spin correlations of the final state fermions, i.e., the decay angles accurately modeled.

An additional optimization of the event generation phase space targets the triboson production with one hadronic W/Z decay. The large branching ratio of vector bosons to quarks causes these processes to contribute about $2/3$ of the number of generated events. The phase-space corresponding to $m_{jj} < 100\,\text{GeV}$ is thus also excluded in the event generation because the electroweak signal for the analysis is concentrated at dijet masses of several hundred GeV.

A final restriction of the sample generation consists in restricting the Z bosons decay to electrons and muons, but not taus. Without such a restriction, $5/9$ of the generated events would feature at least one $Z \to \tau\tau$ decay. In addition to the suppression by the

[2]The event generation of the pure-electroweak signal remains CPU intensive despite these optimizations. Another issue is a poor scaling of the CPU time per event with the number of events to be generated. Central event production was done with only 200 events per job.

35 % branching ratio of leptonic tau decays, these events are highly unlikely to pass the on-shell ZZ selection. The expected contribution of these signal events with Z decays into taus in the on-shell ZZ selection is 0.6 % and these decay modes are thus not included in the simulation.

Without the above optimizations with respect to the default phase space for $ZZ \to 4\ell$ simulation samples it would have been challenging to pursue a multivariate analysis. The training, cross validation, and testing of the BDT requires $\mathcal{O}(100k)$ reconstructed and selected MC events. Finally, one would like to keep the statistical uncertainty of the signal template at a negligible level, particularly in the low-yield but high-purity signal enriched regions.

A second signal sample is generated with the leading-order generator `Phantom` [69]. The PHANTOM prediction includes all diagrams of order α^6 without any on-shell requirements or phase-space restrictions, which allows to perform independent cross check of the phase-space optimizations made in the nominal MG5_AMC sample. The `Phantom` generation is done in the four flavor scheme, i.e., it features massive b quarks which results in a contribution from $H \to b\bar{b}$ decays.

5.1.2 Simulation of the QCD irreducible background

The dominant background in this analysis is the QCD-induced production of the $ZZjj$ final state. At leading order this process features two QCD vertices, making it an $\alpha^4\alpha_s^2$ ($\alpha^2\alpha_s^2$ when excluding the Z decays) process. In addition to the standard $ZZ \to 4\ell$ simulation samples used in CMS multilepton analyses, two new samples are produced specifically for the $ZZjj$ analysis.

A first background sample is a leading-order generation in MG5_AMC, which is obtained via:

```
generate p p > z z j j QCD=2 QED=2, z > l+ l-
```

This LO sample is used to study the background kinematics, its selection efficiency and to train the signal extraction BDT. It is not used in the statistical analysis of the VBS search, which uses a more precise next-to-leading order prediction.

The interference between the signal and the QCD background is evaluated by generating a dedicated sample that includes the electroweak, QCD as well as the interference contributions:

```
generate p p > z z j j QCD=2 QED=4, z > l+ l-
```

The size and kinematic behavior of the interference contribution can then be estimated by the difference between this sample and the sum of the pure signal and background samples. Table 5.1 lists the cross sections for the electroweak signal, the QCD background, and the interference obtained by taking the difference. The interference is positive and amounts to about 0.04 fb or 10 % of the electroweak signal. However, the kinematics are background-like with only a negligible fraction of events at large m_{jj} or $|\Delta\eta_{jj}|$. Both distributions are illustrated in Fig. 5.2, in addition to the spectrum of the BDT used to extract the electroweak signal. The BDT score distributions are shown here to illustrate the negligible impact of the interference in the analysis. The detailed introduction to signal extraction using the BDT is deferred to Section 7.1. The contribution of the interference coincides with the QCD background, and its small size

make it negligible compared to the uncertainties on the total yield predicted in this background-rich region. The interference is thus neglected in the statistical analysis.

Table 5.1: Cross sections of the electroweak (EW) and QCD-induced production of the $4\ell jj$ final state and the interference. Cross sections in fb. The phase space is that of the generation, i.e., $m_{jj} > 100\,\text{GeV}$ and includes the branching ratios for the Z decays to electrons or muons.

σ_{QCD}	σ_{EW}	$\sigma_{\text{sum}} = \sigma_{QCD} + \sigma_{\text{EW}}$	σ_{full}	$\sigma_{\text{full}} - \sigma_{\text{sum}}$
9.335	0.4404	9.7754	9.818	0.0426

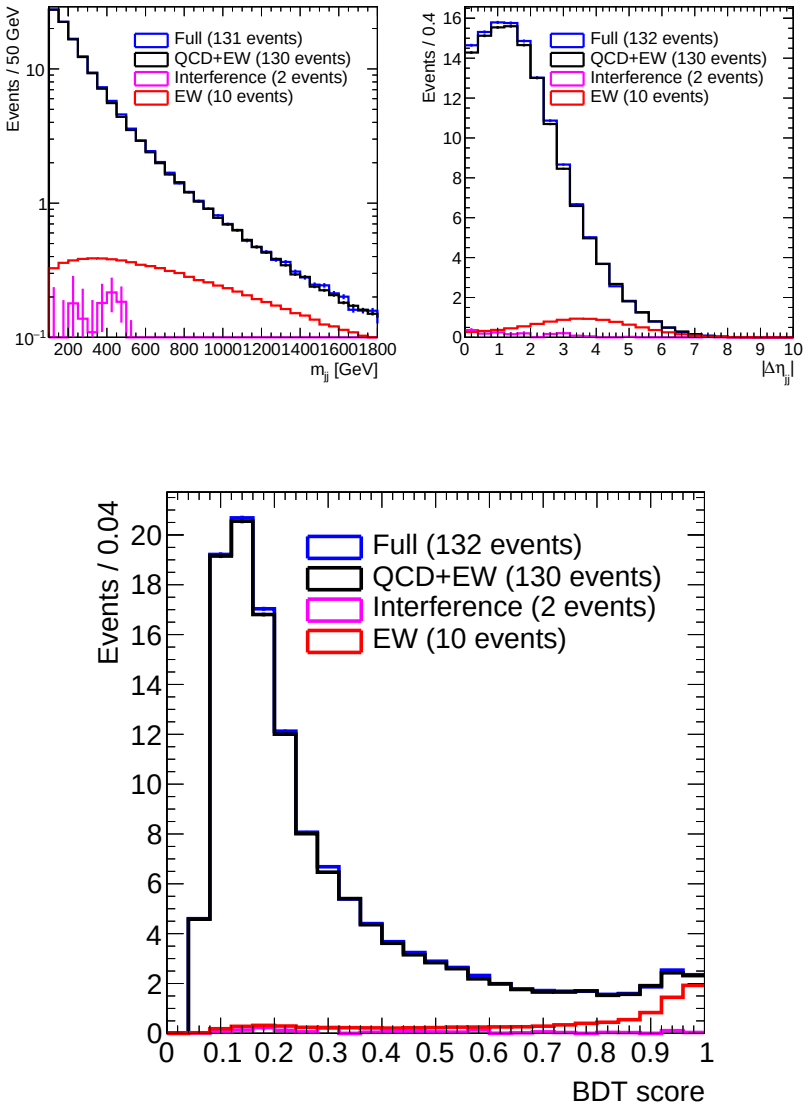


Figure 5.2: Dijet invariant mass (top left), $|\Delta\eta_{jj}|$ separation (top right) distributions and the BDT score distribution (bottom) for the electroweak signal, the QCD background and the interference between the two. For details on the BDT, see Section 7.1.

The second simulation of the QCD background is an next-to-leading order (NLO) prediction also obtained with MG5_aMC. This is the nominal sample for the statistical

analysis, providing NLO accuracy for all observables exploited in this analysis and reducing the scale uncertainties compared to the LO prediction. Three jet multiplicities with zero, one, and two final state partons are merged using the FxFx scheme [70]. The merging scale is set to $q_{\text{cut}} = 30\,\text{GeV}$ and the minimum jet p_{T} cut is $p_{\text{T}}^{\text{jet}} > 15\,\text{GeV}$. The differential jet rate distributions are found to be smooth around the merging scale, validating this choice of merging parameters. The central CMS production system for simulation samples uses so-called *grid-packs* to generate events on the distributed computing infrastructure. These grid-packs include all the code and Monte Carlo integration grids to efficiently generate events. Three such grid-packs are generated for this sample, one corresponding to each jet multiplicity:

```
0-jet grid-pack: generate p p > z z   [QCD]
1-jet grid-pack: generate p p > z z j  [QCD]
2-jet grid-pack: generate p p > z z j j  [QCD]
```

This setup allows to independently generate events for each jet multiplicity and to increase the number of events in the phase-space regions relevant to the $ZZjj$ study. The Z bosons are generated on-shell and their decay to electrons and muons is performed using MADSPIN [71]. MADSPIN describes the natural width of the Z bosons and conserves the spin-correlations between the leptons, similarly to the decay chain syntax at leading-order.

In order for the NLO sample to be useful in the statistical analysis, it has to have a sufficiently low statistical uncertainty. The NLO events generated by MG5_AMC are associated with a sign, which reduces the statistical power. The statistical uncertainty of N events with a fraction of negative weights[3] f is equivalent to $N(1-2f)$ unweighted events. The fraction of negative weights for the QCD samples are around $f \approx 34\,\%$, i.e. the statistical power of the NLO samples are about a third of a LO sample with the same number of generated events. The efficiency of the merging also needs to be considered: in order to avoid double counting due the phase space overlap between the parton shower and matrix element, the merging procedure in the parton shower will reject events. The merging efficiencies for the three multiplicities vary between 40 % and 60 %. Finally, the composition of the events in the $ZZjj$ phase space differs significantly between the three samples, e.g., only 1 % of selected $ZZjj$ events originate from the zero jet sample.

A reliable estimate of the number of events to be requested from the CMS central simulation production had to be established and was obtained from a private production of the above samples. These pre-production samples allowed to perform this estimate and also served to validate the overall setup. With the per-sample efficiencies known, a precise number of events that needed to be generated and reconstructed[4] in the simulation could be given to the CMS central simulation production.

The phase-space optimizations presented for the signal simulation are crucial to the feasibility of the NLO background sample. Without the restriction to on-shell Z bosons and leptonic decays, the number of events to be generated would be $\mathcal{O}(200M)$, i.e., prohibitive for such a specific process. Due to the appreciable uncertainty on the jet energy scale, no requirement on the tagging-jet invariant mass is made for this background sample, since doing so could lead to an underestimate of the yield in the $ZZjj$ selection.

[3] This assumes that the absolute value of the weight is equal for every event.

[4] The bulk of the CPU time per event is spend on the Geant4 simulation and the event reconstruction, rendering the distinction between number of generated and reconstructed events critical.

5.1.3 Simulation of the loop-induced background

Aside from the dominant QCD background mediated by tree-level processes, there is also a gluon loop-induced production process referred to as $ggZZ$[5]. Despite the suppression due to the two additional QCD vertices, it nevertheless contributes to inclusive ZZ production at the 10 % level. This process is generated at LO with MCFM [72] including all off-shell effects and then processed with PYTHIA. The $ZZjj$ phase-space probed by this analysis is covered by the prediction provided by this MCFM +PYTHIA sample, but the jets are simulated by the parton shower and are not part of the hard scattering. To date no event generator is capable of simulating loop-induced $2 \to 6$ scattering processes like $ZZ \to 4\ell$ production with two outgoing partons. The process is part of the next-to-next-to-next-to-next-to-leading (N^4) order correction to ZZ production.

The parton shower configuration for the MCFM samples used in this work differ from the samples used in other multilepton analyses, notably the $H \to ZZ^* \to 4\ell$ analysis. The latter are optimized to provide a better description for the inclusive spectrum of the $H \to ZZ^{(*)}$ transverse momentum by using the *wimpy* shower option in PYTHIA:

```
SpaceShower:pTmaxMatch = 1
```

This instructs the initial-state shower to limit the p_T of the parton shower evolution to the factorization scale of the hard process. The factorization scale in these MCFM samples is set to $\mu_F = m_{4\ell}/2$ and the p_T of the leading parton emission cannot exceed this value. As a consequence, the p_T spectra of the jets are distorted and the jet multiplicity is underestimated by incorrectly rejecting emissions that should have occurred, as will be illustrated in Section 5.2.3. The parton shower simulation used in the $ZZjj$ analysis resort to the default PYTHIA configuration, which allows the parton emissions to populate the entire phase space up to the kinematical limit.

The MCFM + PYTHIA predictions are cross-checked using the MG5_AMC package, which allows to simulate loop-induced processes.

A particularity of the $ZZjj$ process is the existence of both tree-level and quark loop diagrams. In fact the computation of the NLO correction to the tree-level $pp \to ZZjj$ process already features the loop-induced diagrams, which are then squared against the tree-level diagrams. The resulting squared amplitude then features two powers of the strong coupling constant, and these corrections are of the same perturbative order as other NLO diagrams. Imposing that neither of the two amplitudes that are to be squared is at tree-level is also insufficient. This corresponds to the noBorn option in MG5_AMC, and results in the interference of genuine loop-induced Feynman diagrams with NLO-type diagrams. The latter exhibit the usual infrared and ultraviolet divergencies of such diagrams, which would be cancelled by the divergencies of the real emission amplitudes in a full NLO calculation. However, the real emission amplitudes are absent in this leading-order calculation, rendering these combination of diagrams divergent. The NLO-type diagrams are absent if the incoming and outgoing partons are restricted to be gluons, but doing so neglects contributions of valid loop-induced diagrams that yield final state quarks.

The solution adopted in this work and implemented in MG5_AMC is to remove the NLO-type diagrams prior to the phase space integration. This is achieved by rejecting all diagrams where the loop includes non-fermion propagators and diagrams where

[5]This nomenclature obscures the fact that there is a leading-order diagram of $gg \to ZZq\bar{q}$

none of the Z bosons are attached to fermion propagators of the loop. Only a small sample of $\mathcal{O}(10k)$ ggZZ events was generated with this setup because of the long CPU time needed[6]. The leptonic decay of the on-shell Z bosons is done with MADSPIN, however no smearing of the mass is done for loop-induced processes.

5.1.4 Simulation of anomalous gauge couplings

The anomalous quartic gauge coupling (aQGC) operators $\mathcal{O}_{T,0-2}$ and $\mathcal{O}_{T,8,9}$ have been implemented in model files [18] which can be used by MG5_AMC. The process is then generated similarly to the leading-order simulation of the electroweak signal:

```
generate    p p > z z j j QED=5 QCD=0 NP=1
```

The Z boson decay into electrons and muons is handled by MADSPIN. The matrix-element reweighing functionality in MG5_AMC allows to only generate one sample of events for a given configuration of the aQGCs and to obtain the prediction of alternative couplings by modifying the event weights. The method uses event weights W_{new} to reweigh an event corresponding to a different hypotheses of the coupling strength:

$$W_{\text{new}} = W_{\text{old}} \frac{|\mathcal{M}_{\text{new}}|^2}{|\mathcal{M}_{\text{old}}|^2} \tag{5.1}$$

where $\mathcal{M}_{\text{old}}$ is the nominal matrix element and $\mathcal{M}_{\text{new}}$ is the matrix element with the modified coupling strength.

This reduces the number of events that need to be generated and simulated. The default aQGC coupling strength for the event generation is set to a non-zero value in order to bias the number of generated events at large scattering energies and to have sufficient statistics in the high-energy tails. Weights are generated along 2D grids in f_{T8}/Λ^4 and f_{T9}/Λ^4 and along 3D grids in f_{T0}/Λ^4, f_{T1}/Λ^4, and f_{T2}/Λ^4. The coupling strengths are set equal to [0.25, 0.5, 1, 2, 4, 8, 16] TeV^{-4}, including the mixed and negative coupling strength configurations.

5.1.5 Common settings and corrections to the simulation

All nominal simulation samples used in this analysis exploit the NNPDF 3.0 set of parton distribution functions (PDFs) [**?**]. The perturbative order of the PDF is set to match the order of the matrix element calculation.

The PYTHIA 8 [73] package is used for parton showering, hadronization, and the underlying event simulation for all samples, with parameters set by the CUETP8M1 tune [74]. It also handles the merging of the different jet multiplicities.

The CMS event simulation mixes additional interactions with the hard scattering event to model the pileup. The distribution of the number of such additional interactions has to be specified for the production of the samples, based on the expected pileup distribution in data. Once a dataset is recorded, the distribution of additional interactions can be estimated from the data and the simulation is corrected to match the data. Figure 5.3 shows the estimated number of true interactions per event in data and in the

[6] Event generation required about 10 min per event.

MC before and after this correction. The minimum bias cross section used is 69.2 mb and the uncertainties on the pileup profile are obtained by varying this cross section by its uncertainty of 4.6 % .

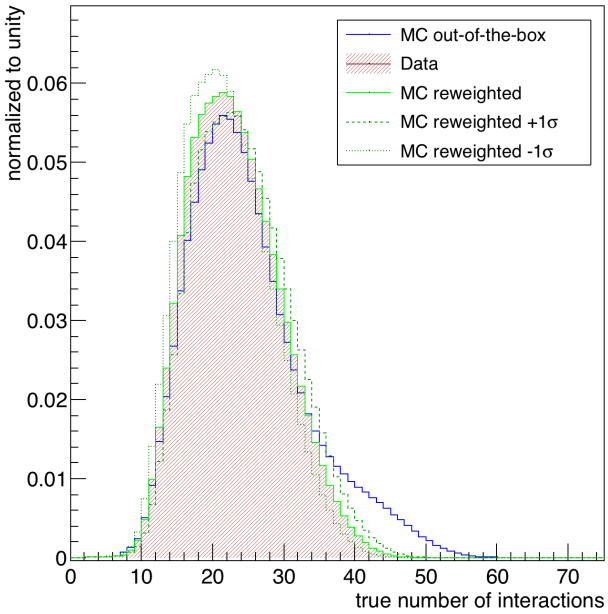


Figure 5.3: Number of true interactions per event extracted from the 2016 dataset and the reweighed simulation sample.

Finally, the simulation is corrected to match the lepton reconstruction and selection efficiencies measured in data. The per-lepton scale factors detailed in Section 3.2.4 and Section 3.3.3 are used to reweight each simulated event based on the flavors and kinematics of the four selected leptons:

$$\mathrm{SF}_{\mathrm{event}} = \prod_{i}^{\mathrm{electrons}} \mathrm{SF}(p_{\mathrm{T}}^{i}, \eta_{\mathrm{SC}}^{i}) \times \prod_{j}^{\mathrm{muons}} \mathrm{SF}(p_{\mathrm{T}}^{j}, \eta^{j}). \tag{5.2}$$

Table 5.2 summarizes all simulated samples used in the analysis along with the cross sections obtained from the respective generator.

Table 5.2: List of signal and background samples used in the analysis. All samples use the RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-* processing and are in the miniAOD format.

Process	Generator	Cross section [fb]	Sample name	Remarks
Signal samples				
$ZZ \to 4\ell$ + 2jets	MG5_AMC (LO)	0.440	ZZJJTo4L_EWK_13TeV-madgraph-pythia8/[1]	used in MVA optimization, $m_{jj} > 100$ GeV
$ZZ \to 4\mu$ + 2jets	Phantom (LO)	0.418	VBFToHiggs0PMContinToZZTo4muJJ_M125_GaSM_13TeV_phantom_pythia8	used to cross-check MadGraph sample
$ZZ \to 4e$ + 2jets	Phantom (LO)	0.418	VBFToHiggs0PMContinToZZTo4eJJ_M125_GaSM_13TeV_phantom_pythia8	used to cross-check MadGraph sample
$ZZ \to 2e2\mu$ + 2jets	Phantom (LO)	0.836	VBFToHiggs0PMContinToZZTo2e2muJJ_M125_GaSM_13TeV_phantom_pythia8	used to cross-check MadGraph sample
Irreducible background samples				
$ZZ \to 4\ell$ + 0jets	MG5_AMC (NLO)	42.1	ZZTo4L_0Jets_ZZOnShell_13TeV-amcatnloFXFX-madspin-pythia8	used for statistical analysis
$ZZ \to 4\ell$ + 1jets	MG5_AMC (NLO)	16.9	ZZTo4L_1Jets_ZZOnShell_13TeV-amcatnloFXFX-madspin-pythia8	used for statistical analysis
$ZZ \to 4\ell$ + 1jets	MG5_AMC (NLO)	8.76	ZZTo4L_2Jets_ZZOnShell_13TeV-amcatnloFXFX-madspin-pythia8	used for statistical analysis
$ZZ \to 4\ell$ + 0,1jets	MG5_AMC (NLO)	1256	ZZTo4L_13TeV-amcatnloFXFX-pythia8	used for cross checks
$ZZ \to 4\ell$ + 2jets	MG5_AMC (LO)	9.34	ZZJJTo4L_QCD_13TeV-madgraph-pythia8	used in MVA optimization, $m_{jj} > 100$ GeV
$ZZ \to 4\ell$	POWHEG (NLO)	1256	ZZTo4L_13TeV_powheg_pythia	
$gg \to ZZ \to 4\mu$	MCFM (LO)	1.59	GluGluToContinToZZTo4mu_DefaultShower_13TeV_MCFM701_pythia8	ggZZ sample with fixed shower
$gg \to ZZ \to 4e$	MCFM (LO)	1.59	GluGluToContinToZZTo4e_DefaultShower_13TeV_MCFM701_pythia8	ggZZ sample with fixed shower
$gg \to ZZ \to 2e2\mu$	MCFM (LO)	3.19	GluGluToContinToZZTo2e2mu_DefaultShower_13TeV_MCFM701_pythia8	ggZZ sample with fixed shower
Irreducible Background Samples				
$t\bar{t}Z \to 4\ell\, 2\nu$	MG5_AMC	662.4	ttZJets_13TeV_madgraphMLM	main sample with sufficient statistics
$t\bar{t}Z \to 4\ell\, 2\nu$	MG5_AMC	0.253	TTZToLLNuNu_M-10_TuneCUETP8M1_13TeV-amcatnlo-pythia8	used for cross checks
WWZ+ jets	MG5_AMC (NLO)	0.1651	WWZ_TuneCUETP8M1_13TeV-amcatnlo-pythia8	
Reducible Background Samples				
Z+jets	MG5_AMC (NLO)	6025.2	DYJetsToLL_M-50_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/	used for data/MC comparison in Z+X control regions
WZ	MG5_AMC (NLO)	5.26	WZJToLLLNu_TuneCUETP8M1_13TeV-amcnlo-pythia8	used data/MC comparison in 2P2F/3P1F control regions
$t\bar{t}$+jets	MG5_AMC (NLO)	831.76	TTJets_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	used data/MC comparison in 2P2F/3P1F control regions

5.2 Generator comparisons

The simulation samples described in Section 5.1 are generated specifically for the $ZZjj$ analysis and their validity needs to be established. To this end, a $ZZjj$ event selection is implemented at the generator level in order to study the kinematics of the leptons, Z bosons, tagging jets, and the correlation between leptons and jets.

Two selections of the $ZZjj$ topology are defined, differing mostly in the object-level requirements. In both selections jets are reconstructed using the anti-k_{T} algorithm with cone size of 0.4 and the invariant mass of the two leading jets, i.e., the tagging jets has to larger than 100 GeV to suppress the triboson contribution. In the rare case were several ZZ boson candidates per event are possible, the one with m_{Z_1} closest to the nominal Z mass is used.

- **Generator selection (GEN)** is chosen to be close to the generation phase-space: leptons must have $p_{\mathrm{T}} > 3\,\mathrm{GeV}$ and $|\eta| < 3.2$, jets must satisfy $p_{\mathrm{T}}^{\ell} > 10\,\mathrm{GeV}$ and $|\eta^{\ell}| < 5.2$. The Z mass window is $40 < m_Z < 140\,\mathrm{GeV}$.
- **Baseline selection (BLS)** is intended to approximate the detector acceptance: electrons must have $p_{\mathrm{T}}^{e} > 7\,\mathrm{GeV}$ and $|\eta^{e}| < 2.5$, while muons satisfy $p_{\mathrm{T}}^{\mu} > 5\,\mathrm{GeV}$ and $|\eta^{\mu}| < 2.4$. Jets must satisfy $p_{\mathrm{T}}^{\mathrm{jet}} > 25\,\mathrm{GeV}$ and $|\eta^{\mathrm{jet}}| < 4.7$. The Z mass window is $60 < m_Z < 120\,\mathrm{GeV}$. About 66 % (0.29 fb) of the generated signal events pass this selection; the background efficiency is 47 % (4.4 fb).

5.2.1 Comparison of the signal predictions

The prediction for the electroweak signal process obtained from the MG5_aMC and PHANTOM fixed-order generators are compared. The cross section prediction of the two generators are reported in Table 5.3. The observed difference of 5 % to 7 % is due to a combination of the different Z decay widths and the different renormalization and factorization scales used in each generator. MG5_aMC uses the leading-order result of the Z decay width ($\Gamma_Z = 2.4414\,\mathrm{GeV}$) due to internal consistency requirements, while PHANTOM uses the world best average published by the Particle Data Group ($\Gamma_Z = 2.4952\,\mathrm{GeV}$ [75]). It was confirmed that using the Particle Data Group value in MG5_aMC decreases the total cross section prediction by 4.8 %. The second driver of the differences in the normalizations arises from the choice of functional form for the dynamic renormalization (μ_R) and factorization (μ_F) scales, which are set equal to one another $\mu = \mu_R = \mu_F$ in each of the two samples. The left panel of Fig. 5.4 compares the shapes of these nominal scales μ. The right panel of Fig. 5.4 shows the impact of varying the nominal scale of the MG5_aMC prediction by a factor of 2 and 1/2. The differences in the scale choice between the two generators is well within the variations. Aside from the difference in the normalization, the predictions on key VBS observables between the two generators are in excellent agreement as illustrated in Fig. 5.5.

The impact of the choice of parton shower program and underlying event tune is assessed by processing the events obtained from the MG5_aMC matrix element simulation at the Les Houches file level with HERWIG [76]. Figure 5.6 shows the BDT score distributions of the nominal MG5_aMC + PYTHIA and the MG5_aMC + HERWIG simulation. A good agreement, in particular in the signal-enriched region, is found. The result of this comparison is taken as a validation of the PYTHIA model.

Table 5.3: Integrated cross section predictions for the electroweak MG5_AMC and PHANTOM samples. The $ZZjj$ and VBS selections are detailed in the following section.

	$ZZjj$ baseline	VBS selection
MG5_AMC [fb]	0.308	0.216
PHANTOM [fb]	0.290	0.202
Rel. difference [%]	5.9	7.2

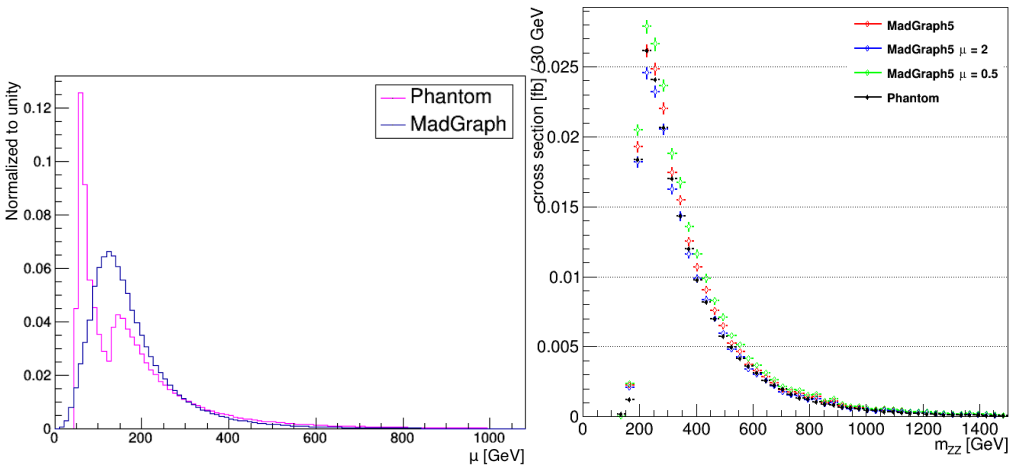


Figure 5.4: Nominal dynamic renormalization and factorization scales ($\mu = \mu_F = \mu_R$) of the electroweak MG5_AMC and PHANTOM samples (left) and scale variations for the MG5_AMC sample (right).

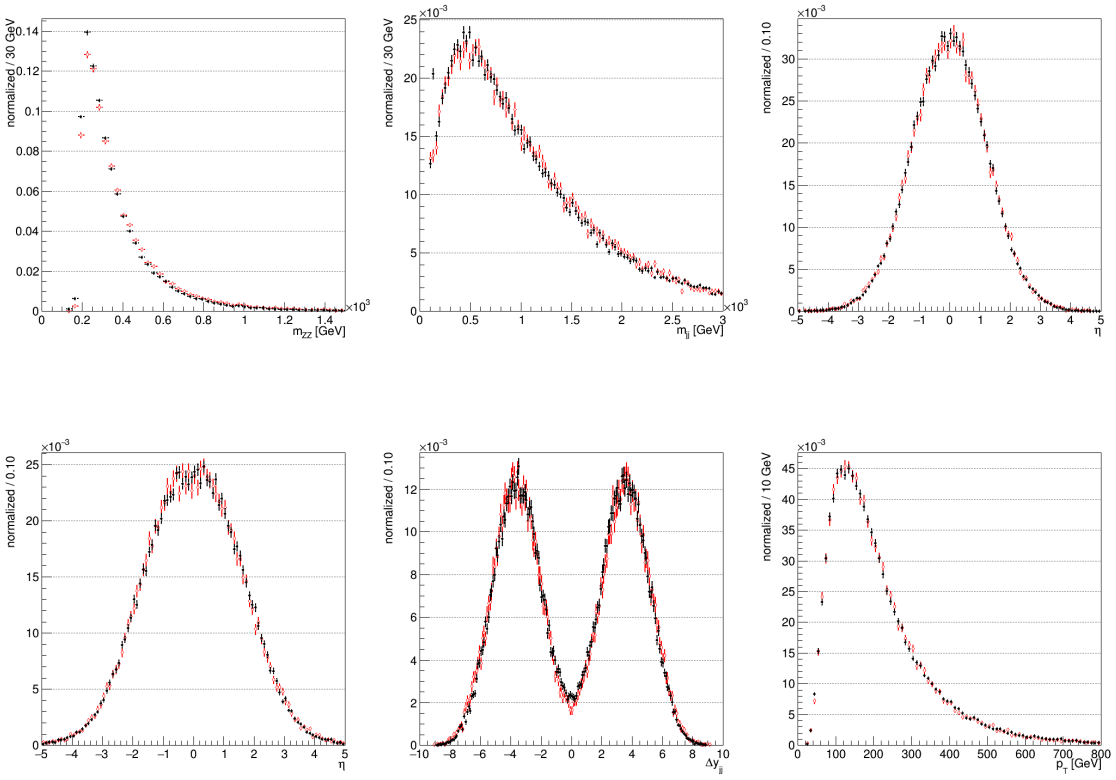


Figure 5.5: Comparison of the kinematics in the electroweak MG5_aMC (red) and PHANTOM (black) samples in the phase space defined by the $ZZjj$ baseline selection at the LHE level. All distributions are normalized to unity.

