

Dipartimento di / Department of

FISICA "Giuseppe Occhialini"

Dottorato di Ricerca in / PhD program FISICA e ASTRONOMIA Ciclo / Cycle XXXIV

Curriculum: FISICA SUBNUCLEARE

# First evidence of VBS in semileptonic decays with $WVjj \rightarrow lvqqjj$ final state and optimization of the CMS electromagnetic calorimeter for Run III

Cognome / Surname VALSECCHI Nome / Name DAVIDE

Matricola / Registration number 785396

Tutore / Tutor: Prof. PIETRO GOVONI

Cotutore / Co-tutor: Dr. ANDREA MASSIRONI

Supervisor: Prof. MARCO PAGANONI

Coordinatore / Coordinator: Prof. MARTA CALVI

ANNO ACCADEMICO / ACADEMIC YEAR 2020/2021

# Abstract

Two main contributions are presented in this thesis: the first evidence in the search for electroweak (EW) vector boson scattering (VBS) in semileptonic decays with  $WVjj \rightarrow lvqqjj$  final state, and the optimization of the CMS electromagnetic calorimeter (ECAL) trigger and local reconstruction for Run III of LHC.

The VBS search uses a data set of proton-proton collisions at 13 TeV, corresponding to an integrated luminosity of  $137 \text{ fb}^{-1}$  collected with the CMS detector during 2016-2018 (Run II). VBS is important because of its strong link with the EW symmetry breaking mechanism (the Higgs mechanism) in the Standard Model. Events are selected requiring one lepton (electron or muon), moderate missing transverse momentum, two jets with a large pseudorapidity separation and a large dijet invariant mass, and a signature consistent with the hadronic decay of a W/Z boson. Events are separated into two categories: either the hadronically decaying W or Z boson is reconstructed as one large-radius jet, or it is identified as a pair of jets with dijet invariant mass close to the bosons mass. The signal strength is measured by fitting the distributions of machine learning based multivariate discriminators, implemented to separate the signal from the backgrounds in each category. The observed EW signal strength is  $\mu_{\text{EW}} = 0.86^{+0.23}_{-0.21} = 0.86 \pm 0.12 \text{ (stat)}^{+0.19}_{-0.17} \text{ (syst)}$ , corresponding to a signal significance of 4.4 standard deviations (5.1 expected). The result reported in this thesis corresponds to the first evidence of vector boson scattering in a semileptonic channel at the LHC.

The large total integrated luminosity delivered by LHC to the CMS detector and the increasing level of simultaneous interactions (pileup) has induced a dynamic and challenging environment for the performance of the ECAL detector during Run II. Ageing effects due to radiation damage have affected both the crystal transparency and the signal pulse shape. The performance of ECAL trigger has been evaluated using these data, and the energy reconstruction algorithms have been adapted to cope with the even more challenging conditions posed by the start of Run III in 2022.

The ECAL trigger on-detector hardware employs fast digital algorithms to precisely measure the energy and timing information of ECAL energy deposits. This thesis describes the optimization of the digital filter weights which improves the robustness of the energy estimation against pileup. Moreover a novel hardware configuration is explored for the first time to help reject anomalous signals or tag out-of-time energy deposits.

The reconstruction of electrons and photons in CMS depends on topological clustering of the energy deposited by an incident particle in different crystals of the electromagnetic calorimeter (ECAL). The presence of upstream material causes electrons and photons to start showering before reaching the calorimeter. Due to the 3.8 T CMS magnetic field, the energy is spread in several clusters around the primary one. It is essential to recover the energy contained in these satellite clusters in order to achieve the best possible energy resolution for physics analyses. Historically, satellite clusters have been associated to the primary cluster using a purely topological algorithm which does not attempt to remove spurious energy deposits from additional

pileup interaction. The performance of this algorithm is expected to degrade during LHC Run III because of the larger average pileup levels and the increasing levels of noise due to the ageing of the ECAL detector. New methods have been investigated in this thesis to exploit state-of-the-art deep learning architectures like Graph Neural Networks (GNN) and Transformer algorithms. These more sophisticated models, implemented for the first time in ECAL, may improve the energy resolution up to 10% and are more resilient to pileup and noise, helping to preserve the electron and photon energy resolution achieved during LHC Runs I and II.

## Contents

|          |                                                                        |           |
|----------|------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction and experimental setup</b>                             | <b>7</b>  |
| 1.1      | The Large Hadron Collider . . . . .                                    | 7         |
| 1.2      | The Compact Muon Solenoid . . . . .                                    | 9         |
| 1.2.1    | Trigger . . . . .                                                      | 11        |
| 1.2.2    | Event reconstruction . . . . .                                         | 12        |
| 1.3      | The Electromagnetic Calorimeter (ECAL) . . . . .                       | 14        |
| 1.3.1    | ECAL layout and geometry . . . . .                                     | 14        |
| 1.3.2    | Lead tungstate crystals and photodetectors . . . . .                   | 15        |
| 1.3.3    | ECAL transparency monitoring and correction . . . . .                  | 16        |
| 1.3.4    | ECAL Trigger and data acquisition system . . . . .                     | 16        |
| 1.3.5    | Off-detector electronics . . . . .                                     | 19        |
| 1.4      | ECAL Trigger Primitive generation . . . . .                            | 20        |
| 1.4.1    | TP generation algorithm . . . . .                                      | 20        |
| 1.4.2    | Calibration . . . . .                                                  | 24        |
| 1.4.3    | Spike reduction . . . . .                                              | 25        |
| 1.4.4    | Run II performance . . . . .                                           | 29        |
| 1.5      | The ECAL offline reconstruction . . . . .                              | 30        |
| 1.5.1    | ECAL local reconstruction and Particle Flow . . . . .                  | 30        |
| 1.5.2    | Rechits . . . . .                                                      | 32        |
| 1.5.3    | PFClustering . . . . .                                                 | 33        |
| 1.5.4    | Mustache SuperClustering . . . . .                                     | 34        |
| 1.5.5    | Run II performance . . . . .                                           | 35        |
| <b>2</b> | <b>Optimization of the CMS Electromagnetic Calorimeter for Run III</b> | <b>37</b> |
| 2.1      | Trigger weights optimization for Run III . . . . .                     | 37        |
| 2.1.1    | Weights granularity . . . . .                                          | 37        |
| 2.1.2    | Optimization for pileup . . . . .                                      | 37        |
| 2.1.3    | Performance validation . . . . .                                       | 42        |
| 2.1.4    | Data - MC comparison . . . . .                                         | 44        |
| 2.2      | Double weights mechanism . . . . .                                     | 47        |
| 2.2.1    | Electronics behavior . . . . .                                         | 47        |
| 2.2.2    | Useful scenarios . . . . .                                             | 49        |
| 2.2.3    | Tests and implementation . . . . .                                     | 51        |
| 2.2.4    | Preliminary performance study . . . . .                                | 54        |
| 2.3      | SuperClustering with machine learning . . . . .                        | 57        |
| 2.3.1    | Truth-matching . . . . .                                               | 57        |
| 2.3.2    | Training dataset . . . . .                                             | 61        |
| 2.3.3    | Graph neural network model . . . . .                                   | 63        |
| 2.3.4    | Loss and training . . . . .                                            | 65        |
| 2.3.5    | Performance . . . . .                                                  | 65        |
| 2.3.6    | Conclusions . . . . .                                                  | 78        |



|          |                                                                                                                       |           |
|----------|-----------------------------------------------------------------------------------------------------------------------|-----------|
| <b>3</b> | <b>Search for vector boson scattering in semileptonic decays with <math>WV \rightarrow lvqqjj</math> final states</b> | <b>81</b> |
| 3.1      | Theoretical introduction . . . . .                                                                                    | 81        |
| 3.1.1    | The Standard Model of particle physics . . . . .                                                                      | 81        |
| 3.1.2    | The Higgs mechanism . . . . .                                                                                         | 84        |
| 3.1.3    | Vector Boson Scattering . . . . .                                                                                     | 85        |
| 3.1.4    | Overview of the recent experimental results . . . . .                                                                 | 86        |
| 3.2      | Analysis strategy . . . . .                                                                                           | 88        |
| 3.2.1    | Signal and background topology . . . . .                                                                              | 89        |
| 3.3      | Data sets and MC simulations . . . . .                                                                                | 93        |
| 3.3.1    | Data sets . . . . .                                                                                                   | 93        |
| 3.3.2    | Signal MC simulation samples . . . . .                                                                                | 93        |
| 3.3.3    | Background MC simulation samples . . . . .                                                                            | 95        |
| 3.3.4    | MC simulation configuration . . . . .                                                                                 | 95        |
| 3.3.5    | Dipole shower simulation comparison . . . . .                                                                         | 100       |
| 3.4      | Object definition and event selection . . . . .                                                                       | 106       |
| 3.4.1    | Object definition . . . . .                                                                                           | 106       |
| 3.4.2    | Events selection and categories . . . . .                                                                             | 106       |
| 3.4.3    | Jet tagging efficiency . . . . .                                                                                      | 108       |
| 3.5      | Backgrounds . . . . .                                                                                                 | 109       |
| 3.5.1    | Top quark background . . . . .                                                                                        | 109       |
| 3.5.2    | W+jets background . . . . .                                                                                           | 116       |
| 3.5.3    | Non-prompt background . . . . .                                                                                       | 123       |
| 3.6      | Quark Gluon Likelihood calibration . . . . .                                                                          | 127       |
| 3.6.1    | Correction strategy . . . . .                                                                                         | 128       |
| 3.6.2    | Application of the QGL morphing . . . . .                                                                             | 129       |
| 3.6.3    | QGL morphing uncertainties . . . . .                                                                                  | 134       |
| 3.7      | Signal discrimination . . . . .                                                                                       | 135       |
| 3.7.1    | Strategy . . . . .                                                                                                    | 135       |
| 3.7.2    | Architecture and training process . . . . .                                                                           | 137       |
| 3.7.3    | DNN input variables . . . . .                                                                                         | 138       |
| 3.7.4    | Training monitoring . . . . .                                                                                         | 144       |
| 3.7.5    | Performance . . . . .                                                                                                 | 147       |
| 3.7.6    | DNN bins kinematic regions . . . . .                                                                                  | 150       |
| 3.8      | W+jets background data-driven estimation . . . . .                                                                    | 151       |
| 3.8.1    | Data/MC discrepancies . . . . .                                                                                       | 151       |
| 3.8.2    | W+jets subcategories . . . . .                                                                                        | 154       |
| 3.8.3    | W+jets closure test . . . . .                                                                                         | 157       |
| 3.8.4    | W+jets rescaling effect . . . . .                                                                                     | 161       |
| 3.8.5    | W+jets NLO vs LO samples . . . . .                                                                                    | 166       |
| 3.9      | Systematic uncertainties . . . . .                                                                                    | 169       |
| 3.9.1    | Experimental systematic uncertainties . . . . .                                                                       | 169       |
| 3.9.2    | Theoretical uncertainties . . . . .                                                                                   | 171       |

|                    |                                                                          |            |
|--------------------|--------------------------------------------------------------------------|------------|
| 3.9.3              | Special treatment for normalization uncertainty for Top and W+jets . . . | 172        |
| 3.9.4              | Impact plots . . . . .                                                   | 173        |
| 3.10               | Results . . . . .                                                        | 176        |
| 3.10.1             | Statistical setup . . . . .                                              | 176        |
| 3.10.2             | Inputs to the likelihood fits . . . . .                                  | 177        |
| 3.10.3             | Measurements . . . . .                                                   | 182        |
| 3.11               | Fit checks . . . . .                                                     | 186        |
| 3.11.1             | W+jets post-fit normalizations summary . . . . .                         | 186        |
| 3.11.2             | Postfit plots . . . . .                                                  | 188        |
| 3.11.3             | Residual discrepancies bias test . . . . .                               | 188        |
| 3.12               | Future perspectives . . . . .                                            | 193        |
| <b>Conclusions</b> |                                                                          | <b>195</b> |
| <b>References</b>  |                                                                          | <b>197</b> |



## Chapter 1

# Introduction and experimental setup

### 1.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a two-ring superconducting accelerator and collider installed in the 27 km long Large Electron-Positron (LEP) [1]. The prime motivation of the LHC is the investigation of electroweak symmetry breaking, for which the Higgs mechanism is presumed to be responsible. The search for the Higgs boson is not the only motivation of the LHC. The accelerator allows the study of the consistency of the Standard Model (SM) at the scale  $\Lambda \simeq 1$  TeV. In addition, precision studies of QCD, electroweak and flavour physics are possible, as well as a variety of searches for exotic particles and forces: these measurements might also open a window onto physics beyond the SM (BSM). The high integrated luminosity that will be recorded during its long operational program makes it possible to study one of the rarer EW processes predicted by the SM: vector boson scattering (VBS), the topic of the search performed in this thesis. Studying the VBS process at the LHC collider is important since it is a fundamental probe of the consistency of the electroweak symmetry breaking mechanism at high energy.

Given the machine design centre of mass energy and luminosity ( $6.5 + 6.5$  TeV and  $\mathcal{L} = \mathcal{O}(10^{34})$  cm<sup>-2</sup>s<sup>-1</sup> for  $p$ - $p$  collisions), the LHC represents an unprecedented challenge from the point of view of technologies and human resources involved. The high beam intensities exclude the use of anti-proton beams, one common vacuum and magnet system for both circulating beams (as it is done in the Tevatron). The LHC is therefore designed as a proton-proton collider with separate magnetic fields and vacuum chambers in the main arcs and with common sections only at the insertion regions where the experimental detectors are located. The peak beam energy in a storage ring depends on the integrated dipole field along the storage ring circumference. Aiming at peak beam energies of up to 7 TeV inside the existing LEP tunnel implies a peak dipole field of 8.33 T and the use of superconducting magnet technology, namely 1600 liquid Helium cooled superconducting magnets made of a Niobium-Titanium compound and with a running temperature of 1.9 K.

The LHC accelerator is served by a complete injection chain starting from the LINAC3 through the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS), as depicted in Fig. 1.1. The bunches of protons are prepared and accelerated to 26 GeV in the Proton Synchrotron (PS). Then they are injected in the Super Proton Synchrotron (SPS) where they are brought to the energy of 450 GeV for the final injection into the LHC. The LHC tunnel has a diameter of 3.7 m, and it is geometrically organized in eight insertion points, flanked by long straight sections and eight arcs (see Fig. 1.2). Four insertion points host collimators and the radio frequency (RF) superconducting cavities that increase the proton energy of 0.5 MeV at every turn. Four large experiments are located in the other four insertion points where the beams are made colliding between each other: CMS (Compact Muon Solenoid), AT-

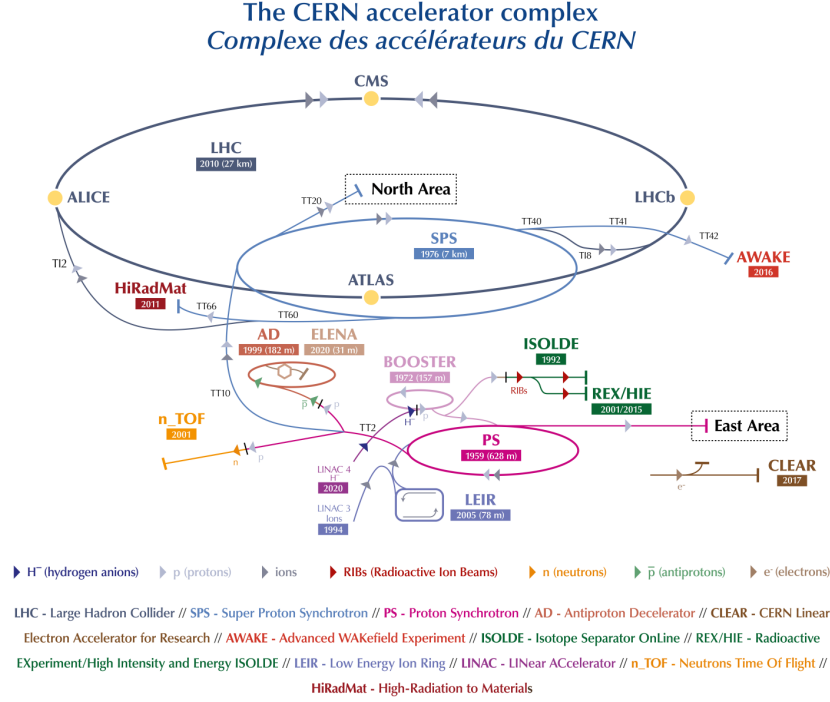


Figure 1.1: CERN accelerators complex

LAS (A Toroidal LHC ApparatuS), ALICE (A Large Ion Colliding Experiment) and LHCb (the Large Hadron Collider beauty experiment). CMS and ATLAS are two general purpose experiments, with complementary features and detector choices.

The LHC beams are divided into bunches containing  $10^{11}$  protons in nominal conditions, distributed with a 25 ns interval between them. The time taken by a bunch to complete a cycle in the LHC tunnel is called an *orbit*. The orbit is 89.1  $\mu$ s long, so it can contain up to 3564 bunches. However, the bunch structure of the beam is not continuous, and it has definite series of gaps for the purposes of synchronization, acquiring calibration data and providing resets to front-end electronics. The rate of events produced via collisions is directly proportional to the cross section of the particles and to the luminosity of the machine. For a Gaussian beam distribution the machine luminosity can be expressed as:

$$L = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta^*} F \quad (1.1)$$

where  $N_b$  is the number of bunches in each beam,  $n_b$  the number of particles per bunch,  $f_{rev}$  the revolution frequency of the beams,  $\gamma_r$  the Lorentz relativistic factor,  $\epsilon_n$  the normalized emittance of the beams, and  $\beta^*$  is the machine betatron function at the interaction point.  $F$  is a geometric factor depending on the beam crossing angle and transverse bunch size RMS.

The amount of data delivered to the experiments is measured by the integrated machine luminosity  $\int_0^t L dt$ . The instantaneous luminosity, determined by the experiments using detectors located at small angles with respect to the interaction points, is not constant during a physics run but decreases due to the degradation of the intensity and emittance of the beams.

The LHC delivered the first proton-proton collisions on the 23<sup>rd</sup> November 2009 at a center of

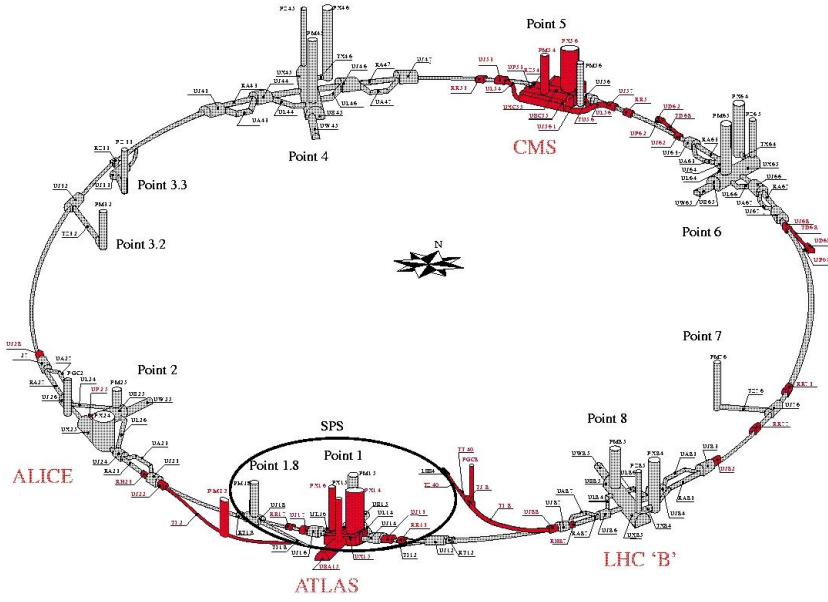


Figure 1.2: Underground schema of the LHC collider

mass energy of 0.9 TeV (450+450 GeV). A new energy record was reached 7 days after, with a collision with beams at 1.18 TeV each, delivering a total integrated luminosity of few  $\mu\text{b}^{-1}$ . During the 2010 run LHC operated at 7 TeV collecting  $45 \text{ pb}^{-1}$ , whereas during the 2011 one a peak luminosity (increasing during the data taking period) of  $\mathcal{L} = 4 \times 10^{33} \text{ cm}^{-2}\text{s}^{-1}$  in  $p$ - $p$  collisions was reached, collecting  $6 \text{ fb}^{-1}$  of data. In the 2012 run LHC had a center of mass energy of 8 TeV and a luminosity  $\mathcal{L} = 6 \times 10^{33}$ , almost flat during the whole year (3/5 of design luminosity). The LHC was shut down on 13 February 2013 for its 3-year upgrade, necessary to improve many aspects of the accelerator, enable collisions at 14 TeV and enhance its detectors and pre-accelerators (the Proton Synchrotron and Super Proton Synchrotron). The LHC restarted in 2015 with protons circulating at an energy of 6.5 TeV, giving a total center-of-mass energy of 13 TeV. During the Run II period (2015-2018) the instantaneous luminosity has increased above the design value up to peaks of  $2.1 \cdot 10^{34} \text{ cm}^{-2}\text{s}^{-1}$  during the 2018 run. Also, the number of simultaneous interaction, or pileup (PU), has increased up to a mean of 40, and peaks of 70 during 2018. At the end of Run II in December 2018, the LHC has delivered more than  $150 \text{ fb}^{-1}$  of integrated luminosity to the experiments to produce scientific results.

During the long shutdown between Run II and Run III (2019-2021) the LHC has started the preparation of its major upgrade called High Luminosity LHC (HL-LHC) that will happen between 2025 and 2027. The centre-of-mass energy will reach 14 TeV and the luminosity will be five to seven times the nominal value ( $10^{34} \text{ cm}^{-2}\text{s}^{-1}$ ). The LHC schedule is shown in Figure 1.3.

## 1.2 The Compact Muon Solenoid

The Compact Muon Solenoid (CMS) is a multipurpose detector currently operating at LHC Point 5. The CMS detector is located  $\sim 100 \text{ m}$  underground, and it has a cylindrical shape with 21.6 m length and 14.6 m diameters. The total weight of the detector is 12500 tons.

The central feature of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. Within the solenoid volume are a silicon pixel and strip



Figure 1.3: Plan for High Luminosity LHC

tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL), each composed of a barrel and two endcap sections. Forward calorimeters extend the pseudorapidity coverage provided by the barrel and endcap detectors. Muons are measured in gas-ionization detectors embedded in the steel flux-return yoke outside the solenoid. A schema of the CMS detector is shown in Fig. 1.4.

CMS uses a right-handed coordinate system, with the origin at the nominal interaction point, the x-axis pointing to the centre of the LHC, the y-axis pointing up (perpendicular to the LHC plane) and the z-axis along the anticlockwise-beam direction. The polar angle  $\theta$ , is measured from the positive z-axis and the azimuthal angle,  $\phi$  is measured in the x-y plane. The pseudorapidity is defined as  $\eta = -\log \tan(\theta/2)$ .

The large proton-proton total cross-section, the high beam intensity and the short spacing between bunches pose very challenging requirements on the detector in terms of radiation-tolerance, online data reduction and time resolution. Moreover, the large multiplicity of the interactions, up to 40 proton-proton collisions per bunch-crossing at nominal luminosity (in Run II), can produce more than 1000 charged particles, thus a very high detector granularity is required.

In order to deal with high track multiplicities, CMS employs 13-14 layers (depending on the region of the detector) of silicon microstrip detectors and 3 additional layers of silicon pixel detectors placed close to the interaction region to improve the measurement of the impact parameter of charged-particle tracks, as well as the position of secondary vertices, produced by the decay of long-living particles as b-quark mesons. The ECAL uses lead tungstate crystals with coverage in pseudorapidity up to  $|\eta| < 3.0$ . A preshower system is installed in front of the ECAL endcaps for improved  $\pi^0/\gamma$  discrimination. The ECAL is surrounded by a brass/scintillator sampling hadron calorimeter with coverage up to  $|\eta| < 3.0$  completed also by "tail-catcher" forward calorimetry. In the region  $|\eta| < 1.74$ , the HCAL cells have widths of 0.087 in pseudorapidity and 0.087 in azimuth ( $\phi$ ). In the  $\eta$ - $\phi$  plane, and for  $|\eta| < 1.48$ , the HCAL cells map

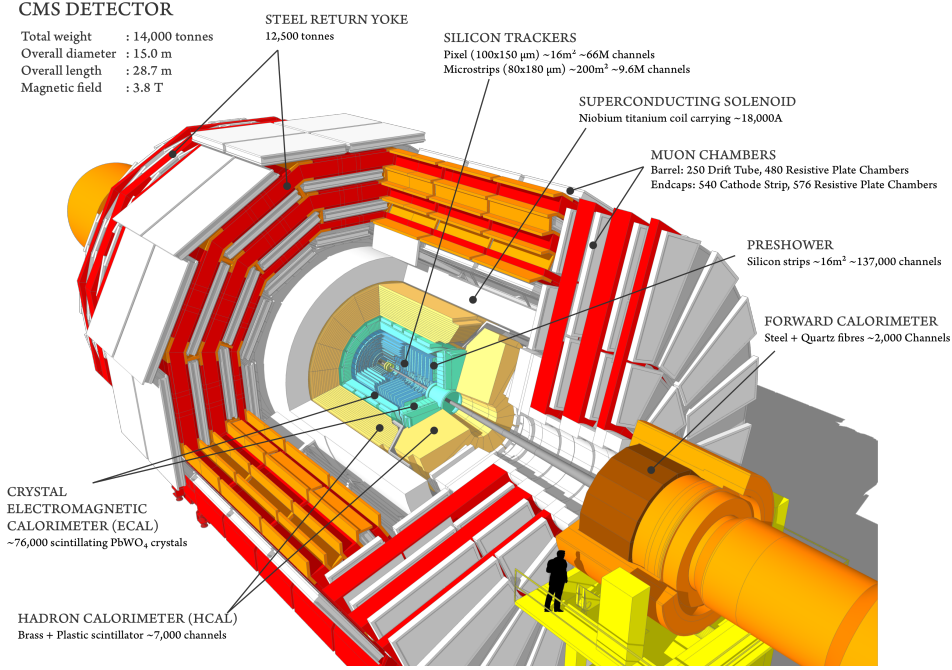


Figure 1.4: The CMS detector overview

on to  $5 \times 5$  arrays of ECAL crystals to form calorimeter towers projecting radially outwards from close to the nominal interaction point. For  $|\eta| > 1.74$ , the coverage of the towers increases progressively to a maximum of 0.174 in  $\Delta\eta$  and  $\Delta\phi$ . Muons are measured in gas-ionization detectors embedded in the steel return yoke. The muons are detected three times: in the inner tracker, after the coil and in the return flux. CMS also has extensive forward calorimetry that ensures full geometric coverage for the measurement of the transverse energy in the event. A more detailed description of the CMS detector can be found in Ref. [2].

### 1.2.1 Trigger

During LHC Run II collisions occurred at up to 40 MHz, whereas the CMS event storage system has been designed to sustain a rate up to 1 kHz, so an effective online reduction system is necessary to achieve the required  $\sim 10^5$  reduction factor. This task is accomplished by the Trigger and Data Acquisition (DAQ) system. Indeed, the cross-sections of processes studied by LHC, such as Higgs and multiboson production, are more than 9 orders of magnitude lower than the elastic pp cross-section, so the trigger system represents the start of the physics event selection process in order to retain only the most interesting data for further offline analysis.

The CMS trigger and DAQ system consists of 4 parts: the detector readout and control electronics, the Level-1 trigger processors, the readout network, and the online event filter system (a processor farm) that runs the software of the High-Level Trigger (HLT). The Level 1 trigger (L1) consists of custom designed, programmable electronics, which uses information from the calorimeters and muon detectors to select events at a rate of up to 100 kHz within a fixed latency of about  $4 \mu\text{s}$  [3]. The High-Level trigger (HLT) consists of a farm of processors running a version of the full event reconstruction software optimized for fast processing, and reduces the event rate to around 1 kHz before data storage [4].



### 1.2.2 Event reconstruction

The global event reconstruction (also called particle-flow event reconstruction [5]) aims to reconstruct and identify each individual particle in an event, with an optimized combination of all subdetector information. In this process, the identification of the particle type (photon, electron, muon, tau, charged hadron, neutral hadron) plays an important role in the determination of the particle direction and energy. Photons (e.g. coming from  $\pi^0$  decays or from electron bremsstrahlung) are identified as ECAL energy clusters not linked to the extrapolation of any charged particle trajectory to the ECAL. Electrons (coming from EW hard scattering processes or B hadron semileptonic decays) are identified as a primary charged particle track and potentially many ECAL energy clusters corresponding to this track extrapolated to the ECAL surface and to bremsstrahlung photons emitted as the particle goes through the tracker material. Muons are identified as tracks in the central tracker consistent with either a track or several hits in the muon system, and associated with calorimeter deposits compatible with the muon hypothesis. Charged hadrons are identified as charged particle tracks neither identified as electrons, nor as muons. Finally, neutral hadrons are identified as HCAL energy clusters not linked to any charged hadron trajectory, or as a combined ECAL and HCAL energy excess with respect to the expected charged hadron energy deposit.

The energy of photons is obtained from the ECAL measurement. The energy of electrons is determined from a combination of the track momentum at the main interaction vertex, the corresponding ECAL cluster energy, and the energy sum of all bremsstrahlung photons associated to the track. The energy of muons is obtained from the corresponding track momentum. The energy of charged hadrons is determined from a combination of the track momentum and the corresponding ECAL and HCAL energies, corrected for the response function of the calorimeters to hadronic showers. Finally, the energy of neutral hadrons is obtained from the corresponding ECAL and HCAL energies.

In the barrel section of the ECAL, an energy resolution of about 1% is achieved for unconverted or late-converting photons in the tens of GeV energy range. The energy resolution of the remaining barrel photons is about 1.3% up to  $|\eta| = 1$ , changing to about 2.5% at  $|\eta| = 1.4$ . In the endcaps, the energy resolution is about 2.5% for unconverted or late-converting photons, and between 3 and 4% for the remaining ones [6]. Muons are measured in the pseudorapidity range  $|\eta| < 2.4$ , with detection planes made using three technologies: drift tubes, cathode strip chambers, and resistive plate chambers. Matching muons to tracks measured in the silicon tracker results in a relative transverse momentum resolution, for muons with  $20 < p_T < 100$  GeV, of 1.3–2.0% in the barrel and better than 6% in the endcaps. The  $p_T$  resolution in the barrel is better than 10% for muons with  $p_T$  up to 1 TeV [7].

The candidate vertex with the largest value of summed physics-object  $p_T^2$  is taken to be the primary pp interaction vertex. The physics objects are the jets, clustered using the jet finding algorithm [8, 9] with the tracks assigned to candidate vertices as inputs, and the associated missing transverse momentum, taken as the negative vector sum of the  $p_T$  of those jets.

For each event, hadronic jets are clustered from these reconstructed particles using the infrared and collinear safe anti- $k_T$  algorithm [8, 9] with a distance parameter of 0.4. Jet momentum is determined as the vectorial sum of all particle momenta in the jet, and is found from simulation to be, on average, within 5 to 10% of the true momentum over the whole  $p_T$  spectrum and detector acceptance. Additional proton-proton interactions within the same or nearby bunch crossings (pileup) can contribute additional tracks and calorimetric energy depositions to the jet momentum. To mitigate this effect, charged particles identified to be originating from pileup vertices are discarded and an offset correction is applied to correct for remaining contributions.

The pileup per particle identification algorithm (PUPPI) [10, 11] is used to mitigate the effect of pileup at the reconstructed particle level, making use of local shape information, event pileup properties, and tracking information. A local shape variable is defined, which distinguishes between collinear and soft diffuse distributions of other particles surrounding the particle under consideration. The former is attributed to particles originating from the hard scatter and the latter to particles originating from pileup interactions. Charged particles identified to be originating from pileup vertices are discarded. For each neutral particle, a local shape variable is computed using the surrounding charged particles compatible with the primary vertex within the tracker acceptance ( $|\eta| < 2.5$ ), and using both charged and neutral particles in the region outside of the tracker coverage. The momenta of the neutral particles are then rescaled according to their probability to originate from the primary interaction vertex deduced from the local shape variable, superseding the need for jet-based pileup corrections [10].

Jet energy corrections are derived from simulation to bring the measured response of jets to that of particle level jets on average. In situ measurements of the momentum balance in dijet, photon + jet, Z + jet, and multijet events are used to account for any residual differences in the jet energy scale between data and simulation [12]. The jet energy resolution amounts typically to 15–20% at 30 GeV, 10% at 100 GeV, and 5% at 1 TeV [12]. Additional selection criteria are applied to each jet to remove those potentially dominated by anomalous contributions from various subdetector components or reconstruction failures.

The missing transverse momentum vector  $\vec{p}_T^{\text{miss}}$  is computed as the negative vector sum of the transverse momenta of all the PF candidates in an event, and its magnitude is denoted as  $p_T^{\text{miss}}$  [13]. The  $\vec{p}_T^{\text{miss}}$  is modified to account for corrections to the energy scale of the reconstructed jets in the event. The PUPPI algorithm[11] is applied also to reduce the pileup dependence of the  $\vec{p}_T^{\text{miss}}$  observable.

### 1.3 The Electromagnetic Calorimeter (ECAL)

The CMS electromagnetic calorimeter (ECAL) [14] is a homogeneous calorimeter made of 75848 lead tungstate ( $\text{PbWO}_4$ ) scintillating crystals, located inside the CMS [15] superconducting solenoid magnet. The ECAL is crucial for the identification and reconstruction of photons and electrons, and for the measurement of jets and of missing transverse momentum. Electrons and photons are typically reconstructed up to  $|\eta| < 2.5$ , the region covered by the tracker, while jets are reconstructed up to  $|\eta| < 3.0$ . The ECAL detector performance is therefore central for the investigation of the electroweak (EW) and Higgs sectors of the standard model: e.g. the  $H \rightarrow \gamma\gamma$  channel provided an important contribution to the Higgs boson discovery due to the high precision of the energy response calibration of the ECAL detector, reaching a better than 1% energy resolution.

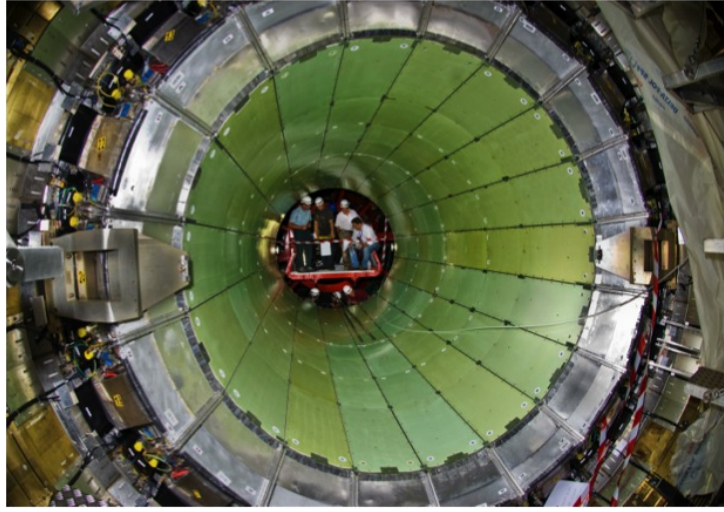


Figure 1.5: The ECAL Barrel during the CMS installation

#### 1.3.1 ECAL layout and geometry

The ECAL detector is divided into a barrel part (EB) covering the region of pseudorapidity  $|\eta| < 1.48$ , and two endcaps (EE), which extend the coverage up to  $|\eta| < 3.0$ . Scintillation light produced in EM showers is detected with avalanche photodiodes (APD) in the barrel and vacuum phototriodes (VPT) in the endcaps. In the forward region the ECAL detector is complemented by a SiPb sampling detector called preshower, located directly before the crystals. This subdetector helps to identify and discrimination prompt photons from  $\pi^0$ .

The EB is made of 36 identical supermodules, 18 in each half barrel, each covering  $20^\circ$  in  $\phi$ . The barrel granularity is 360-fold in  $\phi$  and  $(2 \times 85)$ -fold in  $\eta$ , resulting in a total of 61200 crystals. The truncated-pyramid shaped crystals are mounted in quasi-projective geometry so that their axes make a small angle ( $3^\circ$ ) with respect to the vector from the nominal interaction vertex, both in the  $\phi$  and  $\eta$  projections [14].

The endcaps cover the pseudorapidity range  $1.479 < |\eta| < 3.0$ . The longitudinal distance between the interaction point and each endcap is 314.4 cm. This location takes into account the estimated shift toward the interaction point by 2.6 cm when the 4 T magnetic field is switched on. The endcap consists of identically shaped crystals grouped in mechanical units of  $5 \times 5$  crystals (supercrystals or SC). Each endcap is divided into 2 halves, or Dees, while each Dee comprises 3662 crystals. The crystals are contained in 138 standard SCs and 18 special partial super-crystals on the inner and outer circumference. The SCs are arranged in a rectangular x-y

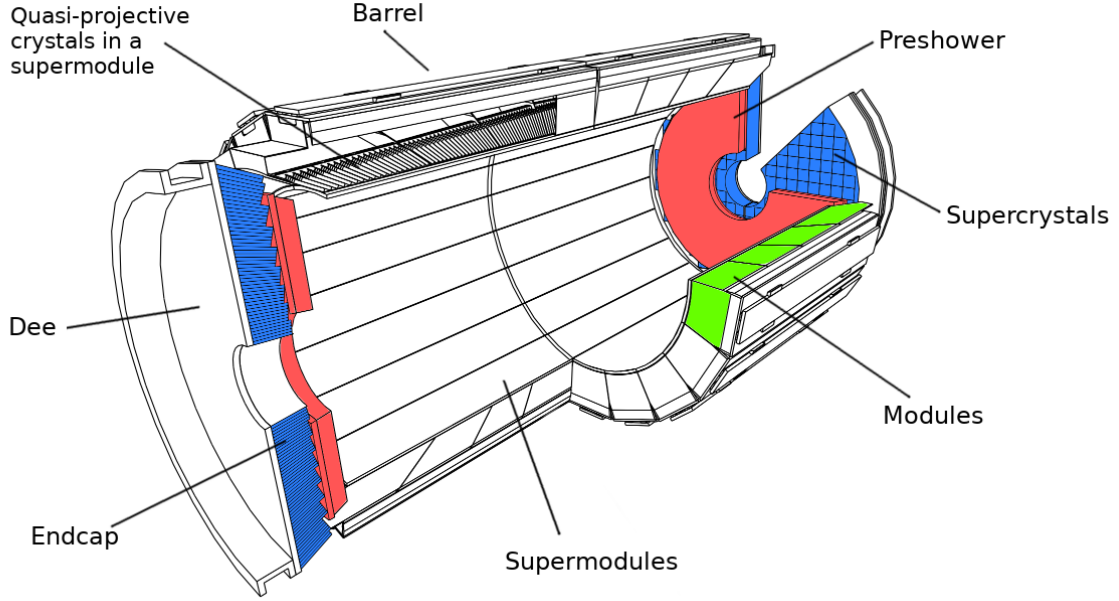


Figure 1.6: ECAL schema and geometry

grid, with the crystals pointing at a focus 1300 mm beyond the interaction point, so that the off-pointing angle varies with  $\eta$ .

### 1.3.2 Lead tungstate crystals and photodetectors

The ECAL sensitive layer consists of lead tungstate ( $\text{PbWO}_4$ ) scintillating crystals. There are 17 different shapes of crystals in the barrel. Their front faces measure  $22 \times 22 \text{ mm}^2$ , their rear faces  $26 \times 26 \text{ mm}^2$  and its length is 23.0 cm corresponding to roughly 25.8 radiation lengths ( $\chi_0$ ). In the endcap the crystals have a front face area of  $28.62 \times 28.62 \text{ mm}^2$ , a rear face of  $30 \times 30 \text{ mm}^2$  and a length of 22.0 cm,  $24.7 \chi_0$ .

The characteristics of the crystals make them an appropriate choice for operation at LHC. The high density ( $8.3 \text{ g/cm}^3$ ), short radiation length (0.89 cm) and small Molière radius (2.2 cm) results in a fine granularity and a compact calorimeter. The Molière radius is a characteristic constant of a material giving the scale of the transverse dimension of the fully contained electromagnetic showers initiated by an incident high energy electron or photon. By definition, it is the radius of a cylinder containing on average 90% of the shower energy deposition: a smaller Molière radius means better shower position resolution, and better shower separation due to a smaller degree of shower overlaps. Incident electrons and photons deposit on average 94% of the energy in the  $3 \times 3$  crystals matrix around the impact point, and 97% in a  $5 \times 5$  matrix.

The crystals emit blue-green scintillation light with a broad maximum at 420 nm. The scintillation decay time is of the same order of magnitude as the LHC bunch time: about 80% of the light is emitted in 25 ns, a characteristic that matches very well the requirements posed by the event pile-up rate expected with high-luminosity operations at the LHC. The light output is relatively low, about 4.5 photoelectrons per MeV are collected in the photodetectors of both barrel and endcap, and it depends heavily on the temperature (2.1% per  $^\circ\text{C}$  at  $18^\circ\text{C}$ ). This requires the use of efficient photodetectors and a cooling system capable of extracting the heat dissipated by the readout electronics and of keeping the crystal temperature stable within  $\pm 0.05^\circ\text{C}$ .

The light readout in the barrel is performed by a pair of APDs (Avalanche PhotoDiodes), each

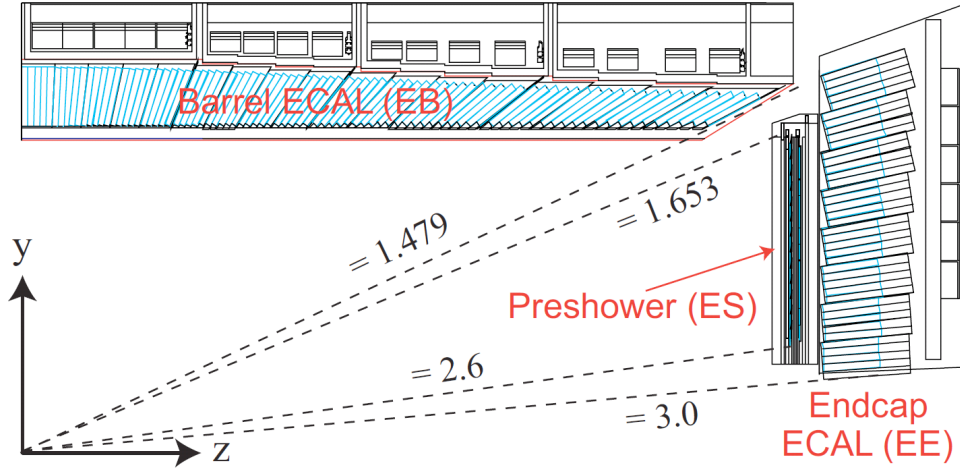


Figure 1.7: ECAL pseudorapidity coverage.

with an active area of  $5 \times 5 \text{ mm}^2$ , operated at gain 50 at a temperature of  $18^\circ\text{C}$  and read out in parallel. The gain stability directly affects the ECAL energy resolution. In the endcaps, the photodetectors are vacuum phototriodes (VPTs), photomultipliers having a single gain stage. They have an anode of very thin copper mesh allowing them to operate in the 4T magnetic field. Each VPT has a diameter of 25 mm and is glued to the back of each crystal.

The crystals have to withstand the radiation levels and particle fluxes anticipated throughout the duration of the experiment. Ionizing radiation combined with impurities in the crystals lattice produce a wavelength-dependent loss of light transmission. At the ECAL operating temperature ( $18^\circ\text{C}$ ) a dose-rate dependent equilibrium of the optical transmission is achieved. In order to preserve the detector resolution the optical transparency of crystals is monitored during data taking using laser pulses injected into the crystals via optical fibers. More details about the transparency monitoring system of ECAL are given in Sec. 1.3.3.

### 1.3.3 ECAL transparency monitoring and correction

As ECAL sustains more radiation from LHC collisions, the crystal transparency is reduced, and the light output measured by the photodetectors for a given deposited energy decreases. The loss of transparency is small in EB, about 10%, but it is more noticeable as  $\eta$  increases, along with radiation dose, in EE, as shown in Figure 1.8. This effect is measured via the ECAL laser monitoring system: a reference laser light is injected into the crystals during the orbit gap (calibration gap) of the LHC orbit with a rate of  $\sim 100 \text{ Hz}$  to calibrate the crystal response. The laser scans different detector regions in a sequence: each crystal is measured once every 40 minutes. The appropriate response correction factors are then computed during the offline reconstruction [16] and used to equalise the energy response of the calorimeter.

### 1.3.4 ECAL Trigger and data acquisition system

The ECAL trigger and data acquisition (DAQ) system hardware is physically divided in two sections connected by copper and optical links: on-detector (or front-end FE) and off-detector electronics. The former is responsible for the signal digitalization, buffering and the production of the primary information for the trigger, called trigger primitives (TP), and consists of radiation tolerant chips located inside of the detector volume. The latter is in charge of the finalization of the trigger primitive calculation, of the event readout, and of the reduction of the full-granularity data. A time frame of 10 consecutive samples is read out every 25 ns, in



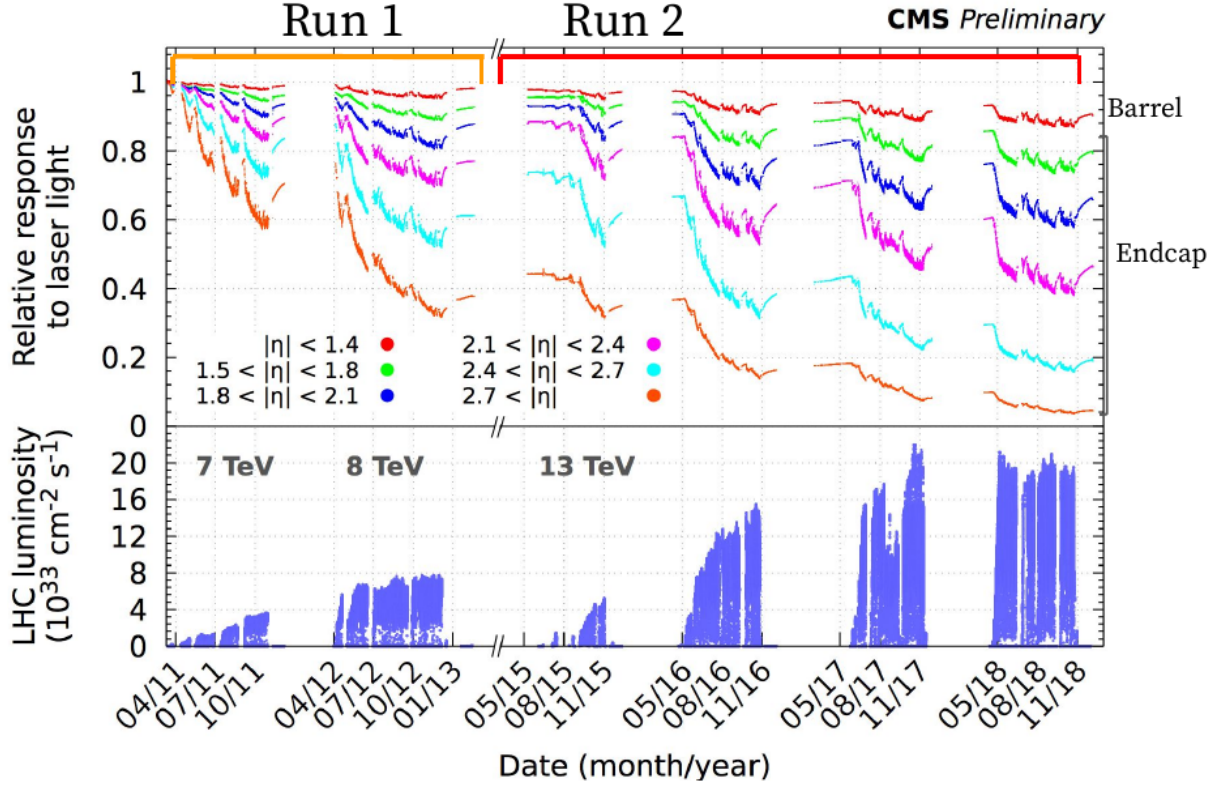


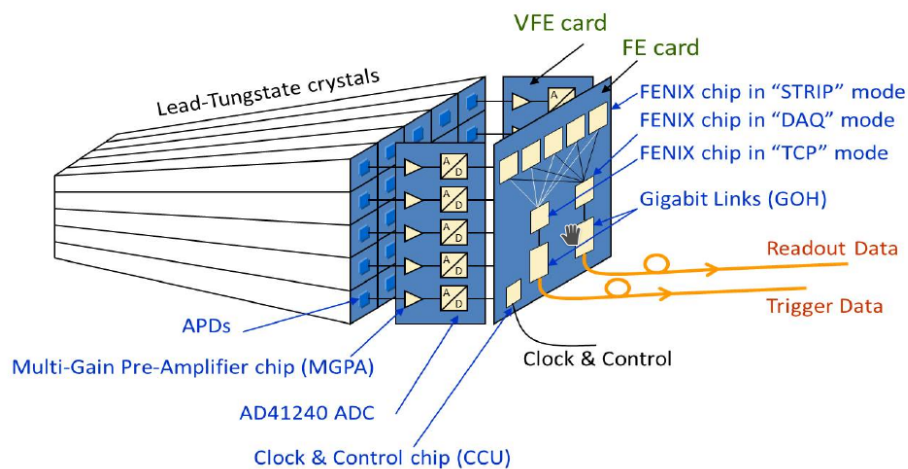
Figure 1.8: Crystal transparency measured by laser along the CMS data-taking history

synchronization with the triggered LHC BX, covering a 250 ns time interval, as depicted in Fig. 1.15. The acquisition window is positioned so that 3 BX are read before the pulse, and the peak of the pulse is on the 6th BX (counting from 1). This choice permits to measure dynamically the pedestal on which the signal pulse is superimposed. The 50 ns rise time of the signal pulse after amplification results from the convolution of the 10 ns decay time of the crystal scintillation emission and the 40 ns shaping time of the pre-amplifier.

#### 1.3.4.1 Front-end electronics

The basic detector unit is the trigger tower (TT). In the ECAL barrel the 25 crystals forming a tower have a squared  $5 \times 5$  topology in  $\eta - \phi$  space, while for the endcaps have a more complicated geometry. Every trigger tower is controlled by a group of boards called the front-end (FE): the front-end stores the full granularity crystal data in buffers waiting for the L1 trigger decision and ready to be sent to the readout off-detector boards. On the other side, the FE is in charge of the TP generation at every bunch crossing which are used in the L1 Trigger to form electron, photon, and jet candidates. Given the high rate of collisions in CMS, one BX every 25 ns, it is not possible to extract the full information from the detector sensors for all events. If we consider the LHC collisions frequency of 40 MHz, 25 ECAL crystals for each TT, and 14 bits of energy information for each crystal, an optical transfer at 14 Gb/s would be needed for each tower, a technology not available when CMS was originally built. Moreover, much larger number of optical fibers would have been needed for each FE card, which poses problems for the radiation hardness of the system. For this reason the trigger primitive information is computed directly in the ECAL FE electronics and only the events selected by the L1 trigger are read out.

Each front-end board consists of a motherboard, 5 Very Front End (VFE) boards and one Front



End (FE) board. Each VFE reads the data from 5 crystals, termed “strip”, and contains for each channel a Multi-Gain Pre-Amplifier (MGPA) with gains (1,6,12), an ADC 40 MHz digitizer, and radiation-hard buffer to send the data to the FE. The MGPA contains 3 gain stages which process the signal in parallel: the gain 12 channel has a least significant bit of  $\simeq 35$  MeV and saturates at approximately 160 GeV, the unity gain determines the upper end of the dynamic range and saturates at approximately 1.7 TeV for the barrel and 3.0 TeV for the endcap [17]. The shaping of the signal is done by a CR-RC network with a shaping time of  $\sim 40$  ns. Then the 3 MGPA analog outputs are passed to the 12-bit multi-channel ADC that digitalizes them at 40 MHz. An integrated logic selects the highest non-saturated signal as output and reports the 12-bit of the corresponding ADC with two bits encoding the ADC channel number (to store the gain used). If the read-out is switched to a lower gain as the pulse grows, the read-out is forced to continue using the lower gain for the next five samples. A schema of the front-end signal processing is shown in Fig. 1.10.

The trigger primitive generation is described thoroughly in the dedicated Sec. 1.4, in which the improvements foreseen for the Run III and developed during this thesis are also reported.

### 1.3.5 Off-detector electronics

The trigger data elaborated by the front end are sent at the LHC clock frequency of 40 MHz to the off-detector electronics, located in the service cavern  $\sim 70$  m from CMS. There, the trigger concentrator card (TCC) collects data from 68 FEs in the barrel, corresponding to a single supermodule, and from up to 48 strips in the endcaps. Four EE TCCs are required to cover a  $20^\circ\phi$  sector. The TCC completes the trigger primitive generation and sends the data to the Level 1 Trigger calorimetry layer through the Synchronization and Link Board (SLB) at each bunch crossing (BX). The synchronization of the on-detector and off-detector boards with the LHC bunches is essential to assign the data to the right bunch-crossing. The SLB is in charge of the correct synchronization procedure that relies on the Bunch Crossing Zero (BC0) broadcast command from the central Trigger Control and Distribution System (TCDS). Moreover, the TCC board is in charge of the classification of trigger towers in high, medium or low interest categories comparing the trigger primitives with configurable thresholds. This data is passed to the Selective Read Processor (SRP) which, based on the ECAL geometry and energy threshold, decides if a FE should be read with or without zero suppression. This is needed to reduce the total event size of ECAL from 40 kB/FED to 2 kB/FED in order to remain within both the DAQ system limitations for the total data rate and the speed limit of the ECAL-DAQ link connection.

Finally, the TCC stores the trigger primitives during the L1 latency for subsequent reading by the DCC board, part of the DAQ system.

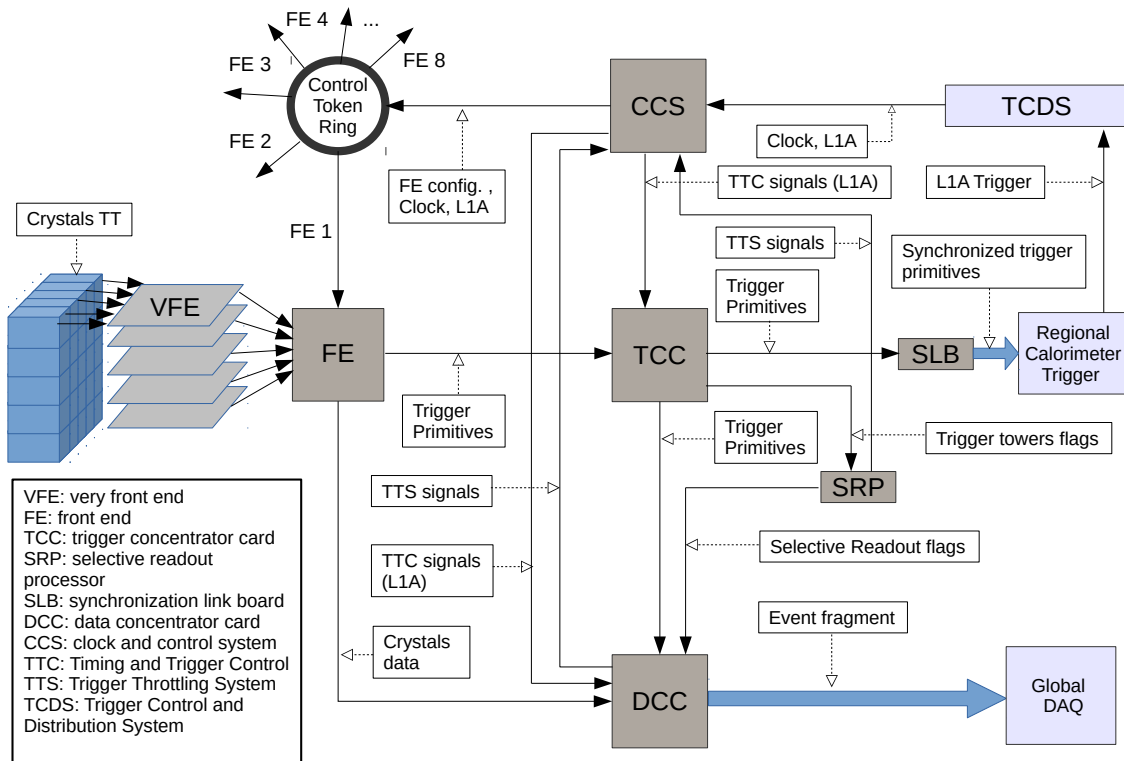


Figure 1.11: ECAL DAQ Hardware schema

The interlocutor of the TCDS system on the off-detector side is the Clock and Control System (CCS) board. It distributes the clock and other commands to several Front End boards using



token rings, dedicated optical links connected in series. The data readout starts when the L1 trigger broadcasts its trigger decision, called a L1A signal. The signal is propagated by the CCS to all the boards. The Data Concentrator Card (DCC) is the main ECAL read-out unit, and it is responsible for collecting crystal data from up to 68 FE boards (one supermodule) in the barrel case. The specific task of the DCC is the readout of the full-granularity crystal data upon a L1A decision (at the average frequency of 100 kHz). Every Front End card receives the L1A signal and sends the raw crystal data corresponding to the right bunch-crossing to the DCCs through optical links. Also, the TCC boards receive the L1A and send their trigger primitives (TP) data for each trigger tower to the DCC boards through the backplane. The DCC merges the crystal data with trigger primitive bits to save all the information elaborated in the trigger path. A schema of the ECAL off-detector electronics and the signals and data path is shown in Fig. 1.11.

## 1.4 ECAL Trigger Primitive generation

ECAL produces trigger primitives (TPs) every 25 ns that are sent to the L1 trigger system to form electrons, photons, jet candidates, and to compute transverse energy sums. These TPs comprise a transverse energy measurement for a matrix of crystals, with one or two feature bits that characterize the energy spread between crystals within the TP, and a bunch-crossing assignment. In total 15 bits are transmitted off-detector to the TCCs with optical fibers at each BX, corresponding to 800 Mb/s of data.