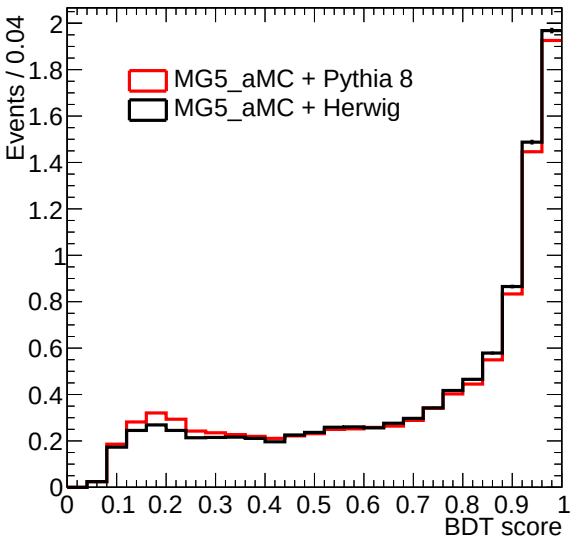


Figure 5.6: Comparison of the impact of the parton showering model in PYTHIA and HERWIG on the MG5_aMC signal simulation of the BDT score sdistribution.

5.2.2 Comparison of the QCD irreducible background predictions

Several predictions of the leading QCD-induced production of the $ZZjj$ final state are available. Specifically, the kinematics for the leading-order, the NLO 0,1 jet merged, and the NLO 0,1,2 jet merged predictions are compared. Figure 5.7 shows the tagging jet observables most relevant to the VBS phase space selection and the output of the BDT. The 0,1 jet NLO sample suffers from low statistics in the phase space probed by the analysis and would not allow to develop a BDT, illustrating the motivation for the phase-space optimizations.

5.2.3 Modeling of the loop-induced background

Figure 5.8 illustrates the poor jet modeling provided by the wimpy parton shower for the η and p_{T} spectrum of the highest-p_{T} jet. Parton emissions in the wimpy shower are limited to the factorization scale $\mu_F = m_{4\ell}/2$. The bulk of the events are concentrated around the ZZ resonance $m_{4\ell} \approx 200\,\mathrm{GeV}$, resulting in a drop at $p_{\mathrm{T}}^{\mathrm{jet}\,1} \approx 100\,\mathrm{GeV}$.

Figure 5.9 and 5.10 show comparisons of the nominal MCFM + PYTHIA predictions with the matrix element simulation of the loop-induced $ZZjj$ production for several VBS observables. The good agreement between the two prediction justifies using the MCFM + PYTHIA simulation in the $ZZjj$ analysis.

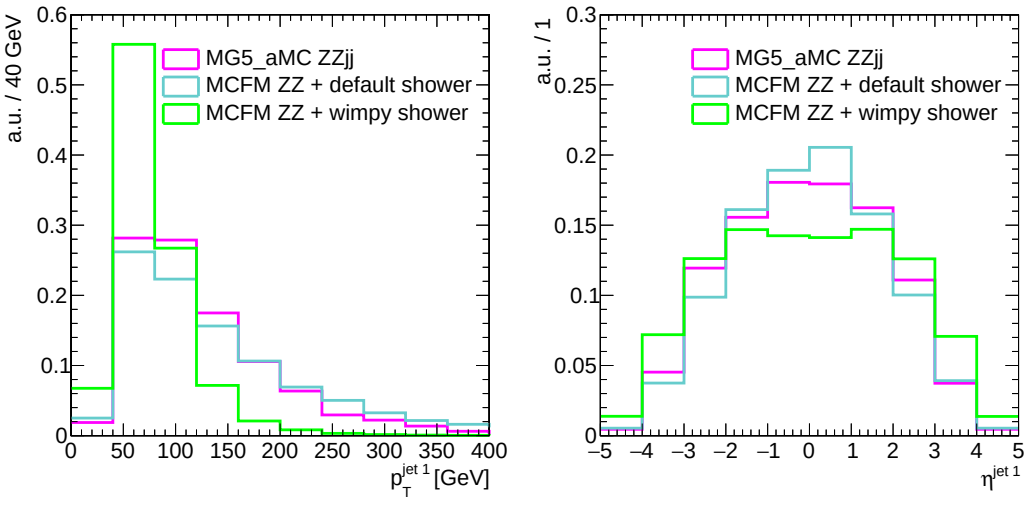


Figure 5.8: Kinematic observables for ggZZ loop-induced ZZ production in the MCFM simulation with regular parton shower (light blue), the 'wimpy' shower (green), and the LO tree-level (brown) for comparison. All histograms are normalized to unity. The Higgs MCFM sample with the wimpy parton shower (green line), a private production of the MCFM sample with a regular parton shower (light blue line), and the MG5_AMC ggZZ plus two jet simulation (brown line) are compared.

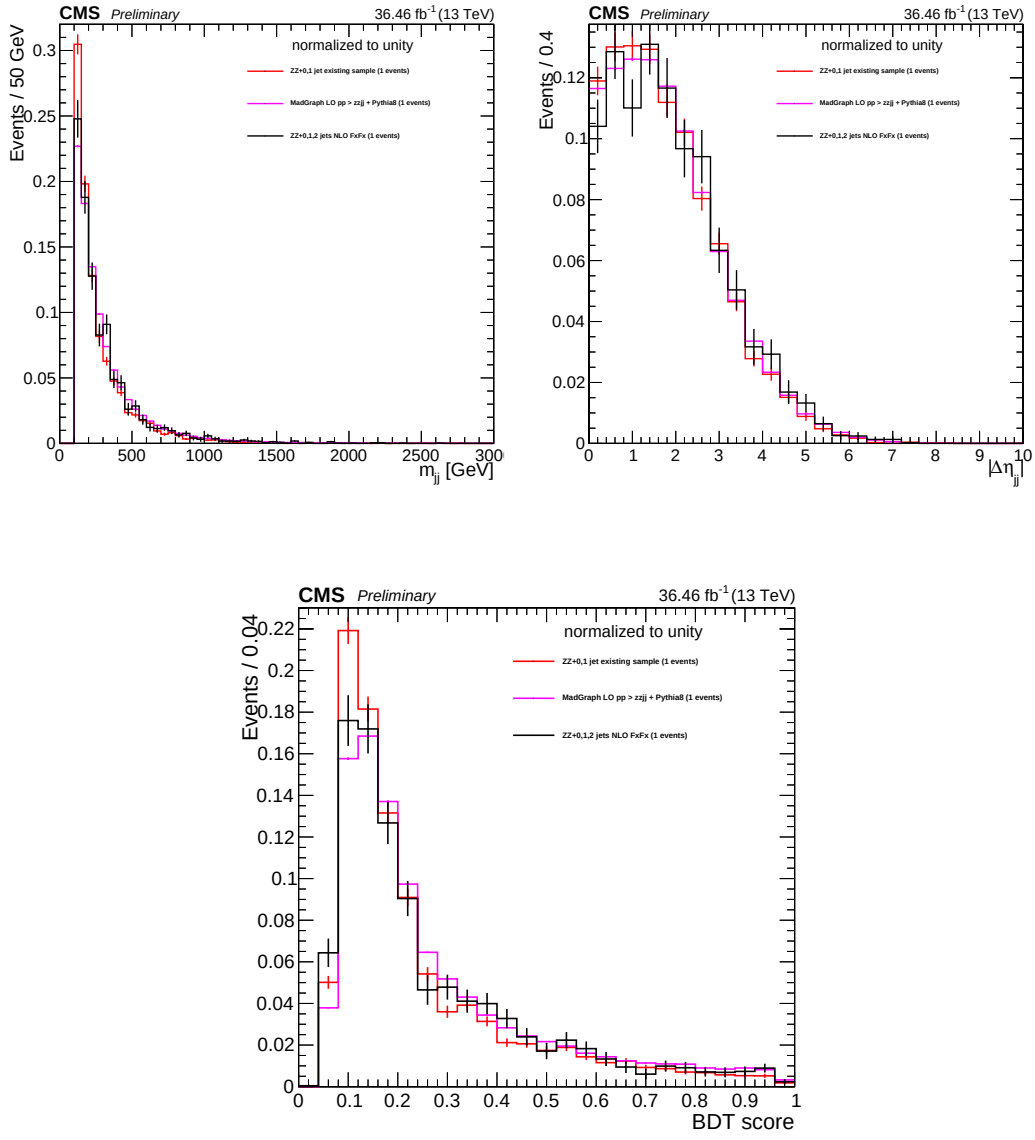


Figure 5.7: Comparison of the VBS kinematics in the dominant QCD process samples in the phase space defined by the $ZZjj$ baseline selection at the generator level. All distributions are normalized to unity. The 0,1 jet NLO sample (red line), the LO QCD $ZZjj$ sample (magenta line), and a private production of the 0,1,2 jet NLO sample (black line) are compared.

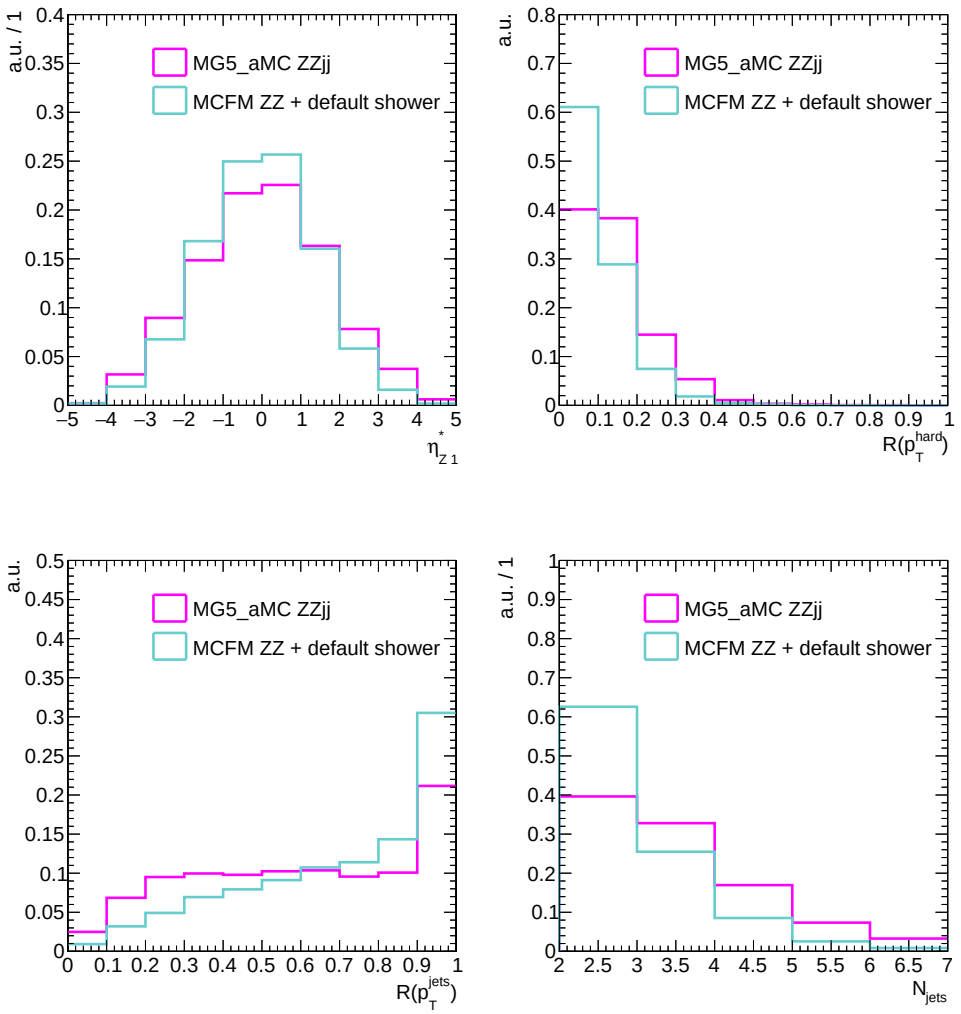


Figure 5.9: Comparison of other VBS observables in the ggZZ loop-induced production of ZZ in the phase space defined by the $ZZjj$ baseline selection at the generator level. All distributions are normalized to unity. MCFM with regular parton showering (light blue line) and the MG5_aMC ggZZ plus two jet simulation with regular parton showering (magenta line) are compared.

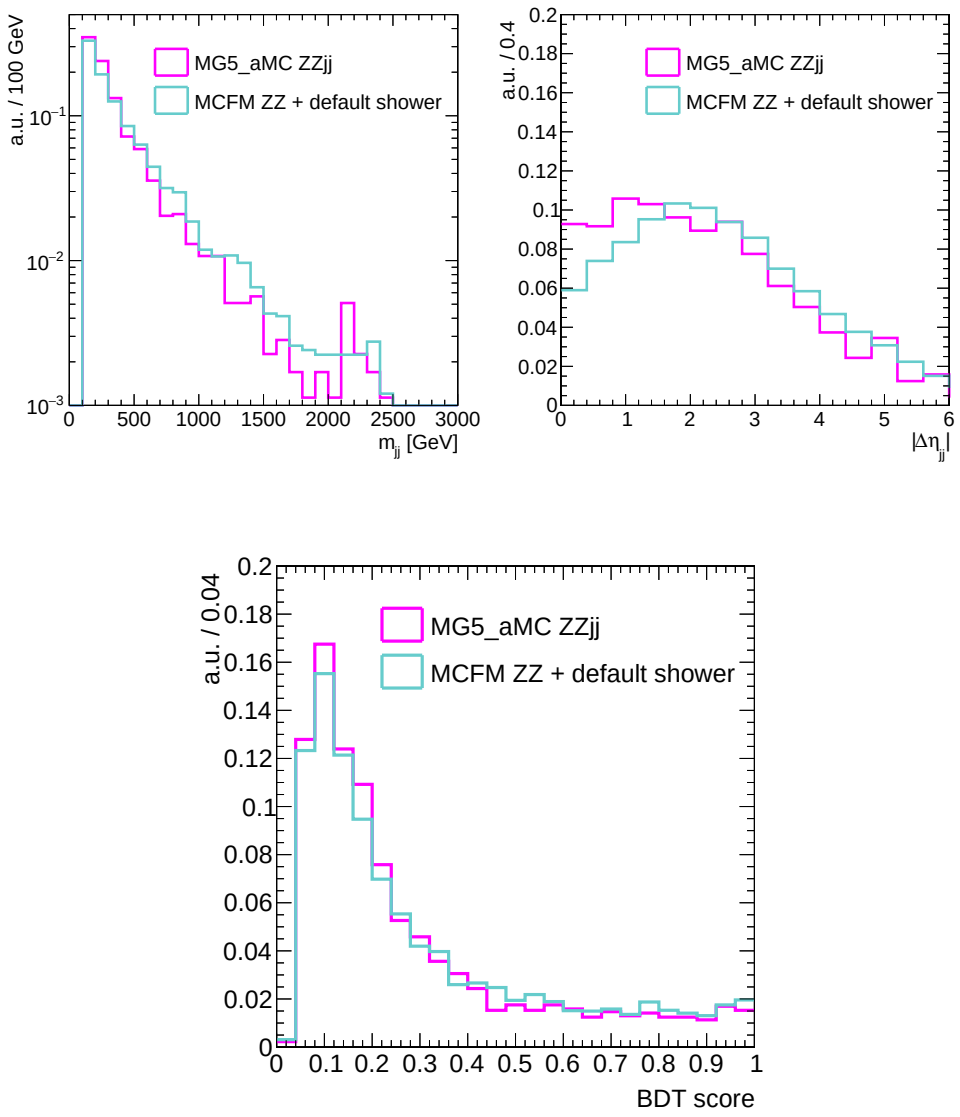


Figure 5.10: Comparison of the VBS kinematics in the ggZZ loop-induced production of ZZ in the phase space defined by the ZZjj baseline selection at the generator level. All distributions are normalized to unity. MCFM with regular parton showering (light blue line) and the MG5_aMC ggZZ plus two jet simulation with regular parton showering (magenta line) are compared.

5.3 Kinematics of the final state and event selection

Understanding the lepton kinematics of the electroweak $ZZjj$ signal is crucial for this multilepton analysis with low event yields. Figure 5.11 (left) shows the transverse momentum of the four leptons (sorted by p_T) in the GEN selection. The distribution of the fourth or softest lepton peaks at approximately 20 GeV, with a significant fraction of leptons at even lower p_T, highlighting the importance of the low-p_T regime for this multilepton analysis. Another important factor for the event acceptance is the pseudorapidity distribution of the most forward lepton in an event, shown in the right panel of Fig. 5.11. Comparing the distributions of the GEN and BLS selection, it can be

concluded that the limited pseudorapidity-coverage for leptons causes an appreciable reduction of the event acceptance.

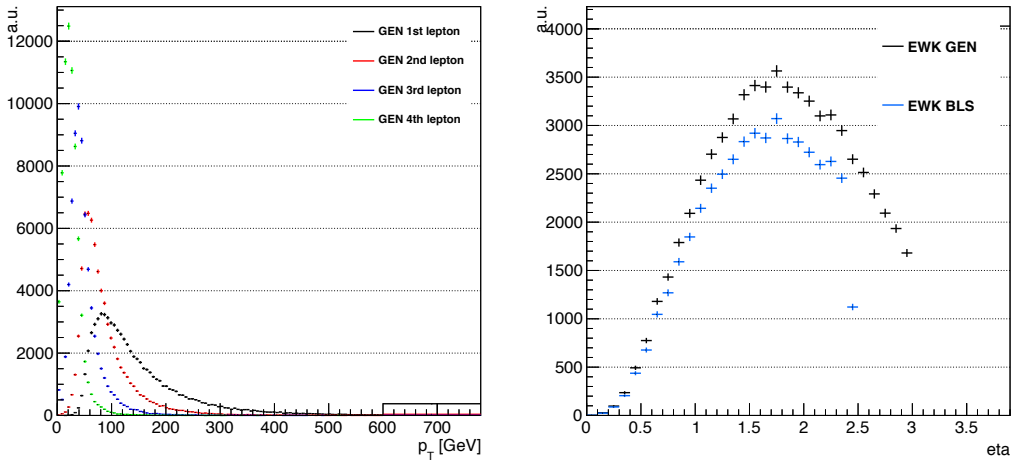


Figure 5.11: Kinematic distributions of the leptons in the GEN and BLS selections from the electroweak signal. The left panel shows the p_{T} of the leptons, sorted by p_{T}. The distribution of the lepton with the largest absolute pseudorapidity ($\max_{\ell}|\eta|$) is shown in the right panel. The last bin contains the overflow.

Figure 5.12 shows the kinematics of the leading Z boson reconstructed from the selected leptons in signal events, as well as the invariant mass of the ZZ system.

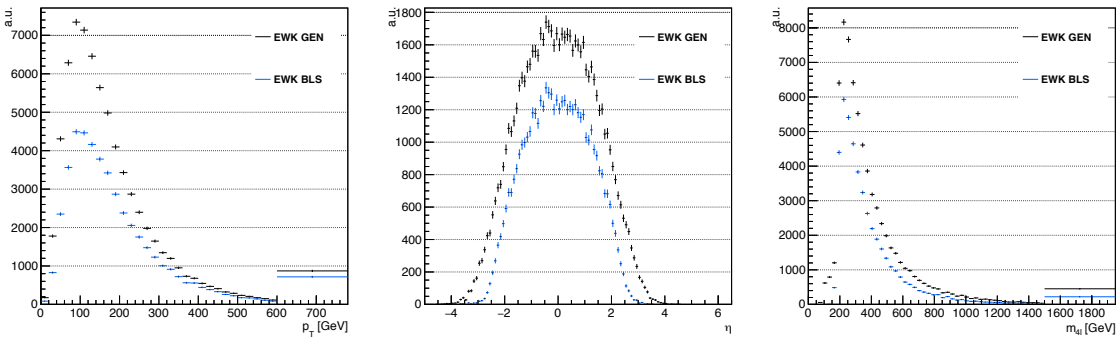


Figure 5.12: Kinematic distributions of the leading Z boson in the GEN and BLS selections from the electroweak signal. The left panel shows the p_{T} of the leading Z boson, the center panel its pseudorapidity, and the right panel shows the invariant mass of the ZZ system. The last bin contains the overflow.

The search for the electroweak production of the ZZ system also requires the presence of the two tagging jets. Figure 5.13 shows the p_{T} and $\max_j |\eta|$ distributions of signal events. While the event selection is not limited by the detector acceptance in η, it is clear that the second tagging jet can be quite soft.

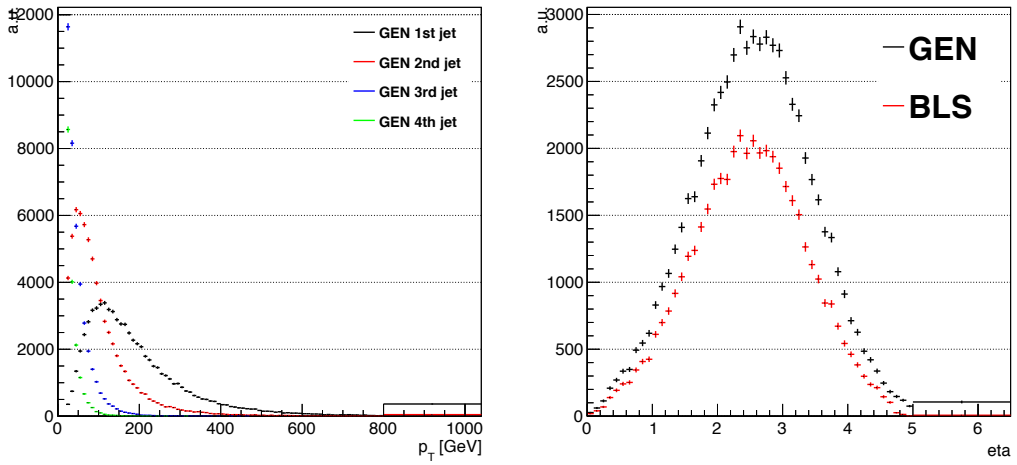


Figure 5.13: Kinematic distributions of the tagging jets of the electroweak signal. The left panel shows the p_{T} of the tagging jets (sorted by p_{T}) and the right panel the largest absolute pseudorapidity of the two jets ($\max_{j_1,j_2} |\eta|$). The last bin contains the overflow.

In order to quantify the effects of the various steps of an $ZZjj$ event selection on both signal and the dominant QCD background, we highlight some of the kinematic differences between the electroweak and the QCD production of the $ZZjj$ final state. As expected from the phenomenology of vector boson scattering, Fig. 5.14 shows that the tagging jets in the electroweak production are more forward, have a larger invariant mass and pseudorapidity separation than in the QCD-induced production.

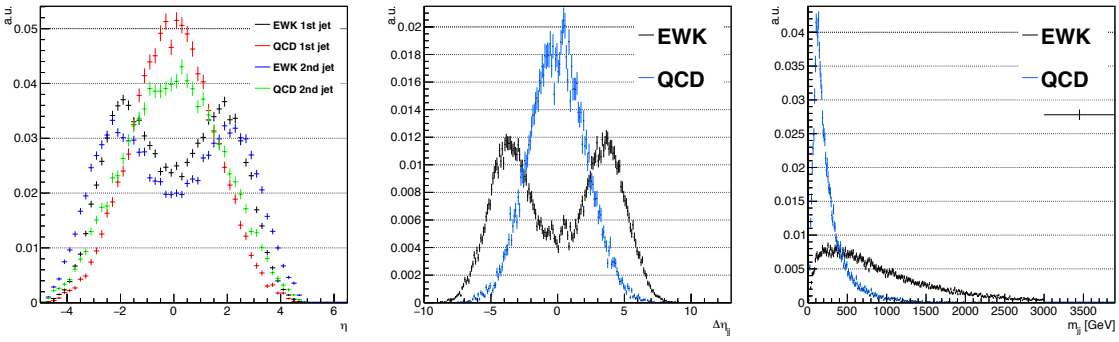


Figure 5.14: Comparison of the tagging jet kinematics of the electroweak and QCD-induced production of the $ZZjj$ final state in the GEN selection. The left panel shows the pseudorapidities of the tagging jets, while the center and right panels show the dijet pseudorapidity separation $\Delta\eta_{jj}$ and invariant mass m_{jj} respectively. The last bin contains the overflow.

The baseline selection implements a selection of the $ZZjj \to \ell\ell\ell'\ell' jj$ final state objects using the kinematic acceptance of the detector for the leptons and jets. Table 5.4 shows the selection efficiencies for a successive application of these requirements. The efficiencies are evaluated from the signal and leading-order QCD background simulation. Starting from all generated events, selecting two pairs of same-flavor opposite-sign

leptons is 100 % efficient. Imposing the lepton p_T thresholds of the baseline selection has a small effect, in contrast to the lepton pseudorapidity acceptance requirements, which reduce the signal efficiency to 76 %. The on-shell Z boson and loose dijet requirements cause efficiency losses of around 5 % each, resulting in an overall signal efficiency for the $ZZjj$ baseline selection of 64 %. The efficiency of the $ZZjj$ selection on the QCD background is 39 %, with a large impact of the dijet selection.

Table 5.5 illustrates variations of the $ZZjj$ baseline selection. Reducing the jet p_T threshold to 20 GeV recovers 3 % signal efficiency, but increases the background efficiency by 14 %, discouraging the use of a lower threshold in the analysis. A significant gain of signal efficiency is possible by extending the electron acceptance. Allowing one electron among the four leptons to come from the pseudorapidity range $2.5 < |\eta| < 3.0$, corresponding to a calorimeter-only electron, increases the signal efficiency by 13 %. This underlines the previous observation on the lepton kinematics (see Fig. 5.11) and the large impact of the lepton acceptance in the baseline selection. Finally, the impact of increasing the lepton p_T threshold to 10 GeV is evaluated, showing minor impact on the baseline selection efficiency of on-shell Z bosons.

Table 5.4: Event selection efficiencies for the electroweak (EW) signal and QCD background simulation. Starting from all generated events, the requirements of the $ZZjj$ baseline selection is successively applied.

Selection	EW Signal	QCD background
All generated events	100 %	100 %
Two pairs of same-flavor opposite-sign leptons	100 %	100 %
+ $p_T^{\mu(e)} > 5\ (7)$ GeV	96 %	96 %
+ $\lvert\eta^{\mu(e)}\rvert > 2.4\ (2.5)$	76 %	68 %
+ $60 < m_{\ell\ell} < 120$ GeV	71 %	65 %
+ $p_T^{\text{jet}} > 25$ GeV, $\lvert\eta^{\text{jet}}\rvert < 4.7$, $m_{jj} > 100$ GeV (= $ZZjj$ baseline selection)	64 %	39 %

Table 5.5: Event selection efficiencies for the electroweak (EW) signal and QCD background simulation. Modifications to the $ZZjj$ baseline selection are compared. The importance of the lepton pseudorapidity acceptance for the overall event selection efficiency is illustrated by the potential improvement by accepting one electron from the pseudorapidity range $2.5 < |\eta| < 3.0$ (calorimeter-only electrons).

Selection	EW Signal	QCD background
$ZZjj$ baseline	64 %	39 %
Min. jet $p_T > 20$ GeV	67 %	53 %
Min. jet $p_T > 30$ GeV	64 %	41 %
One calorimeter-only e	79 %	44 %
$p_T^{\ell} > 10$ GeV	62 %	45 %

Chapter 6

Event selection and reducible background estimation

This chapter presents the online and offline event selection criteria used in the $ZZjj$ analysis. First, the trigger paths exploited in this analysis and the data-driven trigger efficiency measurements are summarized, followed by the description of the ZZ and $ZZjj$ selections. The $ZZjj$ analysis exploits the same ZZ selection as the $H \to ZZ^* \to 4\ell$ analysis, allowing to validate the implementation and streamlining the overall analysis development. Studies using the simulation confirm the high efficiencies of the ZZ and $ZZjj$ selections. Finally, the data-driven estimate of the reducible background is laid out in detail.

6.1 Trigger selection

The trigger strategy used in the $ZZjj$ analysis is identical to the one developed for the $H \to ZZ^* \to 4\ell$ analysis. It relies on the logical OR of single, dilepton and trilepton trigger paths.

The primary triggers require the presence of a pair of loosely isolated leptons, regardless of the lepton flavor. The leading electron (muon) must have $p_T > 23\ (17)$ GeV, and the subleading lepton must satisfy $p_T > 12\ (8)$ GeV. The dilepton triggers furthermore require that the lepton tracks have a minimal longitudinal separation of no more than 2 mm in the transverse plane. Triggers requiring a triplet of low-p_T leptons with no isolation criterion and triggers selecting isolated single-electrons and single-muons with p_T thresholds of 27 and 22 GeV help to recover efficiency. An event is considered for offline analysis if it passes any of these triggers, irrespective of the final state.

Exploiting up to twenty trigger paths in the four lepton analyses results in very high trigger efficiencies of $> 98\,\%$, but poses a challenge to determine this efficiency in data. Performing per-leg efficiency measurements, as is usually done for analyses relying only on single or dilepton lepton triggers, is practically infeasible, given the many correlations between trigger paths and the unavailability of single lepton triggers corresponding to the subleading legs of the trilepton triggers. Instead, the trigger efficiency in data is estimated by exploiting the observed four lepton events that satisfy the offline analysis selection presented in Section 6.2.

6.1.1 Trigger efficiency measurement in data

The central idea is to require that one of the four leptons has fired a single-lepton trigger and to use these leptons as *tags*. The remaining three leptons can then be used as *probes* to evaluate the trigger efficiency. Because any one of the 4 leptons could fire a single lepton trigger, there are up to four tag-probe combinations per event and all of them are considered independently in the denominator of the efficiency in order to avoid any bias in the measurement. Technically, this method relies on a geometrical matching of a reconstructed lepton to the terminal HLT filter object. The latter is a minimal kinematical record of the online object that passed the last filter of a trigger path which resulted in the firing of the trigger. A given probe lepton can be matched to several filter objects corresponding to the different online object definitions used in the trigger, or none if it did not fire any trigger at all. Once the filter objects of the probes are extracted, an attempt is made to build one of the trigger paths used in the analysis. If such a trigger path can be built, the probes would have fired that trigger and the probe is counted in the nominator of the efficiency. It should be noted that it is not possible to use the list of fired trigger paths directly, as this event-level information does not guarantee that it was the probes that caused the trigger to fire.

The above method provides a lower bound on the overall trigger efficiency, i.e., it underestimates the true trigger efficiency. This bias arises for two reason: Firstly, only three of the four leptons per event are exploited, because one of the leptons is lost as the tag. An event could however pass the trigger because the tag and one or two of the probes fulfill the requirements, but these configurations are not probed in the method. The second effect only arises in the $2e2\mu$ final states, where the dilepton paths for the flavor of the tag can never be constructed from the probes and are thus excluded from the measurement. These two biases can be quantified by comparing the efficiency measured via the matching method in the simulation with the efficiency obtained from the HLT simulation. This study was carried out for the $H \to ZZ^* \to 4\ell$ analysis, which exploits the same triggers as the $ZZjj$ analysis, and found trigger efficiencies of 98/98/100% for the $4e/2e2\mu/4\mu$ final states. These numbers are used to correct the expected yield prediction in the simulation and a conservative uncertainty of 2 % is assigned, based on the size of the method bias evaluated from the simulation.