In EB each trigger tower is a square matrix of  $5 \times 5$  crystals, whereas in the EE the geometry is more complex: TP are not formed by tower but by *pseudostrip* (see Fig. 1.12), a group of 5 crystals that belong to the same VFE board, and sent to the TCC cards, which sums the energy of a variable number of pseudostrips following an encoded mapping. This is needed to accommodate for the EE circular geometry while keeping the amount of trigger towers constant by interval of  $\phi$  for the L1 trigger. Moreover, the ECAL crystal are geometrically organized in so called  $\eta$ -rings from the trigger point of view as shown in Fig. 1.13: in EB there are the  $\eta$ -rings  $\in [1, 17]$ , while in EEs the  $\eta$ -rings  $\in [18, 28]$ .

### 1.4.1 TP generation algorithm

The ECAL TPs are computed on-detector by custom ASICs, named FENIX chips, located on the Front End (FE) boards. The readout of the analog amplitudes by the VFE boards has been described in Sec 1.3.4.1. Here we will focus on the algorithm performed in the FENIX chips to form calibrated trigger primitives for each tower (in EB) or pseudostrip (in EE). Figure 1.14 gives a schematic representation of the Trigger Primitive Generation (TPG) algorithm [18]. There are 3 types of FENIX chips located on the FE board (as shown in Fig. 1.9): FENIXes operate in strip mode, TCP mode or DAQ mode. The behavior of the strip and TCP FENIX will be described in the following, whereas the DAQ FENIX is the chip responsible for storing the full crystal information for the readout of triggered events.

#### 1.4.1.1 Response linearization

For each crystal of a pseudo-strip, the VFE board digitizes the amplitude every 25 ns with 12 bits of precision, with 2 additional bits to signal the selected gain. The 5 signals of each pseudostrip must be manipulated in order to be summed consistently: this process is called linearization. The signal baseline, named pedestal (P), is subtracted, and a multiplier is applied to each sample, depending on the gain information, to equalize the amplitudes. Then, a

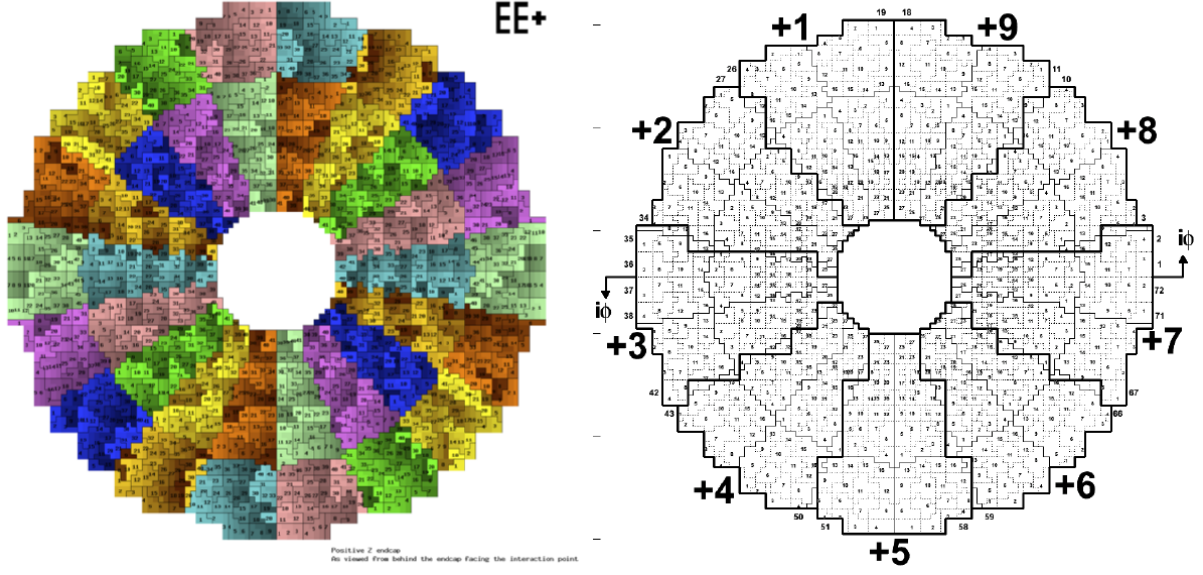


Figure 1.12: ECAL EE trigger tower: on the left the coloring shows each different pseudostrip, on the right the final trigger towers boundaries are shown.

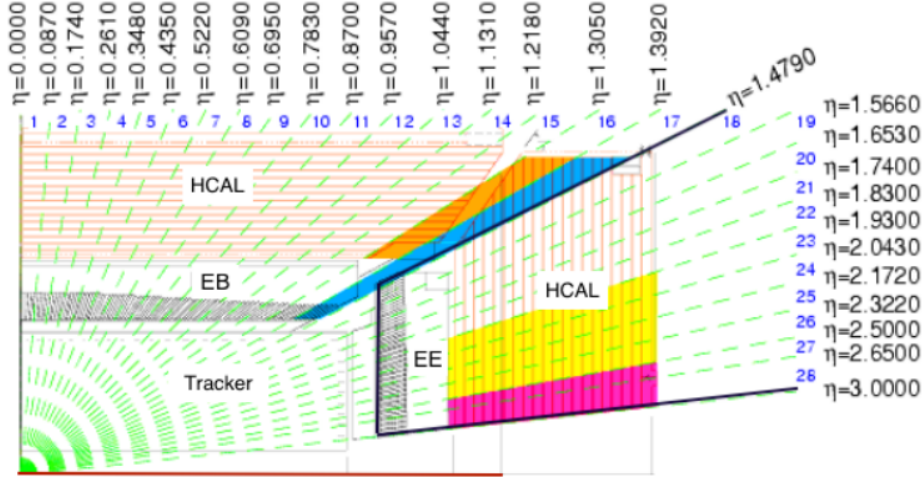


Figure 1.13: Organization of ECAL crystals in  $\eta$ -ring from the trigger point of view.

calibration factor is multiplied to restore the proportionality of the pulse to the energy deposition. The multiplication factor is computed taking into account the transparency correction of the crystal ( $f_{\text{trans}}$ ), the amount of energy corresponding to each bit of the ADC, the correction of the variation of response between crystals (intercalibration IC) and the geometrical factor to compute the transverse energy,  $\sin \theta$ .

$$E_T = (A - P) \times \text{ADCtoGeV} \times \frac{g_{12}}{g_i} \times f_{\text{trans}} \times IC \times \sin \theta \quad (1.2)$$

where  $g_{12}/g_i$  is the gain correction factor relative to gain 12, and it is 12 for gain 1 and 2 for gain 6.

This linearization operation is performed in the strip FENIXs with a 8-bit multiplication and a 4-bit shift (division by powers of 2). Since the multiplication of a 12-bit number with 8-bits gives a 20-bits word, only the 18 most significant bits (MSB) are conserved, using an additional

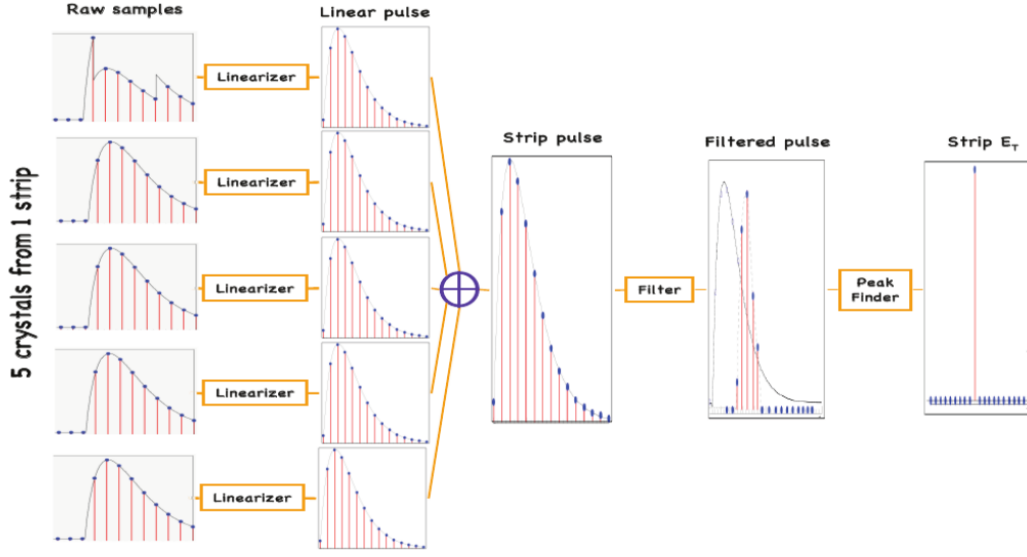


Figure 1.14: ECAL TPG formation algorithm

shift of 2. Mathematically, the transfer function of the linearization is:

$$\text{Linearize}(x) = L(x) = \left[ (x - P) \times \frac{\text{mult}}{2^{\text{shift}+2}} \right] \geq 0 \quad (1.3)$$

where only the integer part of the result is kept,  $x \in [0, 2^{12} - 1]$ ,  $P \in [0, 2^{12} - 1]$ ,  $\text{mult} \in [0, 2^8 - 1]$ ,  $\text{shift} \in [0, 2^4 - 1]$ . The computation of the mult and shift parameters is described in section 1.4.2.

#### 1.4.1.2 Sum

After the linearization, the signals are summed over an 18-bit number for each strip.

$$\text{Sum}(x_1, x_2, \dots, x_5) = \min \left( \sum_{i=1}^5 L(x_i), 2^{18} - 1 \right) \quad (1.4)$$

#### 1.4.1.3 Filter

The role of the next step of the algorithm is to estimate the amplitude of the signal pulse by means of a linear digital filter, easily implementable in the strip FENIX ASICs. The number of weights in the linear filter should be enough to give a precise amplitude estimation of the pulse: the knowledge of the signal shape is approximate and better known around the peak, moreover the probability for the pulse to be distorted by the presence of pileup diminishes when a smaller number of points are used. The optimum resolution is found using 5 consecutive points around the peak[19], as shown in Fig. 1.15. The filter is applied as a weighted sum of the signal amplitudes: using 7-bit signed weights (6-bit value and 1-bit for the sign. Two's complement is then used for negative numbers,  $w < 0 \rightarrow w = 2^7 - |w|$ ), and a resolution of the order 1.5% ( $1/2^6$ ) can be obtained, enough for the trigger primitive precision. Since the weights are distributed in the interval  $[-1, 1]$ , but are encoded with 6-bit values a shift of 6 bits is necessary. The output of this step is an 18-bit number. The filter is applied in a sliding window fashion: at each new BX the pulse is shifted and the filter output is recomputed.

$$\text{Filter}(x) = \max \left( \sum_{j=1}^5 [x_j \times \frac{w_j}{2^6}], 0 \right) \quad (1.5)$$

A constraint is imposed on the weights,  $\sum w_i = 0$ , in order to be insensitive to the pedestal value. The formula to compute pedestal subtracting weights, assuming no noise correlation between samples is:

$$W_{A,i} = \frac{f_i - \frac{\sum_j^N f_j}{N}}{\sum_j^N f_j^2 - \frac{(\sum_j^N f_j)^2}{N}} \quad (1.6)$$

Details about the weights algebraic definition can be found in Ref. [19].

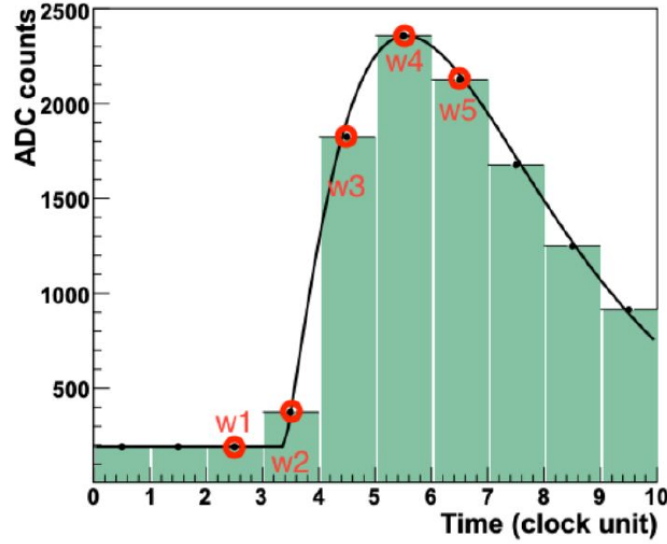


Figure 1.15: ECAL readout window of 10 BX around the peak. Filter weights are applied so that the 4th weight is on the peak.

The weights used currently by ECAL are reported in Tab. 1.1. The first two weights are negative and applied before the rise of the peak to subtract the contribution from the pedestal.

| Subdet | $W_1$     | $W_2$     | $W_3$    | $W_4$    | $W_5$    |
|--------|-----------|-----------|----------|----------|----------|
| EB     | -0.656250 | -0.515625 | 0.250000 | 0.515625 | 0.406250 |
| EE     | -0.562500 | -0.546875 | 0.250000 | 0.484375 | 0.375000 |

Table 1.1: Run I and II TPG weights for EB and EE.

#### 1.4.1.4 Peak-finder

Since the amplitude is computed in a sliding-window fashion, the energy has to be associated to the correct BX for use in the Level 1 Trigger. The filter is optimized so that the maximum amplitude is obtained when the weights are applied in the correct position around the pulse: when the 4th weight is on the peak, the energy is associated to a collision happened in the BX where the 1st weight sits. Therefore, a peak-finder step is implemented which compare each amplitude to the previous and next one to detect a maximum. This algorithm introduces a 1

BX latency in the TP calculation since the amplitude computed in the next BX is needed. The output of the peak-finder is a series of zeroes with some amplitude  $> 0$  representing the energy deposited in that strip in the associated BX 1.14. These amplitudes are coded on 12 bits, the choice of 12 bits among the 18 results from a sliding window that specifies which is the first considered bit, the 11 most significant bits then being selected. If the dynamic range of the signal exceeds 12 bits, there is saturation. During Run I and II the sliding window parameter has been set to 0 to avoid any loss in precision. The final output of the strip amplitude estimation is then:

$$\text{Strip}(x) = \min \left( \left\lceil \frac{\text{peak}(x)}{2^{\text{sliding}}} \right\rceil, 2^{12} - 1 \right) \quad (1.7)$$

#### 1.4.1.5 Tower sum

The  $N$  signals from the pseudostrips making a trigger tower ( $N=5$  strips in the barrel, and  $\leq 5$  in the endcaps) are sent from the strip FENIX to the Trigger Concentrator Processor (TCP) FENIX chip in the case of EB, while they are sent directly to the TCC boards off-detector in EE. In EB, the TCP FENIX sums the pulses and reduces the amplitude a 10-bit precision. Only the 10 maximum significant bits are retained and if the signal is larger than 12-bits, a saturation value is returned.

$$\text{Tower-sum}(X_1, X_1, \dots, X_5) = \min \left( \sum_{i=1}^5 \frac{\text{Strip}(x_i)}{2^2}, 2^{10} - 1 \right) \quad (1.8)$$

#### 1.4.1.6 Compression

The ECAL trigger primitive amplitudes expected as input by the L1 trigger are 8-bit integers. For each trigger tower we have obtained a 10-bit amplitude. To avoid more resolution degradation just applying additional shifts, a flexible compression algorithm is implemented in the form of a Look Up Table (LUT). The LUT consists of a map which associates each of the 1024 ( $= 10$  bits) values to 256 ( $= 8$  bits) values. This operation is performed by the TCC board off-detector, since 1024 values of 8 bits are needed, too much memory for the frontend board. During Run I and II, a linear LUT has been implemented associating 1 bit increment in the output every 4 bit increment in the input. This corresponds in a reduction of  $1/4$  in resolution linearly.

#### 1.4.2 Calibration

The resolution and the range of the ECAL trigger primitives depend on the parameters implemented in the algorithm described above. In particular the linearization step is the one where the pulse read by the amplification and digitization chain needs to be converted in a transverse energy estimation. The main parameters are the *mult*, and *shift*, to be computed for each crystal. The maximum range of the TPs is fixed to  $E_T = 128$  GeV in order to be sensitive the energy range of photons and electrons produced by Higgs and EWK physics processes. Moreover the sensitivity of the digitizer, namely how much energy corresponds to 1 ADC count (also called *ADctoGeV*) is fixed by the electronics to 39 MeV in the barrel and 63 MeV in the endcaps.

To extract suitable values for the parameters given the saturation energy, the overall TPG algorithm can be summarized as:

$$tpg(A) = \text{LUT}^{10b \rightarrow 8b} \left( \frac{1}{2^2} \sum_{\text{strip}=1}^N \left[ \frac{1}{2^{\text{sliding}}} \sum_{\text{xtal}=1}^5 \left[ A_{\text{xtal}}^{\text{strip}} \times \frac{\text{mult}}{2^{\text{shift}+2}} \right] \right] \right) >= 0 \quad (1.9)$$

In order to get a simplified formula for *mult* and *shift*, we can assume to be in the BX of the peak and that the LUT does not change the dynamic range of the signal, but it decreases only the precision. In other words instead of considering the 8-bit signal saturated after the LUT, we consider the 10-bit signal saturated before it.

$$tpg(A) \simeq \sum_{\text{xtal}=1}^{N \leq 25} A_{\text{xtal}} \times \frac{\text{mult}}{2^{\text{shift}+\text{sliding}+4}} \quad (1.10)$$

If the signal is saturated already in a single crystal, we would like to obtain a saturated TP:  $E_T^{\text{max}} = 128 \text{ GeV} = tpg(A^{\text{max}} = 1024 \text{ (10-bit)})$ . Recalling the equation 1.2 for the transverse energy in 1 crystal, one obtains:

$$tpg(A^{\text{max}}) = 1024 \Rightarrow \frac{\text{mult}}{2^{\text{shift}+\text{sliding}}} = 1024 \times 2^4 \times \frac{\text{ADCtoGeV} \times \frac{g_{12}}{g_i} \times f_{\text{trans}} \times IC \times \sin \theta}{E_T^{\text{max}}} \quad (1.11)$$

All the parameters on the right side of the formula are available crystal-by-crystal in the ECAL database. An iterative procedure is implemented to find the best *mult* and *shift* for each crystal. The higher *mult* the higher the precision of the linearizer multiplication factor  $\text{mult}/2^{\text{shift}}$ . Therefore, we start with *shift* = 0, compute *mult* and check if we obtain  $\text{mult} \in [2^7, 2^8 - 1]$ . If the requirement is not satisfied, the *shift* is incremented by 1 until a suitable *mult* is found. The maximum *shift* is  $2^4 - 1$ . If the procedure fails, the *sliding* parameter can be increased by strip and applied in the formula.

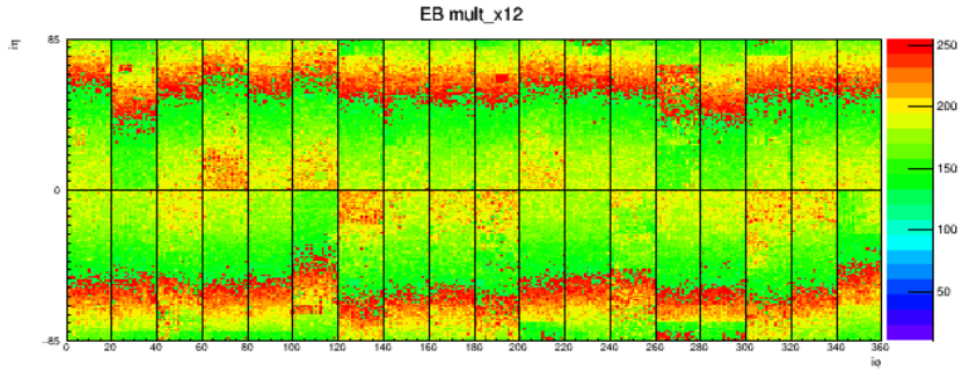
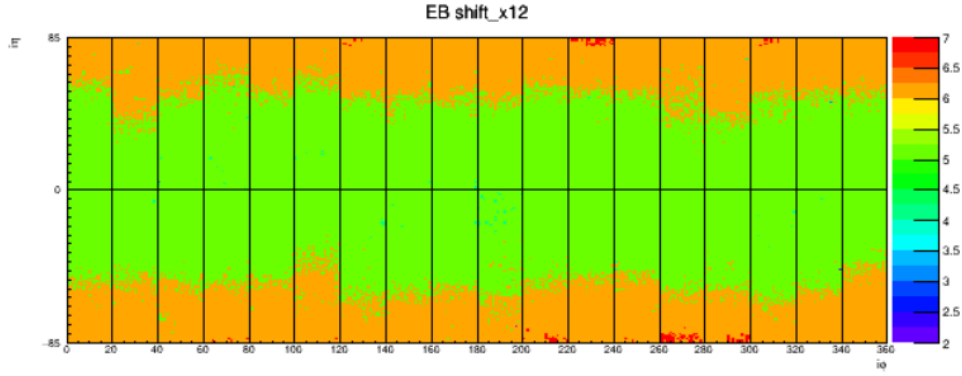


Figure 1.16: *mult* factor for gain 12 in the EB

In Fig. 1.16 and 1.17 the *mult* and *shift* parameters for each crystal in the barrel at the end of 2018 are shown for gain 12: since the crystal transparency decreases with increasing  $|\eta|$  (y-axis) the calibration factor needs to be larger at high  $|\eta|$ .

### 1.4.3 Spike reduction

Anomalous signals, consisting of isolated large signals, have been observed in the ECAL Barrel during LHC proton-proton collisions. These deposits, termed ECAL “spikes”, are observed to occur at a rate that is proportional to the intensity of the colliding beams. As a consequence,

Figure 1.17: *shift* factor for gain 12 in the EB

they must be eliminated in order to prevent large biases in the energies of reconstructed electrons, photons and jets. The spikes are understood to be associated with particles (produced in p-p collisions) striking the APDs and very occasionally interacting with the APD material causing large anomalous signals through direct ionization of the silicon [20]. Fig. 1.18 depicts the possible spike formation processes.

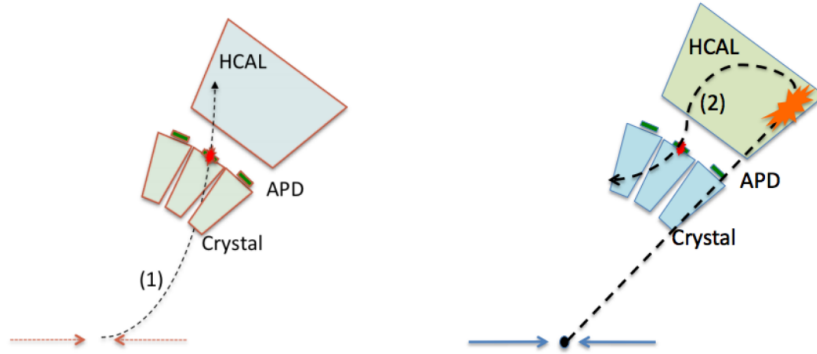


Figure 1.18: Origin of the particles producing spikes in the ECAL EB APD sensors.

Since the characteristics of ECAL spikes are localized high energy signals, they will often satisfy the conditions for triggering electrons and photons in CMS. If untreated, the rate of spikes would be a dominant component of the 100 kHz CMS L1 trigger total rate bandwidth for luminosities above  $\mathcal{L} \sim 10^{33} \text{ cm}^{-2} \text{ s}^{-1}$  such the ones reached in Run I and II of LHC.

Typical electromagnetic (EM) showers resulting from scintillation light have a well-defined spread of energy between adjacent crystals with  $\sim 80\%$  of the energy being deposited in one central crystal and the remaining 20% in the four nearest neighbors (see Fig. 1.19). Spikes on the other hand are almost always contained within a single channel.

The origin of the spikes has been confirmed by detailed Geant4 simulation studies of the APD sensitive layer and in laboratory tests. Unfortunately any permanent solution would require dismounting the ECAL barrel and reworking all the APD sensors (36 supermodules with 1700 crystals each): this would be a too time consuming and delicate operation. Therefore, mitigations have been implemented both online and offline in the L1, HLT triggers, and in the CMS reconstructions, exploiting the precise characteristics of these spurious signals.



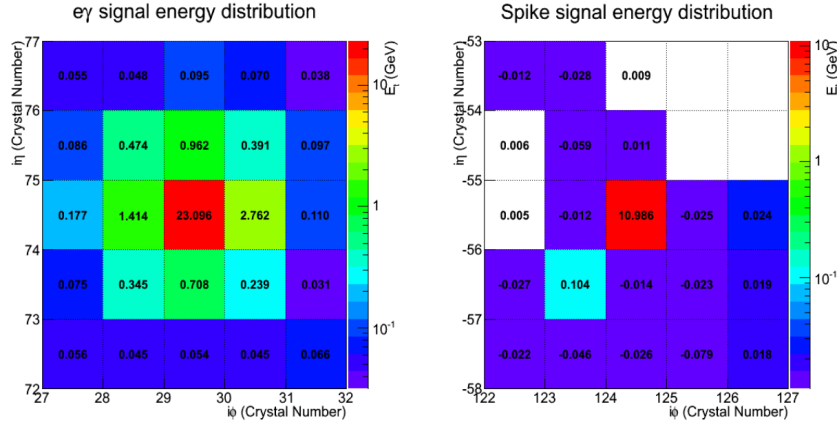


Figure 1.19: Left: typical distribution of the energy in a  $5 \times 5$  matrix of crystals when an electron or photons interacts in the central one. Right: distribution of the energy for a spike occurring in the center of the  $5 \times$  matrix.

#### 1.4.3.1 Online spikes rejection

The current spike mitigating algorithm implemented at the Level-1 trigger is the strip Fine Grain Veto Bit (sFGVB). This uses a feature that exists in the FENIX chip used in the front-end electronics currently installed in the ECAL supermodules [20]. By using topological information, the sFGVB algorithm searches for towers with more than one crystal in a strip above a configurable threshold, and raises a flag when this shower-like topology is observed.

The sFGVB algorithm proceed as follows. For each strip, a 1-bit flag is raised if at least 2 crystals have energy above a configurable threshold in ADC counts. The bits from 5 strips are then ORed together to form the final sFGVB bit. If the overall bit is 0 the tower is spike-like: if the total energy of the tower is higher than a configurable threshold the trigger primitive from the tower is zeroed. This scheme is illustrated in Fig.1.20 for a real shower and for a spike.

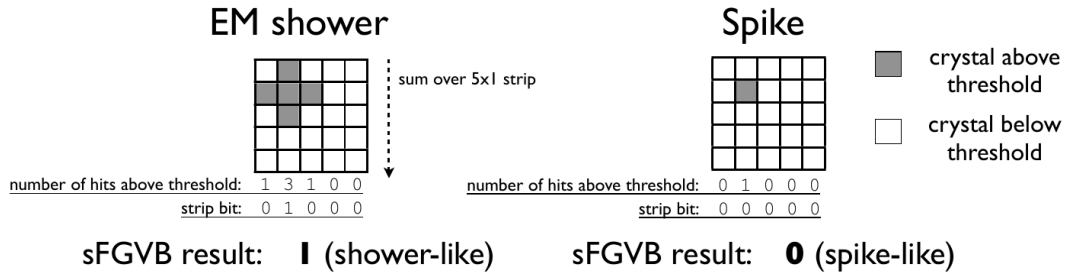


Figure 1.20: Schema of the sFGVB algorithm. On the left an EM shower, on the right a spike.

During Run II the crystal energy threshold was 16 ADC counts, while the tower energy killing threshold was 16 GeV. It has been measured using data that the sFGVB algorithm reject  $> 95\%$  of spikes with transverse energy greater than 16 GeV, with only a small ( $< 0.5\%$ ) effect on the efficiency for triggering well-identified electrons with transverse energy greater than 16 GeV.

#### 1.4.3.2 Offline spikes rejection

In the HLT and offline reconstruction more stringent requirements implemented in software can be imposed to remove spikes from ECAL signals. Two different variables are used: the “Swiss-cross” observable and the pulse timing. The definition of the Swiss-cross is shown in Fig. 1.21.  $E_1$  is the energy deposited in the central crystal channel, while  $E_4$  is the sum of the



energy deposited in its four adjacent crystal channels. Spikes will have a swiss-cross value close to 1 due to  $E_4$  being small as show on the right in Fig. 1.21.

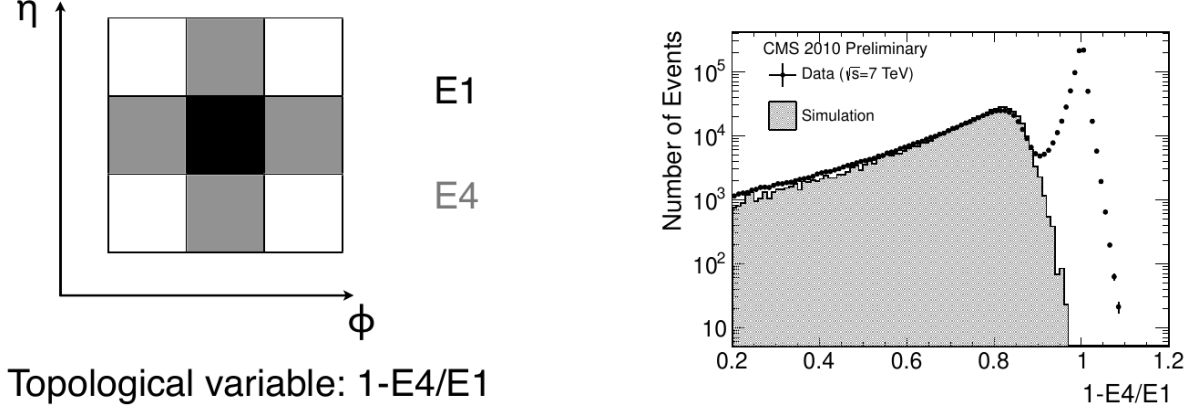


Figure 1.21: On the left the definition of the swiss-cross observable. On the right: distribution of the swiss-cross variable for the highest energy deposit in the event, comparison between data with spikes and simulation, which includes EM deposits only.