6.2 Event selection

The benefits of the clean and fully reconstructed four lepton final state of the ZZ channel targeted by this analysis come at the price of the low $Z \to \ell\ell$ branching ratio of about 6 %. As in other multilepton analyses, it is thus of paramount importance to efficiently select the leptons and then reconstruct the Z and ultimately ZZ candidates. The overall event selection is thus primarily oriented towards maximizing the ZZ reconstruction efficiency, while reducible backgrounds are a lesser concern. The dominant reducible background in this and other multilepton analyses are processes that feature one genuine leptonic decay of a Z boson. The second Z boson candidate is then due to fake leptons that happen to satisfy the same flavor and opposite sign requirements and also feature an invariant mass compatible with the Z boson mass. These kinematic requirements, which are derived from the signal definition, are sufficiently stringent to limit the reducible backgrounds to less than 10 % of the total yield.

The starting point of the event selection for the $ZZjj$ analysis is the selection of the ZZ candidate, which uses the selected leptons as inputs to first construct Z boson can-

didates and then Z boson pairs. Among the selected ZZ events one can then require the presence of extra jets.

ZZ candidate and ZZjj event selection

The input to the ZZ selection algorithm are the selected leptons, i.e., those that pass the selections outlined in Section 3.2.3 and 3.3.2:

- **Kinematic acceptance:** minimum p_{T} of 5 (7) GeV and maximum $|\eta^{\ell}|$ of 2.4 (2.5) for muons (electrons)
- **Impact parameter:** $|\mathrm{SIP}_{3\mathrm{D}}| < 4$, $|d_{xy}| < 0.5\,\mathrm{cm}$, and $|d_z| < 1\,\mathrm{cm}$
- **Lepton identification:** electron MVA and muon selection (Section 3.2.3 and 3.3.1)
- **Lepton isolation:** $\mathcal{I}_{\mathrm{rel}} < 0.35$ (Section 3.2.3 and 3.3.2)

The kinematic restrictions on the leptons are essentially given by the detector acceptance in order to maximize the per-lepton and ultimately the four-lepton selection efficiency. As shown in Section 5.3 the event acceptance is limited by the maximum pseudorapidity of the leptons. Electrons in the barrel-endcap transition regions of the calorimeter are not vetoed in order to maintain the highest ZZ selection efficiency.

Any FSR photon matched to a lepton is implicit in its kinematics, in particular the dilepton invariant mass will include these photons.

An additional electron cleaning is performed, whereby any electron within $\Delta R < 0.05$ of a selected muon is removed. This selection suppresses rare cases where a muon track is matched to the electromagnetic cluster coming from an FSR emission of the muon, giving rise to a fake electron.

The ZZ or four-lepton candidate selection algorithm is adopted from the $\mathrm{H}\to ZZ^{*}\to 4\ell$ analysis and does not require that both Z bosons be on-shell. Having a looser selection has a negligible impact for the on-shell selection efficiency of the algorithm, but using the same algorithm across several analyses eases the synchronization and allows to compare the obtained results.

The sequence is as follows:

1. **Z boson candidates** are constructed from pairs of selected leptons of opposite charge and matching flavor ($e^{+}e^{-}$, $\mu^{+}\mu^{-}$) that satisfy $12 < m_{\ell\ell(\gamma)} < 120\,\mathrm{GeV}$, where the Z candidate mass includes the selected FSR photons, if any.
2. **ZZ candidates** are defined as pairs of non-overlapping Z candidates. The Z candidate with reconstructed mass $m_{\ell\ell}$ closest to the nominal Z boson mass is denoted as Z_1, and the second one is denoted as Z_2. ZZ candidates are required to satisfy the following list of requirements:
   - **Ghost removal** : $\Delta R(\eta,\phi) > 0.02$ between each of the four leptons.
   - **Lepton** p_{T}: Two of the four selected leptons should satisfy $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $p_{\mathrm{T}} > 10\,\mathrm{GeV}$.
   - **QCD suppression**: all four opposite-sign pairs that can be built with the four leptons (regardless of lepton flavor) must satisfy $m_{\ell\ell'} > 4\,\mathrm{GeV}$. Here, selected FSR photons are not used in computing $m_{\ell\ell'}$, since a QCD-induced low mass dilepton resonance (eg. J/ψ) may have photons nearby (e.g. from π^{0} decays).
   - Z_1 **mass**: $m_{\mathrm{Z}1} > 40\,\mathrm{GeV}$

- **Smart cut**: defining Z_a and Z_b as the mass-sorted alternative pairing Z candidates (Z_a being the one closest to the nominal Z boson mass), require NOT($|m_{Za} - m_Z| < |m_{Z1} - m_Z|$ AND $m_{Zb} < 12\,\text{GeV}$). Selected FSR photons are included in m_Z's computations. This cut discards 4μ and $4e$ candidates where the alternative pairing looks like an on-shell Z boson + low-mass $\ell^+\ell^-$.
- **Four-lepton invariant mass**: $m_{4\ell} > 70\,\text{GeV}$

The above selection can lead to several ZZ candidates per event, and an arbitration is needed. Because false ZZ candidates are likely built from extra (fake) leptons, which in turn are more prominent at low p_T, the best ZZ candidate is chosen as the one having the largest scalar p_T sum of the leptons constituting the Z_2 candidate. The analysis uses events in which both Z bosons of the best ZZ candidate are on-shell, i.e., both satisfy $60 < m_Z < 120\,\text{GeV}$. Contrary to the labelling used in the ZZ selection algorithm given above, the bosons are ordered not by the proximity of the measured mass to $m_Z^{PDG} = 91.2\,\text{GeV}$, but by the p_T of the Z bosons in decreasing order.

ZZjj selection

The ZZ event selection is supplemented by the dijet requirement for the VBS search. Specifically, events are required to feature at least two jets with $|\eta| < 4.7$ and $p_T > 30\,\text{GeV}$. The leading and subleading jets are taken as the tagging jets in case of more than two jets in the event. The invariant mass of the tagging jets has to satisfy $m_{jj} > 100\,\text{GeV}$, in order to suppress hadronic WZ decays. No further attempt to enhance the electroweak signal over the irreducible background is made by the event selection in order keep the signal efficiency as high as possible. The discrimination between the signal and large QCD background will be performed by a multivariate classifier and all selected $ZZjj$ events will be considered in the signal extraction. This approach also allows to constrain the normalization of the QCD background and furthermore does not reduce the signal sensitivity by removing any data from the analysis. The significance of the electroweak signal, its signal strength, and the constrains on anomalous quartic gauge couplings are derived in this $ZZjj$ selection.

VBS selection

A VBS-enriched signal region is defined for events that pass the ZZjj selection and also satisfy

- $|\Delta\eta_{jj}| > 2.4$ and
- $m_{jj} > 400\,\text{GeV}$.

This selection is referred to as the VBS selection. Its main usage is to define a simple signal-enriched region for a cut-and-count approach and to enable cross checks of the results.

nVBS selection

A QCD-enriched region is defined by inverting the above VBS selection:

- $|\Delta\eta_{jj}| < 2.4$ or
- $m_{jj} < 400\,\text{GeV}$.

This selection is referred to as the nVBS or not VBS selection. The VBS and nVBS selections are mutually exclusive and collectively exhaustive, i.e. all events passing the $ZZjj$ selection either fall into the VBS or nVBS selection and summing the events in the VBS and nVBS selection recovers all $ZZjj$ events. The main use of this selection is to define a background-enriched control region to cross-check the modeling of the dominant QCD background prior to unblinding, specifically in the observables that are used as an input for the BDT.

6.3 Event selection efficiencies

The event selection is designed to be as efficient as possible given the small reducible background and considering the very low branching ratio of leptonic Z boson decays. In the following we present the selection efficiencies for the electroweak signal in the simulation in order to understand the impact of the main selection steps.

The overall ZZ selection efficiency for the electroweak signal is close to 70 % when summing all final states. The efficiency is significantly higher at about 85 % in the 4μ final state and appreciably lower at about 58 % in the $4e$ final state. Figure 6.1 shows the signal selection efficiencies per final state and the for all final states (left panel), which exhibits a small linear decrease with increasing m_{ZZ}, by about 2 % at $m_{ZZ} = 500\,\text{GeV}$. From the same figure it can be seen that the decrease in efficiency is much smaller in the 4μ final state and most pronounced in the $4e$ final state.

This decrease in efficiency is already present for a selection based on truth-matched leptons without any further identification or isolation requirements and furthermore not observed for events where all electrons are in the barrel part of the detector. This indicates that the decrease in selection efficiency is driven by a change in the pseudorapidity distribution of the electrons: an increase in the scattering energy results in more electrons in the forward region of the detector in which the electron reconstruction efficiency is reduced. This assumption is corroborated by Fig. 6.2, which shows the fraction of events with a specified number of electrons in the barrel: while for $m_{ZZ} = 200\,\text{GeV}$ almost 50 % of $4e$ events have all electrons in the central part of the detector, this number drops to 35 % at $m_{ZZ} = 500\,\text{GeV}$. This trend of a decrease in the 4ℓ reconstruction efficiency is interesting, as the exact opposite is observed for the $H \to ZZ^* \to 4\ell$ decays, where the signal selection efficiency is known to improve with $m_{4\ell}$ because of the increase in electron p_T which increases the event acceptance as well as electron reconstruction and selection efficiency.

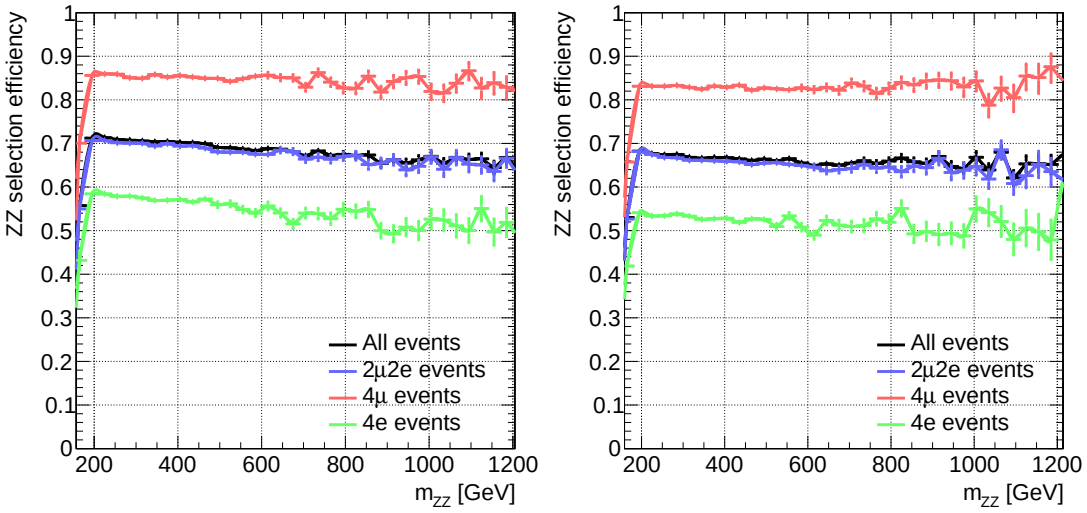


Figure 6.1: Efficiency of the $ZZ \to 4\ell$ selection on the electroweak signal (left) and irreducible QCD background (right) in the simulation. The efficiencies are calculated with respect to all events in the lepton acceptance that pass the ZZ selection at the truth level.

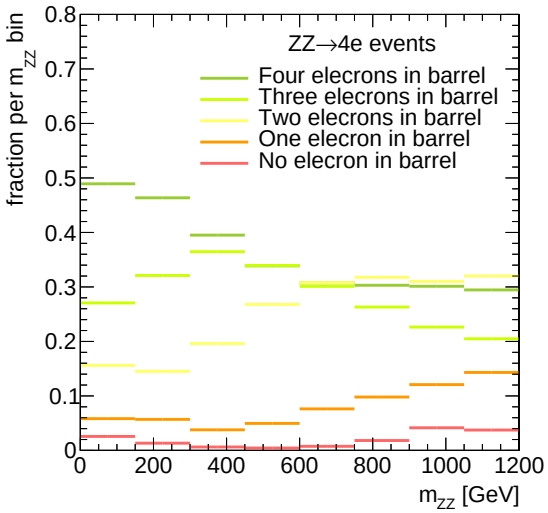


Figure 6.2: Fraction of selected electroweak $4e$ signal events with a specified number of electrons in the barrel part of the detector ($|\eta^e| < 1.48$) in the simulation.

The overall ZZ selection efficiency can be further understood by considering the three lepton selection steps (identification, isolation, and impact parameter) separetely. Figure 6.3 shows their successive impact on the selection efficiency in the 4μ (left) and $4e$ (right) final states. The muon identification has a signal efficiency of almost 100 %, the isolation efficiency is around 93 %, while the impact parameter requirement reduces the efficiency by about 4 %. The signal efficiency of the electron identification is about 6 % or 1.5 % per electron, compatible with the expectation from the per-electron efficiency of 98 % to 99 %. Applying the isolation requirement on top of the identification selection reduces the ZZ efficiency by another 6 % and a significant correlation between both selections can be inferred. The reduction of signal efficiency due to the impact parameter requirement has a larger effect on electrons than on muons.

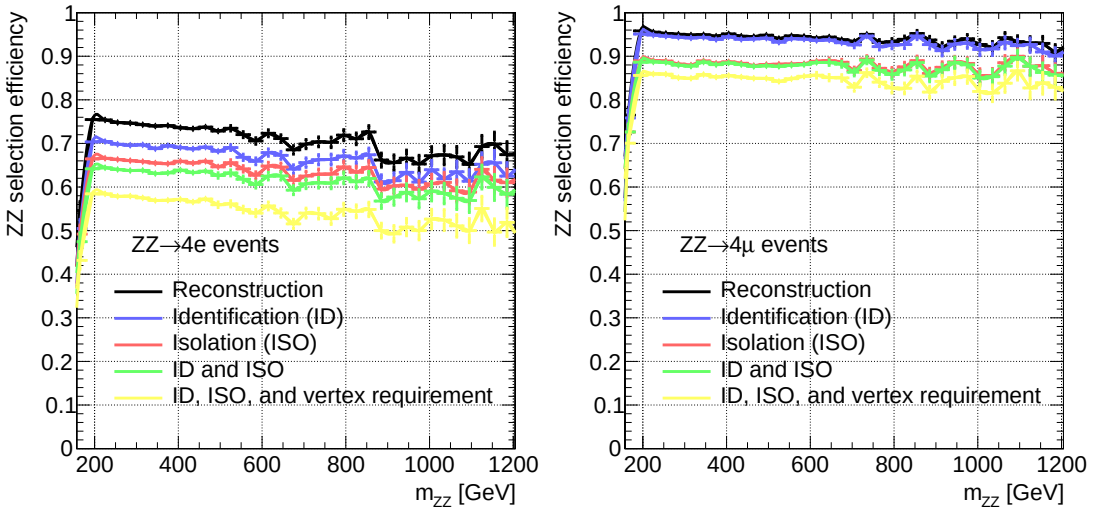


Figure 6.3: Efficiency of the $ZZ \to 4\ell$ selection on the electroweak signal for the 4μ (left) and $4e$ (right) final states for the successive steps of the lepton selection. The efficiencies are calculated with respect to all simulated events in the lepton acceptance that pass the ZZ selection at the truth level.

The final step in the $ZZjj$ selection requireres the presence of the tagging jets with $m_{jj} > 100\,\text{GeV}$. For the electroweak signal this reduces the overall event selection efficiency from around 72 % to 64 %, i.e., a 8 % drop which is driven by the $p_{\text{T}}^{\text{jet}} > 30\,\text{GeV}$ and $m_{jj} > 100\,\text{GeV}$ requirements, as the jet selection efficiency is > 99 % efficient.

6.4 Irreducible non-ZZ backgrounds

Without further kinematic selections the total yield is dominated by the QCD-induced production of $ZZjj$, which contributes about 90 % of the expectation. This contribution is estimated from the simulation which merges the 0,1,2-jet multiplicities at NLO accuracy as described in Section 5.1. As will be shown in Section 7.2, this background can be suppressed by its kinematics and the statistical sensitivity of the analysis is dominated by regions of phase-space where the electroweak signal contributes more than half the total yield. By considering all $ZZjj$ events in the signal extraction fit, the normalization of the QCD component is ultimately constrained by the data in the background-enriched region.

Multilepton processes with four or more leptons originating from non-Z decays in association with jets can also contribute to the $ZZjj$ signal region if the leptons happen to satisfy the ZZ selection. The leading contribution arises from processes that feature one real on-shell Z boson (WWZ + jets, $t\bar{t}Z$ + jets). These processes are sufficiently rare and contribute less than 3 % in the $ZZjj$ event selection. Furthermore, these processes feature background-like jet kinematics. Their contributions are estimated from the simulation. They are furthermore not explicitly considered in the following studies on the $ZZjj$ event acceptance and event selection efficiency.

6.5 Data-driven estimate of the reducible background

The reducible background for the $ZZjj$ analysis arises from processes which contain one or more fake leptons. Here the term fake leptons includes non-isolated leptons coming from decays of heavy-flavor mesons, light-flavor jets misidentified as leptons, and electrons from photon conversions. This background can in principle originate from events where one, two, three or all selected leptons are fake leptons. The latter two cases are highly suppressed by the on-shell Z requirement and the dominant background arises from processes that feature one on-shell Z boson, notably Drell-Yan and WZ production. The reducible background is hence referred to as Z+X.

The Z+X background is estimated by exploiting control regions obtained by inverting the lepton selection and by evaluating the rate of fake leptons to obtain a prediction of the background in the signal region.

6.5.1 Fake rate measurement

The measurement of the fake ratios are defined as the rate at which *loose leptons*[1] pass the lepton selection of the $ZZjj$ analysis. Loose leptons are reconstructed leptons that satisfy the impact parameter requirements ($|\mathrm{SIP_{3D}}| < 4$, $|d_{xy}| < 0.5\,\mathrm{cm}$ and $|d_z| < 1\,\mathrm{cm}$) and that may or may not satisfy the other lepton selection criteria. Because the dominant reducible background arises from Drell-Yann events, a suitable region to measure the fake ratios is defined by selecting events that feature a leptonically decaying Z boson and exactly one of the aforementioned loose leptons.

The Z boson candidate is selected from two same-flavor opposite-sign leptons that pass the lepton selections and satisfy $p_{\mathrm{T}}^1 > 20\,\mathrm{GeV}$ and $p_{\mathrm{T}}^2 > 10\,\mathrm{GeV}$. In addition to the two leptons that form the Z boson candidate, the events are required to have one and only one loose lepton. The invariant mass of this lepton and the opposite-sign lepton from the reconstructed Z boson candidate should satisfy $m_{2\ell} > 4\,\mathrm{GeV}$ to suppress contaminations from low-mass QCD resonances.

The invariant mass of the Z boson candidate is shown in Fig. 6.4, for all $Z + \ell$ events (top row) and for events where the extra lepton satisfies the lepton selection, separately for events where the extra lepton is an electron (left column) or a muon (right column). Here and in the following the distributions only show events where the third lepton is in the central region of the detector, but observations generalize to all events. The mass distribution of events with an extra electron that passes the electron selection (bottom left panel in Fig. 6.4) exhibits an enhancement at masses below the Z peak, a feature not observed for the analogous muon selection. The enhancement is due to FSR conversions and its contamination is removed by requiring that the mass of the Z boson candidate is compatible with the nominal Z mass within $7\,\mathrm{GeV}$, $|m_Z - 91.2\,\mathrm{GeV}| < 7\,\mathrm{GeV}$. Figure 6.5 shows the fake ratios as a function of the Z candidate mass and illustrates the need for the mass requirement.

[1] The selected leptons in multilepton analyses are traditionally referred to as *tight* leptons, a quite misleading nomenclature. For example, the electron identification efficiency in multilepton analyses is $> 98\,\%$, much higher than any of the regular electron ID working points used for CMS analyses.

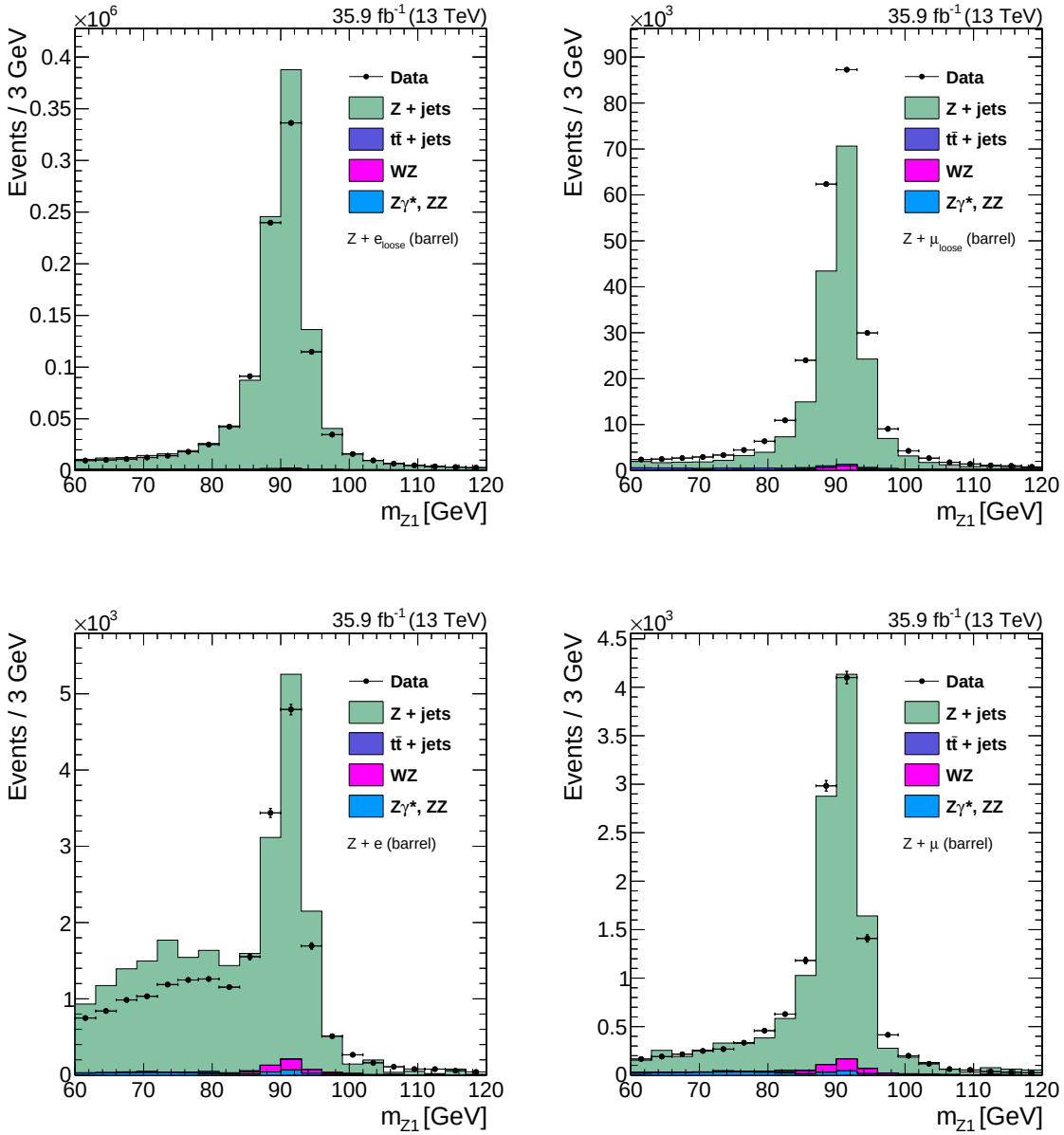


Figure 6.4: Distributions of the Z candidate mass for the $Z + e$ (left) and $Z + \mu$ (right) control regions, as defined in the text. The top row shows all $Z + \ell$ events, while the bottom row shows only events where the third lepton passes the lepton selection. Only events with third leptons in the central region of the detector are shown.

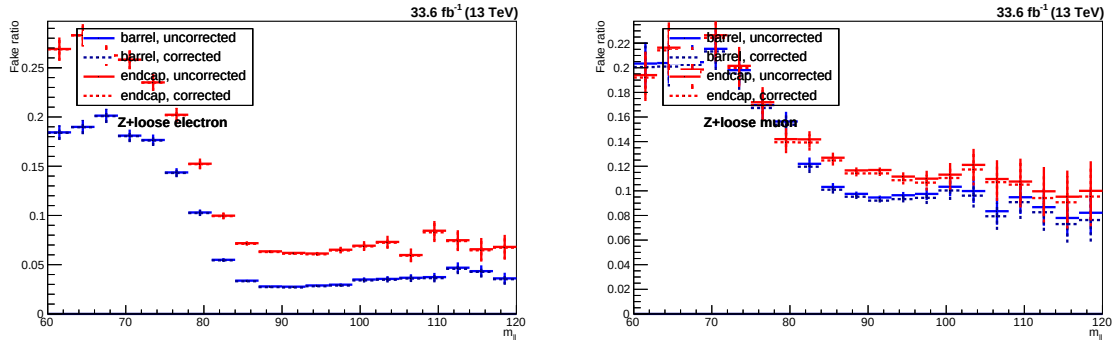


Figure 6.5: Dependence of the measured fake ratio for electrons (left) and muons (right) on the Z boson candidate mass as measured in the 2016 dataset. The impact of the WZ correction is also indicated.

The fake ratio measurement in the $Z + \ell_{\text{loose}}$ region assumes that the third lepton in the event is actually not a prompt lepton. The prediction of the simulation in the bottom row in Fig. 6.4 shows a non-negligible contribution from WZ and ZZ processes. The latter contributes to the selection if one of the four leptons is outside the kinematic acceptance of the detector or if it was not reconstructed. Either case will lead to non-zero $E_{\text{T}}^{\text{miss}}$. Figure 6.6 shows the missing transverse momentum of all $Z + e_{\text{loose}}$ events (left) and those where the loose electron satisfies the electron selection. The prompt lepton contamination from WZ/ZZ contributes at the few-percent level in events where the third lepton passes the lepton selection. A larger relative contamination in $Z + e_{\text{loose}}$ versus $Z + \mu_{\text{loose}}$ events is observed and is due to the lower reconstruction efficiency for electrons. To reduce this prompt lepton contamination, events for the final fake ratio measurement are required to satisfy $E_{\text{T}}^{\text{miss}} < 25\,\text{GeV}$. The residual prompt lepton contamination is negligible at low p_{T} but increases strongly, particularly for muons as shown in Fig. 6.7. The final fake ratios are corrected from this contamination by subtracting the WZ yield from the simulation.

The final electron and muon fake ratios are extracted from events in the $Z + \ell_{\text{loose}}$ control region that satisfy $|m_Z - 91.2\,\text{GeV}| < 7\,\text{GeV}$ and $E_{\text{T}}^{\text{miss}} < 25\,\text{GeV}$. The resulting fake ratios are shown in Fig. 6.8.

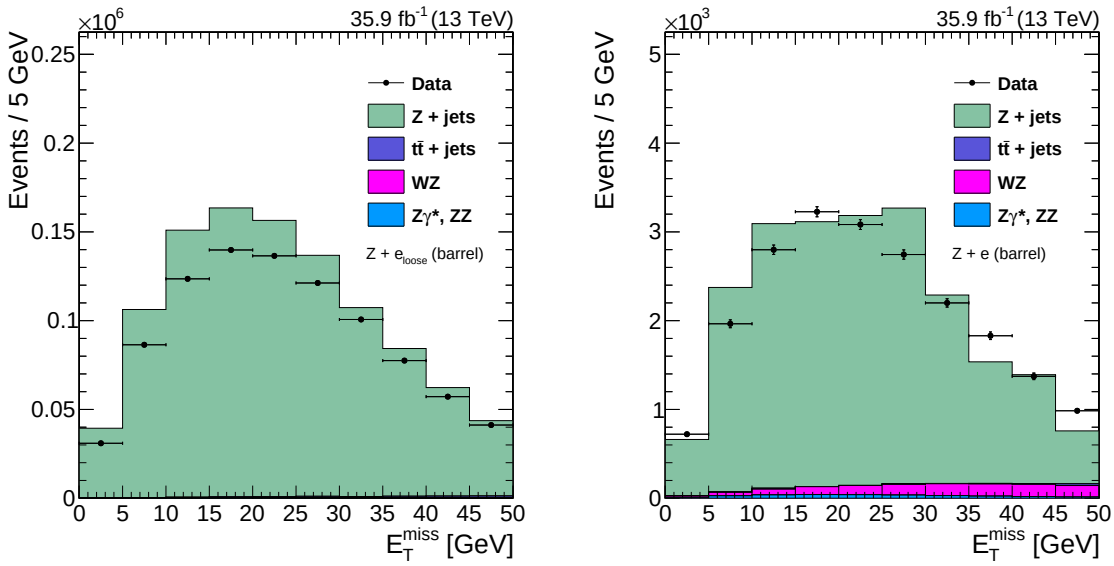


Figure 6.6: Distribution of the missing transverse momentum $E_{\mathrm{T}}^{\mathrm{miss}}$ in events from the $Z + e_{\mathrm{loose}}$ control region, as defined in the text. The left panel shows all events while the right panel only shows events where the loose electron satisfies the electron selection. Only events with third leptons in the central region of the detector are shown.

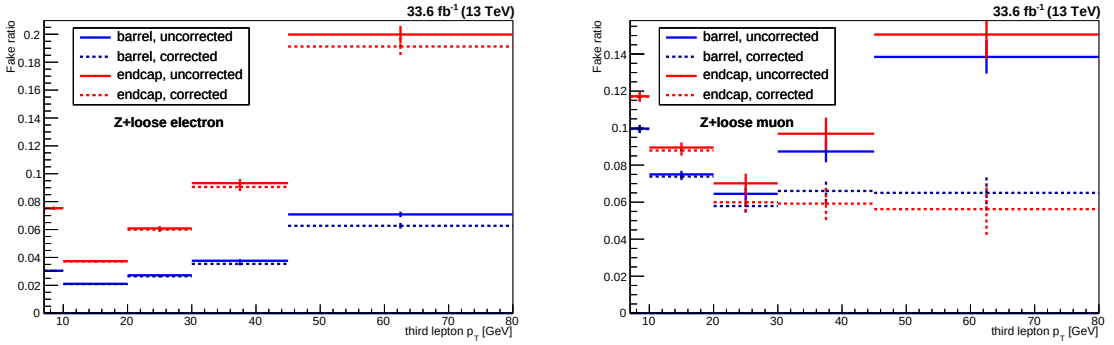


Figure 6.8: Fake ratios measured in the 2016 dataset for probe electrons (left) and muons (right) as a function of the probe p_{T}. The barrel selection includes electrons (muons) up to $|\eta| = 1.479$ (1.2).

The measured fake ratios are stable with regards to the number of reconstructed vertices, a measure of the amount of pileup in the event, as can be seen in Fig. 6.9.