Spikes and EM energy deposits have different signal pulse shapes. Since a spike is produced by a particle directly hitting the APD, the resulting signal does not include the response time of the scintillation light. When the pulse is fitted to extract the timing of the signal, the spikes generally appear “early”, due to the faster rise time of the spike pulse. Figure 1.22 shows the reconstructed timing distribution for ECAL energy deposits with  $E_T > 3$  GeV in both data and simulation. The timing calibration is defined such that a relativistic particle originating from the nominal collisions vertex (which does not include the anomalous signals) is centered at zero. The timing distribution from data shows a significant component of out-of-time signals, due to spikes. A cut on the signal timing of  $\pm 3$  ns is  $> 90\%$  efficient at removing spike pulses, with a negligible impact on EM energy deposits [21].

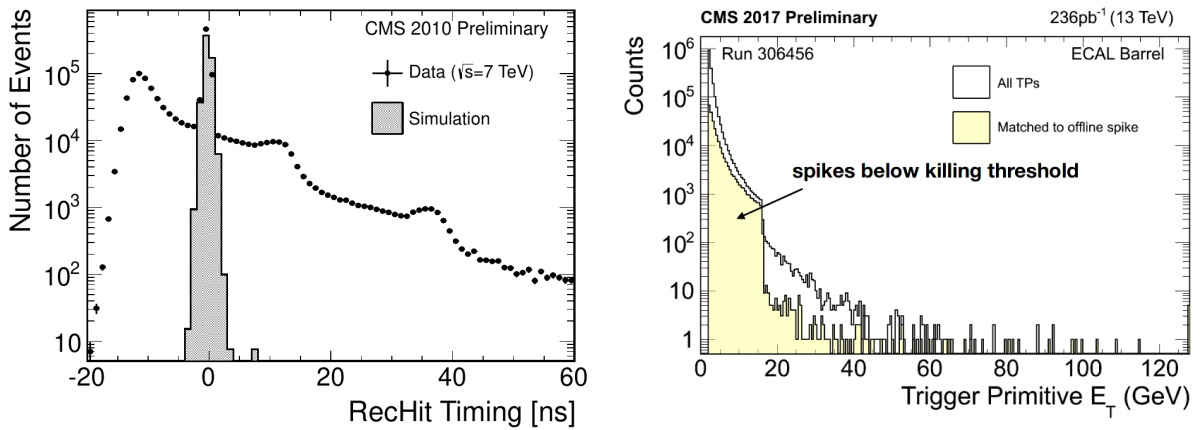


Figure 1.22: On the left: reconstructed timing distribution for ECAL energy deposits with  $E_T > 3$  GeV in both data and simulation. On the right: TP transverse energy distribution from Run 2 data, showing the fraction of TP created by the online trigger that are matched to offline identified spikes.

Figure 1.22 shows the fraction of TPs created by the online trigger that are matched to spikes identified with the offline criteria. Above the sFVGB killing threshold, the amount of spikes reaching the HLT trigger is considerably reduced. Under that threshold the offline algorithm is

essential to reduce the rate of spikes. Since the sFVGB uses pedestal-subtracted amplitudes, a precise knowledge of the ECAL pulse pedestal is needed. Pedestals are monitored and updated periodically during data-taking.

#### 1.4.4 Run II performance

During Run II of the LHC (2015-2018) a greater instantaneous and total integrated luminosity was reached compared to Run 1 (2011-2012). Radiation-induced effects created challenging data taking conditions for ECAL and frequent calibration of the parameters of the ECAL trigger system were necessary to maintain an optimal trigger efficiency and a stable TP energy scale. Multiple time dependent corrections were applied during Run 2 to account for radiation-induced effects: laser response corrections, crystal inter-calibration, pedestal drift corrections, and pulse shape and timing corrections [22]. As a result of these calibrations and corrections, excellent L1 reconstruction efficiencies for electron and photon candidates have been maintained throughout Run II [23] as shown in Figure 1.23.

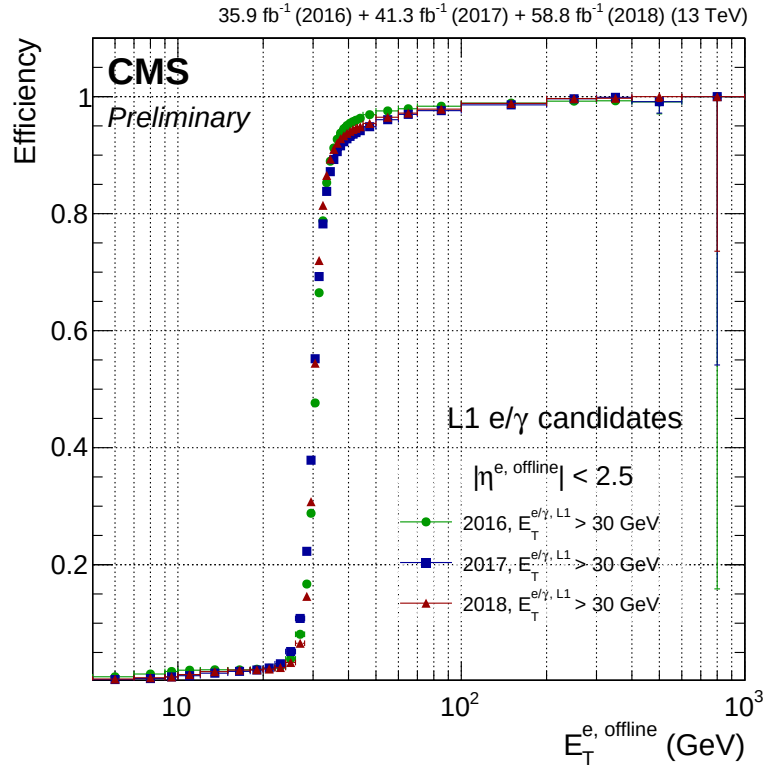


Figure 1.23: Efficiency versus energy for L1 trigger EM candidates during Run II [23].

##### 1.4.4.1 Laser response corrections

A description of the ECAL transparency correction is given in Sec. 1.3.3. The calibration factors have been propagated to the L1 trigger and HLT twice a week during Run II to keep the TP energy scale stable [24].

##### 1.4.4.2 Pulse shape correction

Due to irradiation received by ECAL crystals, the pulse shape of the ECAL electronic signals changes with time. It has been seen that this is strongly dependent on the pseudorapidity of the

crystals, with higher  $\eta$  crystals receiving more radiation leading to more dramatically distorted pulses. If uncorrected, this can impact the bunch crossing assignment of the TPs, leading to substantial trigger inefficiencies [25]. This is compensated for with a change in readout timing, which was performed three times in 2018. The resulting measurements shown in Fig. 1.24 show an increasing time drift of signals in 2017, making regular corrections necessary [26].

#### 1.4.4.3 Spike killer

The spike killer performance affects critically the L1T rate and EM object  $p_T$  thresholds: the sFVGB parameters have been re-optimized for the Run III expected PU profile and noise level, obtaining a  $\times 2$  spikes rejection with minimal loss of EM signal. The crystal energy threshold will be increased to 22 ADC counts, while the tower energy killing threshold stay at 16 GeV. Increasing the crystal threshold more tower will be tagged as spike-like (with isolated deposited energy) since fewer crystals will pass the energy threshold.

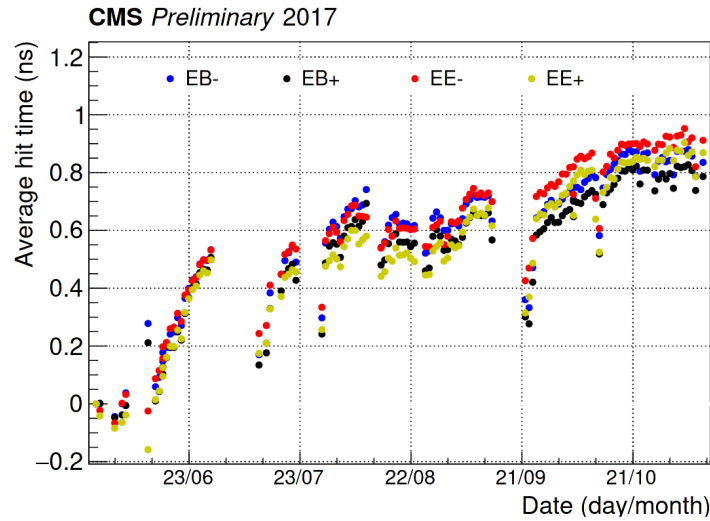


Figure 1.24: Average shift of the pulse timing along 2017 in different regions of the detector. At higher  $\eta$ , the radiation damage induces a more rapid change that needs to be corrected periodically to avoid trigger inefficiencies.

## 1.5 The ECAL offline reconstruction

### 1.5.1 ECAL local reconstruction and Particle Flow

Several algorithms are stacked on top of each other to reconstruct electrons and photons candidates from the measurement of scintillation light in each single crystal in the ECAL detector [27]. The first step builds the ECAL *Rechts*, the measurement of the amount of energy deposited in each crystal at each BX. The second step, called PFClustering, builds the simplest form of energy clusters, looking for crystals with local maxima of energy (*seeds*) and associating the neighbor crystals to them.

A single electron or photon usually leaves more than one cluster of energy in the ECAL detector. The electron, bending in the strong magnetic field of the CMS solenoid (3.8 T) while passing through the Pixel and Tracker detectors (see the material budget distribution in Fig. 1.27), emits bremsstrahlung photons that will leave a trace of small energy clusters in the ECAL detector near the main impact point. The photon, instead, is converted to electron-positron pairs interacting with the several layers of the inner detectors of CMS, thus also depositing multiple clus-

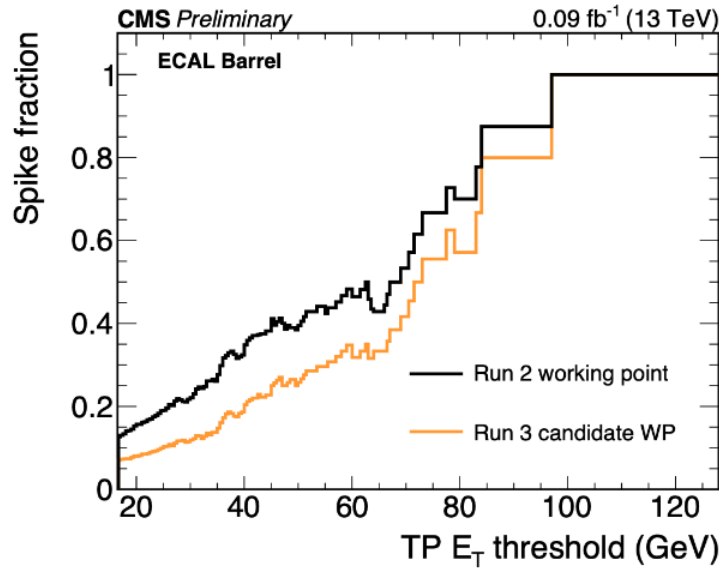


Figure 1.25: Fraction of TPs passing the online spike killer and then identified as spikes offline versus  $E_T$  threshold. Comparison of Run II spike killer working point and candidate for Run III.

ters of energy in ECAL. An additional clustering algorithm, called *SuperClustering* is therefore needed for the electron and photon reconstruction in order to improve the energy resolution by taking into account the energy of the secondary clusters.

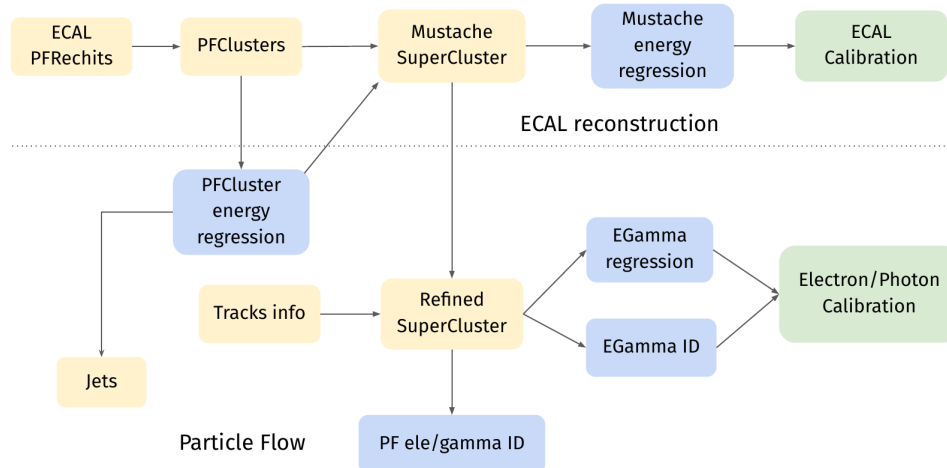


Figure 1.26: Simplified schema of the electron and photon reconstruction sequence in CMS.

The SuperCluster (SC) candidate is built using only ECAL local information, therefore it is the object used for the calibration of the ECAL detector response. The SuperCluster is also one of the inputs of the particle flow CMS global event reconstruction [5] which combines optimally the information from all the sub-detectors. Electrons are identified as a primary charged-particle track linked to ECAL SuperClusters, whereas photons are identified as ECAL deposits not linked to any extrapolated track. A dedicated tracking algorithm, based on the Gaussian Sum Filter (GSF), is used for electrons to estimate the track parameters keeping into account the change in trajectory given by the emission of bremsstrahlung photons [28]. Additional ECAL clusters can be included in the SuperCluster if they are compatible with secondary emissions and photon conversions from the main electron track. The final object is called *refined Super-*

Cluster, and it is the baseline for the formation of electron and photon candidates. Additional criteria are then applied to obtain loosely identified isolated electron and photons, to which an energy regression algorithm, implemented as a semi-parametric BDT, is applied taking into account all the track and ECAL deposits characteristics. A comprehensive description of electron and photon reconstruction during Run II can be found in Ref. [27].

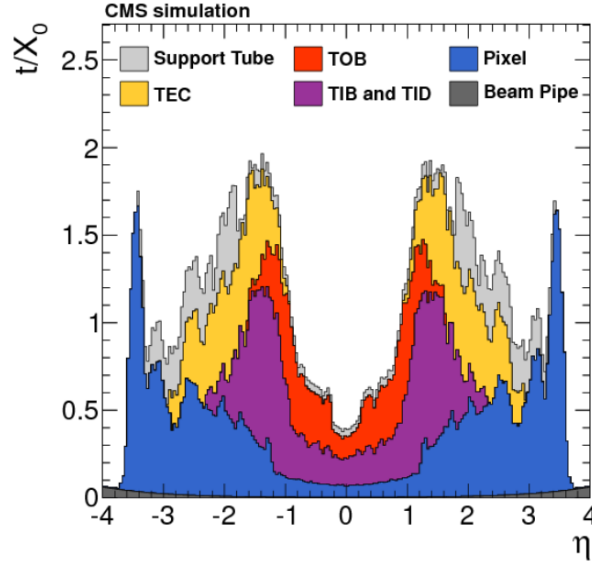


Figure 1.27: Total thickness of the tracker material traversed by a particle produced at the nominal interaction point, as a function of pseudorapidity, expressed in units of radiation length  $X_0$  [29].

### 1.5.2 Rechits

The signals in the ECAL crystals are reconstructed by fitting the signal pulse with multiple template functions to subtract the contribution from out-of-time pileup. The *multifit* [30] algorithm has been introduced in ECAL in 2015 before the beginning of Run II, to cope with the increasing PU level. During Run I, the ECAL rechits were reconstructed with weights, in an analogous way to the TP formation in the frontend (see Sec. 1.4.1.3).

The multifit method uses a template fit with  $N_{BX}$  parameters, comprising one in-time (IT) and up to nine OOT amplitudes, up to five occurring before, and up to four after the IT pulse:  $N_{BX} \in [1, 10]$ . The fit minimizes the  $\chi^2$  defined as

$$\chi^2 = \left( \sum_{j=0}^{N_{BX}} A_j \vec{p}_j - \vec{S} \right) C^{-1} \left( \sum_{j=0}^{N_{BX}} A_j \vec{p}_j - \vec{S} \right)$$

where the vector  $\vec{S}$  comprises the 10 readout samples,  $s_i$ , after having subtracted the pedestal value,  $\vec{p}_j$  are the pulse templates for each BX, and  $A_j$ , which are obtained by the fit, are the signal pulse amplitudes in ten consecutive BXs, with  $A_5$  corresponding to the IT BX. The pulse templates  $\vec{p}_j$  for each BX have the same shape, but are shifted in time by  $j$  multiples of 25 ns. The total covariance matrix  $C$  includes the correlation of noise and the signal between different time samples. At low energy the contribution from the noise is dominant, whereas at higher energies the uncertainty on the pulse shape determination plays a larger role.

The  $\chi^2$  is minimized iteratively for each crystal that is readout in a triggered event (average of 1000 crystals in the EB), and non-zero amplitudes are included iteratively to save time. An example of fitted a pulse shape in EE is shown in Fig. 1.28.

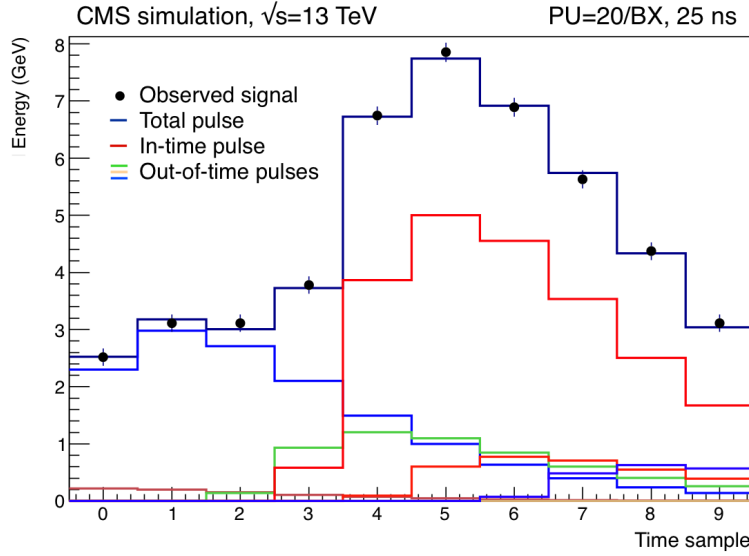


Figure 1.28: Example of fitted amplitudes for a crystal in the ECAL endcap.

### 1.5.3 PFClustering

An electron or photon passing through the ECAL detector deposits energy in more than one crystal: the 94% of the energy is deposited in a  $3 \times 3$  crystals matrix around the impact point, and 97% in the  $5 \times 5$  one. The aim of the PFClustering step is to gather the signal from nearby crystals and interpret them as single particle. The algorithm is also designed to handle properly the overlap of the energy deposits by two different particles, sharing common crystals. The steps of the PFClustering algorithm are briefly described in the following:

**Inputs** The inputs of the PFClustering algorithm are the ECAL rechits passing an energy threshold depending on  $\eta$ , calculated as a  $X \cdot \sigma$  of the noise level. This threshold is called the *PFRechit threshold*.

**Seeding** The PFCluster seeds correspond to the most energetic crystals around which the others will be aggregated. The seeds are identified, in each event, by selecting the crystals having an energy greater than the seeding energy threshold and greater than the energy of their neighbor crystals (in a  $3 \times 3$  matrix).

**Gathering** Starting from each seed, *topological* clusters are formed by aggregating contiguous deposits of energy which pass a *gathering* threshold. Crystals that share at least one corner in common with an already-aggregated crystal are considered as being part of the cluster, and a single topological cluster can embrace different seeds.

**Clusters forming** A topological cluster may comprehend more than one PFCluster: the final step of the algorithm computes for each crystal the fraction of energy associated to each seed. A Gaussian profile of the energy deposition around the seed is assumed. An iterative procedure is implemented: for each crystal its energy fraction along the Gaussian profile is computed. At each iteration, crystals with an energy fraction smaller than  $10^{-7}$

are removed from that PFCluster, and the energy centroid of the PFCluster is updated. The procedure is repeated until the centroids of the PFClusters converge, as shown in Fig. 1.29. At the end of the procedure the final PFCluster and their energy fraction in each crystal are saved.

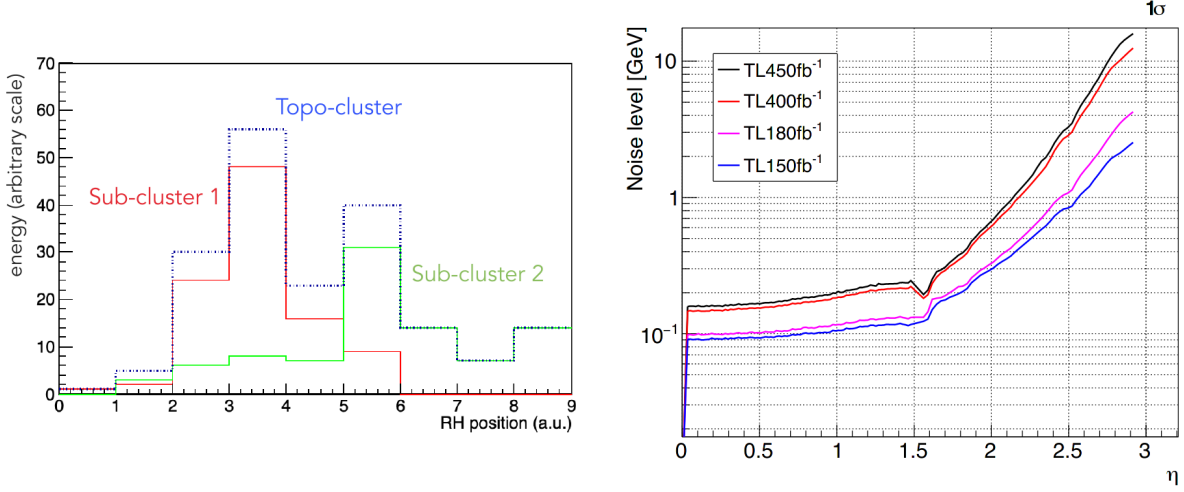


Figure 1.29: On the left: sketch of the clusters forming step in the PFCluster algorithm. The rechits (RH) are projected on a single line to show the energy fraction assigned to each seed. On the right: noise evolution along ECAL for different total integrated luminosities (ageing).

The PFRechit, seeding and gathering thresholds are crucial for the reconstruction efficiency and purity of EM deposits in CMS. Not only do they affect the performance of electron and photon reconstruction, but also jets and missing energy. Therefore, the thresholds have been re-optimized to follow the predicted evolution of the ECAL noise along Run III, see Fig. 1.29. After an intense optimization study, the optimum values have been found in keeping a PFRechit threshold equal to the seeding threshold at  $3\sigma$  ( $4\sigma$ ) of the noise in EB (EE), and dropping the gathering threshold. In practice all the dynamics of the algorithm is driven by the PFRechit filtering. This choice provides the best cluster reconstruction efficiency and purity, maintaining a low probability of fake clusters due to noise in the EE, the region most affected by the detector ageing due to radiation damage.

### 1.5.4 Mustache SuperClustering

The combination of bremsstrahlung radiation from electrons and pair production from photon in the tracker material (see Fig. 1.27), under the influence of the CMS magnetic field, results into electron and photon energy releases which are mostly extended in the transverse  $R - \phi$  plane. If the spray of energy is not suitably collected (clustered) a poor energy resolution would be attained for a non-negligible amount of electrons or photons. The concept of the SuperCluster, narrow in  $\eta$  and large in  $\phi$ , was introduced at the start of CMS, in order to collect all the energy. The algorithm in place during Run I and II, called *Mustache*, consists of purely geometrical criteria: the magnetic field causes the low  $p_T$  constituents of electromagnetic showers to slightly bend in  $\eta$  as they follow a helical path, giving the showers a characteristic mustache shape. In order to recover the energy of the shower, a parabolic-shaped window in  $\Delta\eta$  is used which emulates the bending of the shower in the  $(\eta, \phi)$  plane. The formation of SuperClusters proceed in iterations along the list of all the reconstructed PFClusters in ECAL.

**Input** The list of all the reconstructed PFClusters in ECAL is ordered by transverse energy  $E_T$



**Seeding** In each iteration, the highest energy PFCluster still not assigned to any SuperCluster is taken as a seed if  $E_T > 1$  GeV.

**Mustache** Looping on all the remaining PFClusters, their  $\Delta\eta$  and  $\Delta\phi$  distance from the seed is compared with the area delimited by two  $\eta$  dependent parabolas and a maximum  $\Delta\phi$  window, as shown in Fig. 1.30. The parabolas are parametrized by the  $\eta$  position of the seed and the energy of the cluster. The dynamic  $\Delta\phi$  interval depends instead only on the transverse energy of the cluster.

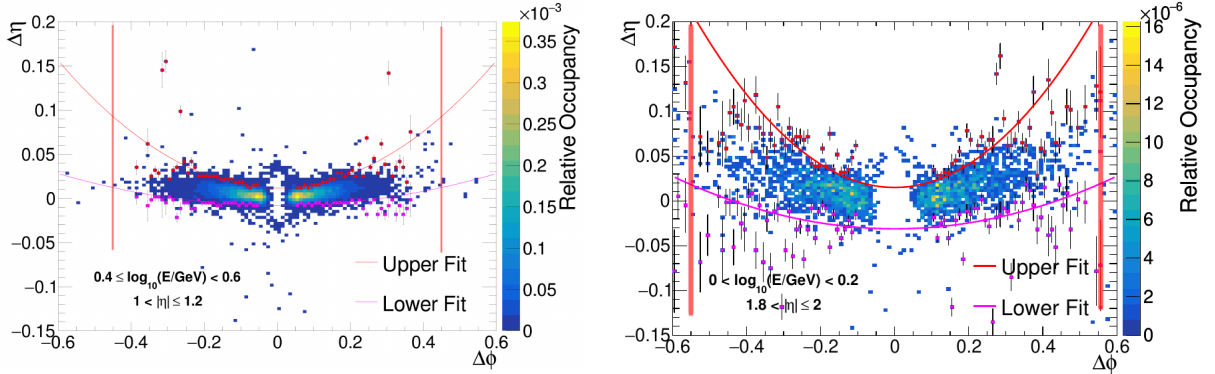


Figure 1.30: Examples of the optimization of parabolas and  $\Delta\phi$  window parameters for the Mustache algorithm. 98% of the EM energy is contained in the Mustache area. The plots are realized accumulating many events in the same  $\eta_{\text{seed}} - E_T^{\text{seed}}$  of the seed bin.

**Iteration** PFClusters inside the Mustache area are associated to that seed and removed from the initial list. A new iteration is performed until no PFClusters with  $E_T > 1$  GeV are available.

The parabolas shapes are parametrized with 10 numbers, whereas the  $\Delta\phi$  window has 4 free parameters. These numbers are optimized in bins of  $\eta_{\text{seed}}$  and  $\log_{10}(E_T^{\text{seed}})$  such that at least 98% of the true EM energy is contained in the Mustache area. Many events are accumulated in each  $\eta_{\text{seed}} - E_T^{\text{seed}}$  bin and fits are performed: Fig. 1.30 shows an example for two different  $\eta$  positions.

### 1.5.5 Run II performance

Deep effort has been put into ECAL response calibration during Run II to obtain the best possible energy resolution from the ECAL local reconstruction. Not only has the crystal transparency been carefully monitored and corrected (see Sec.1.3.3), but also many time-dependent effects due to radiation damage have been understood and corrected. These include drifts in the pedestal values, distortion of the pulse shape, changes in the relative response of crystals to laser light and particles. The resulting performance is shown in Fig 1.31. The energy resolution performance of the electron candidates reconstructed by the PF algorithm are omitted for brevity and can be found in [27].

Electron reconstruction efficiency is defined as the ratio between the number of reconstructed SCs matched to reconstructed electrons and the number of all reconstructed SCs. The electron reconstruction efficiency is computed with a tag-and-probe method using  $Z \rightarrow ee$  [32] events as a function of the electron  $\eta$  and  $E_T$ , and covers all reconstruction effects. This reconstruction efficiency is higher than 95% for  $E_T > 20$  GeV, and is compatible between data and simulation within 2% [27].

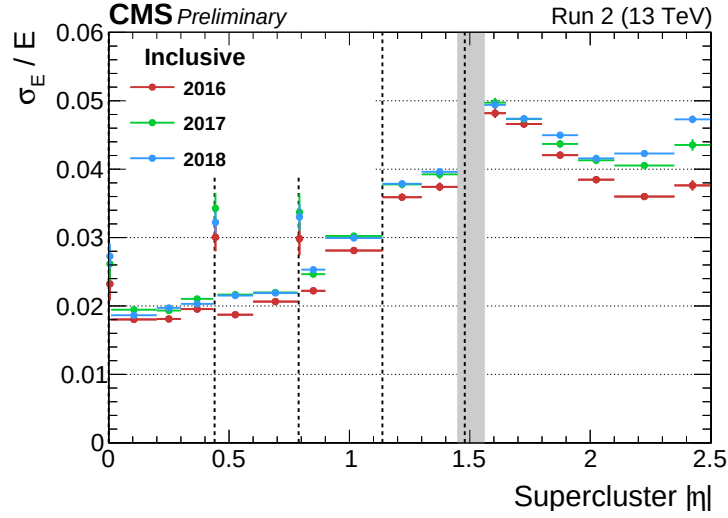


Figure 1.31: Relative electron (ECAL) energy resolution unfolded in bins of pseudorapidity  $\eta$  for the ECAL Barrel and Endcap. Electrons from  $Z \rightarrow ee$  decays are used. The plot compares the resolution achieved after a refined calibration of the data collected at 13 TeV during Run II in 2016, 2017, and 2018. The relative resolution  $\sigma_E / E$  is extracted from an unbinned likelihood fit to  $Z \rightarrow ee$  events, using a Voigtian (Breit-Wigner convolved with Gaussian) as the signal model. A stable ECAL energy resolution is observed over the course of Run 2 despite the increased LHC luminosity and the ageing of the detector [31].