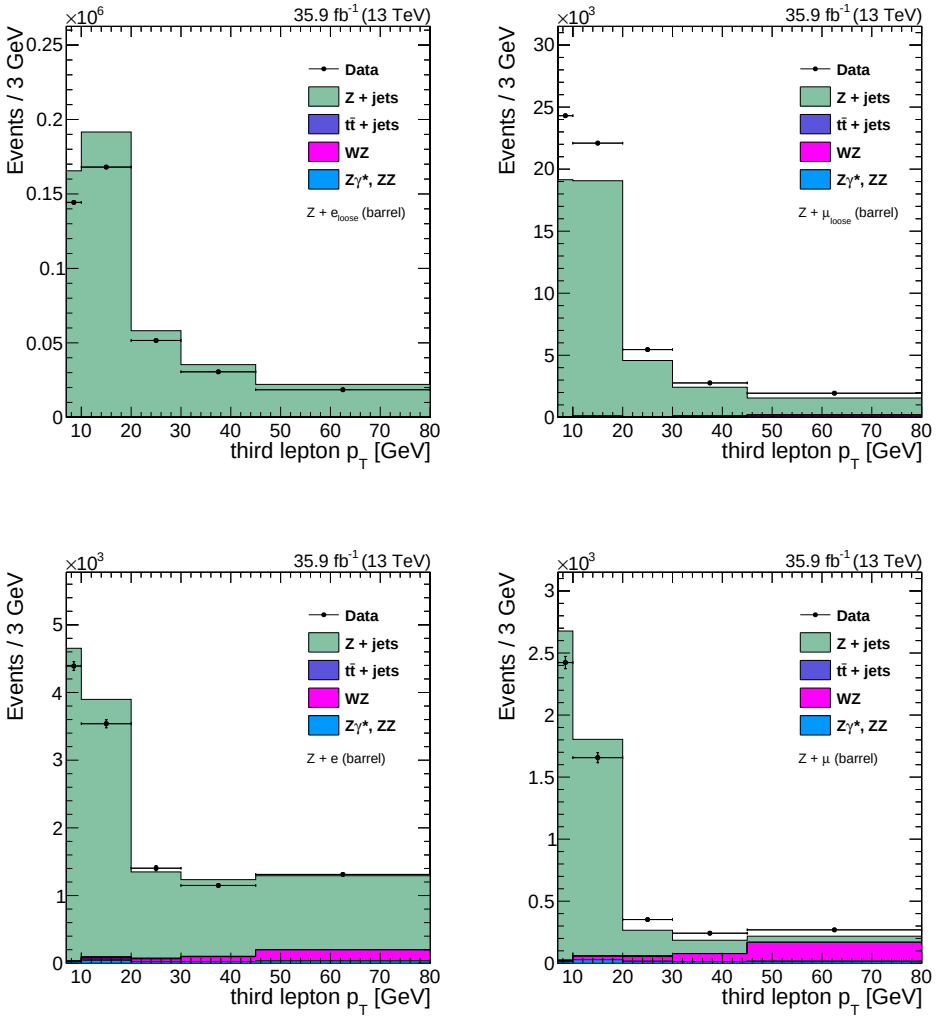


Figure 6.7: Distribution of the third lepton p_{T} in events from the $Z + e_{\mathrm{loose}}$ (left) and $Z + \mu_{\mathrm{loose}}$ (right) control regions, as defined in the text. The top row shows all events, while the bottom row shows only events where the third lepton passes the lepton selection.

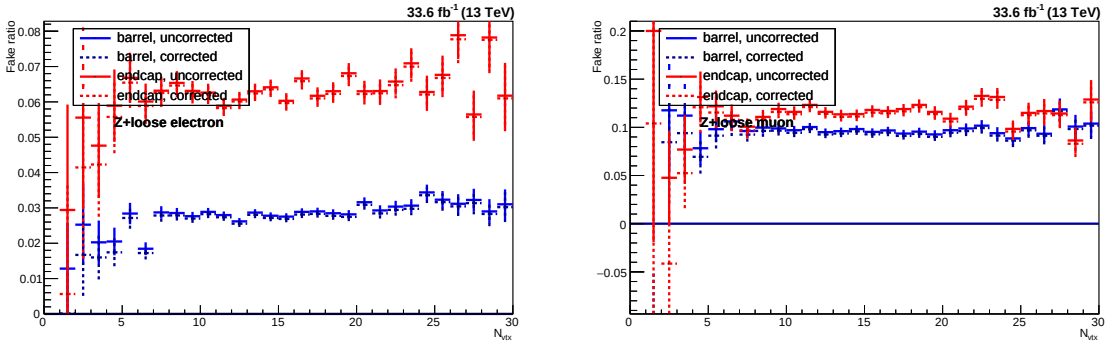


Figure 6.9: Fake rates measured in the 2016 dataset for probe electrons (left) and muons (right) as a function of the number of reconstructed vertices in the event. The barrel selection includes electrons (muons) up to $|\eta| = 1.479$ (1.2).

6.5.2 Control regions and application of fake ratios

The selection of the control regions used to estimate the background in the signal regions proceeds just as the ZZ selection, with the exception of the leptons used in the construction of the Z_2 candidate. Two orthogonal control regions are obtained, one where both leptons of the Z_2 candidate fail the lepton selection, referred to as *2 Prompt + 2 Fail* or *2P2F*. The second control region, called *3 Prompt + 1 Fail* or *3P1F*, is built from Z_2 candidates that feature one loose and one selected lepton. Both control regions are also orthogonal to the signal region.

Care has to be taken to remove the overlap between failing leptons and jets. The anti-matching to selected leptons for jets used in the $ZZjj$ event selection is thus extended to loose leptons in the 2P2F and 3P1F control regions and any jet within $\Delta R < 0.4$ the loose leptons in the control region is discarded.

The 2P2F control region is dominated by processes that intrinsically only feature two prompt leptons: mostly Drell-Yan plus jets with minor contributions from $\mathrm{t\bar{t}}$ and $Z + \gamma$. Figure 6.10 shows the invariant mass distribution of the 2P2F control region for each the four final states and their sum. In the context of these control regions, the final states are labelled first by the flavor of the Z_1 candidate and then by the Z_2 candidate which includes the failing leptons, e.g., the label $2\mu 2e$ refers to events where both muons satisfy the lepton selection and one or both electrons fail the electron selection.

The 3P1F control region is enhanced in events with three prompt leptons, notably WZ production, in addition to the same processes found in the 2P2F region with the difference that one of the fake leptons passes the lepton selection.

The expected number of reducible background events in the 3P1F region, $N^{\mathrm{bkg}}_{\mathrm{3P1F}}$, can be computed from the number of events observed in the 2P2F control region, N_{2P2F}, by weighting each event with the factor $(\frac{f_i}{1-f_i} + \frac{f_j}{1-f_j})$, where f_i and f_j correspond to the kinematics-dependent fake ratios of the two loose leptons:

$$N^{\mathrm{bkg}}_{\mathrm{3P1F}} = \sum (\frac{f_i}{1-f_i} + \frac{f_j}{1-f_j}) N_{\mathrm{2P2F}} \tag{6.1}$$

Figure 6.11 shows the four-lepton invariant mass distribution of the events selected in the 3P1F control sample, together with the expected reducible background estimated from Eq. (6.1).

Would the fake rates be measured in a sample that has exactly the same background composition as the 2P2F sample, the difference between the observed number of events in the 3P1F sample and the expected background predicted from the 2P2F sample would solely amount to the small WZ and $Z\gamma_{\text{conv}}$ contributions. Differences arise because the fake rates used in Eq. (6.1) do not properly account for the background composition of the 2P2F control sample.

The difference seen at low masses in Fig. 6.11 between the distribution from the 3P1F control region and the expected contribution extrapolated from the 2P2F control region using Eq. (6.1), in channels with loose electrons ($4e$ and $2\mu 2e$), is due to photon conversions.

The difference between the 3P1F observation and the prediction from 2P2F is used to recover the missing contribution from conversions - and more generally, to correct for the fact that the fake rates do not properly account for the background composition of the 2P2F sample. More precisely, the expected reducible background in the signal region is given by the sum of two terms:

- **2P2F component:** obtained from the number of events observed in the 2P2F control region, N_{2P2F}, by weighting each event in that region with the factor $\frac{f_i}{1-f_i}\frac{f_j}{1-f_j}$, where f_i and f_j correspond to the fake ratios of the two loose leptons.
- **3P1F component:** obtained from the difference between the number of observed events in the 3P1F control region, N_{3P1F}, and the expected contribution from the 2P2F region and ZZ processes in the signal region, $N^{\text{ZZ}}_{\text{3P1F}} + N^{\text{bkg}}_{\text{3P1F}}$. The $N^{\text{bkg}}_{\text{3P1F}}$ is given by equation 6.1 and $N^{\text{ZZ}}_{\text{3P1F}}$ is the contribution from ZZ which is taken from the simulation. The difference $N_{\text{3P1F}} - N^{\text{bkg}}_{\text{3P1F}} - N^{\text{ZZ}}_{\text{3P1F}}$, which may be negative, is obtained for each (p_{T}, η) bin of the failing lepton, and is weighted by $\frac{f_i}{1-f_i}$, where f_i denotes the fake rate of this lepton.

The full expression for the prediction can be symbolically written as:

$$N^{\text{bkg}}_{\text{SR}} = \sum \frac{f_i}{(1-f_i)}(N_{\text{3P1F}} - N^{\text{bkg}}_{\text{3P1F}} - N^{\text{ZZ}}_{\text{3P1F}}) + \sum \frac{f_i}{(1-f_i)}\frac{f_j}{(1-f_j)}N_{\text{2P2F}} \tag{6.2}$$

which is equivalent to:

$$N^{\text{bkg}}_{\text{SR}} = (1 - \frac{N^{ZZ}_{3P1F}}{N_{3P1F}}) \sum_{j}^{N_{3P1F}} \frac{f^j_a}{1-f^j_a} - \sum_{i}^{N_{2P2F}} \frac{f^i_3}{1-f^i_3}\frac{f^i_4}{1-f^i_4}. \tag{6.3}$$

For channels where the Z_2 candidate is reconstructed from two electrons, the contribution of the 3P1F control region is positive, and amounts to typically 30 % of the total predicted background.

For channels with loose muons (4μ and $2e2\mu$), the 3P1F sample is rather well described by the prediction from 2P2F, as seen in the right panel of Fig. 6.11.

The yields of data-driven estimate of the reducible background in the $ZZjj$ signal region are given in Table 6.1. This estimate is affected by statistical and systematic uncertainties inherent to the method. The former are dominated by the statistics in the

control regions and indicated in Table 6.1. The systematic uncertainty arises from the difference in the background composition in the control region and the $Z + \ell_{\text{loose}}$ region used to estimate the fake ratio. This uncertainty is estimated by calculating the fake ratios on a per-process basis in the simulation and performing the background estimate on each process individually before summing the total. A final systematic uncertainty on the background yield estimate of 30 %*to*40 % is obtained, based on the observed differences in the yield estimates obtained from the two fake ratios.

Table 6.1: Estimated contribution of the reducible background in the $ZZjj$ signal region obtained from the opposite-sign method. The estimate is based on the full 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. The quoted uncertainties reflect the statistical uncertainties in the control regions.

	$4e$	4μ	$2e2\mu$	$2\mu2e$	4ℓ
$ZZjj$ selection	1.68 ± 0.49	1.97 ± 0.43	1.45 ± 0.43	1.93 ± 0.50	7.02 ± 0.96

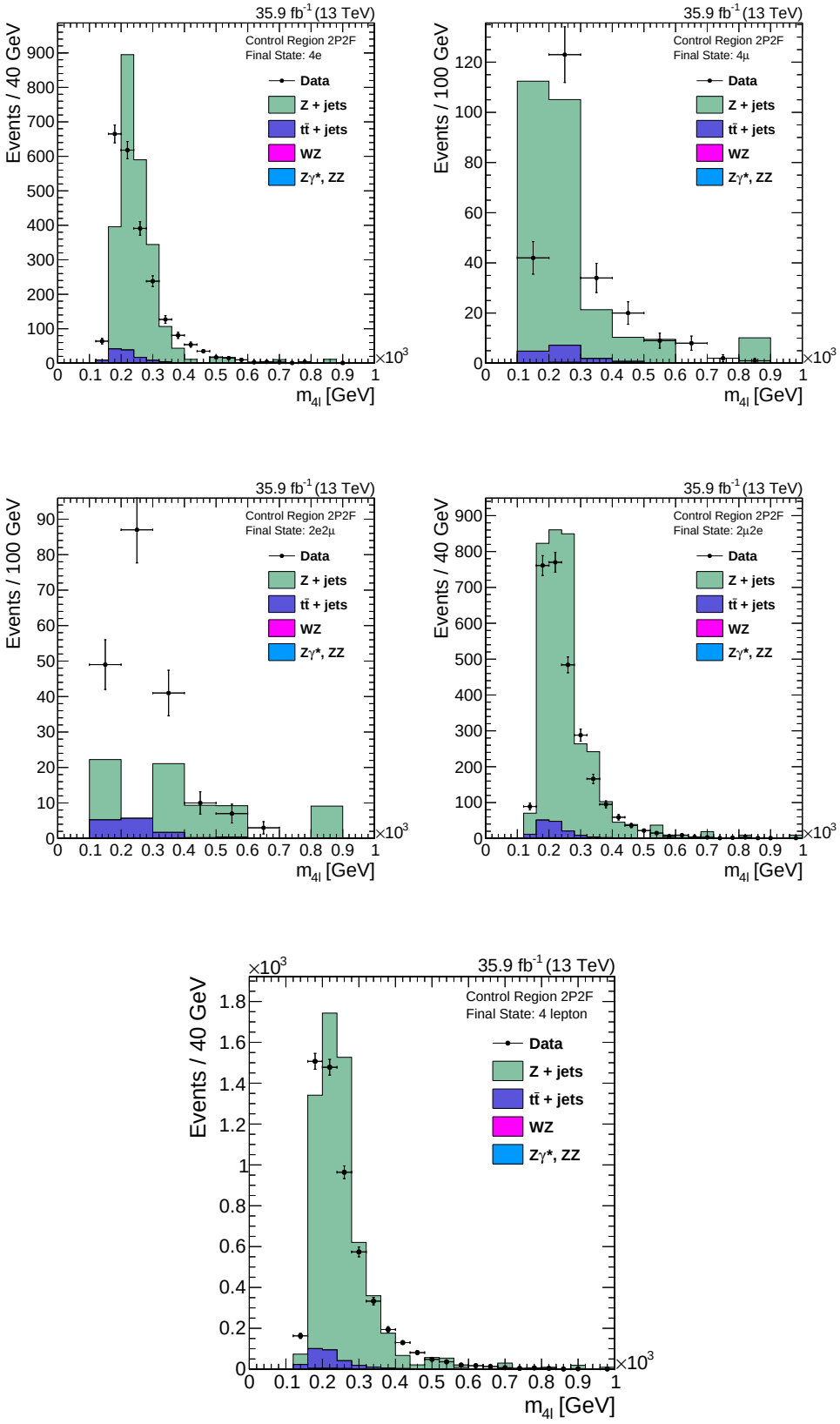


Figure 6.10: Invariant mass distribution of the events selected in the 2P2F control sample, for the $4e$ (top left), 4μ (top right), $2e2\mu$ (center left), $2\mu2e$ (center right), and 4ℓ (bottom) final states. The full 2016 dataset with an integrated luminosity of 35.9 fb^{-1} is used.

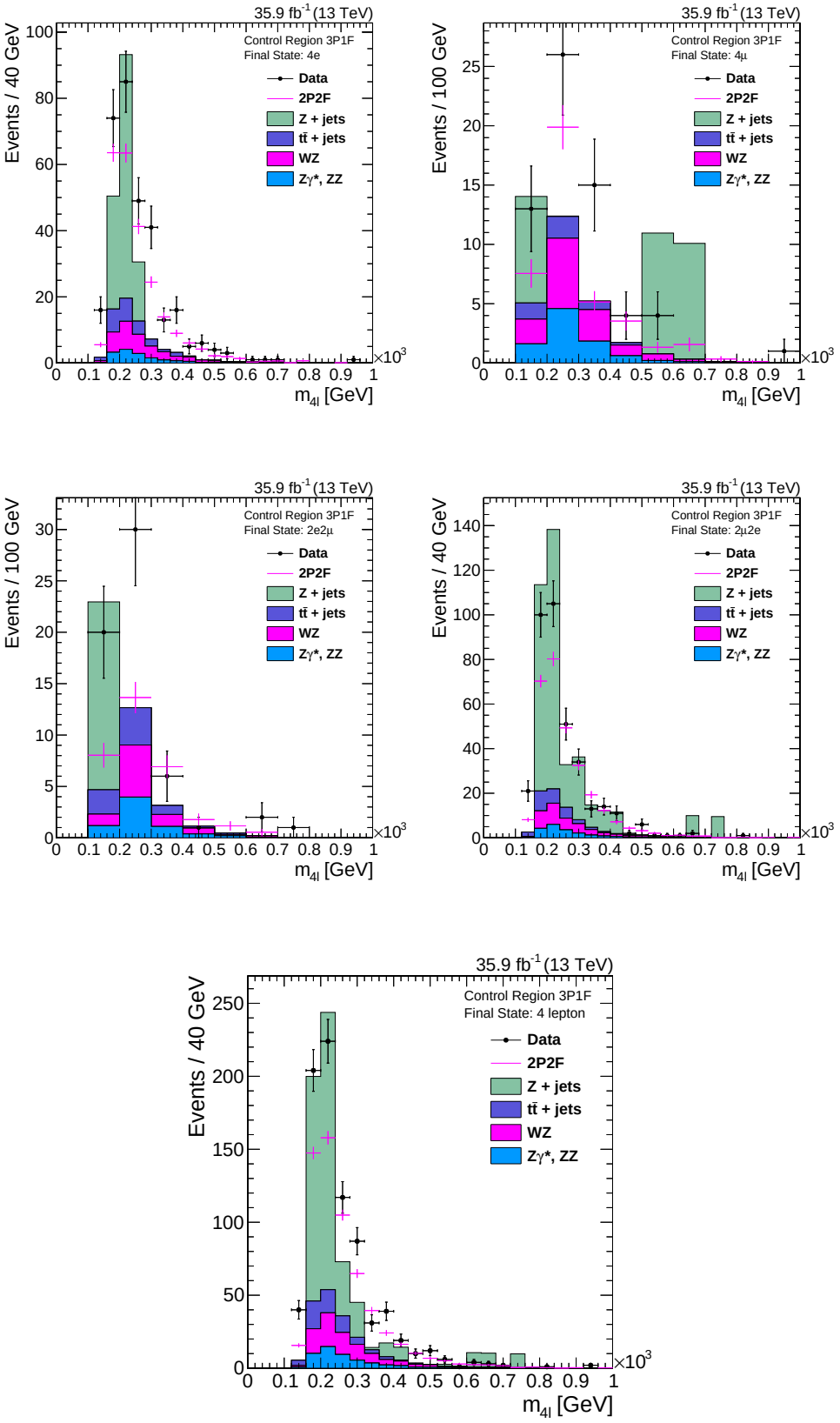


Figure 6.11: Invariant mass distribution of the events in the 3P1F control region, for the $4e$ (top left), 4μ (top right), $2e2\mu$ (center left), $2\mu2e$ (center right), and 4ℓ (bottom) final states. The full 2016 dataset with an integrated luminosity of 35.9 fb^{-1} is used.

Chapter 7

Signal extraction and systematic uncertainties

The electroweak production of the $\ell\ell\ell'\ell' jj$ ($\ell, \ell' = e$ or μ) final state is a sub-femtobarn process which needs to be separated from an irreducible QCD background which has a cross section that is more than one order of magnitude larger. To this end, a multivariate discriminator based on the kinematics of the final state objects is developed and systematically optimized in this chapter. The signal extraction method presented here has been devised to make optimal use of the few expected events while simultaneously reducing the systematic uncertainty on the QCD background normalization. The analysis presented in this thesis is the first multivariate search for vector boson scattering at the LHC.

7.1 Signal extraction strategy

The fully leptonic final state of the ZZ channel investigated in this thesis provides a clean signature and a fully-reconstructed final state. The latter provides a powerful handle to discriminate the irreducible QCD background and the spin correlations of the fermions are sensitive to the longitudinal polarizations of the bosons. However, the branching ratio of a Z boson into electrons or muons amounts to just 6.7 %, and the electroweak production of the $\ell\ell\ell'\ell' jj$ ($\ell, \ell' = e$ or μ) final state is a sub-femtobarn process. The fiducial cross section in the detector acceptance is about 0.3 fb and the efficiency of the $ZZjj$ event selection, which is dominated by the lepton reconstruction and selection efficiencies, is about 65 %. The expected signal yield for the 2016 dataset of 35.9 fb^{-1} is thus only 6.2 events. Considering such signal statistics, it is desirable to exploit the maximum amount of information and to use all events in the statistical analysis of the data.

The dominant background in this analysis is the QCD-induced production of the $ZZjj$ final state. Its cross section is about 15 times larger than the electroweak signal, but it populates a different phase-space. The challenge is thus to identify regions of phase-space with high and constant signal purity. This analysis exploits a multivariate discriminant to construct a one-dimensional distribution which provides such a stratification.

The discriminant used in the analysis is a boosted decision tree (BDT), which has been systematically optimized. The observables exploited in the BDT are chosen with care,

taking into account the reliability of the theory modeling and the experimental accuracy. The performance of the BDT is found to be equal to a discriminator based on matrix elements. The modeling of the QCD background is cross-checked in a QCD-enriched control region, finding good agreement.

The BDT score distributions of the signal and the backgrounds are used in a fit to all $ZZjj$ events observed in the data. The signal extraction fit includes the background-enriched part of the BDT spectrum which allows to constrain the normalization of the irreducible background and to reduce its uncertainty.

7.2 Signal kinematics and cut-based significance estimate

The hallmark signs of the VBS process are the tagging jets which allow to separate the electroweak from the QCD-induced production. In the electroweak production process each quark inside the colliding protons radiates off a vector boson which then interact. The quarks recoil against the vector bosons with little transverse momentum. The signal is thus characterized by two jets in opposite hemispheres of the detector ($\eta_{j_1} \times \eta_{j_2} < 0$) resulting in a large pseudorapidity separation between the jets ($\Delta\eta_{jj}$) and a large dijet invariant mass (m_{jj}). The vector bosons and their decay products tend to be in the central region of the detector, between the tagging jets.

Figure 7.1 shows some of the commonly studied observables related to the separation of VBS processes from the irreducible QCD background. The observables are defined in Table 7.1. The distributions show simulated events in the $ZZjj$ selection that requires two on-shell Z bosons and an invariant mass of the leading and subleading jet to satisfy $m_{jj} > 100\,\text{GeV}$ (see Section 6.2). In Fig. 7.1 and the remainder of this chapter the distributions and numbers on the background only include the leading irreducible QCD background, which contributes about 70 % of the total background. The reducible background, non-ZZ backgrounds and the gluon-loop induced processes are kinematically similar to the leading QCD background and no separate treatment is needed.

A simple selection to enhance the electroweak signal can be defined based on the dijet observables m_{jj} and $\Delta\eta_{jj}$ as these variables provide most of the separation power. The intersection of the signal and background distributions in Fig. 7.1 suggest a reasonable selection of $|\Delta\eta_{jj}| > 2.4$ and $m_{jj} > 400\,\text{GeV}$, which was already defined as the VBS selection in Section 6.2.

The VBS selection results in a signal efficiency of 65 % and a background efficiency of 13 %. The cut-based selection defined this way is close to optimal in terms of the ex-

Table 7.1: Common observables used to select the VBS signal.

Observable	Definition
m_{jj}	invariant mass of the tagging jets
$\Delta\eta_{jj}$	separation of the tagging jets in the η plane
$m_{4\ell}$	invariant mass of the diboson system; the scattering energy $m_{4\ell} = \sqrt{s}$
$\eta^{\star}_{Z_i}$	Zeppenfeld variable: direction of the Z_i boson relative to the tagging jets;
$\eta_{j_1} \times \eta_{j_2}$	product of the pseudorapidities of the tagging jets
$\Delta\varphi(Z_1, Z_2)$	azimuthal angle between the Z bosons

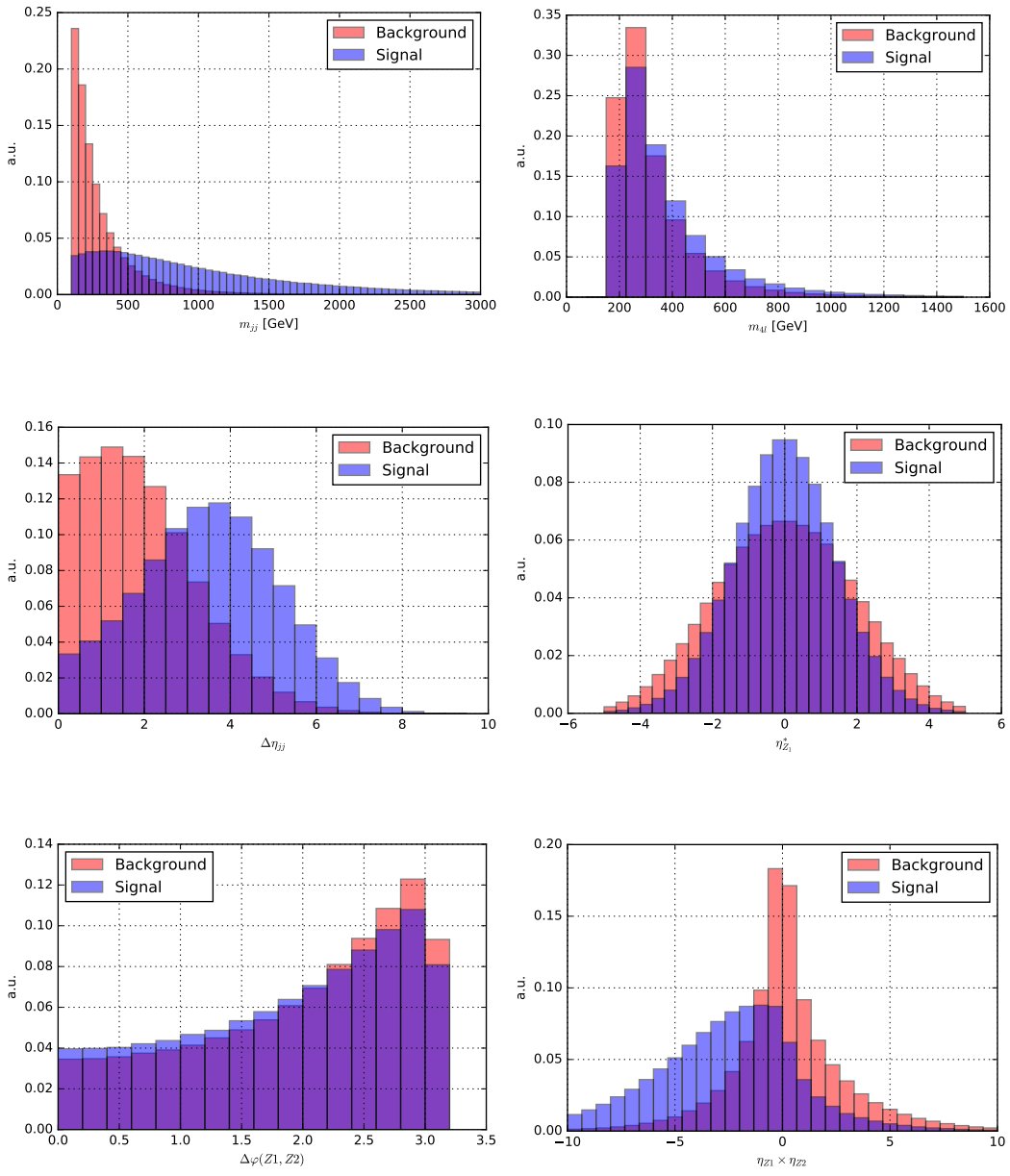


Figure 7.1: Common VBS observables applied to the $ZZjj$ channel as obtained from the simulation. Distributions are shown for events passing the $ZZjj$ selection which requires $m_{jj} > 100\,\text{GeV}$.

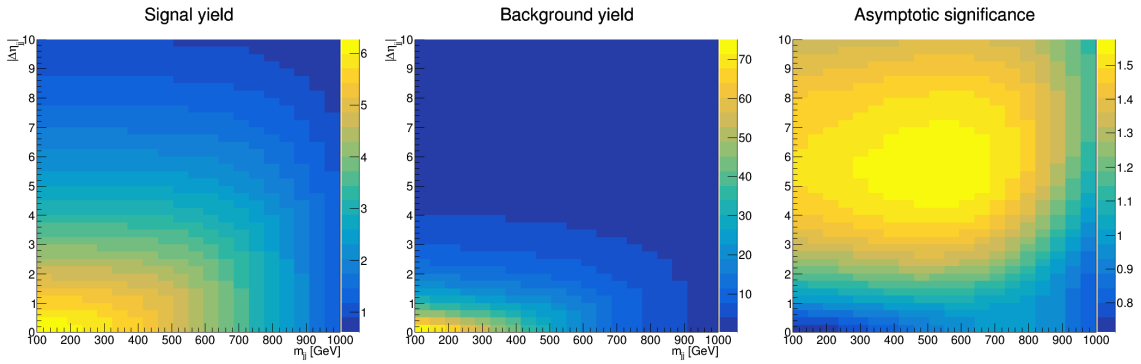


Figure 7.2: Expected yields of the electroweak signal (left), the dominant irreducible background (center), the expected asymptotic significance (right) for $\mathcal{L} = 36.5\,\text{fb}^{-1}$ as a function of the minimum m_{jj} and $|\Delta\eta_{jj}|$ cut values.

pected signal significance of a cut-based selection using only m_{jj} and $|\Delta\eta_{jj}|$. Figure 7.2 shows the expected yields of the electroweak signal (left), the dominant QCD background (center), and the asymptotic significance (right) for $\mathcal{L} = 35.9\,\text{fb}^{-1}$ as a function of the minimum m_{jj} and $|\Delta\eta_{jj}|$ cut values. The asymptotic significance is calculated as

$$s = \sqrt{2[(N_b + N_s)\log(1 + N_s/N_b) - N_s]}, \tag{7.1}$$

where N_s and N_b are the expected number of events for the signal and background respectively.

The cut-based selection defined above achieves an asymptotic significance of 1.3 standard deviations with an expected signal yield of 4 events. Higher significances are feasible, notably by increasing the minimum dijet pseudorapidity separation. A selection of $m_{jj} > 400\,\text{GeV}$ and $|\Delta\eta_{jj}| > 5$ achieves an expected asymptotic significance of 1.6 standard deviations. However, the expected signal yield for such a selection is only 2.4 events and results in a much higher statistical uncertainty. A cut-based analysis with such low expected yields would furthermore raise the issue on whether the result should be interpreted as a measurement or as a limit.

The cut-based VBS selection is used in the analysis to define a fiducial volume and for a cross check of the MVA-based signal significance determination. The complement of this VBS selection, that is events that pass the $ZZjj$ selection but fail either the $m_{jj} > 400\,\text{GeV}$ or the $|\Delta\eta_{jj}| > 2.4$ requirement, is used to define a background-enriched control region to validate the modelling of the QCD background provided by the simulation, see Section 7.3.5.

7.3 Development of the multivariate discriminant

The multivariate discriminant used to separate the electroweak signal from the backgrounds is a gradient boosted decision tree. This section presents the development and optimization of the BDT and the comparison to a discriminant based on matrix elements.

The BDT is trained and optimized using the Python scikit-learn library [77]. A first training of the BDT based on the default hyper-parameters of the scikit-learn BDT im-

plementation and the VBS-sensitive observables illustrated in Fig. 7.1 is used to establish a baseline. Defining a working point that corresponds to a 65 % signal efficiency, this first BDT achieves a background efficiency of 8 %, i.e., an appreciable reduction of the 13 % background efficiency of the VBS cut-based selection.

Having established a first implementation of the BDT, its tunable parameters and the choice of input variables are systematically optimized.

7.3.1 Scan of the hyper-parameters

The construction algorithm of the decision trees and the boosting procedure feature tunable parameters (*hyper-parameters*) that impact the effectiveness of the resulting BDT. The BDT can thus be optimized by varying these hyper-parameters, retraining the BDT, and evaluating the change in a suitable performance metric. The area under the ROC curve is used here and the hyper-parameters are varied on a grid of configurations. Table 7.2 lists the default hyper-parameters, the range of values explored in the grid search for each, and the optimal configuration identified in the parameter scan.

Table 7.2: Hyper-parameters considered in the BDT optimization. Parameter names are those of the scikit-learn library.

Parameter	Default value	Range in grid search	optimal value
n_estimators	100	[800, 1000, 1200]	800
learning_rate	0.1	[0.02, 0.01]	0.01
max_depth	3	[8, 7, 6, 5]	8
min_samples_leaf	1	[800, 1000]	800

7.3.2 Feature selection

Many observables have been proposed in the literature to separate the VBS process from the irreducible backgrounds. The goal of the optimization study is to identify performant observables and to evaluate the correlations between different sets of observables that have been proposed in phenomenological studies on VBS. Table 7.3 lists the variables considered in the optimization of the BDT used in this analysis. The study proceeds by adding each group of variables on top of the variables already listed in Table 7.1 for a BDT with the hyper-parameters determined in the previous subsection.

The second and third group of observables in Table 7.3 are the production and decay angles respectively. The decay angles in particular have been used in the MELA discriminant in the $H \to ZZ^* \to 4\ell$ analysis [78–80] and their definitions are illustrated in the left panel of Fig. 7.3. After a boost into the rest frame of the $ZZjj$ system, one can determine the four-vectors of the incoming partons in the ZZ center of mass frame by exploiting momentum conservation. The production angles are then defined analogously to the decay angles in Fig. 7.3, taking the Z bosons as the incoming particles (replacing the protons) and considering the jets as the fermions. These angles provide an alternative way to parametrize the kinematics of the scattering process and can thus be used to separate the signal from the background. The decay angle ϑ_1 and the production angle ϑ^* in particular are sensitive to the fermion spin and have been

Table 7.3: Observables sensitive to the differences between the electroweak- and QCD-induced production of the $ZZjj$ final state. All observables are considered in the BDT optimization as described in the text.

Observable type	Observable	Definition
classic VBS	m_{jj}	invariant mass of the tagging jets
	$\Delta\eta_{jj}$	separation of the tagging jets in the η plane
	m_{4l}	invariant mass of the diboson system; $m_{4l} = \sqrt{s}$ of the vector boson interaction
	$\eta^{\star}_{Z_i}$	η Zeppenfeld variable; direction of the Z_i boson relative to the tagging jets;
$\eta_{j_1} \times \eta_{j_2}$	product of the pseudorapidities of the tagging jets	
decay angles in the ZZ c.o.m. frame	$cos(\vartheta*)$	angle between the Z_1 boson and the z axis
	$cos(\vartheta_1)$	angle between the fermion and the boost vector of the Z_1 boson
	$cos(\vartheta_2)$	angle between the fermion and the boost vector of the Z_2 boson
	φ	angle between the normal vectors of the decay planes of both Z bosons
	φ_1	angle between the normal vectors of the Z_1 decay plane and zz' plane
production angles in the ZZ c.o.m. frame	$cos(\vartheta*)$	angle between the Z_1 boson and the z axis
	$cos(\vartheta_1)$	angle between the leading jet and the boost vector of the V_1 boson
	$cos(\vartheta_2)$	angle between the subleading leading jet and the boost vector of the V_2 boson
	φ	angle between the normal vectors of the decay planes of both "incoming bosons"
	φ_1	angle between the normal vectors of the V_1 "decay" plane and zz' plane
	q_{V1}	invariant mass of the incoming vector boson V_1
	q_{V2}	invariant mass of the incoming vector boson V_2
hadronic activity	N_j	total number of jets in event
	$\Sigma\lvert p_T\rvert$	scalar sum of non-tagging jet p_{T}
	N_j^c	total number of central ($\lvert\eta\rvert < 2.4$) jets in event
	$\Sigma\lvert p_T^c\rvert$	scalar sum of central non-tagging jet p_{T}
other observables	$\max \eta_{4\ell}$	maximal lepton eta
	$\min \eta_{tj}$	minimal tagging jet η
	$\max \eta_{tj}$	minimal tagging jet η
	η_{tj}	η of leading and subleading jet
	p_{T}^{tj}	p_{T} of leading and subleading jet
	η_{tj}^{sum}	sum of tagging jet η
	$p_{\mathrm{T}}^{\mathrm{product}}$	$m_{jj}/\Delta\eta$
	$\Delta\varphi(Z_1, Z_2)$	angular distance between the Z bosons in the φ plane
	$R(p_{\mathrm{T}}^{\mathrm{hard}})$ or $p_{\mathrm{T}}^{rel.hard}$	$\sum_{Z_{1,2},j_{1,2}} \vec{p}^{\,i}\lvert_{transverse} / \sum_{Z_{1,2},j_{1,2}} p_T^i$
	$R(p_{\mathrm{T}}^{\mathrm{jets}})$ or $p_{\mathrm{T}}^{rel.jets}$	$\sum_{j_{1,2}} \vec{p}^{\,i}\lvert_{transverse} / \sum_{j_{1,2}} p_T^i$
Quark-gluon tagging	$\mathcal{L}_{qg}^i$	Quark-gluon tagger likelihood of both tagging jets

proposed in the literature to separate the longitudinal from the transverse scattering component [81].

A BDT trained using only the production and decay angles as inputs achieves a separation power similar to the cut-based approach as shown in the right panel of Fig. 7.3. However, no appreciable gain is achieved when including these observables along the classic set of VBS observables, indicating the redundancy in information between the two sets. While not exploited in this analysis, these angular variables are suitable to enhance the longitudinal component within the electroweak production and should thus be reconsidered once the electroweak phase-space can be explored in more detail.

The fourth set of observables in Table 7.3 is motivated by the suppression of central hadronic activity due to color decoherence in the electroweak production. Minor improvements are achieved by adding these observables in the BDT training, but the disagreement between different simulations (see Fig. 5.9) discourage using these poorly modeled observables. The fifth set of observables in Table 7.3 includes variables of individual tagging jet kinematics and observables sensitive to the angular correlations between the tagging jets and the vector bosons. The relative transverse momentum fractions $R(p_{\mathrm{T}}^{\mathrm{hard}})$ and $R(p_{\mathrm{T}}^{\mathrm{jets}})$ are found to increase the separation power.

Finally, the scores of a quark-gluon likelihood discriminator for both tagging jets are added to the BDT training. The jets in the electroweak signal are expected to originate from quarks, while the QCD background also features gluon-initiated jets. However, the kinematics of the tagging jets severely limit the effectiveness of quark-gluon likelihood discriminator, which exploits information on the hadron multiplicity in the jet. For tagging jets beyond $|\eta| = 2.4$, i.e., the region relevant for the VBS signal, only the calorimeter and no tracking information is available, which limits the usefulness of

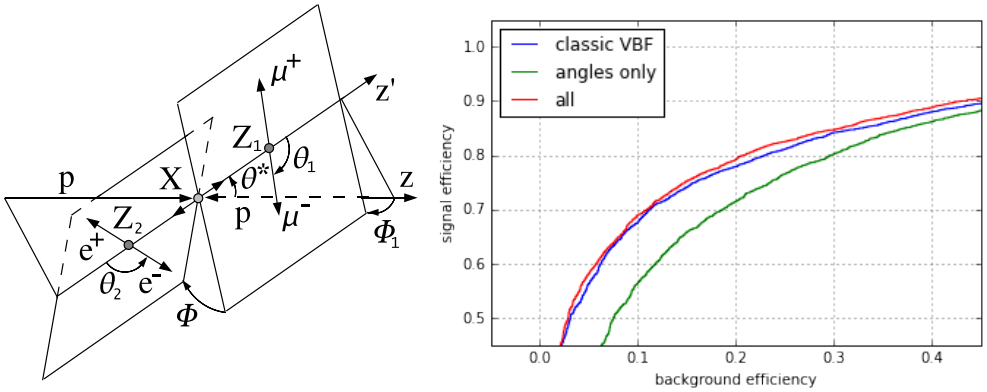


Figure 7.3: Definitions of the decay angles in the $ZZ \to 4\ell$ system (left) [78]. The production angles in the third row of Table 7.3 are obtained by replacing the incoming protons with the outgoing Z bosons and the final state fermions with the hadronic jets. The right panel shows the ROC curves of three BDTs, trained with the "classic" set of VBS variables, only the production and decay angles, and with both sets of variables added as inputs (labelled "all" in the figure).

the discriminator for this study. Including the quark-gluon discriminator values in the BDT only increases the expected significance by about 0.2 standard deviations. Given the considerable modeling uncertainties associated to the quark-gluon tagging, particularly in the forward detector region, these variables are not retained for the final BDT.

To exclude the possibility of separation power coming mostly from the correlations between variables of different sets in Table 7.3, the BDT is also trained with all variables included, resulting in no appreciable gain beyond adding the set of "other variables".

The number of input variables after this first round of optimization was 17. In order to reduce the number of observables to a minimal set, i.e., to include only those variables needed to achieve an optimal separation, this BDT was retrained 17 times, for each training dropping one after the other each input variable. By repeating this procedure of identifying the least performant observable and dropping it, the final list of observables is reduced to 7, summarized in Table 7.4.

Table 7.4: List of input observables retained for the final MVA.

Observable	Definition
m_{jj}	invariant mass of the tagging jets
$\Delta\eta_{jj}$	separation of the tagging jets in the η plane
$m_{4\ell}$	invariant mass of the diboson system
$\eta^{\star}_{Z_1}$	η Zeppenfeld variable of Z_1
$\eta^{\star}_{Z_2}$	η Zeppenfeld variable of Z_2
$R(p_{\mathrm{T}}^{\mathrm{hard}})$ or $p_{\mathrm{T}}^{rel.hard}$	$\sum_{Z_{1,2},j_{1,2}} \vec{p^i}\vert_{transverse} / \sum_{Z_{1,2},j_{1,2}} p^i_{\mathrm{T}}$
$R(p_{\mathrm{T}}^{\mathrm{jets}})$ or $p_{\mathrm{T}}^{rel.jets}$	$\sum_{j_{1,2}} \vec{p^i}\vert_{transverse} / \sum_{j_{1,2}} p^i_{\mathrm{T}}$

The expected significance of the final BDT is about 0.2 standard deviations lower than the BDT trained with all input variables. Given the overall small improvement and considering the appreciable modeling uncertainties introduced by the hadronic ob-

servables related to a third jet veto or the jet related production angles, the final BDT retains only the observables listed in Table 7.4.

7.3.3 Signal separation based on matrix elements

Following the optimizations described in the proceeding sections, it could still be possible that a better separation of the signal and background could be achieved by a better tuning the BDT parameters or by adding some observables not considered during the feature selection. To test this hypothesis, a simple matrix element-based discriminator (MED) is developed.

The MED is based on the value of the squared matrix elements (ME) of the electroweak signal $\mathcal{M}_{\mathrm{EW}}$ and the main QCD background $\mathcal{M}_{\mathrm{QCD}}$. The matrix elements are evaluated using the four-vectors of the selected leptons and tagging jets, with no smearing of their momenta nor taking into account any other detector effects. No integration or averaging on the phase-space is necessary as the final state is fully reconstructed. The flavor of the incoming and outgoing partons is unknown and the different ME are weighted by the parton distribution functions, where the momenta of the incoming partons are determined assuming momentum conservation and considering only the selected leptons and tagging jets as the final state particles. The technical implementation of this ME-based discriminant relies on matrix elements provided by MG5_aMC, which allows to export the Fortran code to calculate the ME. This code is used with a Python wrapper and interfaced with LHAPDF [82] for the PDF evaluation.

The final discriminant is taken as the log-ratio of $\mathcal{M}_{\mathrm{EW}}$ and $\mathcal{M}_{\mathrm{QCD}}$. Figure 7.4 compares the performance of the MED to the BDT. The ROC curves of the two discriminators are very similar, indicating that the BDT is optimal.

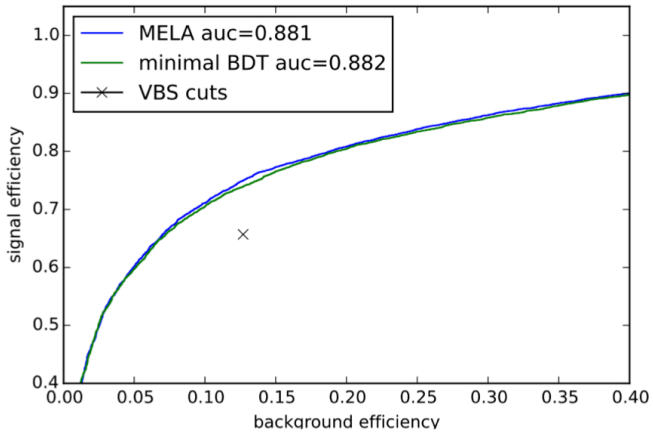


Figure 7.4: ROC curves of the BDT and the matrix element discriminator developed for this analysis, evaluated on the signal and leading-order background simulation. The efficiencies of the cut-based selection are indicated by the marker.

7.3.4 Final multivariate discriminant

The discriminant ultimately used in the analysis is the BDT trained with the hyper-parameters listed in Table 7.2 and using the input variables listed in Table 7.4.

In using a BDT, care needs to be take to avoid the bias coming from over-training, i.e., a difference in performance between the test and training datasets. Figure 7.5 shows the BDT output distributions for the training and test dataset of the BDT as well as the corresponding ROC curves. The BDT exhibits some minor overtraining. To avoid any residual effect in the statistical analysis, the signal templates are based on simulated events from the test set, i.e., on events that have not been used during the BDT training.

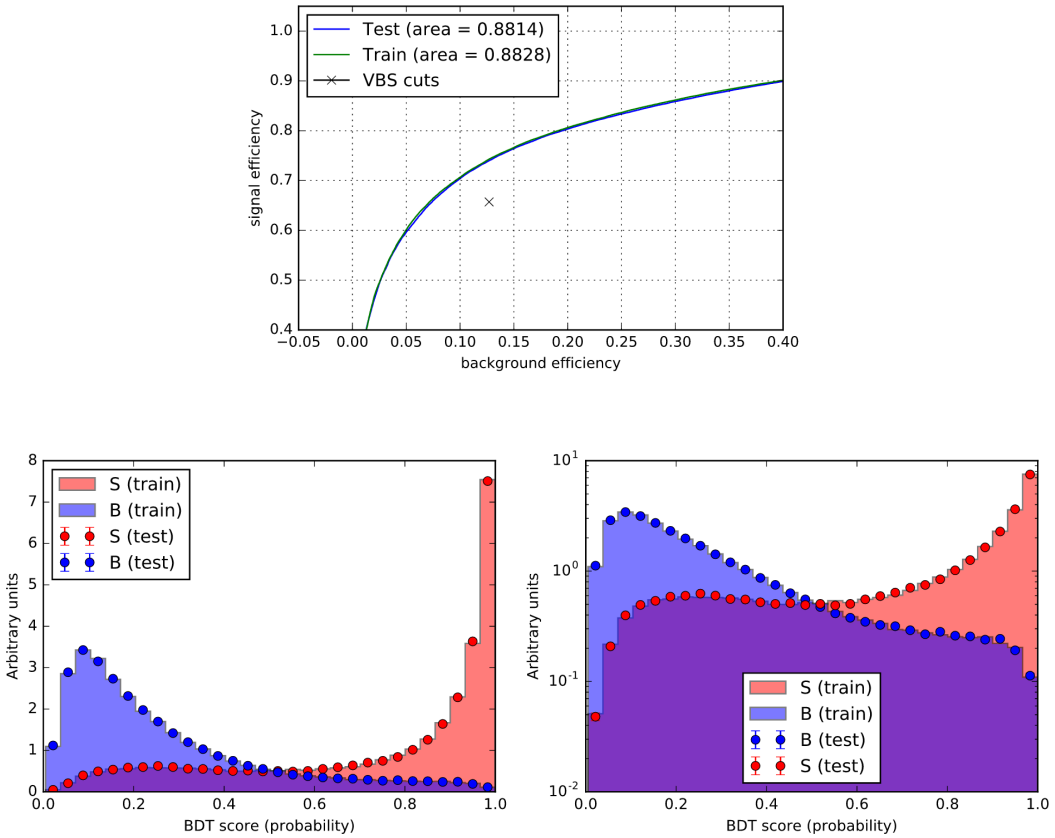


Figure 7.5: ROC curves (top row) and the BDT output distributions for the test and training sets (bottom row) from the simulation.

The final significance depends on the chosen binning of the BDT distribution and the functional form of the BDT score transformation. The latter takes the weighted output of the decision trees in the ensemble and projects it onto a given range. The TMVA default of $2/(2 - e^{-2x}) - 1$, which projects unto the range [-1, 1], is suboptimal as it requires a highly nonlinear binning in order to separate the most signal-enriched bins. The simple logistic transformation $1/(1 - e^{-x})$ provides much better results.

The final binning is then determined by requiring the statistical uncertainty on the background template in the most signal-rich bin to be less than 5 %. Figure 7.6 shows the resulting BDT spectrum with the optimized binning.

To illustrate the phase-space selected by the BDT, a working point that has the same

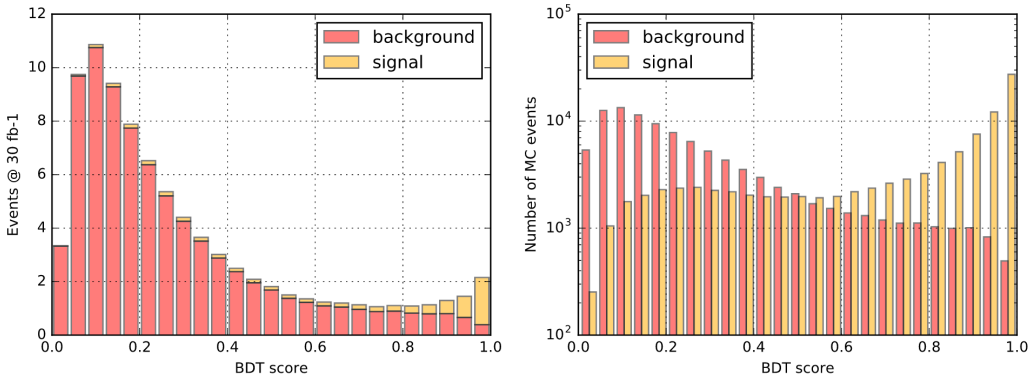


Figure 7.6: BDT distributions with the final binning for the signal and leading QCD background.

signal efficiency as the cut-based VBS selection (65 %) is defined. Figure 7.7 and Fig. 7.8 show the distributions of all events for the $ZZjj$ selection and for those selected by this working point of the BDT. As expected, the BDT selects the phase-space of large dijet masses and large dijet pseudorapidity separations with central Z bosons.

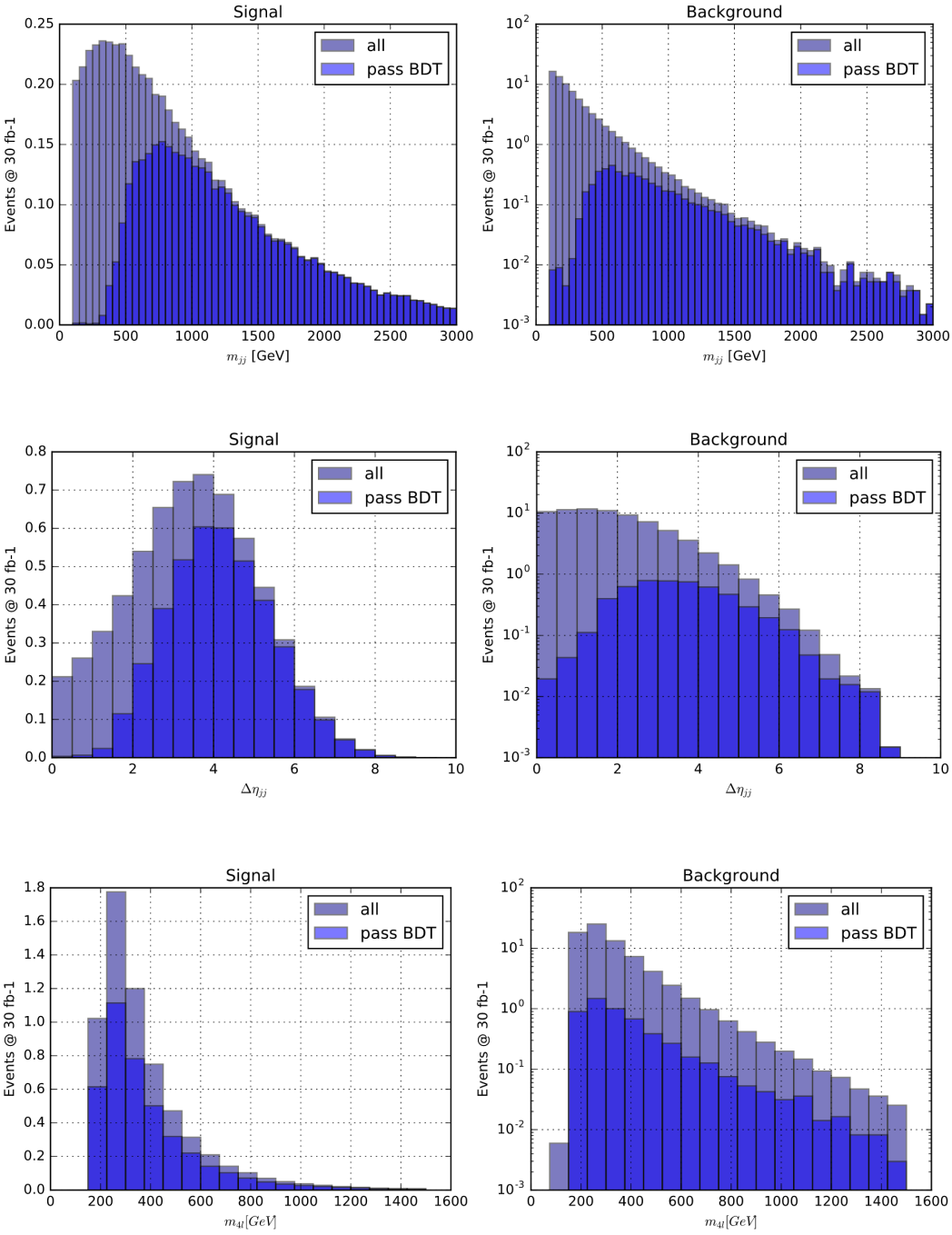


Figure 7.7: Distributions of the observables used in the BDT to select the electroweak signal for the signal (left) and the QCD background (right) in the leading-order simulation. The light blue histogram shows all simulated events in the $ZZjj$ selection and the dark blue histogram shows only events that are selected by the 65 % working point of the BDT described in the text.

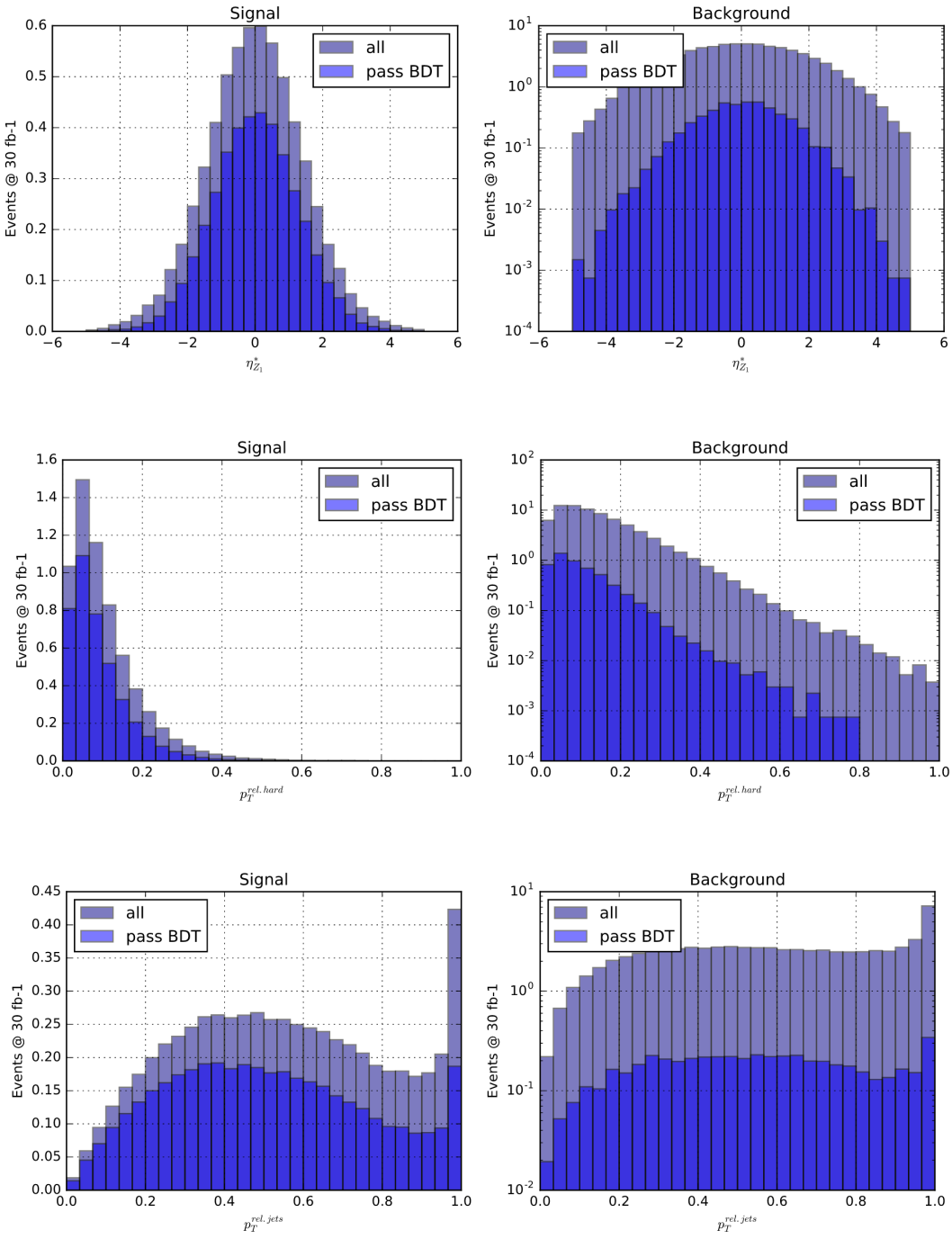


Figure 7.8: Continuation: Distributions of the observables used in the BDT to select the electroweak signal for the signal (left) and the QCD background (right) in the leading-order simulation. The light blue histogram shows all simulated events in the $ZZjj$ selection and the dark blue histogram shows only events that are selected by the 65 % working point of the BDT described in the text.

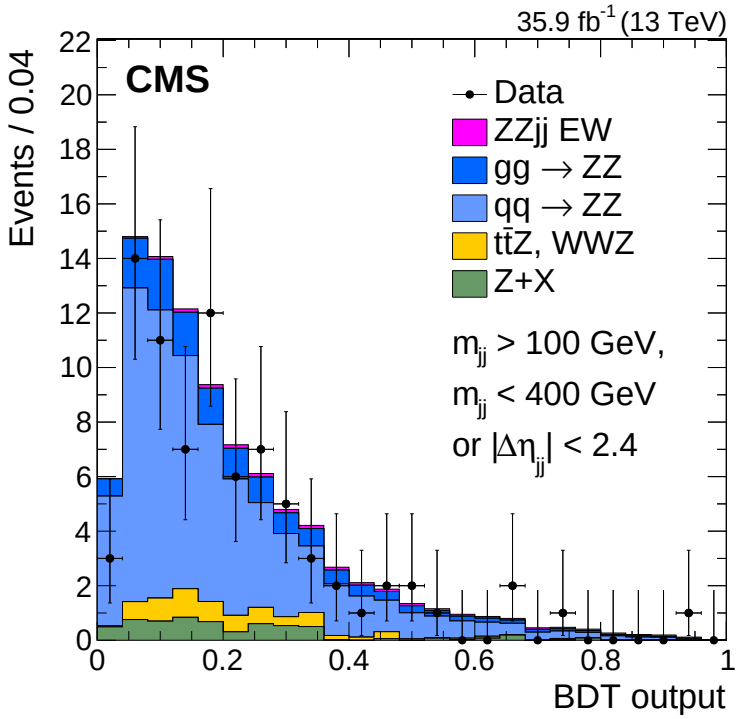


Figure 7.9: Distribution of the BDT score observed in the 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. All events satisfying the nVBS selection with $m_{jj} > 100\,\text{GeV}$ and ($m_{jj} < 400\,\text{GeV}$ or $|\Delta\eta_{jj}| < 2.4$) are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

7.3.5 Validation of background model and BDT in data

The cut-based nVBS event selection that requires $m_{jj} > 100\,\text{GeV}$ and ($m_{jj} < 400\,\text{GeV}$ or $|\Delta\eta_{jj}| < 2.4$) is used to define a background-enriched control region. The nVBS selection efficiently excludes the signal phase-space while maintaining sufficient statistics to compare the QCD modeling with the observed data.

Figure 7.9 shows the BDT spectrum with the 2016 data corresponding to an integrated luminosity of 35.9 fb^{-1}. The expectation from the simulation and the data-driven estimate of the reducible backgrounds are also shown. Figure 7.10 compares the same data and expectations for observables related to the Z boson (including the BDT input $m_{4\ell}$), while Fig. 7.11 shows the kinematics of the tagging jets. The remaining VBS observables exploited in the BDT are shown in Fig. 7.12.

The statistical analysis is performed on all events in the $ZZjj$ selection, i.e., the sum of the VBS and nVBS selections.