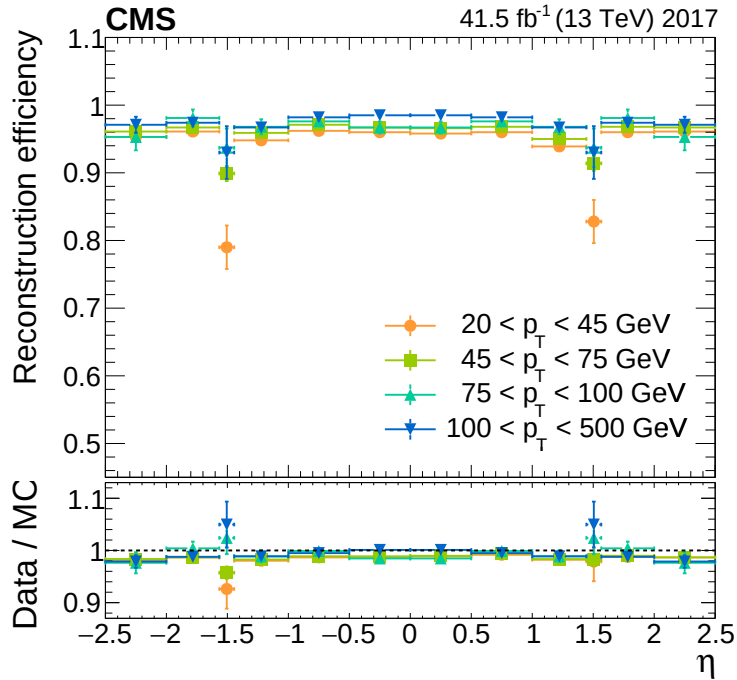


Figure 1.32: Electron reconstruction efficiency versus  $\eta$  in data (upper panel) and data-to-simulation efficiency ratios (lower panel) for the 2017 data taking period. The vertical bars on the markers represent the combined statistical and systematic uncertainties. The region  $1.44 < |\eta| < 1.57$  corresponds to the transition between the barrel and endcap regions of ECAL and is not considered in physics analyses.

## Chapter 2

# Optimization of the CMS Electromagnetic Calorimeter for Run III

### 2.1 Trigger weights optimization for Run III

One set of weights for the barrel and one for the endcaps are currently used by the on-detector FENIX chips to estimate the TP signal amplitudes. These weights were derived using the ideal pulse shape for undamaged crystals measured during the ECAL test beams in 2005. These weights have been used throughout Run I and Run II without changes. The weights are no longer optimal because of radiation-induced changes to the pulse shapes and timing, especially at high  $\eta$  in EE, that induce large biases in the TP energy estimate. To keep the excellent trigger performance reached so far and cope with the further pulse shape distortions due to detector ageing and increased out-of-time pileup (OOT-PU) foreseen during Run 3, a re-optimization of the TPG weights has been performed.

#### 2.1.1 Weights granularity

The TPG weights are applied on-detector on a strip-by-strip basis, and this is the maximal granularity that can be reached in the weights optimization and application. Fig. 2.1 shows the bias in the amplitude reconstruction, defined as  $(E_{\text{TP}}/E_{\text{True}}) - 1$  where  $E_{\text{True}}$  and  $E_{\text{TP}}$  are the true and reconstructed amplitudes respectively, for different weights configurations considering a signal-only scenario (no out-of-time PU) in the endcaps. The violet line shows the large positive bias created by the current weights due to the radiation-induced changes in pulse shapes. The blue line shows the improvement obtained by using updated pulse shapes, precisely measured for each crystal in September 2018 at the end of Run 2 shifting the acquisition time by small steps and fitting the pulse shape, to obtain one averaged set of weights for EE. Increasing the granularity of the weights up to the strip level (red line) decreases further the bias in the energy reconstruction [22, 33].

#### 2.1.2 Optimization for pileup

The ECAL trigger uses 5 pulse samples (as shown in figure 1.15) to reconstruct the amplitude, therefore it can be biased by PU interactions happening before or after the signal BX, called out-of-time (OOT) PU. The effect of OOT PU varies along  $\eta$  and it is more pronounced in the innermost rings of the endcaps. Moreover, the position of the signal within LHC proton bunch trains changes the average OOT PU effect on the pulse shape. For example, the LHC schema in late 2017, trains with 8 colliding proton bunches followed by 4 empty ones (8b4e), shown in Fig. 2.2, with its continuously varying OOT PU is the most challenging for the ECAL trigger. Moreover, this type of configuration could still be part of the filling scheme plans for Run III.

A standalone Monte Carlo (MC) simulation has been developed to model the ECAL digitized

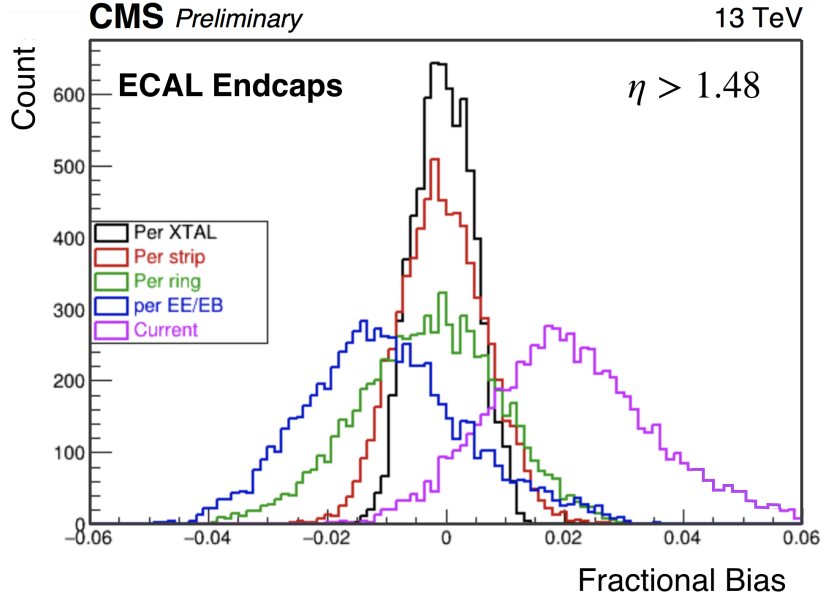


Figure 2.1: Fractional bias in amplitude reconstruction ( $E_{TP}/E_{True} - 1$ ) for current weights (violet), updated pulse shapes weights (blue), and weights optimized with finer granularity:  $\eta$  ring (green), strip (red), and crystal (black) [33].

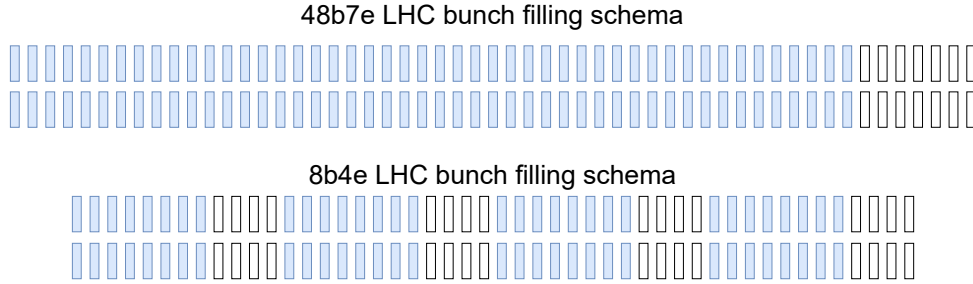


Figure 2.2: LHC bunch filling schemas used during Run II. The blue boxes represent the bunches containing colliding protons, whereas the empty white boxed represent pauses in the collisions. On top: the 48 colliding, 7 empty (48b7e) schema used in 2018; on the bottom: the 8 colliding, 4 empty (8b4e) schema used in late 2017.

samples for different pulse shapes, signal energies, pileup, and LHC colliding bunch train structures. With this setup it is possible to derive optimal weights for various data-taking conditions and to study the effect of different sets of weights on configurable scenarios. The development of this tool has been fundamental for the optimization of ECAL trigger primitive performance for Run III.

### 2.1.2.1 Pulse simulation

The goal of the standalone MC simulation is to produce realistic ECAL digitized samples in order to optimize the weights for different PU conditions and test their application. In the past, a similar optimization has been carried out without considering the PU [18]. Given the strong  $\eta$  dependence of the PU energy contribution and ECAL pulse shape, the simulation is performed separately for each ECAL strip, and then aggregated and analyzed by  $\eta$ -ring.

**Pulse shape** The pulse shape of each ECAL crystal is the first ingredient of the MC. The analytical form of the pulse shape which best approximates the data is given in Eq. 2.1. The

parameters  $\alpha$ ,  $\beta$ , the time  $t_{\max}$  corresponding to the maximum of the pulse, and the pedestal  $P$  have been measured precisely for each crystal by fitting the pulse shape formula on data points acquired during special runs where the timing of the ECAL acquisition is shifted in small steps of 1 ns to observe the full pulse precisely.

$$f(x) = \begin{cases} P + A \left(1 + \frac{x-t_{\max}}{\alpha*\beta}\right)^\alpha \cdot e^{(x-t_{\max})/\beta} & \text{if } t \geq (t_{\max} - \alpha * \beta) \\ P & \text{otherwise} \end{cases} \quad (2.1)$$

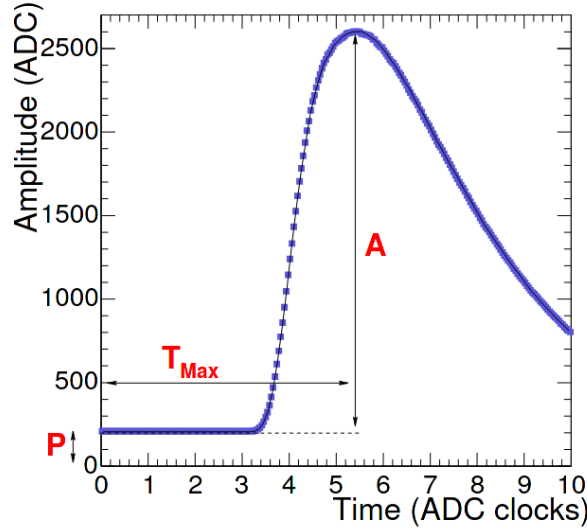


Figure 2.3: Analytical form of the ideal ECAL pulse shape, depending on the  $\alpha$  and  $\beta$  parameters.  $t_{\max}$  is the time corresponding to the maximum of the pulse.  $P$  is the pedestal of the pulse.

**Pileup simulation** The MC considers as inputs (see Fig. 2.4): the p.d.f. of the energy deposited by PU in different  $\eta$  rings at each collision extracted from the full Geant4 CMS simulation of the PU, the desired PU level, and the structure of the LHC train of colliding bunches. The LHC setup consists of a series of BXs with colliding protons, followed by a number of empty bunches, with the pattern repeated along the orbit, as shown in Fig. 1.2. The configurations used during Run II were: *8b4e* in late 2017 with 8 colliding bunches and 4 empty BX, and *48b7e* in 2018 with 48 colliding bunches and 7 empty ones. In order to simulate correctly the effect of the OOT-PU on the signal, the energy deposits from PU from a sufficient amount of time before the signal energy deposit must be considered. Therefore, two trains are simulated and the signal pulse is superimposed over the BXs of the second train, to have a realistic OOT-PU contribution. The averaged contribution from PU is shown in Fig. 2.5: the PU contribution from *8b4e* schema varies quickly since the ECAL pulse covers more than 10 BXs, whereas in the *48b7e* one the PU contamination is more stable. From the ECAL trigger point of view the *48b7e* configuration is preferred.

The MC simulation proceeds as follow:

**Setup** Inputs preparation:

- Take an ECAL strip at  $\eta = \eta_1$ , take the pulse shape parameters of the corresponding crystals.
- Take a LHC train structure (1 = colliding BX, 0 = empty BX): i.e. *8b4e* = 11111111000011111111

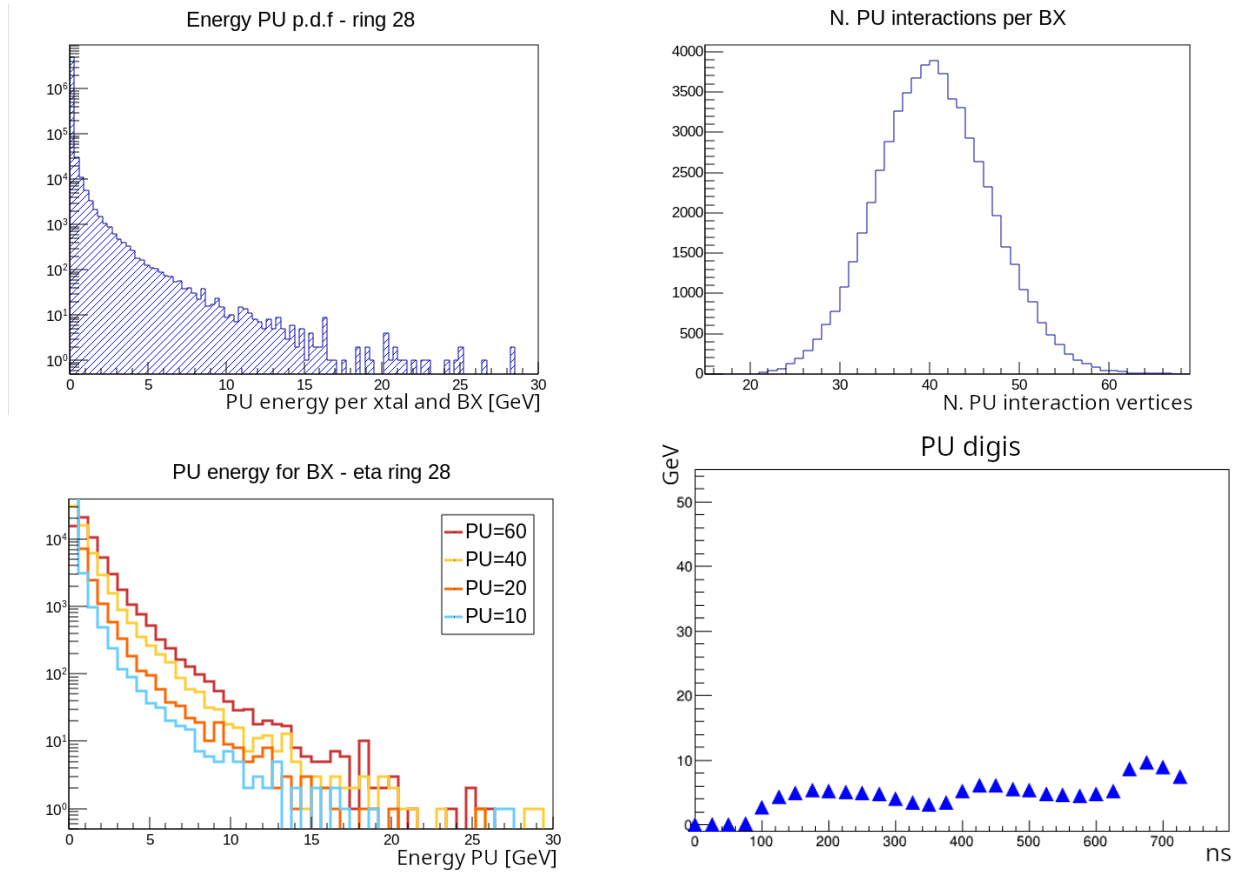


Figure 2.4: PU simulation ingredients and output. Upper left: p.d.f. of PU energy deposits by crystal and BX. Upper right: Poisson p.d.f. of number of PU interactions by BX given a mean PU level of 40. Lower left: distribution of total energy deposited by PU in  $\eta$ -ring 28 for different PU levels. Lower right: final PU pulse simulation.

- Take a mask to superimpose the signal only in the desired position: i.e. only in the second train mask = 00000000000011111111
- Take a desired mean PU level: i.e.  $N_{PU} = 50$
- Take a signal energy distribution of choice. Flat or following a realistic p.d.f.

**PU train** For each BX in the LHC train schema:

- Sample from a  $Poisson(N_{PU})$  the number of PU vertices in the current BX
- For each interaction, sample the deposited energy from the PU deposition p.d.f.
- Create an ECAL pulse with the total PU energy deposited in the crystals of the strip
- Save the pulse to be superimposed later
- Move to the next BX in the train and repeat

**Total PU** When the PU pulses have been simulated for all the BXs, the pulses are summed to create the total PU contribution

**Signal** For each BX where the signal is not masked

- Sample the signal energy from the desired p.d.f

- Create the signal pulse for each crystal in the strip and sum them
- Superimpose the total PU pulse
- Simulate the acquisition window of 10 BX: save only the amplitudes around the peak as they would be saved by the ECAL DAQ.

The resulting dataset consists of standard ECAL acquisition windows of 10 BX around the peak of the signal. The position of the signal along the LHC train is also saved for later analysis.

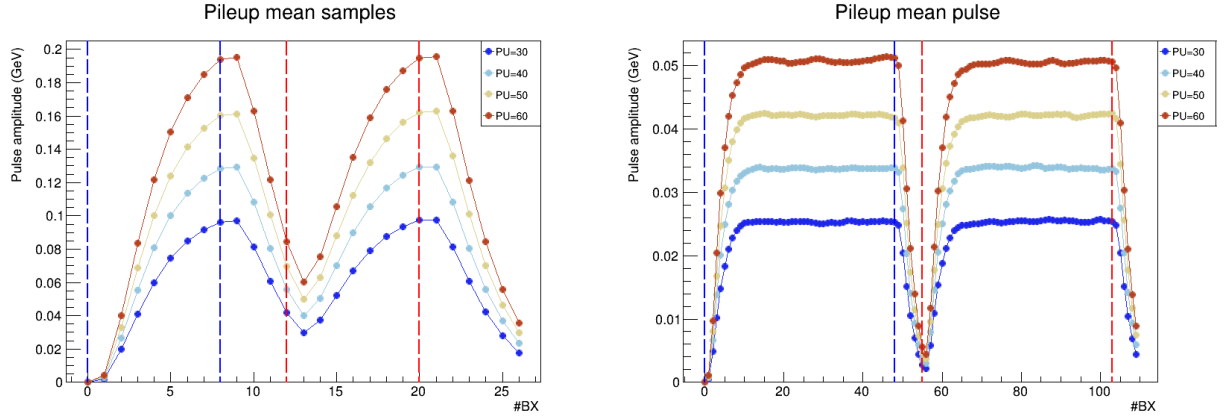


Figure 2.5: Mean pulse amplitude given by PU-only contribution. On the left: *8b4e* LHC configuration, on the right: *48b7e*. Two subsequent trains are shown: the PU pulse contribution varies a lot in the *8b4e* configuration because the ECAL pulses are longer than 8 BXs.

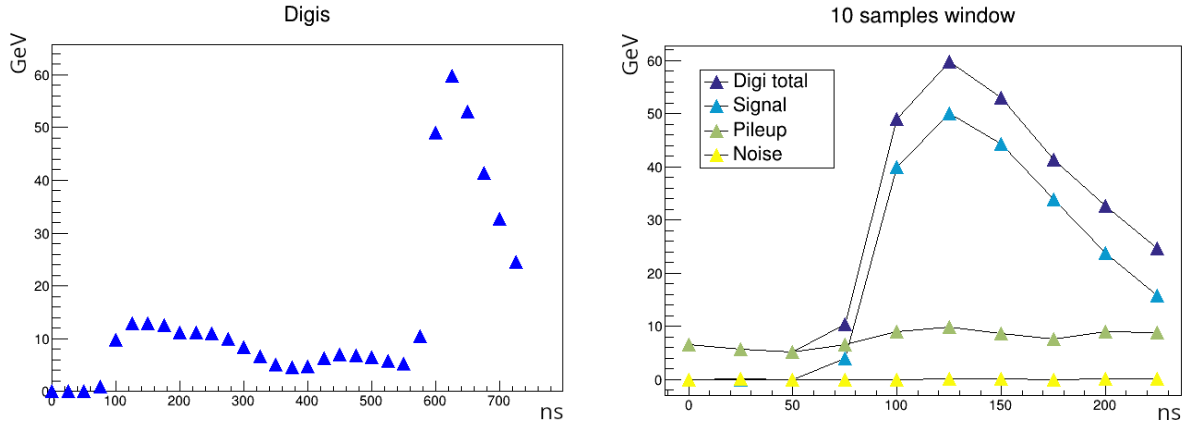


Figure 2.6: On the left: total pulse resulting from the PU and signal simulation. On the right: final ECAL acquisition window of 10 BX around the signal.

### 2.1.2.2 Weights optimization

The PU simulation is repeated for many events for different PU levels, signal amplitudes and detector regions. For each pulse, the optimal weights are derived with the analytical procedure described in Ref. [19]: since the signal pulse is distorted by the PU, each event will return slightly different optimal weights. The weights obtained with this procedure are then analyzed for different  $\eta$ -rings, signal amplitudes and PU levels. Fig. 2.7 shows the distribution of the 1st and 4th weight for signals of  $E_T = 10$  GeV in the last 3  $\eta$ -rings in the EE ( $2.3 < |\eta| < 3.0$ ) for different PU levels. The optimal weight for each configuration is then extracted as the average of the weight distribution. Larger PU distorts more the pulse, making the distribution of the optimal weight larger, but always following a gaussian shape with long tails.



**Pileup dependence** Fig. 2.8 shows a summary of the change of each weight in  $\eta$ -ring 28 for  $E_T = 10$  GeV signal. The change of each weights from the  $PU = 0$  weights (no distortion) is plotted in units of the electronics precision for the weight values  $1/64$ . With higher PU the first 2 weights become more negative because a larger energy from OOT-PU pulses has to be removed from the signal. The other weights become more positive to counterbalance this and give more importance to the BX where the peak is located. This effect is highly  $\eta$  dependent: Fig. 2.9 displays the change in the value of the 4th weight for  $PU = 30$  and  $PU = 50$  in the EE, compared to weights computed with no pileup. This dependence in the EB is much less important since the average energy deposited by PU is smaller.

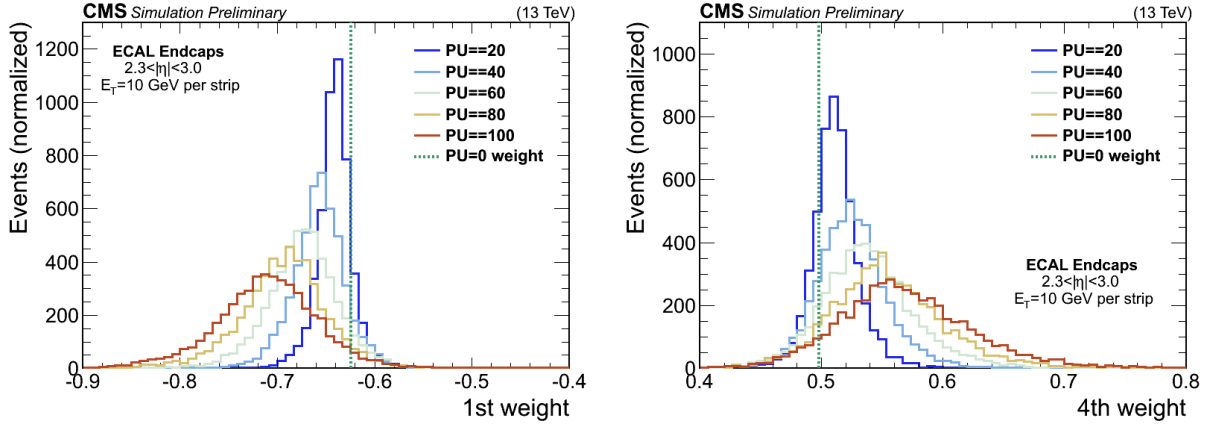


Figure 2.7: Distribution of the optimal weights extracted from each pulse in different PU conditions. On the left the 1st weights, on the right the 4th weight.

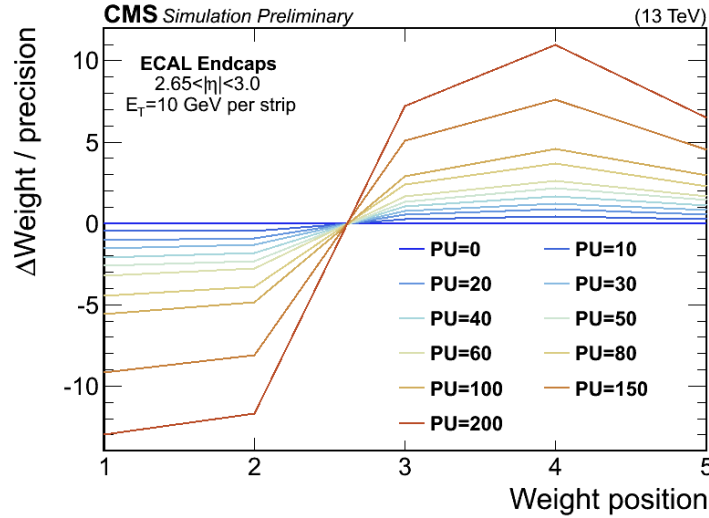


Figure 2.8: Difference between PU optimized weights computed for different PU level w.r.t the  $PU = 0$  ones. The difference is plotted in units of the electronics precision ( $1/64$ ).

### 2.1.3 Performance validation

Two sets of weights have been identified as the main candidates for the Run III data-taking period: one set optimized for low energy 2 GeV and one for high energy 30 GeV, both considering  $PU = 50$ . For the computation of these weights a flat LHC schema is used, where PU is present in 20 BX before the signal BX. Whereas for the evaluation of the weights performance, the two different LHC bunch filling schema have been simulated and compared. The performance of

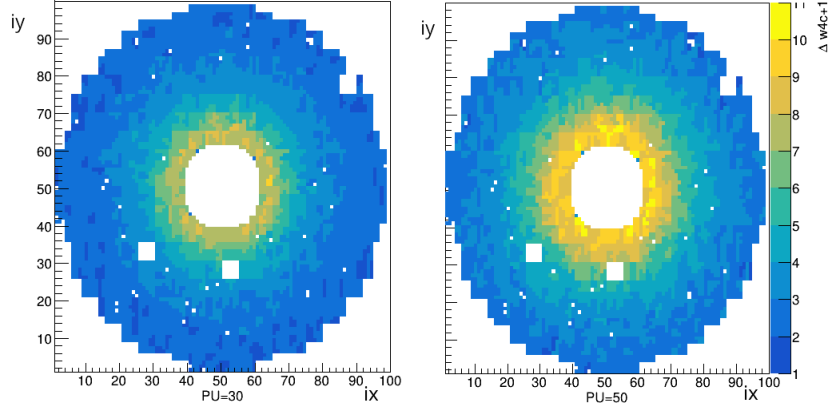


Figure 2.9: 4th weight variation along EE for  $PU = 30$  and  $PU = 50$  w.r.t the  $PU = 0$  optimal weights.

each set of weights is evaluated in terms of amplitude reconstruction bias  $E[(E_T^{\text{reco}}/E_T^{\text{true}})] - 1$  and fractional spread  $\text{RMS}[E_T^{\text{reco}}/E_T^{\text{true}}]/(E[(E_T^{\text{reco}}/E_T^{\text{true}})])$ . Fig. 2.10 shows the amplitude reconstruction bias in the region  $\eta\text{-ring} \in [26, 28]$  for small and large energy signals. In both cases, weights optimized for the correct energy regime deliver better performance.

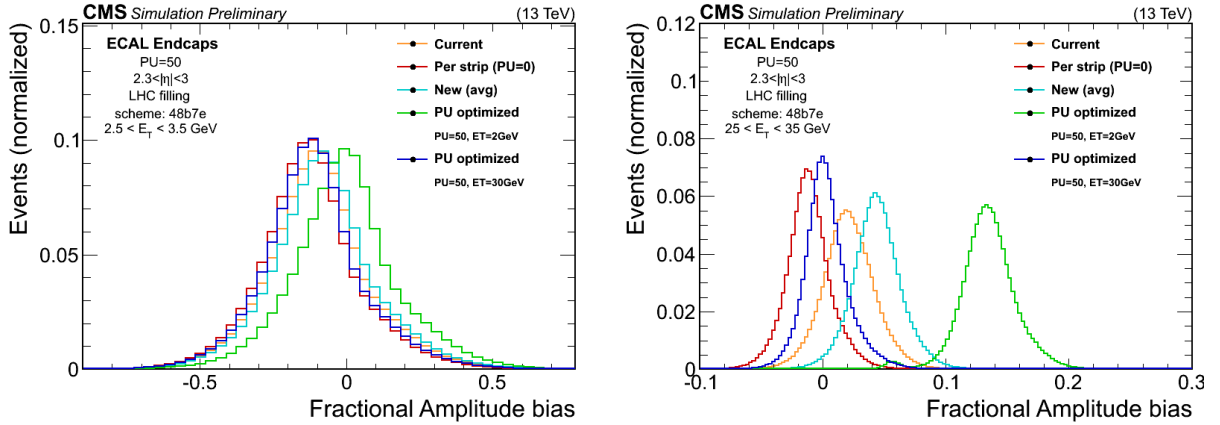


Figure 2.10: Amplitude reconstruction bias  $E[(E_T^{\text{reco}}/E_T^{\text{true}})] - 1$  for different set of weights. On the left 3 GeV signal, on the right 30 GeV signal, both with  $PU = 50$ .

In order to have a realistic scenario the TP spectrum from a collision Run in 2018 is used as a p.d.f. to generate events in the *48b7e* configuration and compare the performance of different set of weights, see Fig 2.11.

Fig. 2.12 shows on the left the TP bias and on the right the TP resolution obtained with PU optimized weights in different  $E_T$  bins for the  $\eta\text{-rings} \in [26, 28]$ . Weights optimized for low energy signals show an improvement in bias and resolution in the bulk of the TP distribution, for  $E_T < 5$  GeV.

Looking instead at the TP performance along the LHC train, inclusively in energy, the TP bias and resolution for different weights are shown in Fig. 2.13. The TP bias for low-energy-optimized weights is worse than the bias obtained with large-energy weights, but the TP resolution is 20% better, given the fact that the bulk of the TP distribution is  $E_T < 5$  GeV.

The two set of candidate weights have been also applied to different PU and energy conditions using the fast MC to assess the mean bias in the TP energy. The result is shown in the Fig. 2.14. Since the TP energy scale is calibrated by the L1 trigger in a subsequent step, the improvement

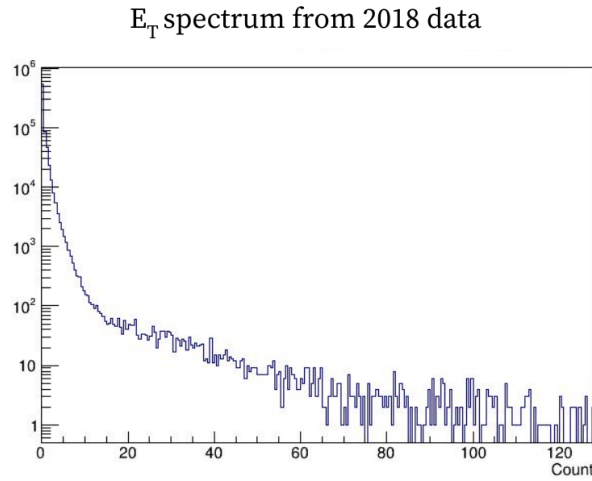


Figure 2.11: TP spectrum taken from a collision Run during 2018, and used as a p.d.f. for a realistic TP simulation for the validation of the weights.

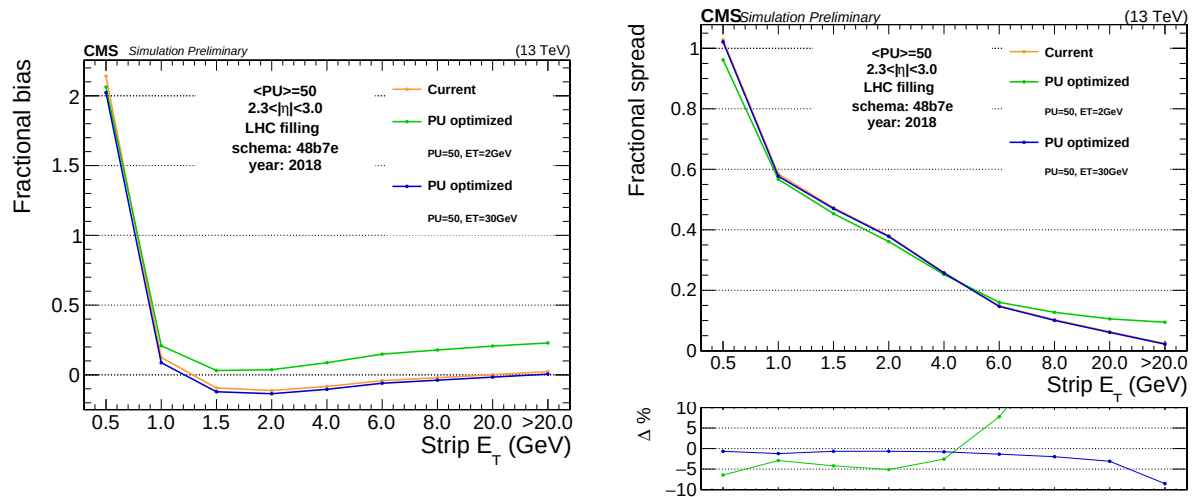


Figure 2.12: TP bias and resolution in bins of  $E_T$  obtained with Run II (current) and PU optimized weights. The simulation is performed for  $\eta$ -rings  $\in [26, 28]$  with a  $48b7e$  LHC schema.

in resolution is favorable even if it can introduce a bias in the TP energy. Therefore, the low-energy-optimized set of weights is a good candidate to improve considerably the TP resolution in the high  $\eta$  region of the EE: this improvement has the potential to help the resolution of L1 triggers for jets and missing energy. Further validation is still on-going to finally assess the potential effect on the L1 object resolution and on L1 rates before the start of Run III.

### 2.1.4 Data - MC comparison

The candidate sets of weights have been also applied on data from Run II thanks to a process called re-emulation, where the TP computation is applied to the real digitized sampled acquired from ECAL in a CMS software (CMSSW) package. The goal is to validate the standalone fast simulation of the ECAL pulses described in Sec. 2.1.2.1 comparing the TP performance in data and in the MC. The detector conditions and LHC schema used in the data are replicated carefully in the pulse simulation, as well as the TP spectrum.

Fig. 2.15 shows the TP bias along the BX of the LHC  $8b4e$  schema used in 2017, whereas Fig. 2.16 shows the TP resolution. On the left of the plots the performance on re-emulated data is shown,

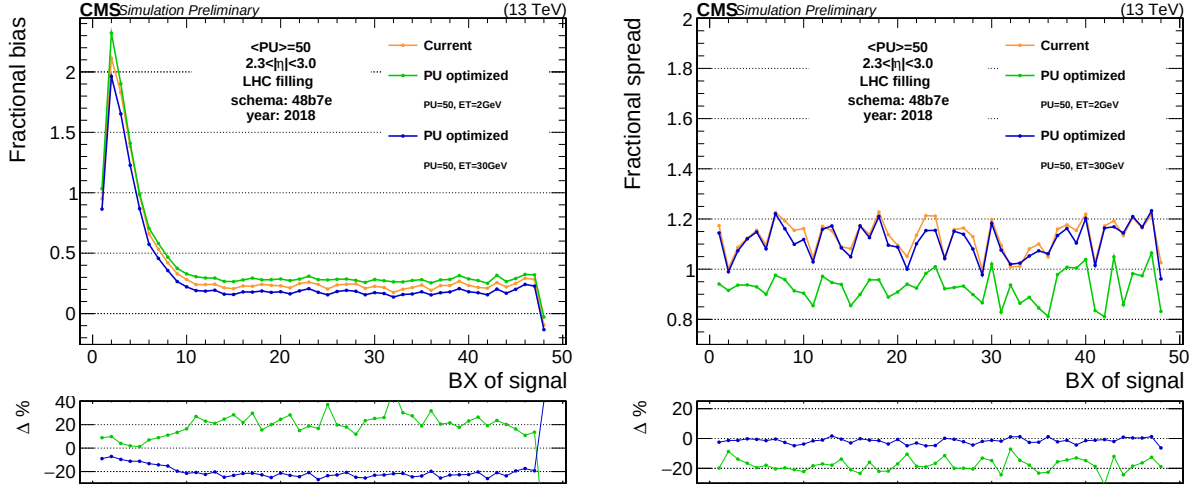


Figure 2.13: TP bias and resolution along the LHC train obtained with Run II (current) and PU optimized weights. The simulation is performed for  $\eta$ -rings  $\in [26, 28]$  with a  $48b7e$  LHC bunch filling schema.

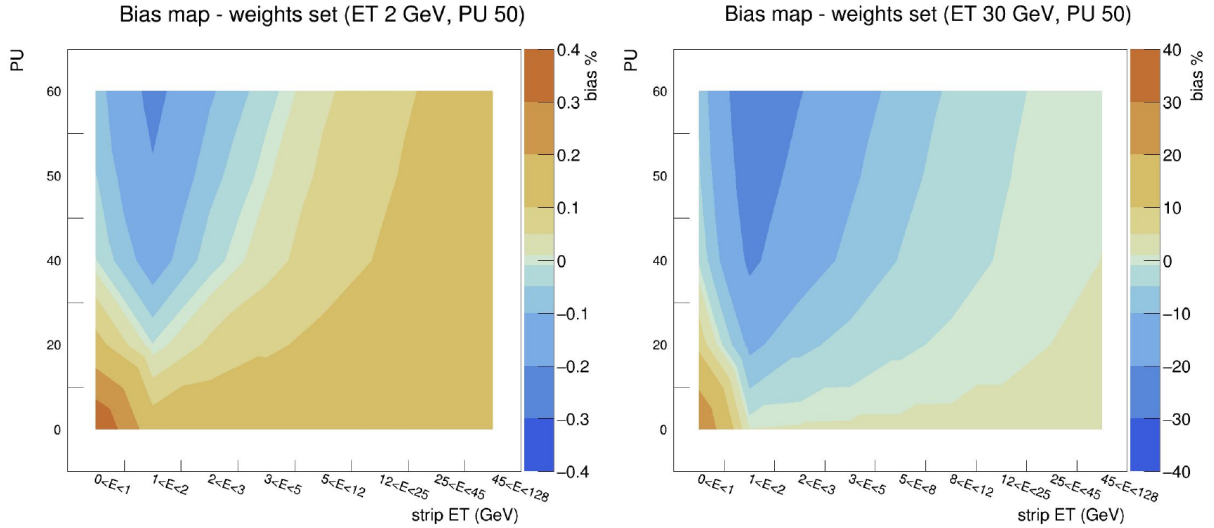


Figure 2.14: Bias in the TP energy for different PU and signal amplitude using the low-energy weights (left) or the high-energy one (right) in the  $\eta$ -ring  $\in [26, 28]$ .

on the right the one computed with the fast MC. Although, the absolute values for the bias and resolution do not match perfectly, the trend and the relative performance of different set of weights are the same. The scale differences have been understood as coming from the approximation implemented in the simulation, missing intercalibration and more realistic noise simulation, and residual discrepancies between the energy and BX distribution in simulation and data.

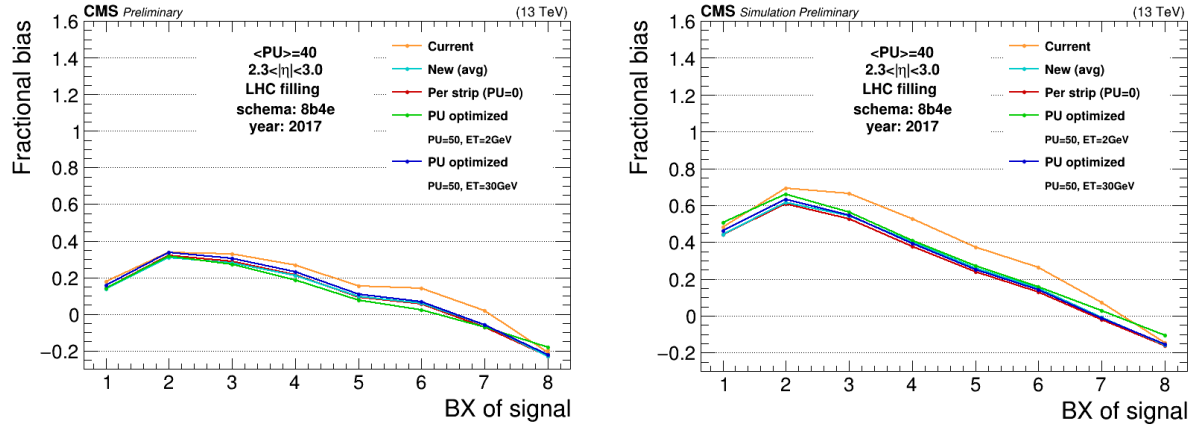


Figure 2.15: Comparison of TP bias with different set of weight on re-emulated data (left) and standalone MC (right).

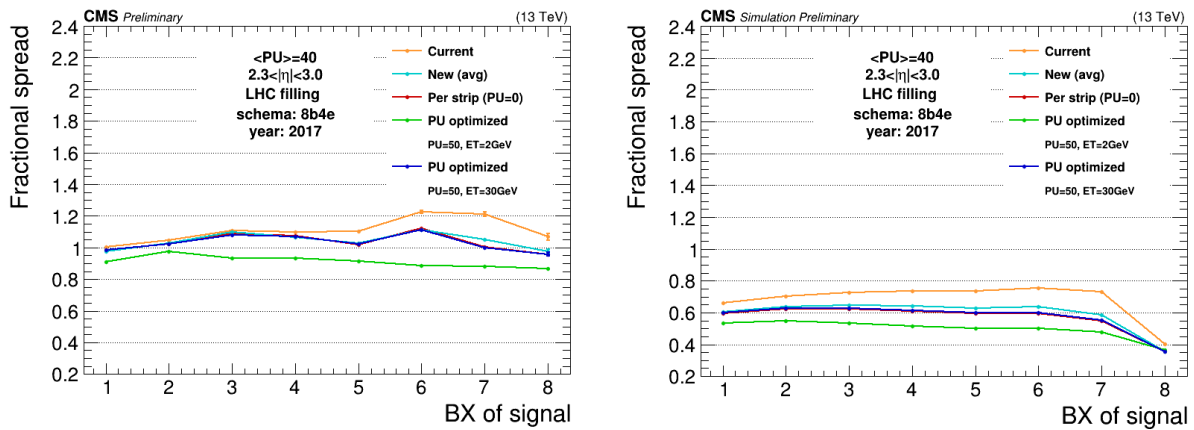


Figure 2.16: Comparison of TP resolution with different set of weight on re-emulated data (left) and standalone MC (right).

## 2.2 Double weights mechanism

The ECAL front-end electronics has been described in detail in Sec. 1.3.4.1, as well as the formation of the trigger primitives (TP) in the FENIX chip in Sec. 1.4.1. The TP algorithm is implemented in ASICs, therefore the algorithms cannot be changed, but only configured with the linearization parameters and trigger weights (see Sec. 1.4.2). During the long shutdown between Run II and III, a previously unused configuration of the trigger electronics has been re-discovered and investigated: the presence of a second filter applied in parallel to the ECAL pulses. This behavior has a limited set of options, but it is flexible enough to accommodate some interesting scenarios, like improved spike killer or out-of-time pulse tagging. In this section this functionality of the electronics will be described in detail, as well as the possible applications during Run III to improve the performance of the ECAL trigger system.

### 2.2.1 Electronics behavior

The double weights mechanism involves both the strip FENIX and the TCP FENIX (in EB). The behavior of these chips is configured by the CCS board via a set of predefined registers.

#### 2.2.1.1 Strip FENIX

After the linearization of the signal from each crystal and the formation of the strip sum, two independent filters are present in the strip FENIX chip, as well as two independent peak finders. A schema of the strip FENIX can be found in Fig. 2.17. For clarity the first filter is called *even*, the second is called *odd*, but they are completely inter-exchangable. During Run I and II the weights of the odd filter have always been 0. Each filter consists of a set of up to 6 weights and they compute an amplitude in a sliding window fashion, as described in detail in Sec. 1.4.1.3. Each peak finder can be independently switched on and off.

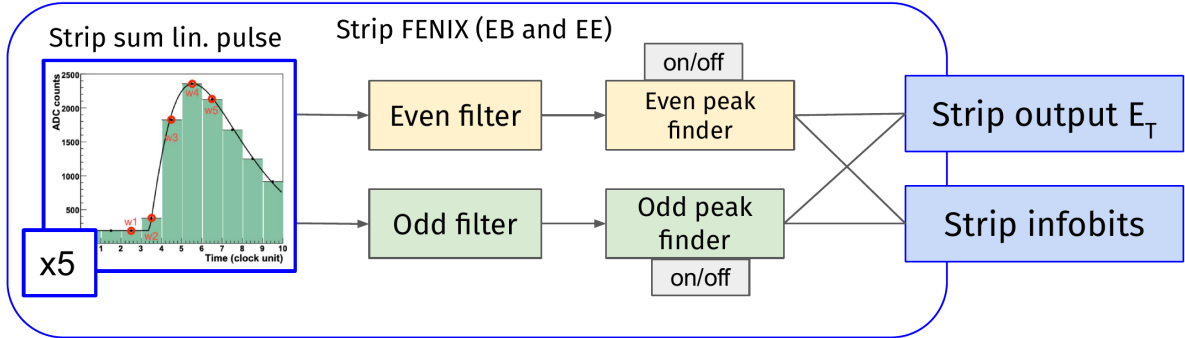


Figure 2.17: Schema of the strip FENIX inner working with 2 parallel filters and peak finders.

However, the output of the strip FENIX is a single 12-bit amplitude and 2 flags, called *infobits*. The output of the strip FENIX is configurable:

- even amplitude
- odd amplitude
- even+odd amplitude
- the largest between even and odd amplitudes

One of the infobits is the strip FGVB used by the spike killer (see Sec. 1.4.3.1). The other infobit can be configured to return 1 if the odd amplitude is larger than the even amplitude.

### 2.2.1.2 TCP FENIX

In the barrel the TCP FENIX receives the 12-bit amplitudes and 2 infobits from each strip, performs the tower sum and compression (see Sec.1.4.1.5). In reality, two separate energy sums are computed in the TCP FENIX: an even and odd sum. The chip can be configured to use the infobit 2 from the strip TCP as an odd energy flag: if infobit 2 is ON the corresponding strip energy is put in the odd energy sum, if it is OFF in the even total sum. During Run I and II the TCP FENIX has always been configured to ignore the infobit 2 from the strip FENIX, hence all the strip energy has been summed in a single number. The output of the TCP FENIX can be configured to be:

- even energy sum
- the larger between the even and odd energy sums
- the sum of the even and odd energy sums

The TCP FENIX transmits 2 infobits to the TCC boards off detector. The first infobit is the OR of the strip FGVB (sFVVBG, see Sec.1.4.3.1) and it is used for the spike killing. The second infobit can be configured to alternately have 3 meanings:

**FGVB** A Fine Grain Veto Bit computed at tower level by the TCP FENIX. It is ON if the maximum energy of a pair of adjacent strips is  $\max(E_T^{\text{pair}}) > X \cdot E_T^{\text{tot}}$  with the fraction  $X$  configurable in the frontend. This bit is currently used in the L1T for discrimination between EM and jet candidates.

**Gap flag** The infobit 2 is ON in the BX where the TCP FENIX receives the signal (from the CCS board) that the orbit gap is starting. This is used to analyze and correct the synchronization of all the trigger towers.

**Odd energy flag** The infobit 2 is ON when the odd energy sum is larger than the even energy flag.

The configurations of the output and infobits are independent.

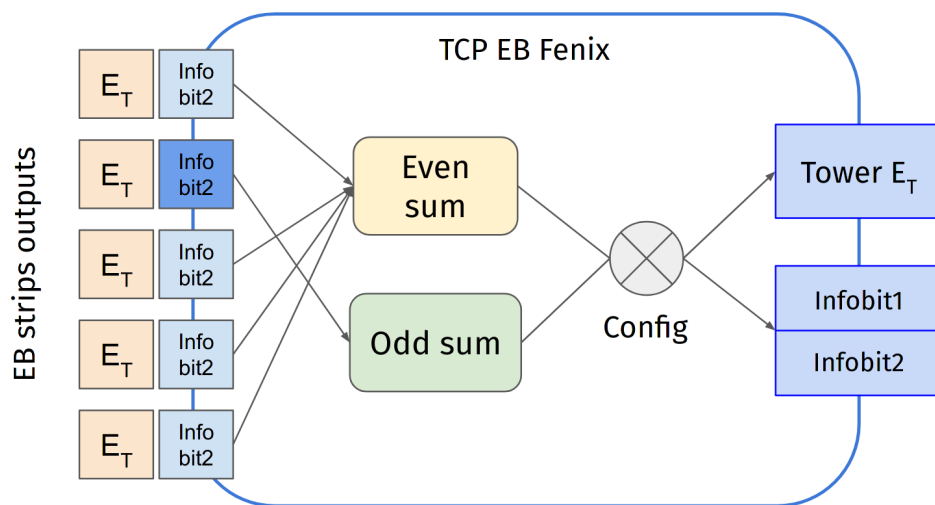


Figure 2.18: Schema of the TCP FENIX inner working with even and odd strip energy sums.



### 2.2.2 Useful scenarios

The capability of comparing the output of the two filters via flags and separate energy sums makes possible the implementation of two scenarios for the double weights mechanism: reduce the spike contamination or tag the out-of-time signals.

#### 2.2.2.1 Spike killing

The spike timings distribution in ECAL has a long positive tail, see Fig. 2.19. The time 0 corresponds to the time of arrival of particles in front of ECAL from the collision point. The long tail is not given by particles directly interacting in the APD sensors, but by secondary particles, typically neutrons, generated elsewhere in the detector and drifting for 10s of ns before interacting close to the barrel APDs (see Fig. 1.18 for a sketch of the process). The online spike killer exploits just the energy topology to zero out towers with single crystal high energy deposits, whereas the timing information is used just offline to remove spikes with an offset of  $\pm 3$  ns from the expected timing. Timing discrimination included in the online trigger would help to remove more spikes in an orthogonal way. The double weights mechanism can be tuned to zero out the energy from the strips where the pulse is out-of-time. More specifically the FENIXs need to be configured as follows:

- **Weights:** the odd weights have to be optimized for a positive time shifted signal. Doing so the odd amplitude will be larger of the even amplitude if the pulse is time shifted.
- **Strip FENIX:**
  - Output the largest between odd and event energy.
  - Use the infobit to signal if the odd amplitude is larger than the event one.
- **TCP FENIX:**
  - Use the infobit from the strip FENIX to put odd energy from strip flagged as odd in the odd sum.
  - Output the even energy sum, effectively removing the energy from the strips where the pulse is flagged as out-of-time.
  - Two configurations are possible for the TCP FENIX oENIX is always the even energy, doing so the total TP output is the nominal (even) one.
    - **Flagging** Output a flag if the odd sum is larger than the event sum: in this way the zeroing of the energy can be monitored offline.
    - **Non-flagging** Output the FGVB bit as infobit2 as in the standard configuration: in this way the zeroing can be monitored only with re-emulation of the trigger algorithm.

This algorithm has been applied the standalone MC, simulating spikes with the timing p.d.f. shown in Fig.2.19, and the tagging efficiency for different weights delays have been computed. Optimizing the weights for a pulse delayed by  $+X$  ns, late spikes around that timing are correctly removed. A careful optimization has to be performed to avoid loosing efficiency in the triggering of EM signals. More details about the candidate set of odd weights and its application on data will be given in Sec. 2.2.4.

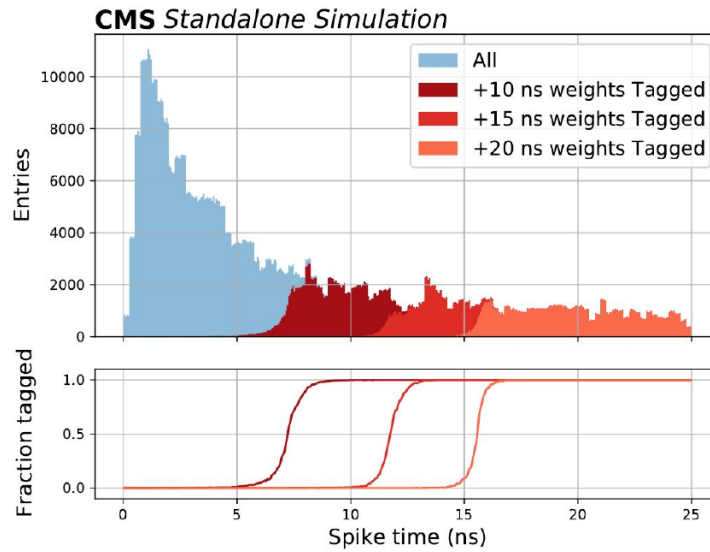


Figure 2.19: Timing distribution for spikes in EB.  $t = 0$  corresponds to the time of arrival of particles in front of the ECAL detector. The plot shows the fraction of spikes tagged by the double weights mechanism for 3 different set of odd weights optimized for different pulse time shifts.

### 2.2.2.2 Out-of-time signal tagging

A similar approach can be used to tag the presence of out-of-time energy without affecting the TP energy output. This configuration can be of interest for searches of exotic heavy or long-lived beyond the standard model (BSM) particles. The configuration of the FENIX chips is only slightly different:

- **Weights:** the odd weights have to be optimized for a positive time shifted signal. Doing so the odd amplitude will be larger than the even amplitude if the pulse is time shifted.
- **Strip FENIX:**
  - Output the even energy.
  - Use the infobit to signal if the odd amplitude is larger than the even one.
- **TCP FENIX:**
  - Use the infobit from the strip FENIX to put energy from strip flagged as odd in the odd sum.
  - Output the sum of the even and odd energy sums. Since the output of the strip FENIX is always the even energy, the total TP output is the nominal (even) one.
  - Output a flag if the odd sum is larger than the event sum: in this way if there is an delayed pulse, the information can be propagated to the L1 trigger.

In order to exploit this configuration, the FGVB bit of the tower has to be replaced by the out-of-time bit from the double weights mechanism. Investigations are ongoing in collaboration with the L1T groups to understand the current use of the FGVB bit and the possibility to remove it to make space for the out-of-time pulse flag.

### 2.2.3 Tests and implementation

#### 2.2.3.1 Hardware tests

The double weights mechanism has been tested in the ECAL frontend electronics in the ECAL DAQ test system located in building 904 at CERN. The setup consists of 4 fully equipped EB towers and 4 EE ones, connected to the full chain of off-detector boards, reduced to a single DCC-CCS-TCC module. The towers are equipped with FE and VFE boards, but there are no crystals or photodetectors connected. Realistic pulses are needed in order to fully test the behaviour of the different configurations of the FENIX chips. The VFE cards can be configured in a special way in order to inject a series of ADC counts into the strip FENIX to emulate the presence of a real signal. The flexibility of this approach permits injecting time shifted pulses and to analyze the FENIX output with different double weights configurations.

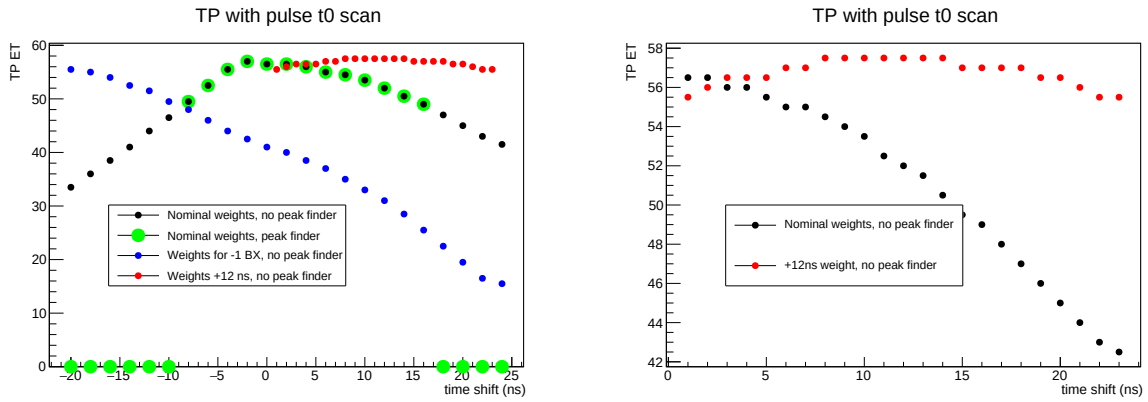


Figure 2.20: Comparison of the effect of different set of weights for time shifted pulses with and without the peak finder.

A signal of amplitude 50 GeV is injected in the VFE card varying the time shift from -20 ns to +25 ns in steps of 2 ns. Three FENIX configurations are compared: nominal weights (with and without the peak finder active), weights optimized for a pulse shifted of -25 ns (1BX), and weights optimized for +12 ns time shift. Fig. 2.20 on the left shows the comparison between the nominal weights and the BX-1 weights. The peak finder assigns the energy in the central BX only when the pulse is within  $(-10, +17)$  ns: the comparison with the blue dots from the BX-1 weights shows that the energy would have been assigned to the BX before if the time shift had been  $< -10$  ns. Nominal weights are also compared with +12 ns weights without the peak finder: from +3 ns the amplitude from the time shifted weights is larger than the nominal one, see a zoom in Fig. 2.20 on the right. The maximum TP occurs at approximately +12.5 ns with the time-shifted weights, as expected.

Having checked the dependence of the amplitude on the shift of the pulse in the hardware for different weights, the double weights mechanism has been set in the FENIX to zeroing and flagging mode. The odd filter is configured with +12 ns weights and the FENIX is configured to zero out the energy if the odd energy in the strip is larger than the even one. Fig. 2.21 shows the result of the test: the FENIX output is zeroed on the left because of the peak finder, and on the right due to the odd weights. Checking the TCP FENIX infobits reveals that the infobit2 flag is correctly set when odd energy is larger than even energy for those pulses.