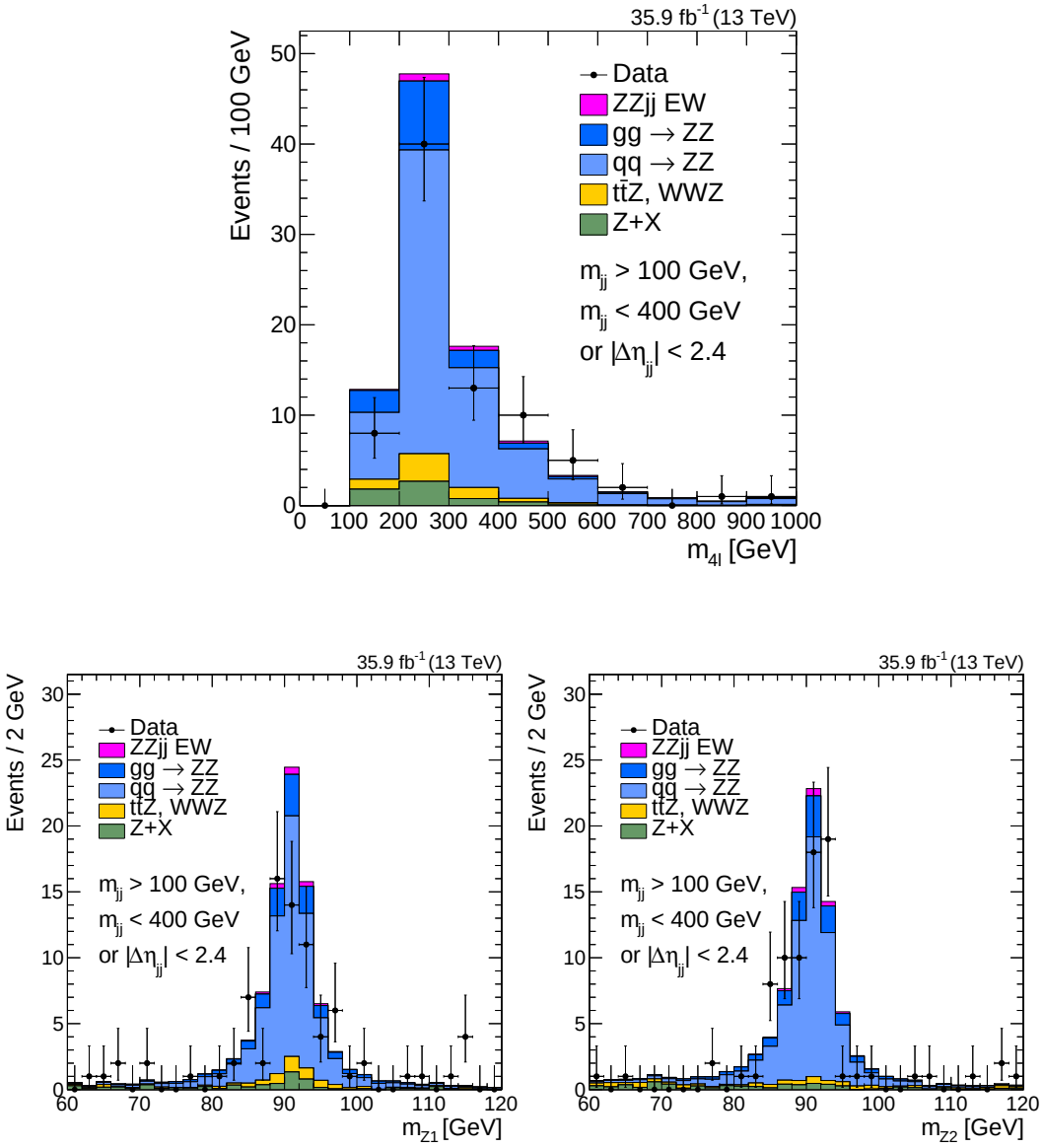


Figure 7.10: Distributions of the invariant masses of the Z boson pair (top), the leading (bottom left), and subleading (bottom right) Z boson observed in the 2016 dataset with an integrated luminosity of $35.9\,\mathrm{fb}^{-1}$. All events satisfying the nVBS selection with $m_{jj} > 100\,\mathrm{GeV}$ and ($m_{jj} < 400\,\mathrm{GeV}$ or $|\Delta\eta_{jj}| < 2.4$) are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

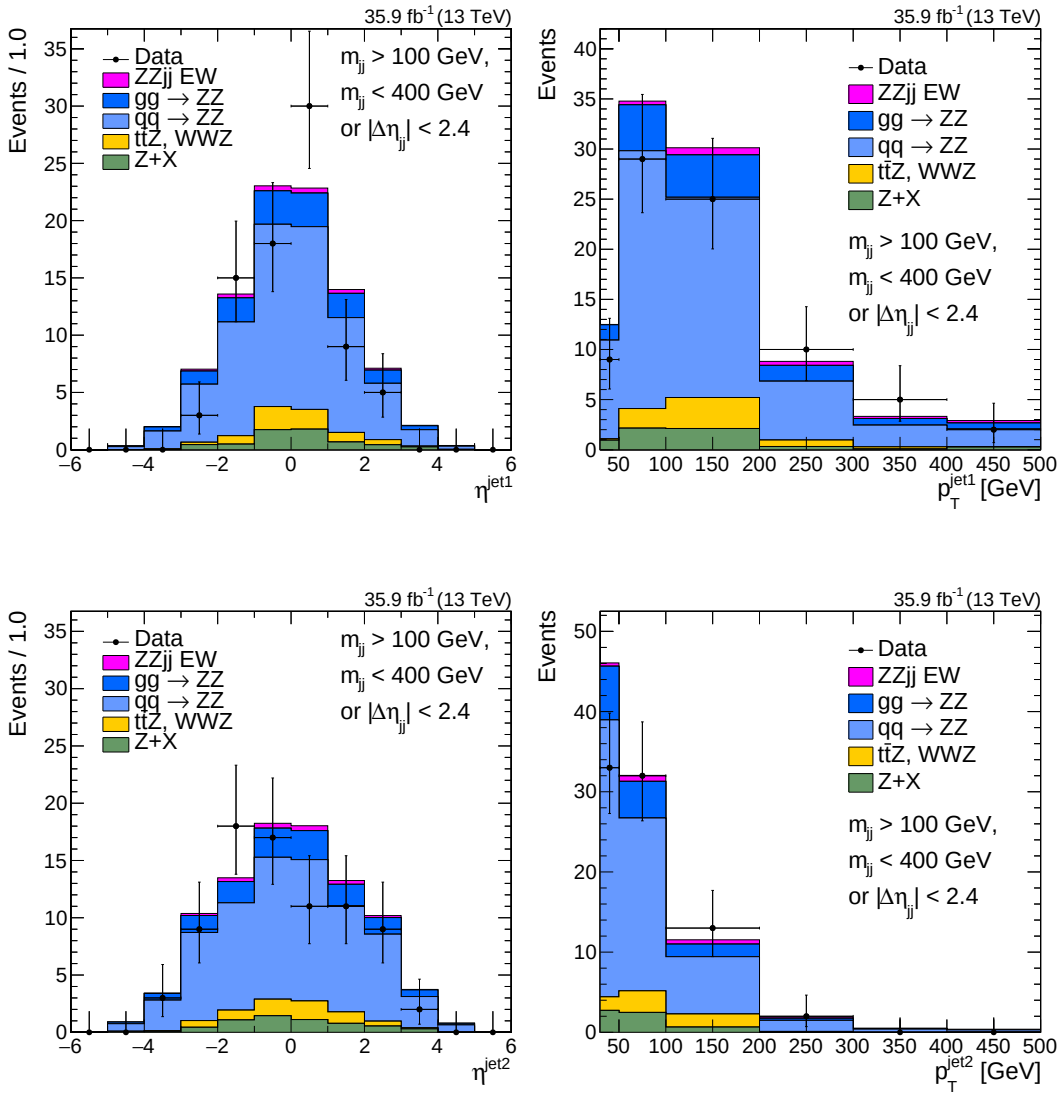


Figure 7.11: Distributions of the leading (top row) and subleading tagging jet kinematics observed in the 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. All events satisfying the nVBS selection with $m_{jj} > 100\,\mathrm{GeV}$ and ($m_{jj} < 400\,\mathrm{GeV}$ or $|\Delta\eta_{jj}| < 2.4$) are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

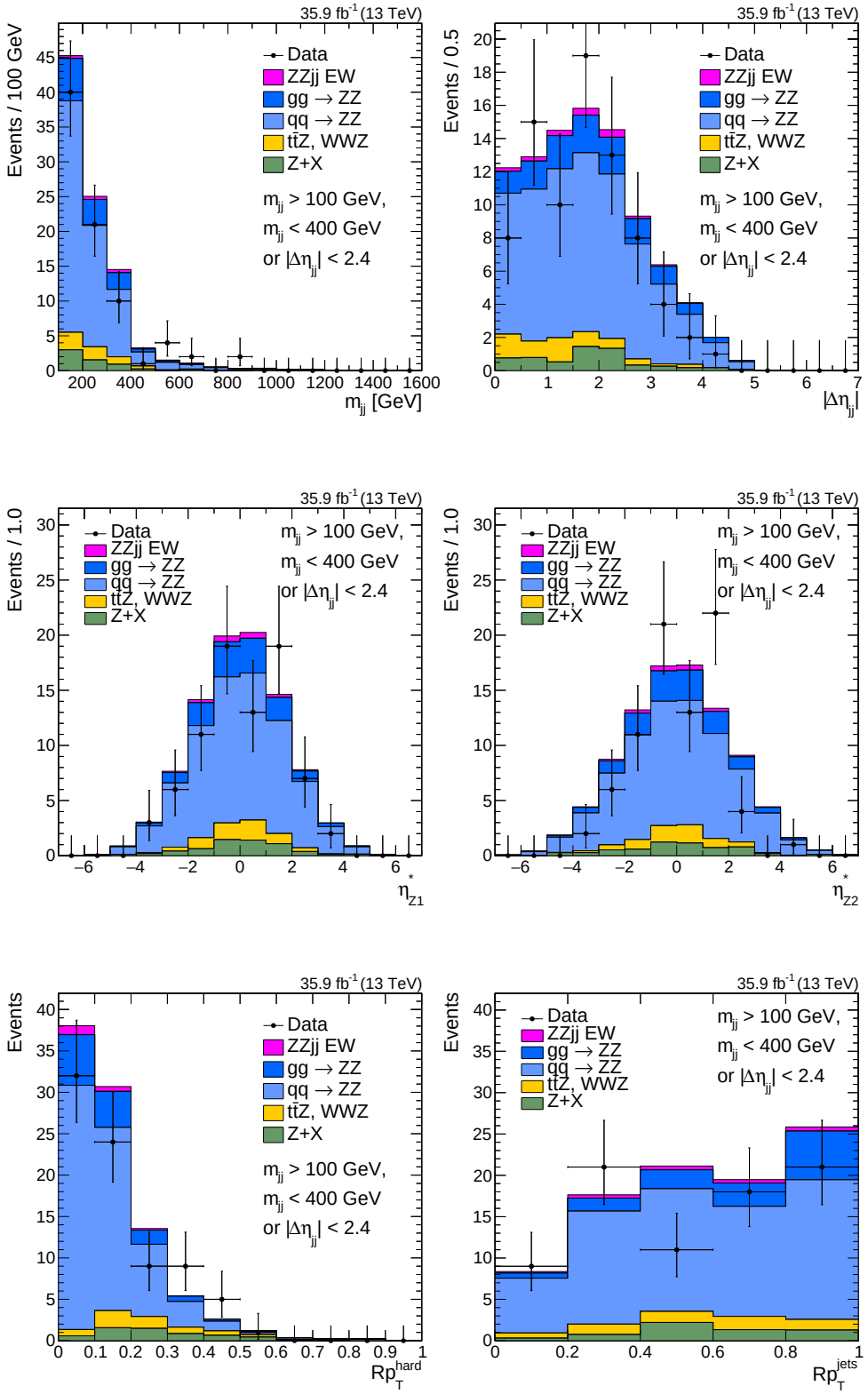


Figure 7.12: Distributions of the VBS observables used in the BDT observed in the 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. All events satisfying the nVBS selection with $m_{jj} > 100$ GeV and ($m_{jj} < 400$ GeV or $|\Delta\eta_{jj}| < 2.4$) are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

7.3.6 Data-simulation comparison of the BDT input variables

Observables relating to the ZZ system (Fig. 7.13) and the kinematics of the tagging jets (Fig. 7.14) in the $ZZjj$ event selection used in the statistical analysis are presented. The invariant mass of the ZZ system and the observables in Fig. 7.15 enter the BDT calculation.

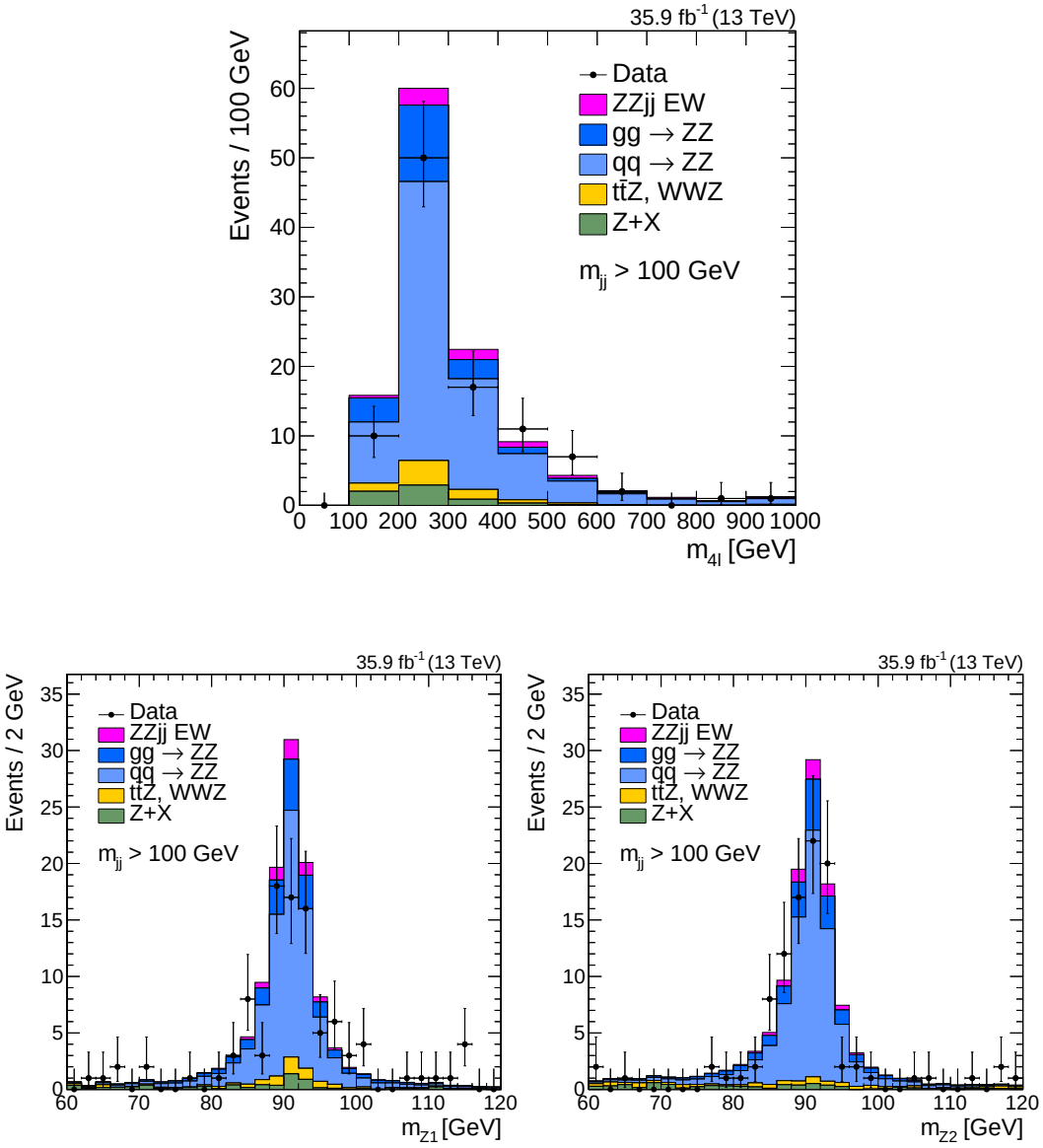


Figure 7.13: Distributions of the invariant masses of the Z boson pair (top), the leading (bottom left), and subleading (bottom right) Z boson observed in the 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. All events satisfying the $ZZjj$ selection with $m_{jj} > 100$ GeV are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

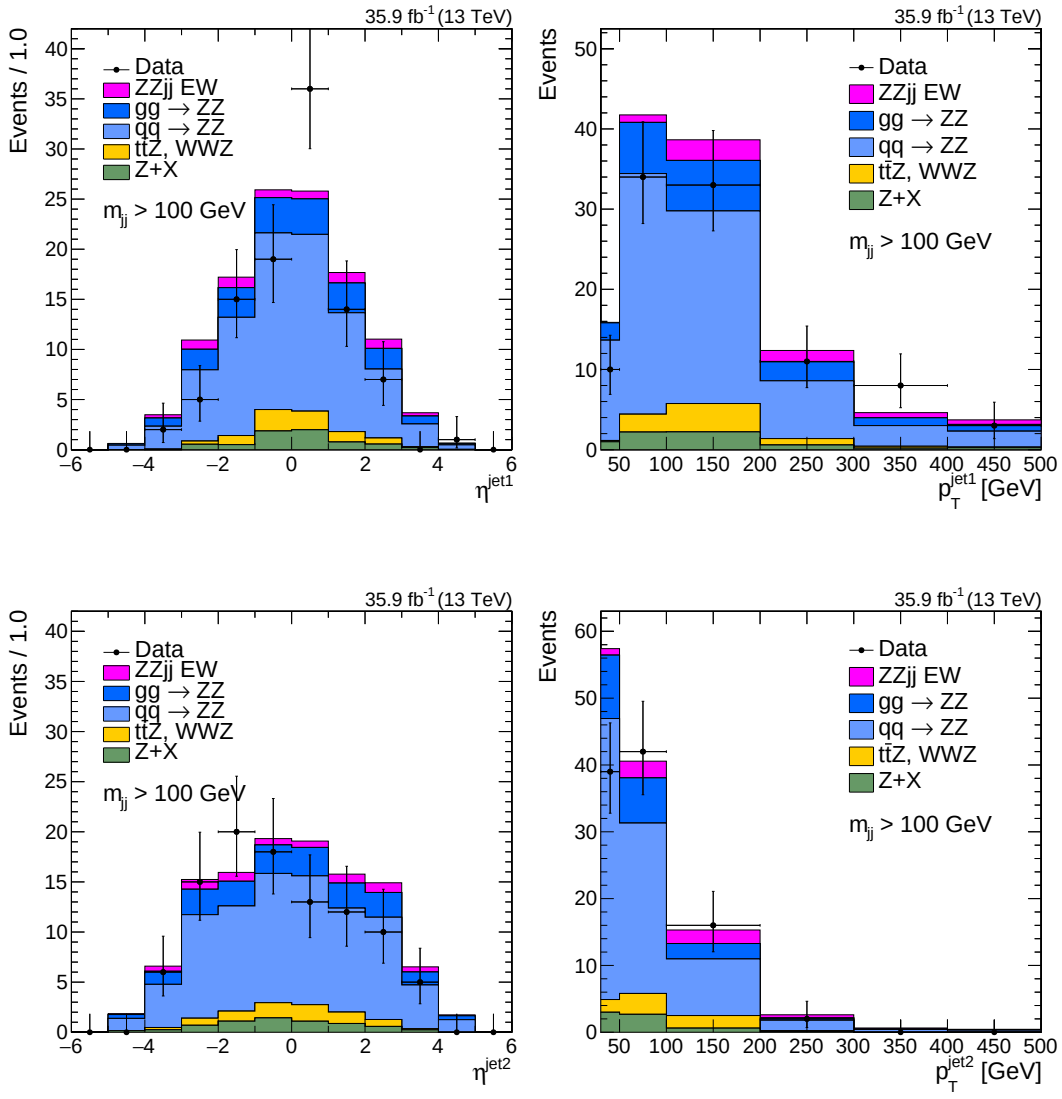


Figure 7.14: Distributions of the leading (top row) and subleading tagging jet kinematics observed in the 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. All events satisfying the $ZZjj$ selection with $m_{jj} > 100\,\text{GeV}$ are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

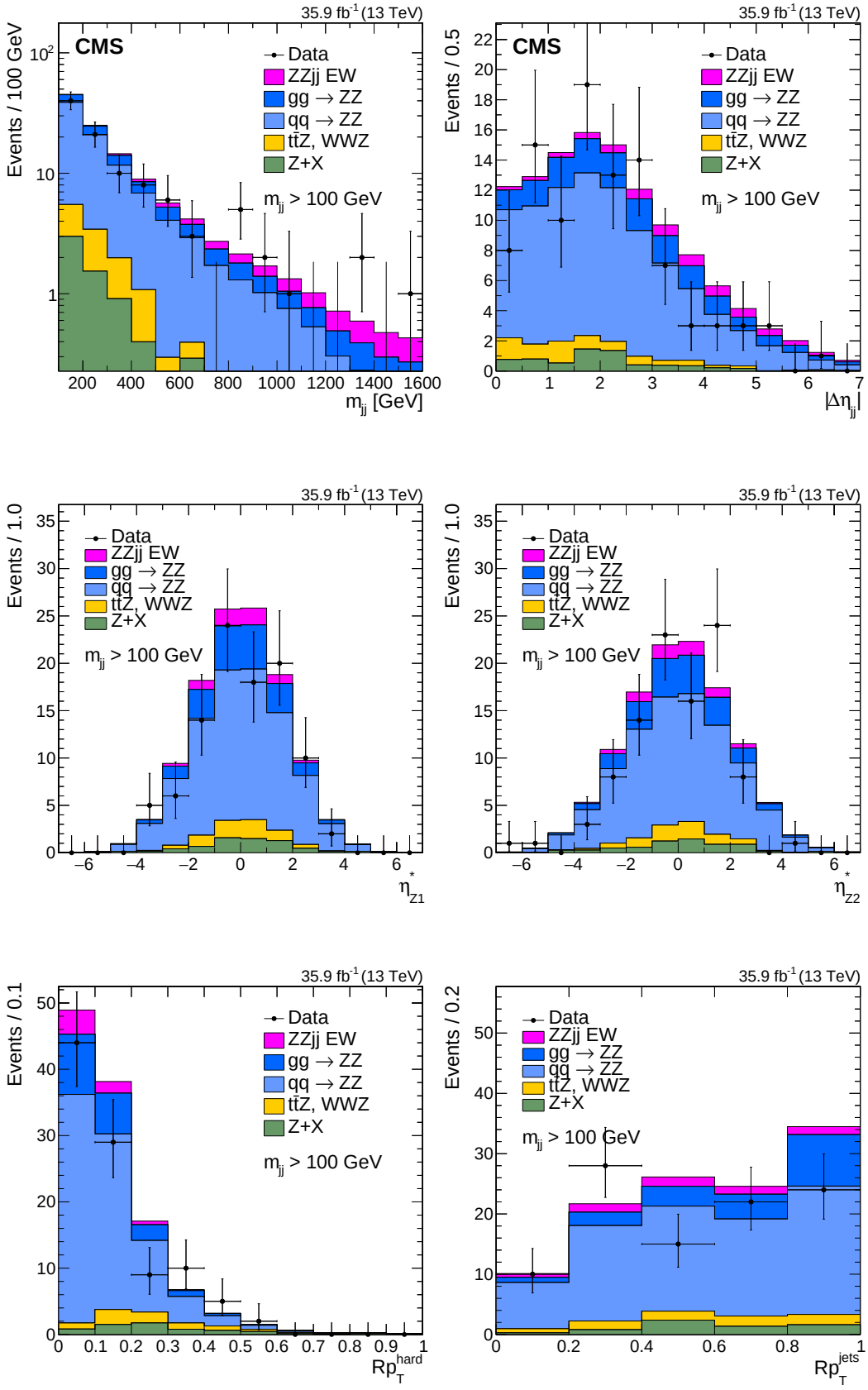


Figure 7.15: Distributions of the VBS observables used in the BDT observed in the 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. All events satisfying the $ZZjj$ selection with $m_{jj} >$ 100 GeV are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

7.4 Systematic uncertainties

The systematic uncertainties detailed below are taken into account in the statistical model. For the MVA-based signal extraction, both the shape and yield variations are considered, and the resulting BDT output distributions are used in the fit. The largest experimental uncertainty is the uncertainty in the jet energy scale, which distorts the shape of the m_{jj} distribution exploited in the BDT and causes an uncertainty in the prediction of the yields as events migrate in or out of the $ZZjj$ selection due to the tagging jet requirements. The uncertainty in the normalization of the loop-induced background and the uncertainty in the signal shape and yield are the leading theory uncertainties.

7.4.1 Theory uncertainties

The estimation of the theory uncertainties follows the established prescriptions for all processes except for the sub-leading loop-induced QCD production.

Theoretical uncertainties are estimated by simultaneously varying the renormalization and factorization scales, up and down by factors of two and one-half with respect to the nominal values. The top row of Fig. 7.16 shows the effect of the scale variations for the processes most relevant to the VBS search: the dominant QCD background and the electroweak signal. The former exhibits a mild increase of the relative scale uncertainty from 8 % in the background-like to around 12 % in the signal-like region. The impact of the scale choice in the next-to-leading order prediction is greatly reduced with respect to the leading order prediction. As a pure electroweak process, the VBS signal exhibits an overall low dependence on the scale choice even for a leading-order calculation. There is however a pronounced kinematic dependence of the scale choice in the signal, which peaks at 10 % in the most signal-like bin of the BDT distribution.

Uncertainties related to the choice of the PDF and the value of the strong coupling constant are evaluated following the `PDF4LHC` [83, 84] prescription and using the `NNPDF` [85] PDF sets. The resulting template variations are show in the bottom row of Fig. 7.16 with rather small impacts on the shapes and a 3 (8) % variation of the background (signal) yield.

7.4.2 Experimental uncertainties

The uncertainty in the integrated luminosity of the 2016 data sample is 2.5 % [55]. The uncertainty in the trigger efficiency is evaluated by measuring the trigger efficiency in data, which is found to be larger than 98 %, while the efficiency in the simulation is larger than 99 %. A systematic uncertainty of 2 % is assigned. Uncertainties arising from lepton reconstruction and selection efficiencies are taken from the respective tag-and-probe measurements described in Section 3.2.4 and amount to about 6/4/2 % in the 4e/2e2μ/4μ final states. The impact of the uncertainty in the pileup modeling is estimated by varying the minimum bias cross section by its uncertainty of $\pm$4.6 %, and is shown in the top row of Fig. 7.17. It results in a minor source of uncertainty of less than 2 %. The uncertainty in the data-driven reducible background estimate is 40 %.

The jet energy scale (JES) uncertainty is estimated by simultaneously varying the p_{T} of all selected jets in the event by their respective per-jet uncertainty. The $ZZjj$ event selection is repeated, using the modified jet momenta to select the tagging jets. For

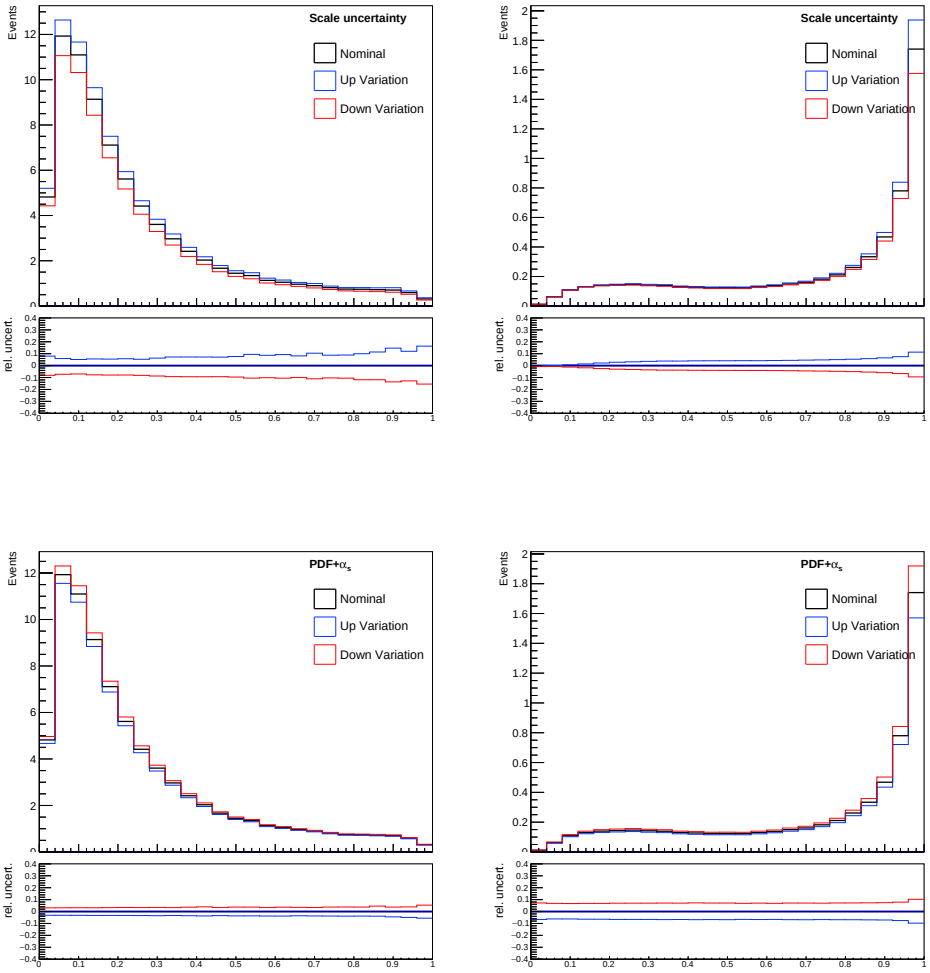


Figure 7.16: Systematic uncertainties due to the variation of the default factorization and renormalization scales (top row) and the systematic uncertainties due to the PDF+α_S variations (bottom row). The left column shows the leading QCD background and the right column the electroweak signal.

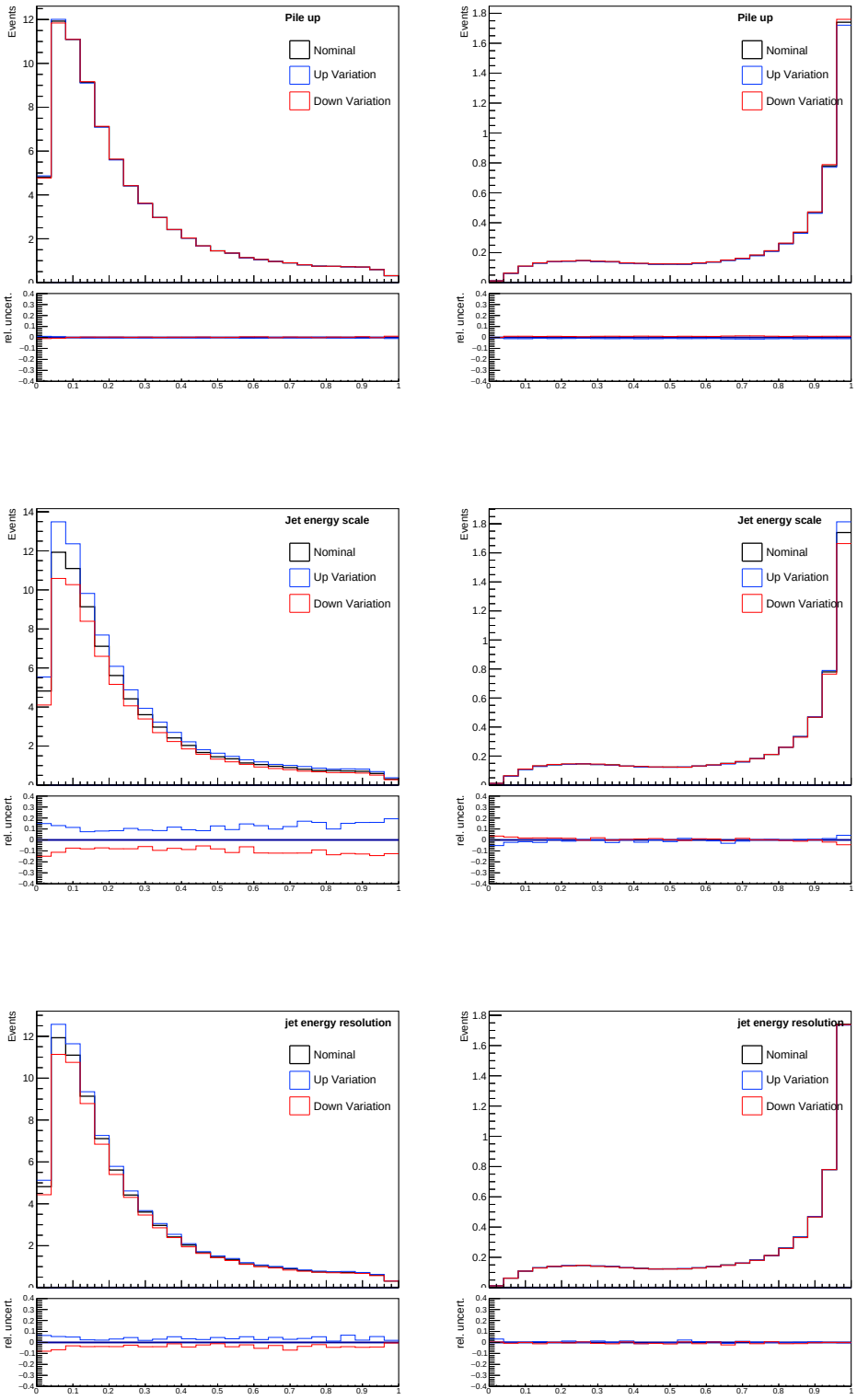


Figure 7.17: Systematic uncertainties due to the minimum bias cross section variations in the pileup reweighing (top row), jet energy scale uncertainty (center row), and jet energy resolution uncertainty (bottom row). The non-loop-induced QCD background is shown in the left and the electroweak signal in the right column.

events that satisfy the $ZZjj$ requirements, the BDT score is recalculated. The resulting distributions thus include the shape and yield variations and are shown in the center row of Fig. 7.17 for the electroweak signal and and the dominant QCD background. The JES uncertainty changes the yield of the background prediction by about 12 %, while having a much smaller impact on the signal. The background yield depends on the minimum jet p_{T} threshold and the JES uncertainty results in simulated events entering or exiting the $ZZjj$ selection.