#### 2.2.3.2 Implementation in CMS software

After the feasibility of the double weights scenario has been demonstrated in hardware, the new mechanism has been implemented in the ECAL DAQ software and configuration database.

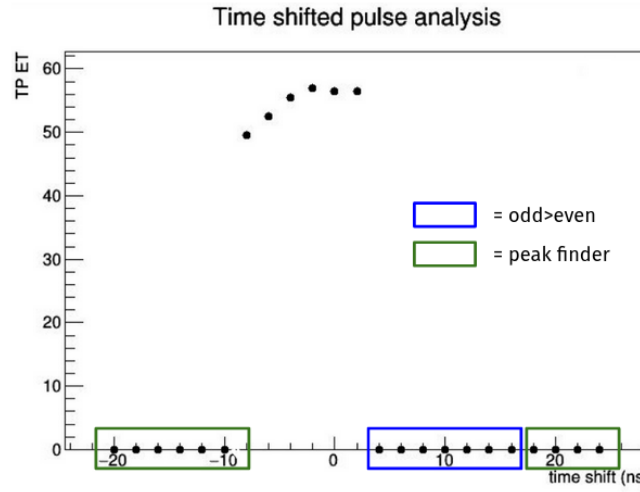


Figure 2.21: TP output obtained in the double weights mechanism in zeroing and flagging mode. The odd weights are configured with weights optimized for a +12 ns pulse. The even amplitude is correctly zeroed when the odd energy is larger than even energy and the TCP FENIX infobit2 is ON.

It has been implemented in the CMS C++ software trigger emulator and in the conditions database. Having the possibility to re-emulate the double weights mechanism in Run II data is fundamental to assess the performance of odd weights candidates before the actual Run III data taking starts.

The software changes have been validated during the commissioning periods in 2021, confirming the double weights mechanism inner working also on the actual detector hardware. The LHC pilot beam in October 2021 will be useful from the ECAL trigger side to experiment with the use of the double weights using actual collision data.

### 2.2.3.3 Weights numerical optimization

Finding the best candidate for the odd weights in spike killing mode requires looking for a balance between spike rejection efficiency and loss of EM signal efficiency. Multiple factors should be taken into account: the energy spectrum of the signal and of the spikes, the timing p.d.f. of the signal and spikes, and the PU contamination. The fast MC developed for the weights optimization with PU (see Sec. 2.1.2.1) is used to produce many pulses for both EM signal and spikes, using a realistic TP spectrum and the appropriate timing p.d.f.. Finding the best odd weights candidates, maximizing both signal efficiency and spike rejection, is then rephrased as a numerical optimization problem: a loss function is defined and the gradient descent algorithm is implemented to find the best weights values.

- **Inputs of the algorithm:**
  - N set of digis  $D_i$  for signal pulses
  - N set of digis  $S_i$  for spike pulses
  - Even weights for in-time amplitude  $W_i^e$
- **Target of the algorithm:** find the best set of weights  $W_i^o$  so that
  - for signal pulses:  $\sum_i^5 W_i^e \cdot D_i > \sum_i^5 W_i^o \cdot D_i$
  - for spike pulses:  $\sum_i^5 W_i^e \cdot S_i < \sum_i^5 W_i^o \cdot S_i$
- **Procedure:**

- Define a loss function as :

$$\mathcal{L} = \lambda_{\text{sig}} \left[ \sum_i^5 (W_i^o - W_i^e) \cdot D_i \right] + \lambda_{\text{spike}} \left[ \sum_i^5 (W_i^e - W_i^o) \cdot S_i \right] + \lambda_{\text{norm}} \left[ \sum_i^5 (\max(|W_i^o|, 1) - 1) \right]$$

The loss is larger if the odd weights gives larger energy than even weights on signal pulses and the opposite on spike pulses. A loss component is added to keep the absolute values of each odd weight less than 1 (for encoding it in the FENIX chip). The  $\lambda_{\text{sig}}$  and  $\lambda_{\text{spike}}$  hyperparameters regulates the relative importance of the signal efficiency versus the spike efficiency.

- Include the constraint  $\sum_{i=1}^5 W_i^o = 0$  by optimizing only 4 weights and taking the last as  $W_5^o = -\sum_{i=1}^4 W_i^o$
- Optimize the weights value using the stochastic gradient descent technique. Split the pulses in batches of N signal pulses and N spike pulses, compute the loss function on those pulses and its gradient w.r.t the odd weights. Then modify the odd weights in the direction of maximum negative gradient. The parameter  $\eta$  is called the learning rate and it regulates how much each weight is changed at each step during the optimization. Repeat the process until a minimum is reached.

$$W_i^o = -\eta \nabla_{W^o} \left[ \sum_{N_{\text{sig}}} \mathcal{L}(D_i) + \sum_{N_{\text{spike}}} \mathcal{L}(S_i) \right]$$

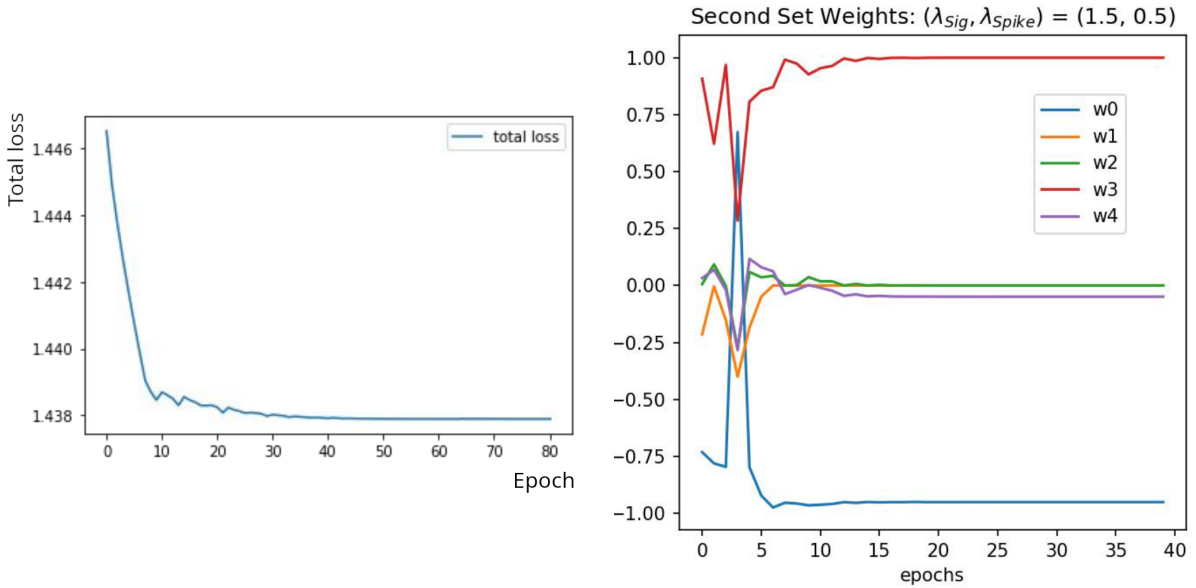


Figure 2.22: Variation of the loss and weights  $W_i^o$  during the optimization.

The procedure above has been implemented with the PyTorch[34] library automatic gradient computation and applied on pulses obtained from the fast MC. More importance has been given to the signal efficiency than to the spike rejection. The optimization technique can still be

improved a lot by introducing further regularization in the training. Moreover its advantage is that the training can be guided further, by giving more importance to specific regions of the signal spectrum by applying weights to the pulses in the gradient descent. A preliminary optimization has been performed considering and the resulting weights are shown in Fig. 2.22.

## 2.2.4 Preliminary performance study

### 2.2.4.1 Spike killing

The double weights candidate has been applied on real CMS data from 2018 via software re-emulation. Events selected by the *ZeroBias* trigger have been picked for analysis in order to have an unbiased sample of ECAL TPs. Spikes are discriminated in data by the offline timing and swiss-cross selections (see Sec. 1.4.3.2). If the timing requirement  $\pm 3$  ns of the rechits associated with the TP fails the TP is categorized as *severity 3*, if instead the the Swiss-cross cut fails, it is categorized as *severity 4*. The zeroing configuration is emulated to check the amount of TPs zeroed by the mechanism. The original TP energy without the double weights mechanism is retrieved from the original data.

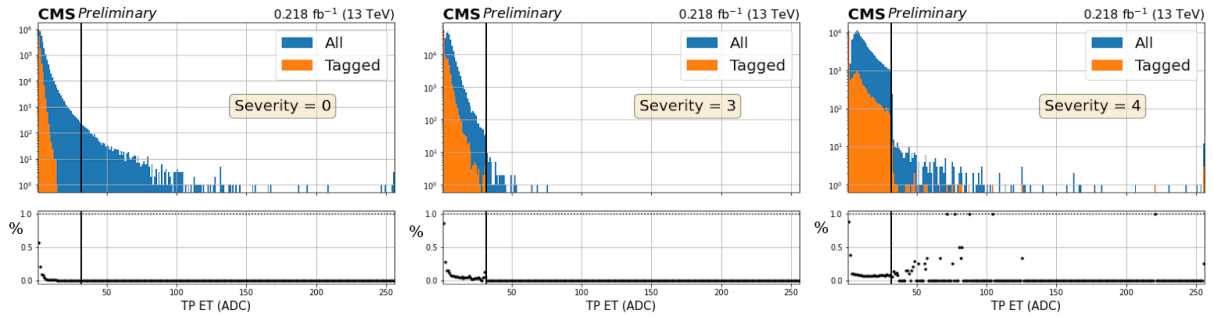


Figure 2.23: Fraction of TP in data that are zeroed by the double weights mechanism when re-emulated with the zeroing configuration. Signal on the left, TPs with failed time cut (severity 3) in the center, TPs with failed swiss-cross (severity 4) on the right.

Fig. 2.23 displays the fraction of signal and spikes TPs identified by the double weights mechanism along the TP spectrum. The 32 ADC (16 GeV) cut imposed by the online spike killer (see Sec.1.4.3.1) is clearly visible in the spectrum of Severity 3 and 4 TPs (offline spikes). Around 10% of the spikes at low energy are zeroed by the double weights, as well as a non-negligible fraction of signal TPs below 10 ADC (5 GeV).

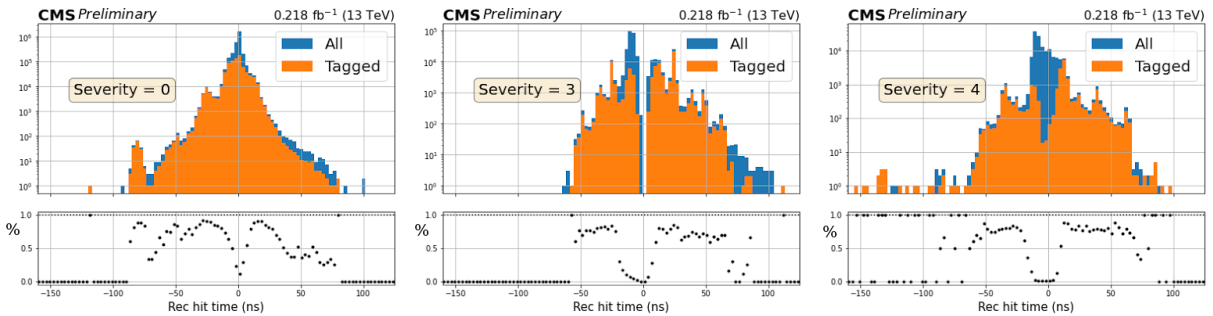


Figure 2.24: Fraction of TP in data that are zeroed by the double weights mechanism when re-emulated with the zeroing configuration. Signal on the left, TPs with failed time cut (severity 3) in the center, TPs with failed swiss-cross (severity 4) on the right.

A timing measurement is linked to each TP by looking at the timing of the highest energy associated rechit. The zeroing efficiency of the double weights mechanism is then analyzed in

terms of TP timing: results are shown in Fig. 2.24. TP are identified as out of time and zeroed with high efficiency as soon as the time shift is  $\pm 10$  ns, in agreement with the fast MC studies.

This preliminary study demonstrates the possible advantage of the double weights mechanism in killing spikes in the online trigger, therefore reduce the rate contamination of the L1T from spikes. The loss of signal efficiency will be studied carefully to find an optimal working point.

### 2.2.4.2 Tagging

The double weights mechanism in tagging mode has been tested for the first time on a real physics event during the beam commissioning period in October 2021. During the LHC beam commissioning, beam splash event are recorded by CMS. These special events happen when the LHC beam is made to colliding against the graphite collimators near the CMS experimental cavern. The interaction of the proton beam ( $E = 450$  GeV) with the collimator produces a spray of high energy secondary particles, but only muons pass through the 1m thick collimator and reach the CMS detector. The shower of highly energetic muons are recorded by the ECAL and HCAL calorimeters, that trigger the acquisition of the event, an example is shown in Fig. 2.25.

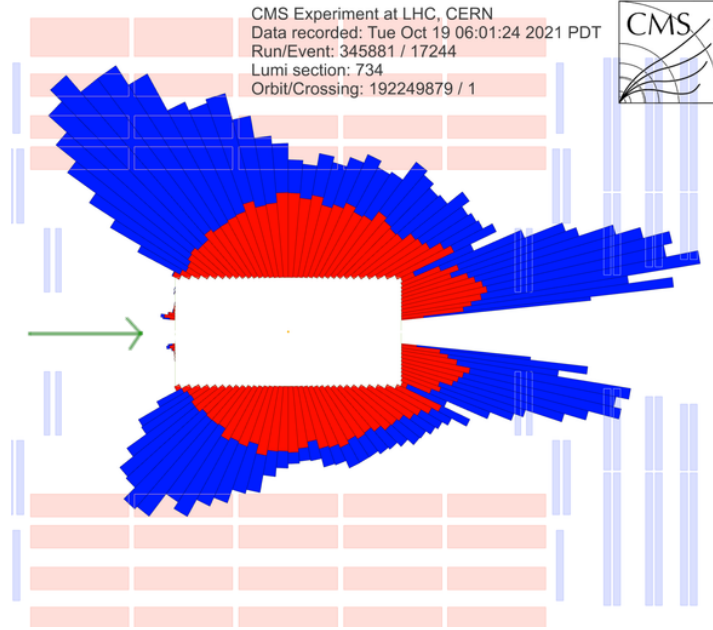


Figure 2.25: Event display of a beam splash event in CMS, recorded in October 2021 beam commissioning. ECAL towers are shown in red, while HCAL ones in blue. The muons are coming from the left in the picture.

This type of event has a characteristic timing profile from the ECAL point of view, since particles are coming from one side of the CMS detector and the readout timing is instead synchronized to detect particles spreading from its center. Therefore, the signal appears as largely out-of-time in ECAL, as shown by the plot in Fig. 2.26. Double weights in tagging mode, keeping the TP energy unchanged (as explained in Sec.2.2.2.2), have been loaded in the ECAL frontend: from re-emulation studies the candidate set of weights should tag pulses which are out-of-time of  $\sim 10, 15$  ns, as shown in Fig. 2.24. The infobit2 of the TCP FENIX is received by the TCC and sent to the L1T: in this configuration the bit is not the FGVB but the double weights odd larger than event weight. Fig. 2.27 shows that the infobit2 is fired in the region of the detector at  $i\eta_{L1} = 9$  corresponds to  $i\eta_{ECAL} = 50$ , a region where the timing of the pulses from the splash is in the expected range. This is the first successful test of the double weights on real physics events.



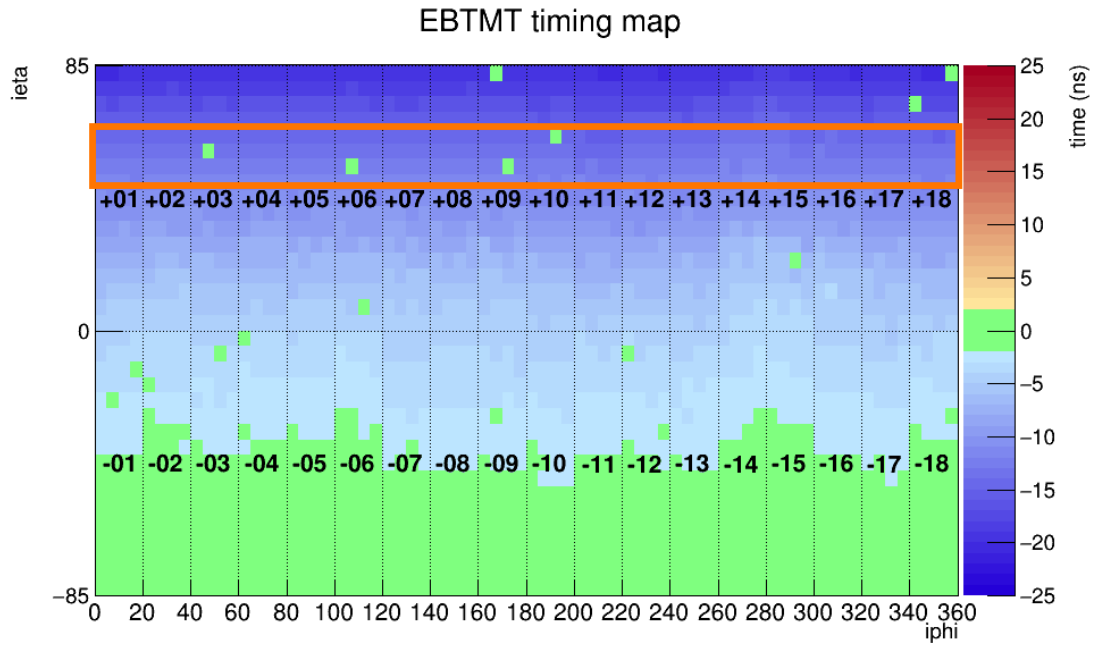


Figure 2.26: Timing of the ECAL signals in the barrel in the beam splash event.

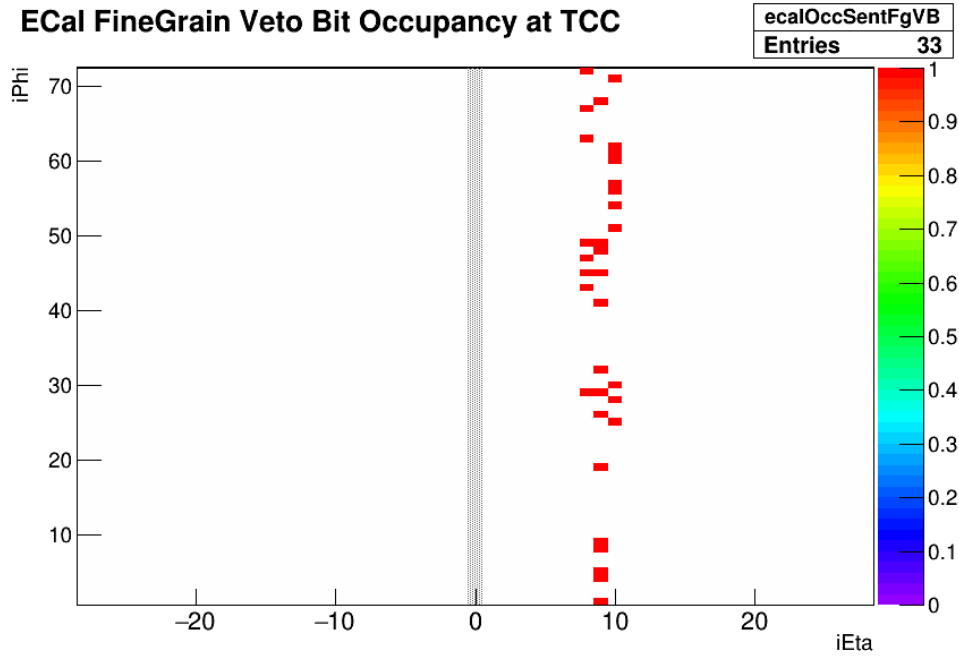


Figure 2.27: TCP FENIX infobit2 configured in the double weights tagging mode in a splash event. The region of the detector at  $i\eta_{L1} = 9$ , corresponding to  $i\eta_{ECAL} = 50$ , which is out of time by  $\sim 10$  ns in this event, is flagged as out-of-time by the infobit.

## 2.3 SuperClustering with machine learning

The Mustache SuperClustering is a purely geometrical algorithm, that does not attempt to identify and remove noise and pileup clusters from the aggregated PFClusters. Every cluster falling the  $\Delta\eta - \Delta\phi$  Mustache region around an energetic seed cluster is added to the electron (or photon) candidate. The increase of the pileup level and the expected increase of noise contamination during Run III, makes the development of a more performant SuperClustering algorithm based on supervised machine learning (ML) methods desirable. Particular interest is aimed at graph neural networks (GNN) [35–37], a novel class of supervised models suitable for problems involving a variable number of objects and their relation.

### 2.3.1 Truth-matching

The first necessary ingredient to build a SuperCluster algorithm based on supervised machine learning methods is the definition of an optimal true association between the seeds and the PFClusters in the event. This is required in order to understand which PFClusters contains the energy coming from electrons and photons and which are coming from PU and noise. The CMS full MC simulation based on Geant4 is used to produce events with realistic noise and PU interactions also simulated and superimposed onto the signal energy deposits. The reconstruction of simulated events is carried on up to the PFClusters level.

#### 2.3.1.1 CaloParticle - PFCluster matching

A new object called *CaloParticle*, tracks the EM shower produced by particles in the MC simulation, and stores the information about the amount of energy, called *simEnergy*, deposited in each ECAL crystal, in the form of *simHits*. SimHits do not correspond perfectly to rechits, given the inefficiencies of energy reconstruction, but they are a good proxy for the deposited energy generated by a particle detected in ECAL. This information allows a precise matching between each PFCluster and the particle defined at the Geant4 level (also called GEN level), taking fully into account effect from the energy deposition and reconstruction of the calorimeter, and it performs much better than the usual position matching based on the  $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$  relative to the GEN-level particles. Different matching strategies have been studied, and the one showing the best performance consists of the following:

- For each **PFCluster**:
  - For each CaloParticle  $k$  having at least one energy deposit (*simHit*) in the crystals of the PFCluster, compute the matching score as:

$$\text{score}_k = \frac{\sum_{xtals} E_{simHit}^k}{E_k^{tot}}$$

where  $E_{simHit}^k$  is the amount of energy deposited in the simulation by the CaloParticle in the specific crystal and  $E_k^{tot}$  is the total deposited energy of the CaloParticle in the detector. This score is also called *simFraction*.

- Sort the CaloParticles by the maximum score
  - Assign the PFCluster to the CaloParticle with the highest score.
- For each **CaloParticle**:
  - keep the list of the associated PFClusters with the criterion described above
  - sort the PFClusters by maximum score.
  - the highest score PFCluster is called *caloSeed*.

CaloParticles are defined both for electrons and photons and for in-time and out-of-time PU particles. In order to save computational costs, all the particles simulated from PU are aggregated under a single special PU CaloParticle. PFClusters not assigned to a CaloParticle, signal or pileup, are considered to be produced by noise. Figure 2.28 shows a sketch of the simHits deposited by a CaloParticle in PFClusters, and the computation of the association score.

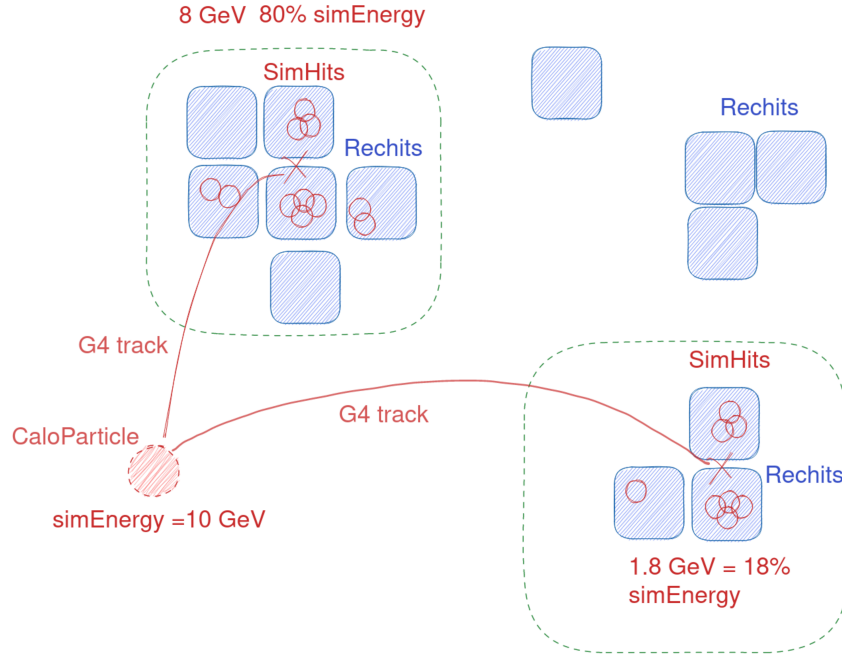


Figure 2.28: Sketch of the simHits deposited by a CaloParticle in the PFClusters and the computation of the association score.

### 2.3.1.2 PFCluster - seed matching

Given a seed PFCluster, the goal of the truth association is to define which PFClusters have to be collected in order to get the best possible energy resolution. Once the association of each PFCluster to CaloParticles or noise has been performed, a PFCluster is associated to a nearby seed only if it is associated to the same CaloParticle, and if its score is larger than a  $\zeta(E_T^{\text{seed}}, \eta^{\text{seed}})$  threshold, which depends on the position and energy of the seed. These thresholds are optimized collecting many seeds (1M) and analyzing all the PFClusters associated to the same CaloParticle in several bins of  $E_T^{\text{seed}}$  and  $\eta^{\text{seed}}$ . A scan on the value of  $\zeta$  is performed, computing for each value the total energy associated to the seed. The parameter of interest is the energy resolution, taken as the central interval containing 68% of the events. Fig. 2.29 shows several  $\zeta$  scans for seeds with  $E_T \in [6, 10] \text{ GeV}$ : the minimum energy resolution defines the best  $\zeta$ . A different optimization is performed over electrons and photons, in order to achieve the best possible truth association for each kind of SuperCluster. The final maps of  $\zeta(E_T^{\text{seed}}, \eta^{\text{seed}})$  are shown in Fig. 2.30.

### 2.3.1.3 PU cleaning

The simFraction score thresholds are optimized considering only the association to the signal CaloParticles. Additional information is available thanks to the PU CaloParticles: the amount of simEnergy associated to PU is known for each PFCluster. The fraction of energy from signal,

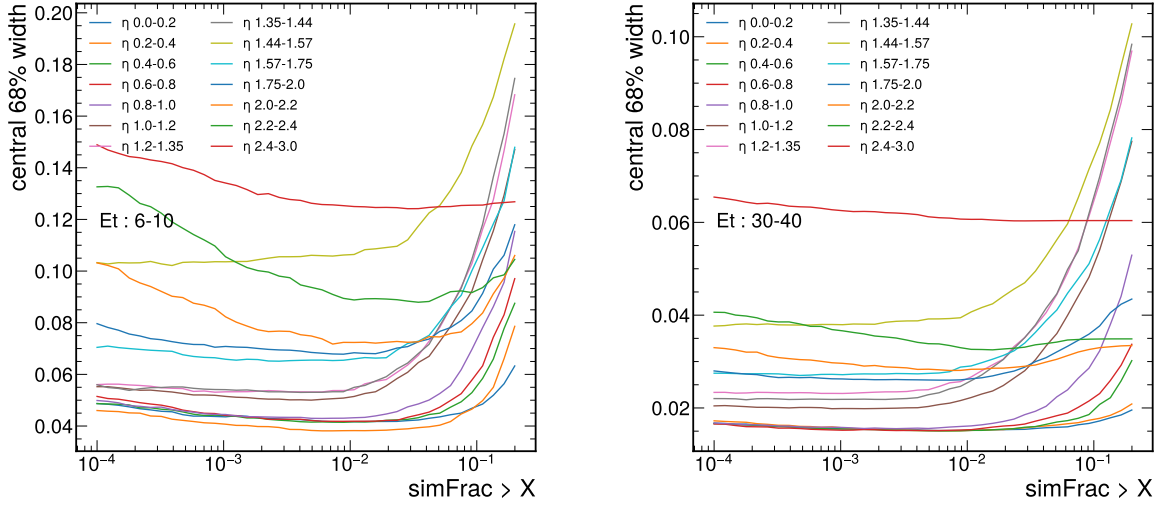


Figure 2.29: Scan of the association score to find the best energy resolution in several  $\eta^{\text{seed}}$  bins for a seed of  $E_T \in [6, 10]$  GeV (left) and  $E_T \in [20, 30]$  GeV (right). The energy resolution is taken as the central interval containing 68% of the events.