The jet energy resolution (JER) in the simulation is corrected to match the data using the hybrid method [86]. The uncertainty in the JER scaling factor is propagated to the input jet collections, the tagging jets are re-selected, and the BDT score is recalculated, analogously to the treatment of the JES uncertainty.The bottom row of Fig. 7.17 shows the variations due to the JER uncertainty which are about 5 % on the background and less than 2 % on the signal, hence comparable to the JES uncertainty.

Chapter 8

Statistical analysis and results

The statistical analysis and the physics results presented in this thesis are based on the 2016 dataset of proton–proton collisions. Owing to the exceptional performance of the LHC and the CMS detector, the integrated luminosity available to this analysis is 35.9 fb^{-1}. Two independent results are extracted from this dataset. First, the signal strength of the electroweak production of the $ZZjj$ final state is measured using the BDT presented in the previous chapter. The measured signal strength is interpreted as a fiducial cross section and the data are used to reject the background-only hypothesis. A second analysis constrains anomalous quartic gauge couplings in an effective field theory approach.

8.1 Search for electroweak production of ZZjj

The search for the electroweak production of the $ZZjj$ final state is carried out in the $ZZjj$ event selection which requires two on-shell Z bosons and two jets with $m_{jj} > 100\,\text{GeV}$, see Section 6.2 for the details. The expected yields of the signal and background processes as well as the observed number of events are listed in Table 8.1. The table also lists the yields in the cut-based VBS selection as an illustration.

Table 8.1: Signal and background yields for the $ZZjj$ selection and for the illustrative VBS signal-enriched selection that requires $m_{jj} > 400\,\text{GeV}$ and $|\Delta\eta_{jj}| > 2.4$.

Selection	$\text{t}\bar{\text{t}}\text{Z}$ and WWZ	QCD $ZZjj$	Z+X	Total bkg.	EW $ZZjj$	Total expected	Data
$ZZjj$	7.1 ± 0.8	97 ± 14	6.6 ± 2.5	111 ± 14	6.2 ± 0.7	117 ± 14	99
VBS	0.9 ± 0.2	19 ± 4	0.7 ± 0.3	20 ± 4	4 ± 0.5	25 ± 4	19

The determination of the signal strength for the electroweak production, defined as the ratio of the measured cross section to the Standard Model expectation $\mu = \sigma/\sigma_{\text{SM}}$, is based on the BDT described in Section 7.3.4. Figure 8.1 shows the BDT output distribution of the observed data for events satisfying the $ZZjj$ selection. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

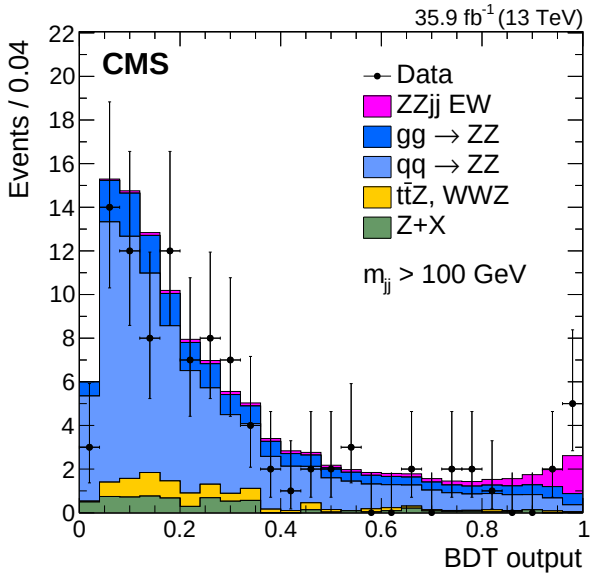


Figure 8.1: Distribution of the BDT output for the events observed in the 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. All events satisfying the $ZZjj$ selection with $m_{jj} > 100$ GeV are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms.

The signal strength is determined from a maximum-likelihood fit of the statistical model to the observed data. This allows to simultaneously extract the signal strength from the signal-enriched part of the BDT output distribution (BDT output ≈ 1) and to constrain the yield of the irreducible QCD background from the background-enriched part of the distribution (BDT output ≈ 0)). The expected distribution for the signal and reducible backgrounds are taken from the simulation while the reducible background is estimated from data. The systematic uncertainties described in Section 7.4 are included as nuisance parameters in the fit. The shape and normalization of the distribution of each process is allowed to vary within its respective uncertainties. The fit is performed using the RooFit and RooStat packages, where the test statistics is the profiled log-likelihood [87].

The statistical model was checked to not exhibit any bias under the signal plus background and background-only hypotheses, by sampling *toy experiments* from the corresponding model. The statistical model is fit to the BDT output distribution obtained from each toy experiment and the signal strength is determined. Figure 8.2 shows the distribution of the signal strength of these toy experiments. The median signal strength of the toy experiments from the signal plus background and background-only models are found to be 1 and 0 respectively, indicating that the statistical procedure is unbiased.

The model consistency is furthermore validated by performing a fit to an artificial dataset that reproduces the signal plus background model perfectly, the *Asimov dataset*. As expected, the signal strength of the Asimov dataset is found to be unity. The top plot of Fig. 8.3 shows the *impact plot* of the Asimov dataset, which lists the nuisance parameters of the model, sorted by their effect on the measured signal strength. The right column lists the change in the signal strength if the nuisance parameter in question is shifted by one standard deviation, the *impact* of the systematic uncertainty on

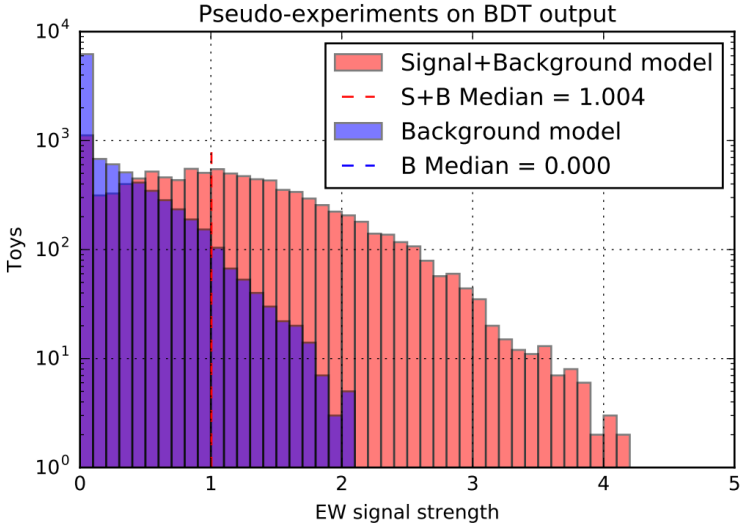


Figure 8.2: Distribution of the electroweak signal strengths obtained by sampling the signal plus background and background-only models (toy experiments). The medians at 1 and 0 indicate that the fit of the BDT output distribution is unbiased for the signal plus background and background-only hypotheses.

Table 8.2: Yields of the signal and background processes before (pre-fit) and after (post-fit) the fit described in the text.

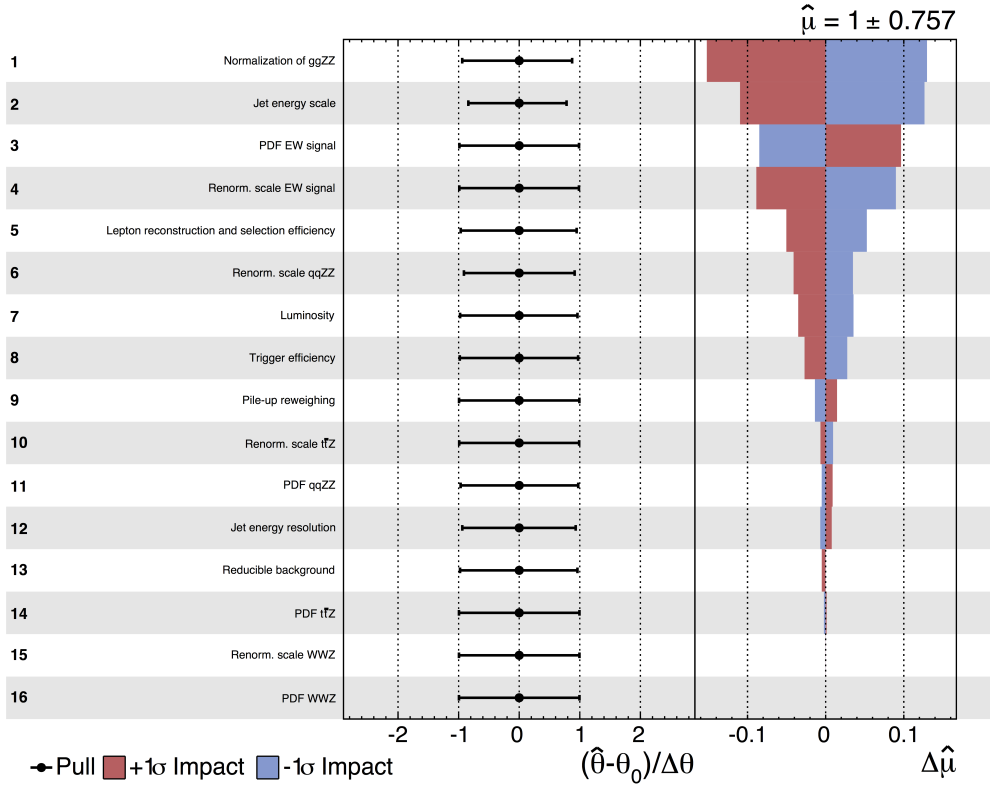
	$t\bar{t}Z$ and WWZ	QCD $ZZjj$	Z+X	Total bkg.	EW $ZZjj$	Total expected	Data
pre-fit	7.05 ± 0.97	97 ± 14	6.4 ± 2.5	111 ± 14	6.21 ± 0.73	117 ± 14	99
post-fit	6.77 ± 0.90	83.6 ± 8.0	6.2 ± 2.4	96.6 ± 8.4	8.6 ± 4.5	105.1 ± 9.5	–

the measurement. The left column shows the pull of the nuisance parameter, i.e., the difference between the pre- and post-fit central values of the nuisance divided by the pre-fit uncertainty. The error bars on the pulls correspond to the post-fit uncertainty on the nuisance and indicate whether the data is able to constrain the nuisance beyond the model specification. This is indeed the case for the renormalization scale uncertainty on the leading QCD background (labeled "Renorm. scale qqZZ " in Fig. 8.3), whose pull uncertainty is slightly reduced. This is due to the fact that the statistical uncertainty on the QCD background yield is lower than the scale variations of about 20 %.

The signal strength determined from the fit of the BDT output distribution to the observed data is $\mu = 1.39^{+0.72}_{-0.57}\,(\text{stat})\,^{+0.46}_{-0.31}\,(\text{syst}) = 1.39^{+0.86}_{-0.65}$. The background-only hypothesis is rejected at 2.7 standard deviations, where 1.6 standard deviations are expected.

The yields and uncertainties of the signal and background processes of the statistical model before and after the template fit are listed in Table 8.2, and the impact distributions of the fit to the data are shown in the bottom panel of Fig. 8.3.

The measured signal strength is used to determine a fiducial cross section of the electroweak production. The fiducial volume is almost identical to the selections imposed at the reconstruction level, the only difference being the looser lepton thresholds of $p_{\mathrm{T}}^{\ell} > 5\,\text{GeV}$ and $|\eta^{\ell}| < 2.5$. The generator-level lepton momenta are corrected by adding the momenta of generator-level photons within $\Delta R(\ell, \gamma) < 0.1$. The kinematic selection of Z bosons and the final $ZZjj$ candidate then proceeds as in the



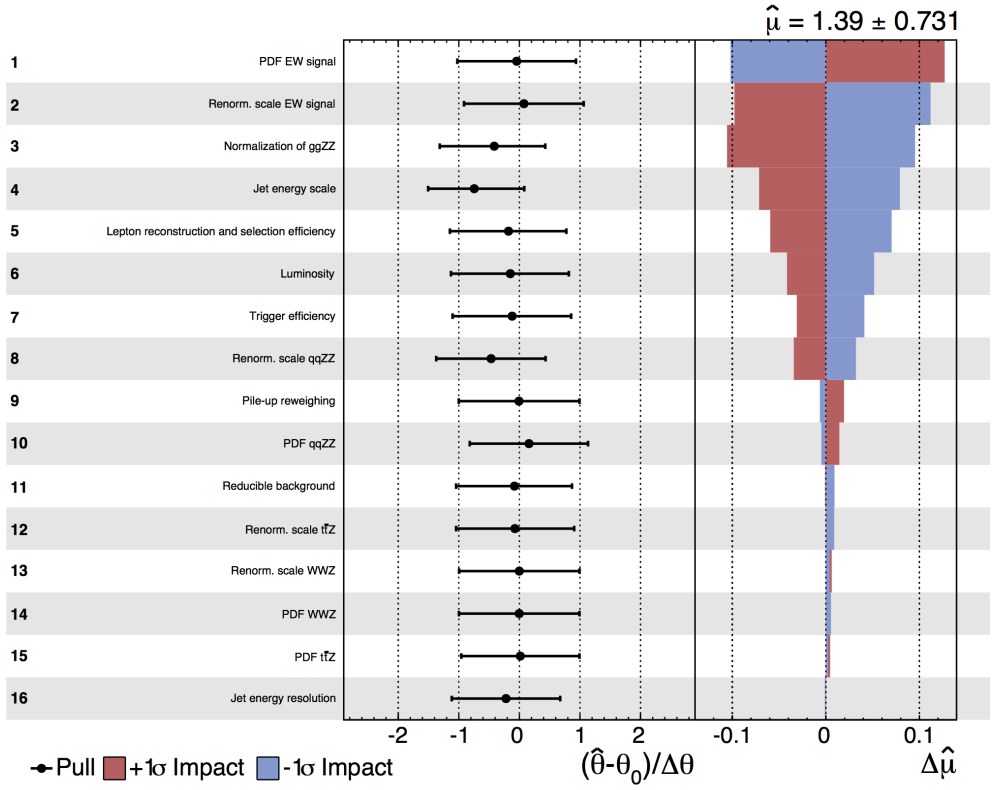


Figure 8.3: Distribution of the impact and pull distributions for the template fit to the Asimov dataset (top) and the observed data (bottom).

Table 8.3: Kinematic properties of the observed events in the signal-enriched region BDT score > 0.9.

$m_{4\ell}$ [GeV]	m_{Z1} [GeV]	m_{Z2} [GeV]	m_{jj} [GeV]	$\lvert\Delta\eta_{jj}\rvert$	$\eta^{\star}_{Z_1}$	$\eta^{\star}_{Z_2}$	BDT score
365.8	91.4	101.1	844.1	3.4	-0.7	0.0	0.97
325.1	93.1	96.3	1332.9	5.2	0.0	-1.8	0.98
263.8	91.9	88.0	829.7	2.2	-0.5	1.1	0.94
562.8	93.7	88.0	947.3	2.8	0.6	0.6	0.93
248.8	91.5	89.2	1340.9	5.4	-0.5	0.2	0.98
375.2	89.4	98.5	1052.5	3.8	0.7	-0.2	0.96
482.1	95.0	95.6	1543.1	4.8	-1.6	2.5	0.99

reconstruction-level selection. The observed signal strength corresponds to a fiducial cross section of $\sigma_{\text{fid}}(\text{EW pp} \to ZZjj \to \ell\ell\ell'\ell' jj) = 0.40^{+0.21}_{-0.16}\,(\text{stat})\,^{+0.13}_{-0.09}\,(\text{syst})$ fb, compatible with the SM prediction of $0.29^{+0.02}_{-0.03}$ fb.

The kinematic properties of the observed events with VBS-like kinematics, identified by requiring a BDT output above 0.9, are listed in Table 8.3. Figure 8.4 visualizes one of the signal-like events with a large tagging jet invariant mass and pseudorapidity separation ($m_{jj} > 1.3$ TeV and $|\Delta\eta_{jj}| = 5.4$).

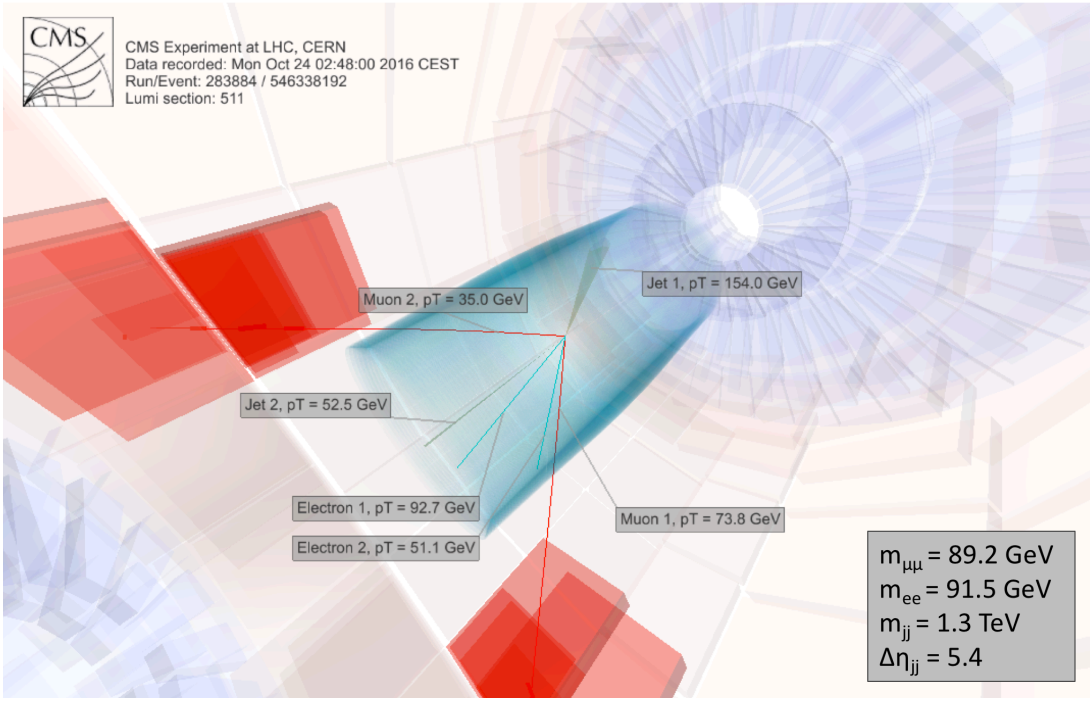


Figure 8.4: Event display of an observed $ZZjj$ event with two energetic electrons (light blue lines), two energetic muons (red lines), and two hadronic jets (dark green cones). The presence of two opposite-sign same-flavor lepton pairs with mass close to the nominal Z boson mass, of two hadronic jets in opposite hemispheres of the detector with a large pseudorapidity separation, as well as the absence of hadronic activity in the central region of the detector, are indicative of the electroweak production of two Z bosons and two jets.

8.2 Limits on anomalous quartic gauge couplings

The events in the $ZZjj$ selection are used to constrain anomalous quartic gauge couplings in an effective field theory approach. The $ZZjj$ channel is sensitive to the neutral current operators $\mathcal{O}_{T,8}$ and $\mathcal{O}_{T,9}$, as well as the operators $\mathcal{O}_{T,0,1,2}$, which increase the production cross section at large masses of the ZZ system. Limits on the couplings f_{T_i}/Λ^4 are derived based on the invariant mass of the diboson system.

The expected distributions for different values of the couplings are obtained using the reweighing feature of the MG5_aMC package as detailed in Section 5.1.4. A semi-analytic description of the expected m_{ZZ} distribution as a function of the aQGC couplings is obtained by fitting quadratic functions to the ratio of the aQGC and standard model yields in each m_{ZZ} bin. Figure 8.5 illustrates the procedure by showing the predicted yield ratio in m_{ZZ} bins for the discrete parameter points and the result of the fit of a parabola for the operator $\mathcal{O}_{T,8}$. As expected, the quadratic function provides a good model for the yield ratio as a function of the coupling. The last m_{ZZ} bin, which includes all events above 1.2 TeV, provides the most statistical power to the limit setting. Figure 8.6 shows the parametrizations for the operators $\mathcal{O}_{T,0,1,2}$ and $\mathcal{O}_{T,9}$.

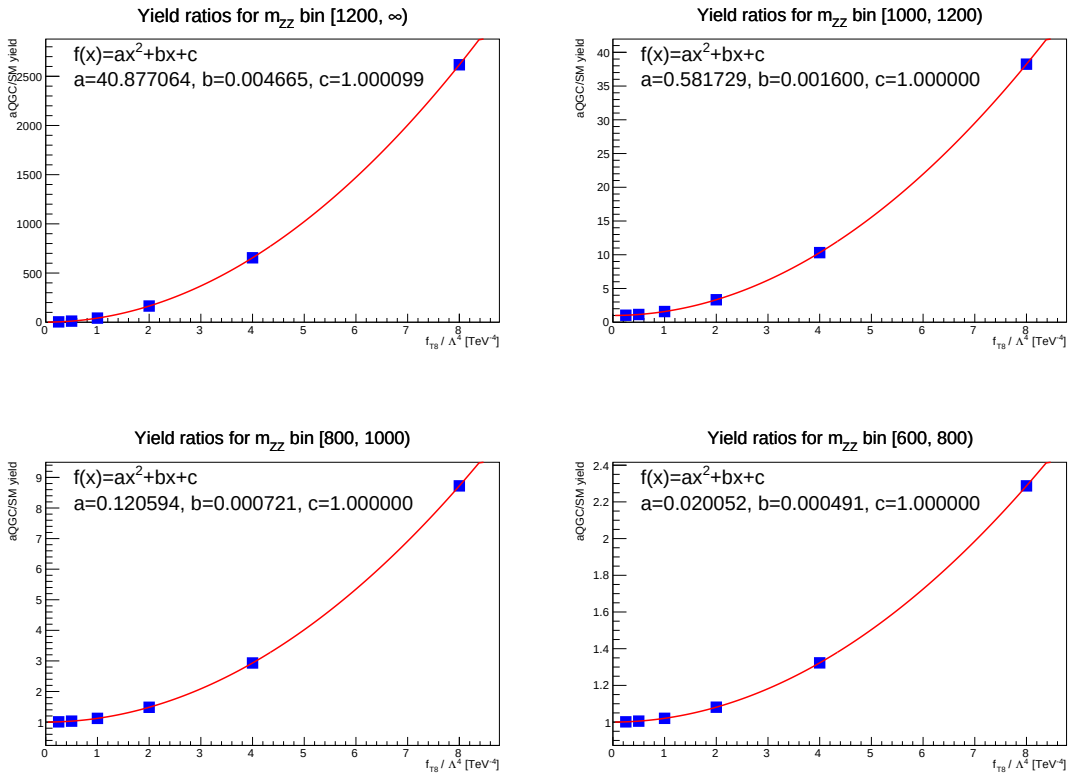


Figure 8.5: Yield ratios of the discrete operator couplings obtained from the reweighing and the result of the fit of the quadratic function for the interpolation. Markers indicate the discrete couplings f_{T_8}/Λ^4 obtained from the reweighting and the fitted quadratic interpolation. Only the highest mass bins are shown, as these are the most sensitive to the anomalous couplings.

Figure 8.7 (left) shows the expected m_{ZZ} distribution for the SM and two aQGC scenarios as well as the observed events in the 2016 data.

Confidence levels on the operator couplings are derived using the RooStat tool. The

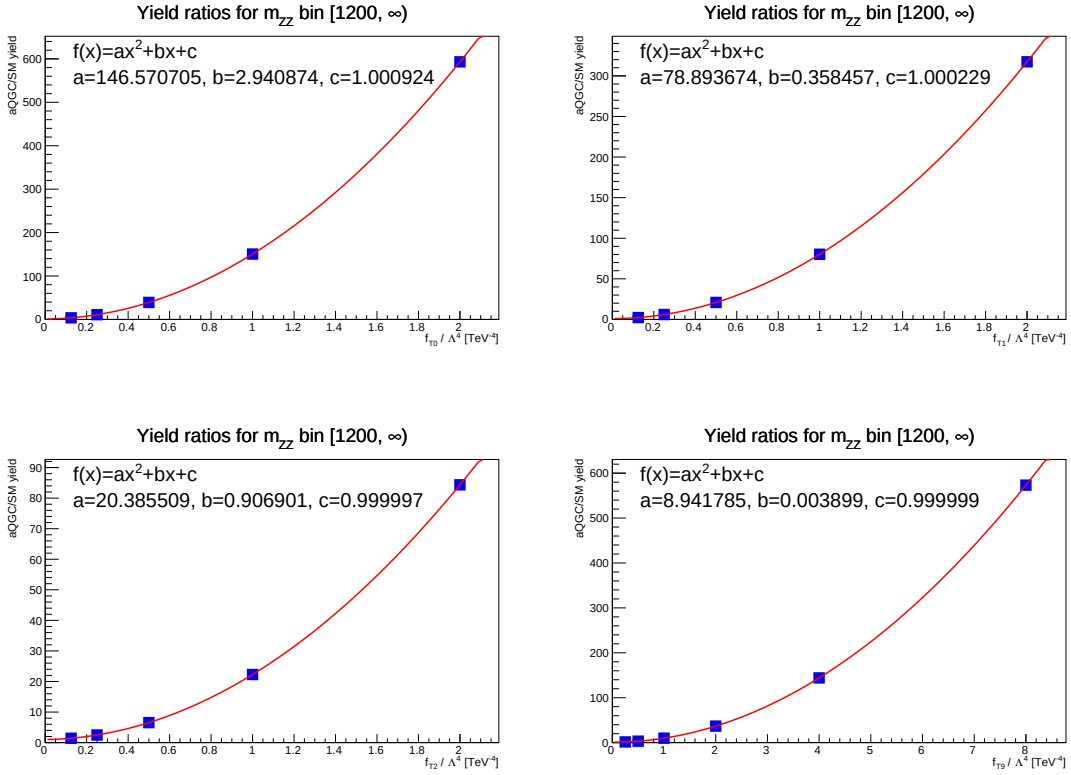


Figure 8.6: Yield ratios of the discrete operator couplings obtained from the reweighing and the result of the fit of the quadratic function for the interpolation. Shown is the last m_{ZZ} bin of the distribution for the f_{T_0}/Λ^4 (top left), f_{T_1}/Λ^4 (top right), f_{T_2}/Λ^4 (bottom left), and f_{T_9}/Λ^4 (bottom right) operators.

test statistics t is the same log-likelihood ratio used for the determination of the significance of the electroweak signal, again with all systematic uncertainties profiled as nuisance parameters. The confidence levels are determined using Wilk's theorem and the assumption that the likelihood approaches a χ^2-distribution with one degree of freedom. The 95 % confidence level is then determined by finding the coupling strength that yields a likelihood ratio of $t = 3.84$.

Table 8.4 lists the individual confidence level (CL) obtained by setting the other coupling to zero as well as the unitarity limit. The later is determined using the form factor tool that is part of the VBFNLO package [19]. In the calculation, the coupling strength is set to the observed limit and the cut-off scale Λ at which the scattering amplitude would violate unitarity is reported as the unitarity limit. Care is taken to account for the difference in the operator definitions for the MG5_aMC model given in Ref. [18] and those used in the VBFNLO framework.

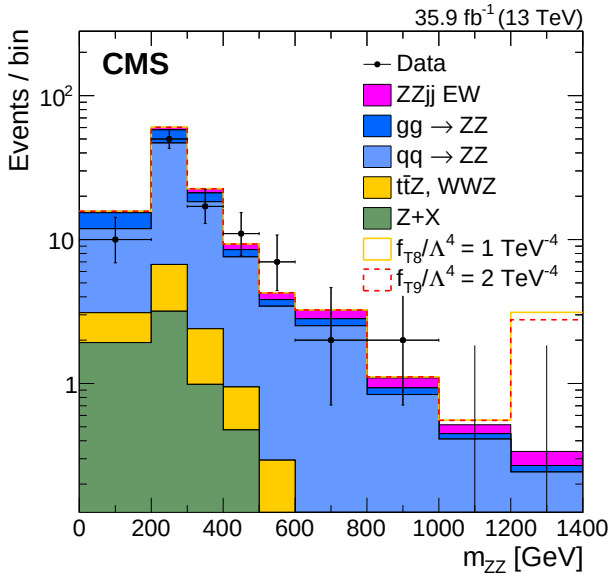


Figure 8.7: Distribution of the diboson invariant mass observed in the 2016 dataset with an integrated luminosity of 35.9 fb^{-1}. All events satisfying the $ZZjj$ selection with $m_{jj} > 100$ GeV are considered. The expected distributions obtained from the simulation and the data-driven estimate of the reducible background are shown as stacked histograms. Two aQGC coupling hypothesis are also shown. The last bin includes all contributions with $m_{ZZ} > 1200$ GeV.

Table 8.4: Observed and expected lower and upper 95 % confidence levels on the couplings of the quartic operators $\mathcal{O}_{T,0,1,2}$, as well as the neutral current operators $\mathcal{O}_{T,8}$ and $\mathcal{O}_{T,9}$. The unitarity limit are also listed. All coupling units are in TeV^{-4}, the unitarity limits are in TeV. Operator definitions are those of [18].

Coupling	Exp. lower	Exp. upper	Obs. lower	Obs. upper	Unitarity limit
f_{T_0}/Λ^4	−0.53	0.51	−0.46	0.44	2.4
f_{T_1}/Λ^4	−0.72	0.71	−0.61	0.61	2.4
f_{T_2}/Λ^4	−1.4	1.4	−1.2	1.2	2.4
f_{T_8}/Λ^4	−0.99	0.99	−0.84	0.84	2.8
f_{T_9}/Λ^4	−2.1	2.1	−1.8	1.8	2.9

Conclusions

This thesis presented the first study of vector boson scattering in the $ZZjj$ channel [24]. Both Z bosons are identified by their leptonic decay, resulting a clean experimental signature with low reducible backgrounds. A multivariate analysis of 35.9 fb^{-1} of data allowed for the measurement of a signal strength of $\mu = 1.39^{+0.72}_{-0.57}\,(\text{stat})\ ^{+0.46}_{-0.31}\ (\text{syst}) = 1.39^{+0.86}_{-0.65}$, which is compatible with the standard model prediction. The background-only hypothesis is rejected with a significance of 2.7 standard deviations, where 1.6 standard deviations are expected. The signal strength is converted into a fiducial cross section $\sigma_{\text{fid}}(\text{EW pp} \to ZZjj \to \ell\ell\ell'\ell' jj) = 0.40^{+0.21}_{-0.16}\,(\text{stat})\ ^{+0.13}_{-0.09}\,(\text{syst})\,\text{fb}$, providing the first measurement of this quantity and continuing the remarkable success of the standard model as an effective description of the fundamental interactions, see Fig. 8.8.

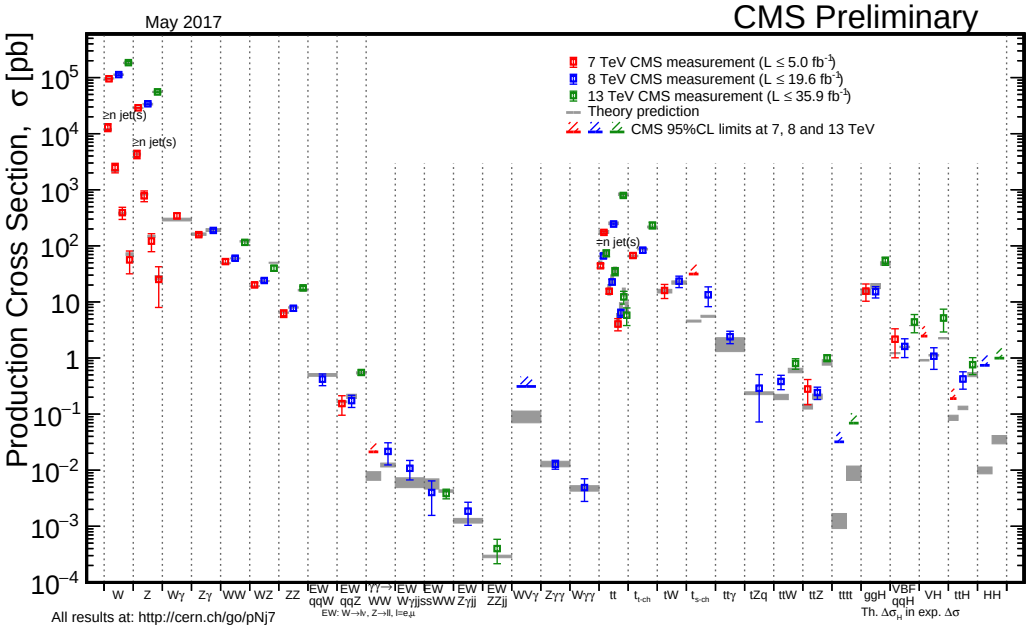


Figure 8.8: Summary of standard model cross section measurements performed by the CMS Collaboration [88]. Based on public results, including this thesis work.

This thesis furthermore reports constraints on physics beyond the standard model in the form of 95 % confidence limits on anomalous quartic gauge couplings. The reported limits on the coupling coefficients $f_{\text{T},8}$, $f_{\text{T},9}$, and $f_{\text{T},0}$ are the most stringent to date, while those on $f_{\text{T},1,2}$ are competitive, see Fig. 8.9. The reported limits are significantly more stringent than than expected from sensitivity studies, exceeding already those projected for the full Run II dataset of 300 fb^{-1} [89]. The improvement in the experimental sensitivity is a consequence of the inclusiveness of the phase-space used for the anomalous coupling search.

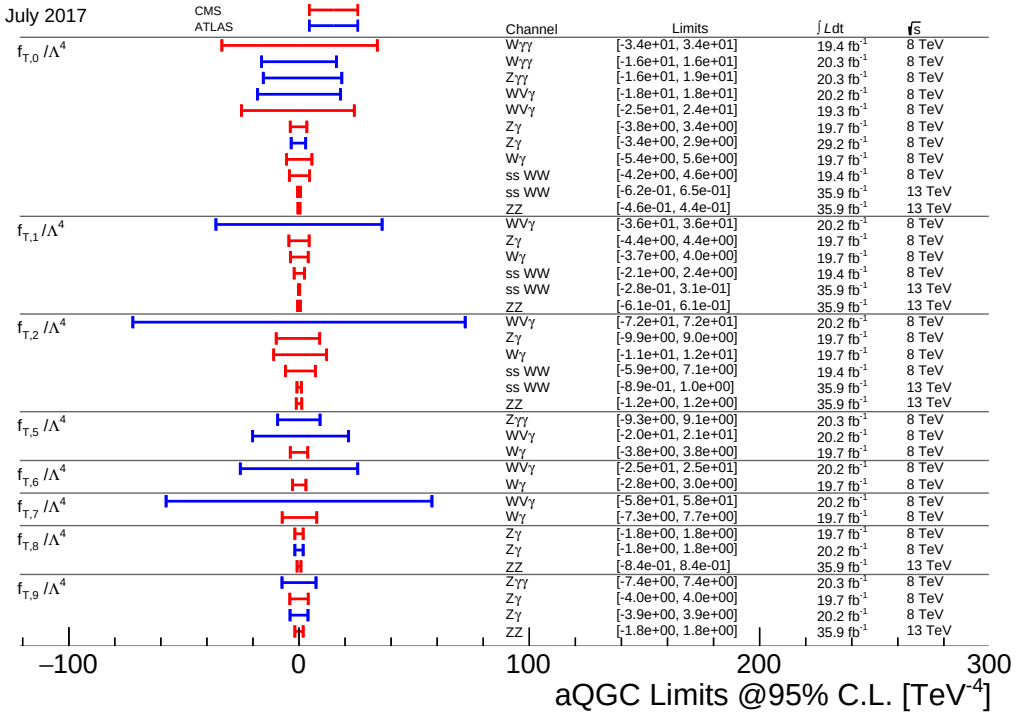


Figure 8.9: Summary of the limits on anomalous quartic gauge couplings for operators that involve only the electroweak field strength tensors [90]. The limits reported in this thesis work are labelled as "ZZ".

The analysis of massive vector boson scattering (VBS) presented in this thesis work is part of a larger effort to understand the breaking of the electroweak symmetry. The study of VBS complements the measurements of the production and decay rates of the Higgs boson and probes the Goldstone boson nature of the longitudinal polarizations in high-energy interactions of massive gauge bosons. The increased center-of-mass energy and large integrated luminosities provided by the LHC in the first year of the run II enabled the first observation of a VBS process using the WW channel, but the number of signal events available for analysis is still low. The study of VBS in the $ZZjj$ channel presented in this thesis is one of the first analysis of VBS and the first VBS analysis to employ a multivariate technique to extract the signal. Compared to a traditional cut-and-count approach, the use of the multivariate discriminant increases the sensitivity by about 20 % and demonstrates the benefit of these methods for future VBS measurements.

To understand the interplay between the electroweak and scalar sector via VBS, the longitudinal polarizations of the gauge bosons will need to be inferred. The fully leptonic final state of the $ZZjj$ channel considered in this thesis allows for a complete and unambiguous reconstruction of all production and decay angles, which permits to separate the longitudinal from the dominating transverse polarization. However, the study of the longitudinal polarization in VBS will be a challenge, even with the integrated luminosity of $3\,\mathrm{ab}^{-1}$ that is projected for the end of the HL-LHC [91]. The study of these rare processes will benefit from novel experimental techniques, which allow to enhance the statistics available for physics analysis.

One such approach specific to the fully leptonic $ZZjj$ channel is to exploit the kinematic constraints provided by the on-shell Z bosons to further relax the object selection or to consider leptons that are not covered by the standard object reconstruction. The

former approach was implemented during this PhD as a contribution to the search of high-mass resonances in the $ZZ \to 4\ell$ channel [92]. The relaxed electron selection developed for this analysis allowed to increase the signal selection efficiency, particularly at high-mass. The new event category based on these ZZ candidates from relaxed selection electrons allowed a significant improvement on the constraints on additional high-mass resonances. The approach is well suited for a search of a narrow resonance peak, but challenging to implement in a multivariate analysis as implemented in the $ZZjj$ VBS study.

A similar idea is to exploit electrons outside the acceptance of the tracker, i.e., to include electrons that are only reconstructed in the calorimeter. The study of the lepton kinematics in the simulation demonstrated that the signal acceptance could be increased by about 20 % if one such electron were permitted in the selection of the ZZ candidate. This makes a compelling physics case for the anticipated $300\,\mathrm{fb}^{-1}$ of data that will be collected before the phase II upgrade to the CMS tracker. The planed upgrades of the CMS detector for phase II will extend the lepton and jet reconstruction capabilities in the forward region and VBS analyses in particular will benefit from these extensions.

Finally, the VBS studies are currently carried out as standard model measurements with no consideration for the related analyses at the Higgs boson pole mass. Combining such measurements of vector boson fusion and scattering would allow to constrain the Higgs boson couplings to vector bosons in a model-independent way and to confront the experimental data with the standard model prediction with increased sensitivity.

References

[1] Wikipedia, "Illustration of fundamental particles in the standard model". Wikimediacommons:https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg. Accessed: 2017-06-15.

[2] S. L. Glashow, "Partial-symmetries of weak interactions", *Nuclear Physics* **22** (1961) 579, doi:10.1016/0029-5582(61)90469-2.

[3] S. Weinberg, "A model of leptons", *Phys. Rev. Lett.* **19** (1967) 1264, doi:10.1103/PhysRevLett.19.1264.

[4] A. Salam, "Weak and electromagnetic interactions", in *Proceedings of the eighth Nobel symposium*, p. 367. 1968.

[5] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons", *Phys. Rev. Lett.* **13** (1964) 321, doi:10.1103/PhysRevLett.13.321.

[6] P. W. Higgs, "Broken symmetries and the masses of gauge bosons", *Phys. Rev. Lett.* **13** (1964) 508, doi:10.1103/PhysRevLett.13.508.

[7] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global conservation laws and massless particles", *Phys. Rev. Lett.* **13** (1964) 585, doi:10.1103/PhysRevLett.13.585.

[8] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* **716** (2012) 1, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[9] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett. B* **716** (2012) 30, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[10] ATLAS and CMS Collaborations, "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV", *JHEP* **08** (2016) 045, doi:10.1007/JHEP08(2016)045, arXiv:1606.02266.

[11] 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.* **38** (1977) 883, doi:10.1103/PhysRevLett.38.883.

[12] 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* **16** (1977) 1519, doi:10.1103/PhysRevD.16.1519.

[13] J. Brehmer, "Polarised WW Scattering at the LHC", Master's thesis, U. Heidelberg, ITP, 2014.

[14] M. Rauch, "Vector-Boson Fusion and Vector-Boson Scattering", arXiv:1610.08420.

[15] D. Rainwater, R. Szalapski, and D. Zeppenfeld, "Probing color singlet exchange in $Z + 2$-jet events at the CERN LHC", *Phys. Rev. D* **54** (1996) 6680, doi:10.1103/PhysRevD.54.6680, arXiv:hep-ph/9605444.

[16] S. Weinberg, "Phenomenological Lagrangians", *Physica A* **96** (1979) 327, doi:10.1016/0378-4371(79)90223-1.

[17] C. Degrande et al., "Effective field theory: A modern approach to anomalous couplings", *Annals Phys.* **335** (2013) 21, doi:10.1016/j.aop.2013.04.016, arXiv:1205.4231.

[18] O. J. P. Éboli, M. C. Gonzalez-Garcia, and J. K. Mizukoshi, "$pp \to jj\, e^{\pm}\mu^{\pm}\nu\nu$ and $jj\, e^{\pm}\mu^{\mp}\nu\nu$ at $\mathcal{O}(\alpha_{\rm em}^6)$ and $\mathcal{O}(\alpha_{\rm em}^4\alpha_{\rm s}^2)$ for the study of the quartic electroweak gauge boson vertex at CERN LHC", *Phys. Rev. D* **74** (2006) 073005, doi:10.1103/PhysRevD.74.073005, arXiv:hep-ph/0606118.

[19] K. Arnold et al., "VBFNLO: A parton level Monte Carlo for processes with electroweak bosons", *Comput. Phys. Commun.* **180** (2009) 1661, doi:10.1016/j.cpc.2009.03.006, arXiv:0811.4559.

[20] ATLAS Collaboration, "Evidence for Electroweak Production of $W^{\pm}W^{\pm}jj$ in pp Collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector", *Phys. Rev. Lett.* **113** (2014) 141803, doi:10.1103/PhysRevLett.113.141803, arXiv:1405.6241.

[21] CMS Collaboration, "Study of Vector Boson Scattering and Search for New Physics in Events with Two Same-Sign Leptons and Two Jets", *Phys. Rev. Lett.* **114** (2015) 051801, doi:10.1103/PhysRevLett.114.051801, arXiv:1410.6315.

[22] CMS Collaboration, "Observation of electroweak production of same-sign W boson pairs in the two jet and two same-sign lepton final state in proton-proton collisions at 13 TeV", CMS Physics Analysis Summary CMS-PAS-SMP-17-004, 2017.

[23] CMS Collaboration Collaboration, "Measurements of differential cross sections and search for the electroweak production of two Z bosons produced in association with jets", CMS Physics Analysis Summary CMS-PAS-SMP-16-019, 2017.

[24] 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", arXiv:1708.02812. Submitted to *Phys. Lett. B*.

[25] ATLAS Collaboration, "Measurements 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", *Phys. Rev. D* **93** (2016) 092004, doi:10.1103/PhysRevD.93.092004, arXiv:1603.02151.

[26] ATLAS Collaboration, "Search for anomalous electroweak production of WW/WZ in association with a high-mass dijet system in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector", *Phys. Rev. D* **95** (2017) 032001, doi:10.1103/PhysRevD.95.032001, arXiv:1609.05122.

[27] CMS Collaboration, "Measurement of electroweak-induced production of Wγ with two jets in pp collisions at $\sqrt{s}$ = 8 TeV and constraints on anomalous quartic gauge couplings", *JHEP* **06** (2017) 106, doi:10.1007/JHEP06(2017)106, arXiv:1612.09256.

[28] 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", *Phys. Lett. B* **770** (2017) 380, doi:10.1016/j.physletb.2017.04.071, arXiv:1702.03025.

[29] 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", (2017). arXiv:1705.01966. Submitted to *JHEP*.

[30] CMS Collaboration, "Measurement of electroweak production of two jets in association with a Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary CMS-PAS-SMP-16-018.

[31] CMS Collaboration, "Measurement of the hadronic activity in events with a Z and two jets and extraction of the cross section for the electroweak production of a Z with two jets in pp collisions at $\sqrt{s}$ = 7 TeV", *JHEP* **10** (2013) 062, doi:10.1007/JHEP10(2013)062, arXiv:1305.7389.

[32] ATLAS Collaboration, "Measurement of the electroweak production of dijets in association with a Z-boson and distributions sensitive to vector boson fusion in proton-proton collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector", *JHEP* **04** (2014) 031, doi:10.1007/JHEP04(2014)031, arXiv:1401.7610.

[33] CMS Collaboration, "Measurement of electroweak production of two jets in association with a Z boson in proton–proton collisions at $\sqrt{s} = 8$ TeV", *Eur. Phys. J. C* **75** (2015) 66, doi:10.1140/epjc/s10052-014-3232-5, arXiv:1410.3153.

[34] CMS Collaboration, "Measurement of electroweak production of a W boson and two forward jets in proton-proton collisions at $\sqrt{s} = 8$ TeV", *JHEP* **11** (2016) 147, doi:10.1007/JHEP11(2016)147, arXiv:1607.06975.

[35] ATLAS Collaboration, "Measurements of electroweak Wjj production and constraints on anomalous gauge couplings with the ATLAS detector", *Eur. Phys. J. C* **77** (2017) 474, doi:10.1140/epjc/s10052-017-5007-2, arXiv:1703.04362.

[36] CMS Collaboration, "Study of exclusive two-photon production of W^+W^- in pp collisions at $\sqrt{s} = 7$ TeV and constraints on anomalous quartic gauge couplings", *JHEP* **07** (2013) 116, doi:10.1007/JHEP07(2013)116, arXiv:1305.5596.

[37] CMS Collaboration, "Exclusive photon-photon production of muon pairs in proton-proton collisions at $\sqrt{s} = 7$ TeV", *JHEP* **01** (2012) 052, doi:10.1007/JHEP01(2012)052, arXiv:1111.5536.

[38] CMS Collaboration, "Evidence for exclusive $\gamma\gamma \to W^+W^-$ production and constraints on anomalous quartic gauge couplings in pp collisions at $\sqrt{s} = 7$ and 8 TeV", *JHEP* **08** (2016) 119, doi:10.1007/JHEP08(2016)119, arXiv:1604.04464.

[39] ATLAS Collaboration, "Measurement of exclusive $\gamma\gamma \to W^+W^-$ production and search for exclusive Higgs boson production in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector", *Phys. Rev. D* **94** (2016) 032011, doi:10.1103/PhysRevD.94.032011, arXiv:1607.03745.

[40] L3 Collaboration, "Study of the $W^+W^-\gamma$ process and limits on anomalous quartic gauge boson couplings at LEP", *Phys. Lett. B* **527** (2002) 29, doi:10.1016/S0370-2693(02)01167-X, arXiv:hep-ex/0111029.

[41] D0 Collaboration, "Search for anomalous quartic $WW\gamma\gamma$ couplings in dielectron and missing energy final states in $p\bar{p}$ collisions at $\sqrt{s}$ = 1.96 TeV", *Phys. Rev. D* **88** (2013) 012005, doi:10.1103/PhysRevD.88.012005, arXiv:1305.1258.

[42] CMS Collaboration, "Search for $WW\gamma$ and $WZ\gamma$ production and constraints on anomalous quartic gauge couplings in pp collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. D* **90** (2014) 032008, doi:10.1103/PhysRevD.90.032008, arXiv:1404.4619.

[43] ATLAS Collaboration, "Evidence of W Production in pp Collisions at $\sqrt{s} = 8$ TeV and Limits on Anomalous Quartic Gauge Couplings with the ATLAS Detector", *Phys. Rev. Lett.* **115** (2015) 031802, doi:10.1103/PhysRevLett.115.031802, arXiv:1503.03243.

[44] CMS Collaboration, "Measurements of the pp $\to W\gamma\gamma$ and pp $\to Z\gamma\gamma$ cross sections and limits on anomalous quartic gauge couplings at $\sqrt{s} = 8$ TeV", arXiv:1704.00366. Submitted to *JHEP*.

[45] C. De Melis, "The CERN accelerator complex. Complexe des accélérateurs du CERN",. General Photo.

[46] T. Mc Cauley, "Collisions recorded by the CMS detector on 7 May 2016 at the start of the year's physics run". CMS Collection.

[47] T. Sakuma and T. McCauley, "Detector and Event Visualization with SketchUp at the CMS Experiment", *J. Phys. Conf. Ser.* **513** (2014) 022032, doi:10.1088/1742-6596/513/2/022032, arXiv:1311.4942.

[48] V. Halyo, P. LeGresley, and P. Lujan, "Massively Parallel Computing and the Search for Jets and Black Holes at the LHC", *Nucl. Instrum. Meth. A* **744** (2014) 54, doi:10.1016/j.nima.2014.01.038, arXiv:1309.6275.

[49] CMS Collaboration, "CMS Physics: Technical Design Report Volume 1: Detector Performance and Software", Technical Design Report, 2006.

[50] CMS Collaboration, "ECAL Laser monitoring till end of 2016 and ECAL phi-symmetry", CMS Detector Performance Note CMS-DP-2017-003.

[51] CMS Collaboration, "Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV", *JINST* **8** (2013) P09009, doi:10.1088/1748-0221/8/09/P09009, arXiv:1306.2016.

[52] CMS Collaboration, "Schematic overview of the CMS HCAL from the XDAQ project". http://xdaq.web.cern.ch/xdaq/setup/images/HCAL.png. Accessed: 2017-06-15.

[53] CMS Collaboration, "The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s} = 7$ TeV at the LHC", *JINST* **8** (2013) P11002, doi:10.1088/1748-0221/8/11/P11002, arXiv:1306.6905.

[54] S. Regnard, "Measurements of Higgs boson properties in the four-lepton final state at $\sqrt{s} = 13$ TeV with the CMS experiment at the LHC.". PhD thesis.

[55] CMS Collaboration, "CMS Luminosity Measurements for the 2016 Data Taking Period", CMS Physics Analysis Summary CMS-PAS-LUM-17-001, 2017.

[56] CMS Collaboration, "CMS luminosity summary plot". https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults. Accessed: 2017-06-15.

[57] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV", *JINST* **10** (2015) P06005, doi:10.1088/1748-0221/10/06/P06005, arXiv:1502.02701.

[58] H. Bethe and W. Heitler, "On the Stopping of Fast Particles and on the Creation of Positive Electrons", *Proceedings of the Royal Society of London Series A* **146** 83, doi:10.1098/rspa.1934.0140.

[59] R. Frühwirth, "A Gaussian-mixture approximation of the Bethe-Heitler model of electron energy loss by bremsstrahlung", *Computer Physics Communications* **154** 131, doi:10.1016/S0010-4655(03)00292-3.

[60] CMS Collaboration, "Electron Reconstruction within the Particle Flow Algorithm", CMS Analysis Note AN2010-034 (internal).

[61] CMS Collaboration, "Electron and photon performance in CMS with the full 2016 data sample.", CMS Detector Performance Summary CMS-DP-2017-004, 2017.

[62] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", *JINST* **7** (2012) P10002, doi:10.1088/1748-0221/7/10/P10002, arXiv:1206.4071.

[63] CMS Collaboration, "Measurements of properties of the Higgs boson in the four-lepton final state at $\sqrt{s} = 13$ TeV", CMS Analysis Note AN2016-442 (internal).

[64] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *Eur. Phys. J. C* **72** (2012) 1896, doi:10.1140/epjc/s10052-012-1896-2, arXiv:1111.6097.

[65] CMS Collaboration, "Jet energy scale and resolution performances with 13 TeV data", CMS Detector Performance Summary CMS-DP-2016-020, 2016.

[66] A. Rogozhnikov et al., "New approaches for boosting to uniformity", *JINST* **10** (2015) T03002, doi:10.1088/1748-0221/10/03/T03002, arXiv:1410.4140.

[67] G. Louppe, M. Kagan, and K. Cranmer, "Learning to Pivot with Adversarial Networks", arXiv:1611.01046.

[68] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[69] A. Ballestrero et al., "PHANTOM: A Monte Carlo event generator for six parton final states at high energy colliders", *Comput. Phys. Commun.* **180** (2009) 401, doi:10.1016/j.cpc.2008.10.005, arXiv:0801.3359.

[70] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO", *JHEP* **12** (2012) 061, doi:10.1007/JHEP12(2012)061, arXiv:1209.6215.

[71] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk, "Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations", *JHEP* **03** (2013) 015, doi:10.1007/JHEP03(2013)015, arXiv:1212.3460.

[72] J. M. Campbell and R. K. Ellis, "MCFM for the Tevatron and the LHC", *Nucl. Phys. B Proc. Suppl.* **205-206** (2010) 10, doi:10.1016/j.nuclphysbps.2010.08.011, arXiv:1007.3492.

[73] T. Sjostrand, S. Mrenna, and P. Z. Skands, "A Brief Introduction to PYTHIA 8.1", *Comput. Phys. Commun.* **178** (2008) 852, doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820.

[74] CMS Collaboration, "Event generator tunes obtained from underlying event and multiparton scattering measurements", *Eur. Phys. J. C* **76** (2016) 155, doi:10.1140/epjc/s10052-016-3988-x, arXiv:1512.00815.

[75] Particle Data Group, "Review of Particle Physics", *Chin. Phys. C* **40** (2016) 100001, doi:10.1088/1674-1137/40/10/100001.

[76] M. Bahr et al., "Herwig++ Physics and Manual", *Eur. Phys. J. C* **58** (2008) 639, doi:10.1140/epjc/s10052-008-0798-9, arXiv:0803.0883.

[77] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python", *J. Machine Learning Res.* **12** (2011) 2825, arXiv:1201.0490.

[78] Y. Gao et al., "Spin determination of single-produced resonances at hadron colliders", *Phys. Rev. D* **81** (2010) 075022, doi:10.1103/PhysRevD.81.075022, arXiv:1001.3396. [Erratum: *Phys. Rev. D* **81** (2010) 079905 doi:10.1103/PhysRevD.81.079905].

[79] S. Bolognesi et al., "Spin and parity of a single-produced resonance at the LHC", *Phys. Rev. D* **86** (2012) 095031, doi:10.1103/PhysRevD.86.095031, arXiv:1208.4018.

[80] I. Anderson et al., "Constraining anomalous HVV interactions at proton and lepton colliders", *Phys. Rev. D* **89** (2014) 035007, doi:10.1103/PhysRevD.89.035007, arXiv:1309.4819.

[81] G. Brooijmans et al., "Les Houches 2013: Physics at TeV Colliders: New Physics Working Group Report", arXiv:1405.1617.

[82] A. Buckley et al., "LHAPDF6: parton density access in the LHC precision era", *Eur. Phys. J. C* **75** (2015) 132, doi:10.1140/epjc/s10052-015-3318-8, arXiv:1412.7420.

[83] M. Botje et al., "The PDF4LHC Working Group Interim Recommendations", arXiv:1101.0538.

[84] S. Alekhin et al., "The PDF4LHC Working Group Interim Report", arXiv:1101.0536.

[85] NNPDF Collaboration, "Parton distributions for the LHC run II", *JHEP* **04** (2015) 040, doi:10.1007/JHEP04(2015)040, arXiv:1410.8849.

[86] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *JINST* **12** (2017) P02014, doi:10.1088/1748-0221/12/02/P02014, arXiv:1607.03663.

[87] ATLAS and CMS Collaborations, LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011", CMS-NOTE-2011-005; ATL-PHYS-PUB-2011-11, 2011.

[88] CMS Collaboration, "CMS summary plot of standard model cross sections". https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsCombined. Accessed: 2017-07-03.

[89] C. Degrande et al., "Studies of Vector Boson Scattering And Triboson Production with DELPHES Parametrized Fast Simulation for Snowmass 2013", 2013. arXiv:1309.7452.

[90] CMS Collaboration, "CMS summary plot of anomalous quartic gauge coupling limits". https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSMPaTGC. Accessed: 2017-06-15.

[91] CMS Collaboration, "Prospects for the study of vector boson scattering in same sign WW and WZ interactions at the HL-LHC with the upgraded CMS detector", CMS Physics Analysis Summary CMS-PAS-SMP-14-008, 2016.

[92] CMS Collaboration, "Measurements of properties of the Higgs boson and search for an additional resonance in the four-lepton final state at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary CMS-PAS-HIG-16-033, 2016.

université PARIS-SACLAY | ÉCOLE DOCTORALE PHENIICS

Title: Electron studies and search for vector boson scattering in events with four leptons and two jets with the CMS detector at the LHC

Key words: *Vector boson scattering, CMS experiment, LHC collider, standard model, electrons*

Abstract:

This thesis reports the first experimental investigation into vector boson scattering (VBS) in the ZZ channel, where both Z bosons are required to decay into electrons or muons and are accompanied by at least two hadronic jets ($ZZjj \to \ell\ell\ell'\ell' jj$, where $\ell, \ell' = e$ or μ). VBS is a key process in elucidating the physics of electroweak symmetry breaking and the role of the recently discovered Higgs boson. This study analyses 35.9 fb^{-1} of proton-proton collisions collected with the CMS experiment at the CERN Large Hadron Collider at a center-of-mass energy of 13 TeV. A multivariate analysis technique is exploited to separate the electroweak signal from the QCD irreducible background and to measure the signal strength μ,i.e., the ratio of the observed number of events to the standard model expectation. The observed signal strength is $\mu = 1.39^{+0.86}_{-0.65}$ which excludes the background-only hypothesis at 2.7 standard deviations (1.6 standard deviations expected). Limits on physics beyond the standard model are derived in terms of anomalous quartic gauge couplings in the effective field theory approach, providing the most stringent constraints to date on the couplings for the operators $\mathcal{O}_{T,8}$ and $\mathcal{O}_{T,9}$.

Multilepton analyses like the search for VBS in the ZZ channel depend on the ability to efficiently reconstruct and identify the final state leptons. This work presents the optimizations on the multivariate electron identification algorithms used in the first data at 13 TeV in 2015. A study on extending the use of tracking information in the MVA resulted in the reduction of the non-prompt electron background by up to 50 %. Monitoring changes to the reconstructed electron objects and continuous optimizations allowed to improve the performance of the electron MVA ID algorithms, despite the harsher pileup conditions in the 2016 data. The electron efficiency measurements performed for the 2016 multilepton analyses in CMS are also documented.

Titre : Identification des électrons et mise en évidence de la diffusion de bosons massifs dans les événements à quatre leptons et deux jets avec le détecteur CMS auprès du LHC

Mots-clés : *diffusion de bosons massifs, expérience CMS, collisioneur LHC, modèle standard, électrons*

Résumé :

Cette thèse présente la première étude expérimentale de la diffusion de bosons massifs (VBS) dans le canal ZZ où les deux bosons Z se désintègrent en muons ou en électrons et sont associés à deux jets hadroniques ($ZZjj \to \ell\ell\ell'\ell' jj$, avec $\ell, \ell' = e$ ou μ). VBS constitue un processus clé dans la compréhension de la physique de la brisure de la symétrie électrofaible et du rôle du boson de Higgs découvert en 2012. Cette étude exploite 35.9 fb^{-1} de collisions proton-proton enregistrées avec le détecteur CMS au Grand collisionneur des hadrons (LHC) à $\sqrt{s} = 13$ TeV. Une analyse multivariée est utilisée pour séparer le signal électrofaible du bruit de fond irréductible QCD et pour mesurer la force du signal μ, définie comme le quotient des taux d'événements observés et attendus. La force du signal observée est de $\mu = 1.39^{+0.86}_{-0.65}$, excluant l'hypothèse de l'absence de signal à 2.7 écarts-types (1.6 écart-types attendu). Des limites sur la physique au-delà du modèle standard sont placées sur les couplages quartiques anomaux dans le cadre de la théorie des champs effective, fournissant les limites les plus strictes sur les couplages des opérateurs $\mathcal{O}_{T,8}$ et $\mathcal{O}_{T,9}$.

Les analyses multi-leptons telles que la recherche du processus VBS dans le canal ZZ reposent sur la capacité de reconstruire et identifier de façon efficace les leptons de l'état final. Cette thèse présente les optimisations de l'algorithme multivarié d'identification des électrons utilisé dans les premières données à 13 TeV en 2015. L'exploitation des variables liées aux traces des électrons permet de réduire de 50 % le bruit de fond des électrons non prompts. Grâce au suivi des changements dans la reconstruction des électrons et à une optimisation continue des algorithmes, les performances d'identification des électrons ont été préservées dans la prise de données de 2016, malgré l'empilement plus sévère. Les mesures d'efficacité de sélection des électrons effectuées pour les études multi-leptons de CMS sont également passées en revue.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France