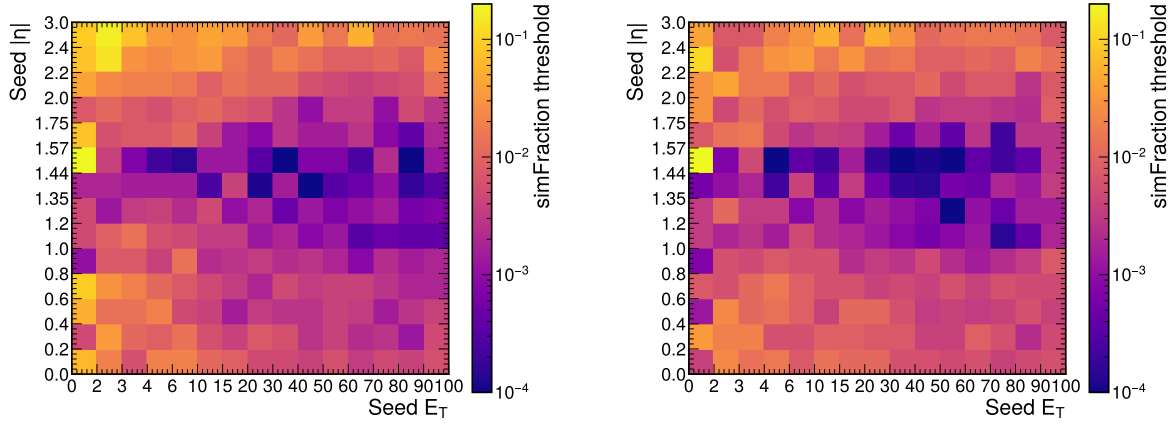


Figure 2.30: Maps of the optimum simFraction score thresholds  $\zeta(E_T^{\text{seed}}, \eta^{\text{seed}})$  for electrons (left) and photons (right).

$En_{\text{sim}}^{\text{signal}}/En^{\text{tot}}$ , is compared with the one from PU,  $En_{\text{sim}}^{\text{PU}}/En^{\text{tot}}$ , for all the PFClusters associated to a seed before and after the simFraction selection in Fig. 2.31. The plot shows that a large amount of clusters removed by the simFraction thresholds are already the ones with small simEnergy from signal and large simEnergy from PU. A scan is also performed to find the best energy resolution for different values of  $En_{\text{sim}}^{\text{PU}}/simEn_{\text{sim}}^{\text{signal}} < \lambda(E_T^{\text{seed}}, \eta^{\text{seed}})$ , after the simFraction thresholds have been applied. The scan is shown in Fig. 2.32: the energy resolution does not depend strongly on this cut, but a safe working point has been chosen. Cluster are not associated to a seed if the associated simEnergy from PU is larger than the simEnergy from the signal CaloParticle.

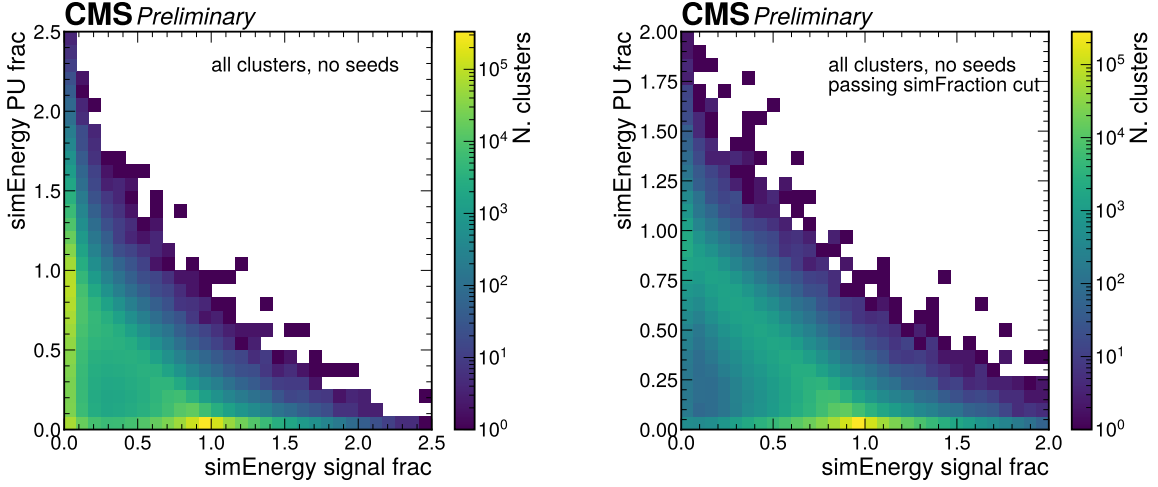


Figure 2.31: Comparison of  $\text{simEnergy}$  from the signal  $En_{\text{sim}}^{\text{signal}}/En^{\text{tot}}$ , and  $\text{simEnergy}$  from the PU for all the clusters (left), and only the clusters associated to a seed and passing the  $\text{simFraction}$  threshold (right).

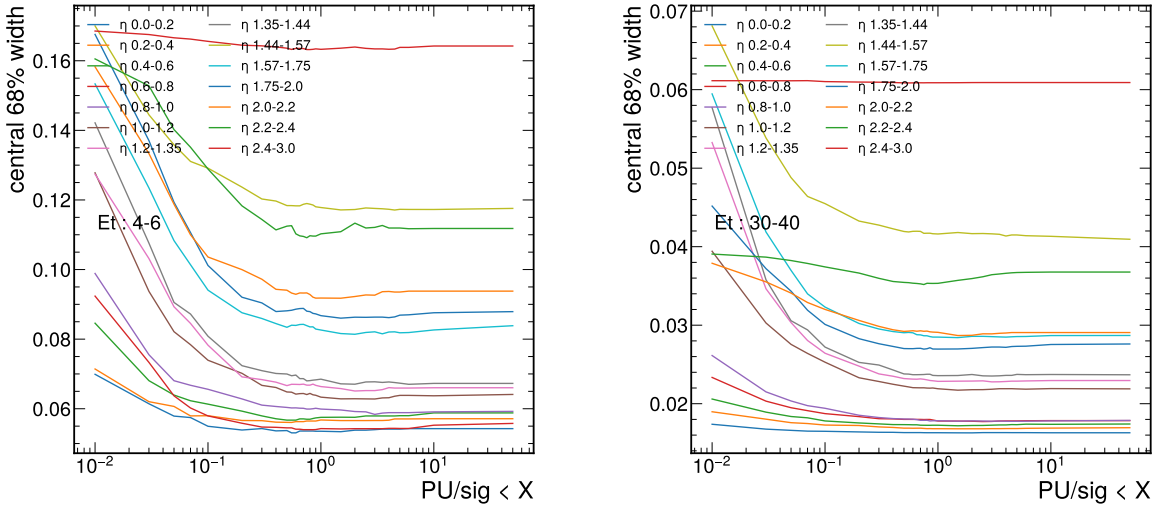


Figure 2.32: Scan of the optimal threshold  $En_{\text{sim}}^{\text{PU}}/En_{\text{sim}}^{\text{signal}} < \lambda(E_T^{\text{seed}}, \eta^{\text{seed}})$  to obtain the best energy resolution performed after the application of the  $\text{simFraction}$  threshold.

### 2.3.1.4 Ideal performance

The ideal energy resolution obtained by collecting all the clusters passing the  $\text{simFraction}$  threshold and PU cleaning, is compared with the raw energy resolution obtained with the Mustache SuperClustering. Fig. 2.33 shows the ideal resolution improvement w.r.t the Mustache for different transverse energy of the GEN-level particle  $E_T^{\text{GEN}}$  and  $\eta^{\text{seed}}$ : this is the maximum improvement reachable by any supervised machine learning model trained using the described truth association. The ideal resolution is also compared directly between electron and photons: the ratio is shown in Fig. 2.34 in bins of  $E_T^{\text{seed}}$  and  $\eta^{\text{seed}}$ . The ideal resolution is better for photons at high energy in the barrel, while electrons are better reconstructed than photons in the region at high  $|\eta|$  in EB. In both cases the maximum improvement is reachable in the region

$|\eta| \in [0.7, 1.479]$  of the barrel, where the material in front of ECAL (see Fig. 1.27) has its peak and electrons and photons are more fragmented. At higher  $\eta$  in the EE, the Mustache becomes narrower and does not include many spurious clusters, therefore the ideal improvement in that region is smaller.

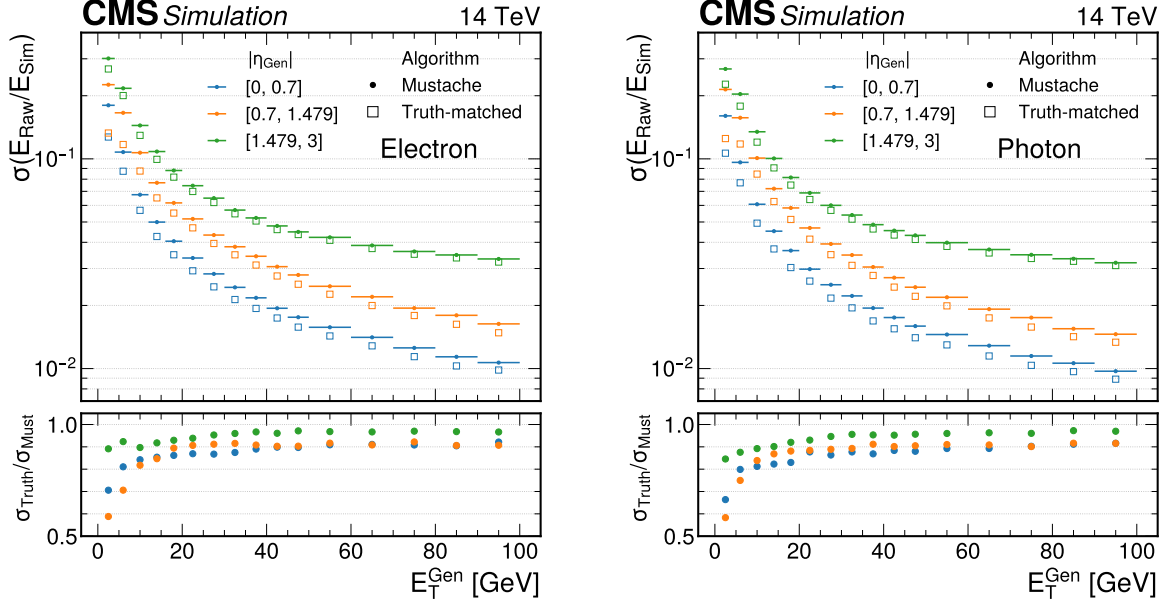


Figure 2.33: Maximum improvement in energy resolution w.r.t the Mustache SuperCluster using the truth association. The resolution is analyzed versus the GEN energy of the electron/photon particle.

### 2.3.2 Training dataset

The dataset used in the training of the SuperClustering ML algorithm consists of a set of samples, each corresponding to a seed and to the list of PFClusters around it. The maximum dimension of the detector window in which clusters are considered around a seed is chosen to be always larger of the Mustache window, and following the natural width of the electron and photon decay areas in the detector. The formula defining the window is given in Eq.2.2, and the  $\eta^{\text{seed}}$  dependent parameters are shown in Fig. 2.35.

$$\begin{cases} \Delta\eta(\text{seed}, \text{cl}) \leq \Delta\eta^{\text{max}}(\eta^{\text{seed}}) \\ \Delta\eta(\text{seed}, \text{cl}) \geq \Delta\eta^{\text{min}}(\eta^{\text{seed}}) \end{cases} \quad \eta^{\text{seed}} > 0$$

$$\begin{cases} \Delta\eta(\text{cl}, \text{seed}) \leq \Delta\eta^{\text{max}}(\eta^{\text{seed}}) \\ \Delta\eta(\text{cl}, \text{seed}) \geq \Delta\eta^{\text{min}}(\eta^{\text{seed}}) \end{cases} \quad \eta^{\text{seed}} < 0$$

$$\Delta\phi(\text{seed}, \text{cl}) < \Delta\phi^{\text{max}}(\eta^{\text{seed}})$$

(2.2)

(2.3)

The preparation of the dataset is described in the following and sketched in Fig. 2.36:

- Events with electrons and photons are simulated with the CMS full MC and CaloParticle information is stored in the event.
- Each PFCluster is associated to a signal or PU CaloParticle following the criterion described in 2.3.1.1. If it is not associated it comes from noise.

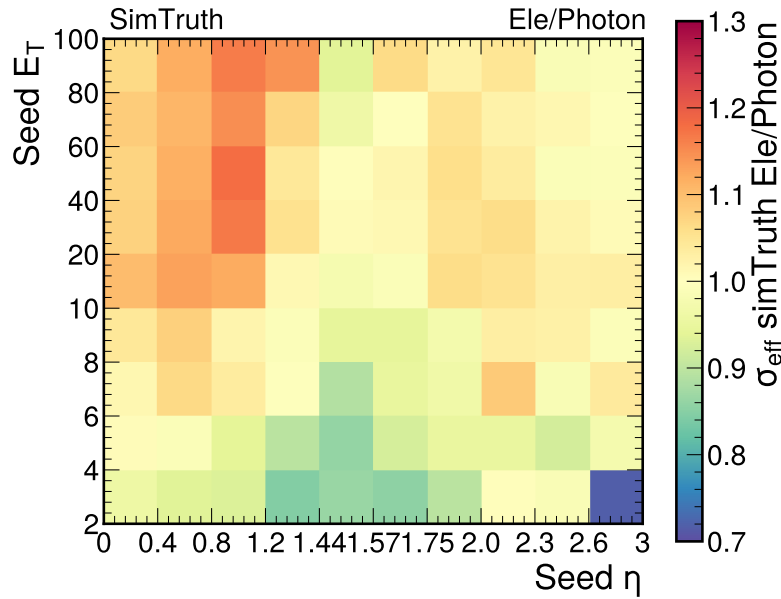


Figure 2.34: Comparison of ideal energy resolution for electrons and photons.

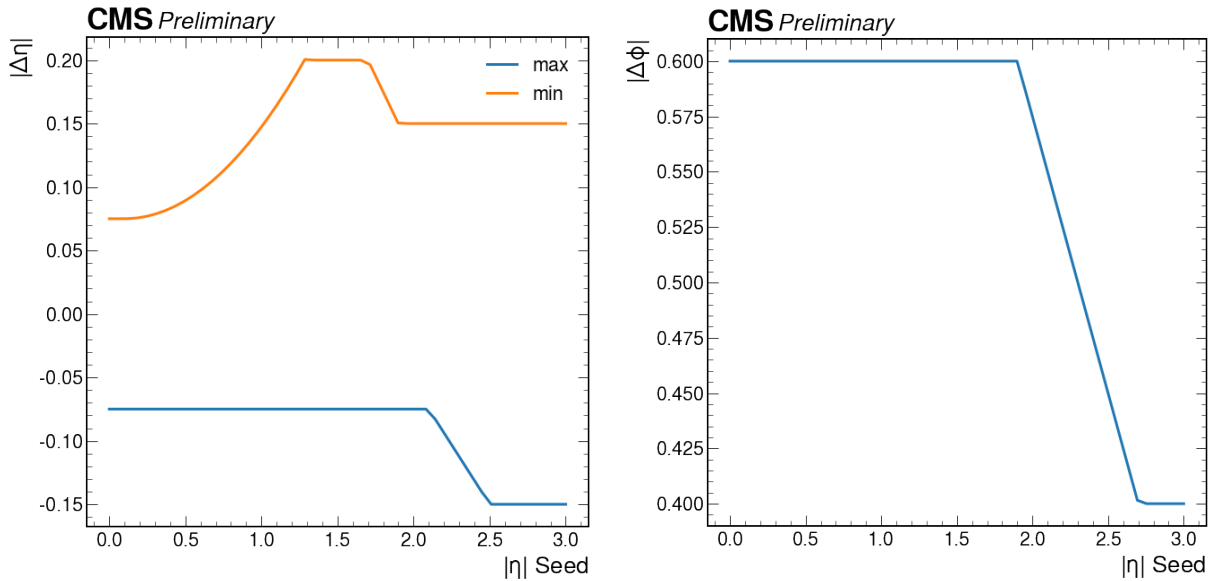


Figure 2.35: Parameters defining the detector window dimension used to select PFClusters around a seed:  $\Delta\eta^{\max}(\eta^{\text{seed}})$  and  $\Delta\eta^{\min}(\eta^{\text{seed}})$  on the left,  $\Delta\phi^{\max}(\eta^{\text{seed}})$  on the right.

- The list of PFClusters is ordered by  $E_T$ .
- Iterate on the PFClusters with  $E_T > 1$  GeV. For each cluster:
  - If the cluster is associated with a CaloParticle with at least 1% of score, then it is saved as *calo-matched*. Otherwise it is discarded.
  - Check if the cluster is already contained in the window defined by another seed. In that case exclude the PFCluster.



- If the cluster is not in another window, create a window around it.
- Save all the clusters inside the window of the current seed.
- **Truth association:** for each cluster in the window around the current seed:
  - If the cluster is not associated to the same CaloParticle of the seed, label it as *out-the-SC*.
  - If the simFraction of the cluster is larger than the simFraction threshold  $\zeta(E_T^{\text{seed}}, \eta^{\text{seed}})$  and  $En_{\text{sim}}^{\text{PU}} < En_{\text{sim}}^{\text{signal}}$ , then label it as *in-the-SC*.

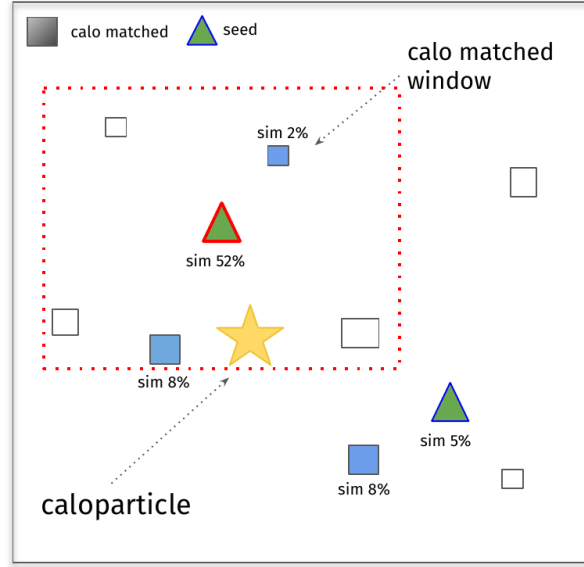


Figure 2.36: Schema reproducing the training dataset preparation: windows are created around energetic seeds and the truth matching with the CaloParticle is used to define the training labels.

### 2.3.2.1 Model inputs

The information available for each training sample are:

- **Cluster information:**
  - Energy and position  $(i\eta, i\phi, iz)$  of the cluster, number of crystals, seed flag
  - Information relative to the seed:  $\Delta\eta(\text{seed}), \Delta\phi(\text{seed}), \Delta E_T(\text{seed}), \Delta En(\text{seed})$
  - (optional) PFCluster shower shapes variables: swiss-cross,  $\eta$ -width,  $\phi$ -width,  $\sigma_{i\eta-i\phi}$
- **Rechits information:** for each cluster  $i$  the list of rechit information is used as input:

$$[(i\eta, i\phi, iz, En)_0, (i\eta, i\phi, iz, En)_1, \dots, (i\eta, i\phi, iz, En)_{N_{\text{rechits}}}]^i$$

- **Window information:** The minimum, maximum and average of  $(E_T, En, \Delta\eta(\text{seed}), \Delta\phi(\text{seed}), \Delta E_T(\text{seed}), \Delta En(\text{seed}))$  of the clusters in the window.

### 2.3.3 Graph neural network model

The ML model architecture chosen to tackle the SuperClustering problem involves Graph Neural Networks (GNN) [35–37] and Self-Attention (SA) layers [38, 39]. This class of models permits working on a set of related objects and can infer properties about the single elements,

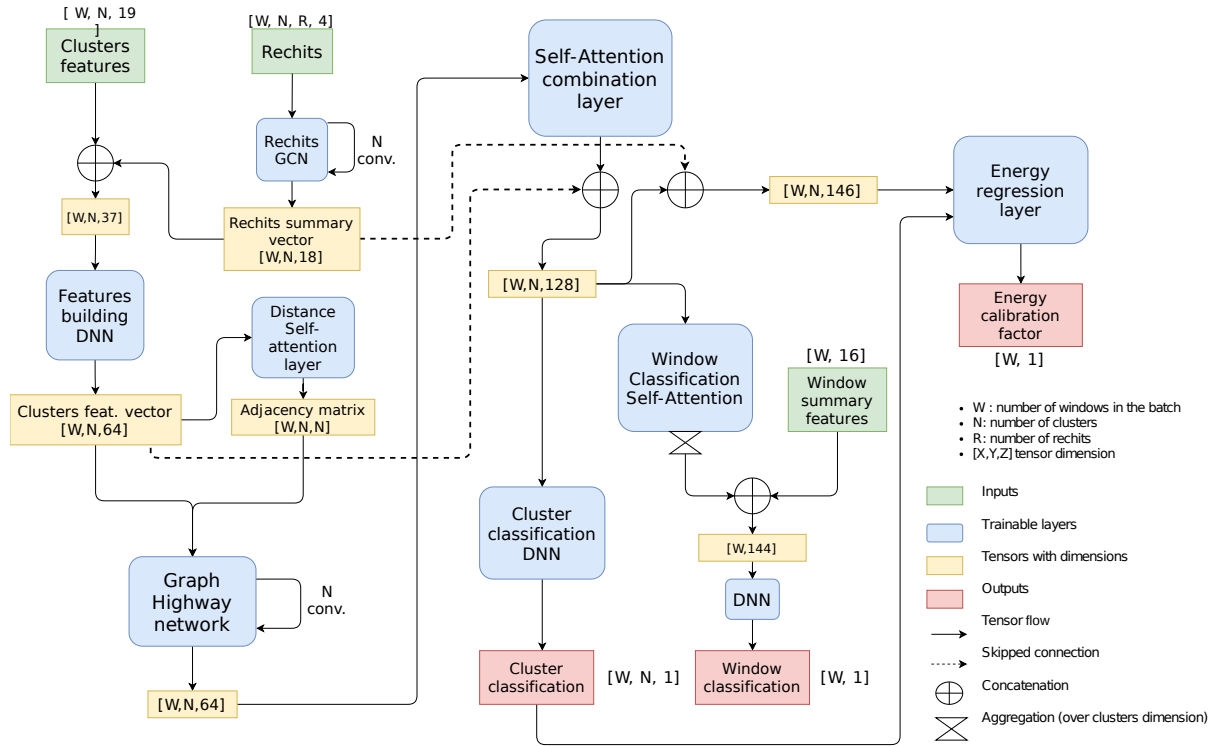


Figure 2.37: Schema of the DeepSC model: the models inputs, layers and tensors flowing through them are depicted in the schema.

or the overall set. One of the advantages over other architecture is that the number of objects in the set can be different for each sample. This type of architecture have been recently explored with discrete success for more complex problem as a ML approximation of the PF algorithm[40]. The *DeepSuperCluster* (DeepSC) model has been trained to perform three tasks:

**SuperClustering** Identify which PFCluster should be associated to the seed cluster to build the optimal SuperCluster. This is a cluster-by-cluster classification problem.

**Energy regression** Since all the information about the clusters and rechits is given to the model, the estimation of the energy correction needed to restore the generator-level electron energy can be included as one of the outputs of the same model. Currently, this regression is trained separately on the Mustache SuperClusters in order to have the best possible energy for the ECAL calibration.

**Window identification** Another piece of information that can be inferred by the model is the type of window created by the seed: distinguishing between an electron, photon, jets or noise candidate may be useful for the next steps in the Particle Flow (PF) reconstruction. At the moment this identification is part of the PF algorithm and exploits track information, but having it in the previous step of the chain may be useful for future improvements.

The GNN implemented in the current model is based on Graph Highway Networks (GHN) [41], a special type of graph convolutional networks [42]. The model has been fully implemented in

TensorFlow library version 2.3 [43]. Fig. 2.37 represents the layers in the model and the tensors passing through them. The ML model is conceptually organized in four steps: encoding, graph building, graph elaboration, decoding (for each output):

**Encoding** In the encoding step the input variables of each cluster and their rechits are manipulated to build-up information used directly by the next layers. A GHN layer is applied on the rechits of each cluster and then the resulting vectors are summed to yield a single vector for each cluster representing rehit information. Then a simple feature-extracting DNN is applied on each cluster input vector and rehit latent vector to get 64 features for each cluster. .

**Graph building** In order to build a graph the adjacency matrix  $A_{ij}$  defining the distance between each point (cluster) has to be defined. In this model the distance is dynamically constructed [35] assigning to each cluster a 3D-space coordinates vector with a SA layer and computing the Euclidean distance. This way the relative importance of the interaction between the cluster, i.e. their distance in the graph, is learned automatically during the training.

**Graph elaboration** Once each cluster has a features vector and is part of the adjacency matrix, a GHN layer is applied to perform graph convolutions and a SA layer is applied to extract another features vector for each cluster. The application of the graph network implies a passage of information between the clusters.

**Decoding** The final step is extracting the output information from the latent space. First, a simple DNN network is applied to each cluster features vector to get the classification output (in or out the SC). Then, all the clusters features vectors are combined with a different SA layer and then aggregated in a single vector to obtain the window classification. An energy regression layer can be applied in the same fashion.

### 2.3.4 Loss and training

The DeepSC model is optimized minimizing a loss defined with two components. A binary cross-entropy loss for clusters classification (in or out of the SC), and a Softmax multi-class classification loss for the window classification. The inclusion of an energy regression layer has been postponed as the next step development. The total dataset contains 2M electron-matched windows and 2M photon-matched ones: they have been split in 3 datasets for training, validation and testing in the proportion 50%,30%,20%. The training is performed with stochastic gradient descent and the Adam[44] optimizer, with a batch size of 512 windows. Dropout and weights regularization are implemented to avoid overtraining and help the model to converge faster. The model is evaluated on the validation dataset at the end of each iteration on the training dataset (epoch) to monitor the convergence of the optimization. The training is stopped when the validation loss does not increase for 10 epochs. The testing dataset is instead used for the selection of the best model between different trainings: the current best candidate is a rather model with only  $\sim 380k$  weights. The training time is around 10 hours on a NVidia Tesla GPU. Fig. 2.38 shows the trend of the losses along the training epochs, until convergence.

### 2.3.5 Performance

The performance of the new DeepSC algorithm is compared with the Mustache SuperClustering in terms of energy resolution and PU dependence [45]. The energy resolution will be analyzed in the next section for electrons and photons for different seed  $E_T$  and  $\eta$  value and for

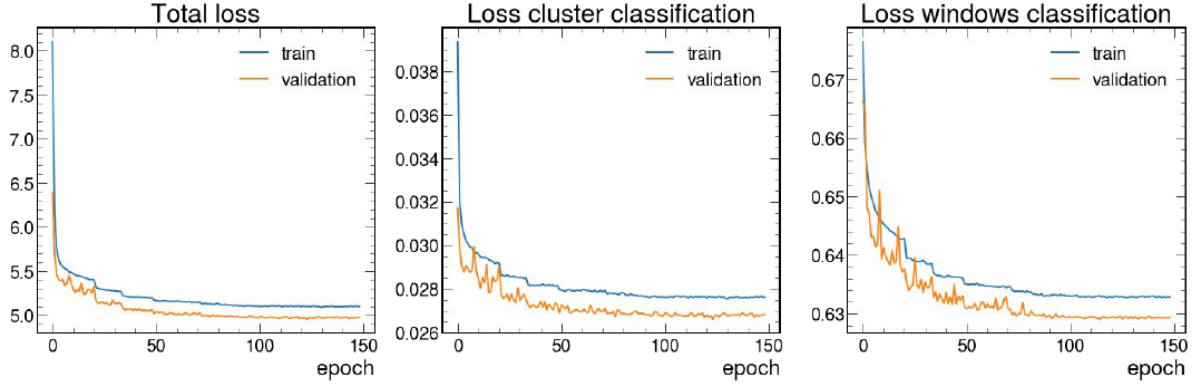


Figure 2.38: Training and validation losses along the training epochs.

different numbers of clusters in the window. The recall (or efficiency) and purity in the selection of the energy and number of clusters are compared between the two algorithms. Finally, the performance of the window classification output will be shown.

### 2.3.5.1 Energy resolution

The energy resolution is computed by analyzing the distribution of the energy scale ( $\sum_i^{N_{cls}} En_i \cdot SC(i) / simEn^{calo}$ , where  $SC(i)$  is 1 if the  $i$ -th cluster has been selected by the SuperClustering algorithm, 0 otherwise. The resolution is given by  $\sigma_{eff} = [\text{quantile}(0.84) - \text{quantile}(0.16)] / 2$ . The energy is always considered as the uncorrected one, no regression is applied either to the Mustache SC or to the DeepSC one.

Figure 2.39 shows the comparison of the energy resolution between each algorithm and the ideal performance given by the truth matching (as defined in Sec. 2.3.1.4). The DeepSC is clearly closer to the energy resolution in all regions of seed  $E_T$  and  $\eta$ . Moreover, Fig. 2.40 shows the energy scale of each algorithm compared to the ideal matching: the DeepSC on average collects the correct amount of energy, whereas the Mustache is clearly contaminated by PU and noise at low energy.

The energy resolution obtained with the two algorithms is also directly compared: in Fig. 2.41 the performance is compared in bins of GEN-level particle  $E_T^{Gen}$  and  $\eta$ , in Fig. 2.42 in bins of seed  $|\eta|$  and  $E_T^{Gen}$ , whereas in Fig. 2.43 in bins of  $E_T^{Gen}$  and total number of clusters in the window. The largest improvement is obtained at low energy in the barrel where the material budget in front of ECAL increases, and the contamination of the PU in the Mustache degrades more its resolution. Moreover, the DeepSC performs considerably better in windows with a lot of clusters as shown in Fig. 2.43.

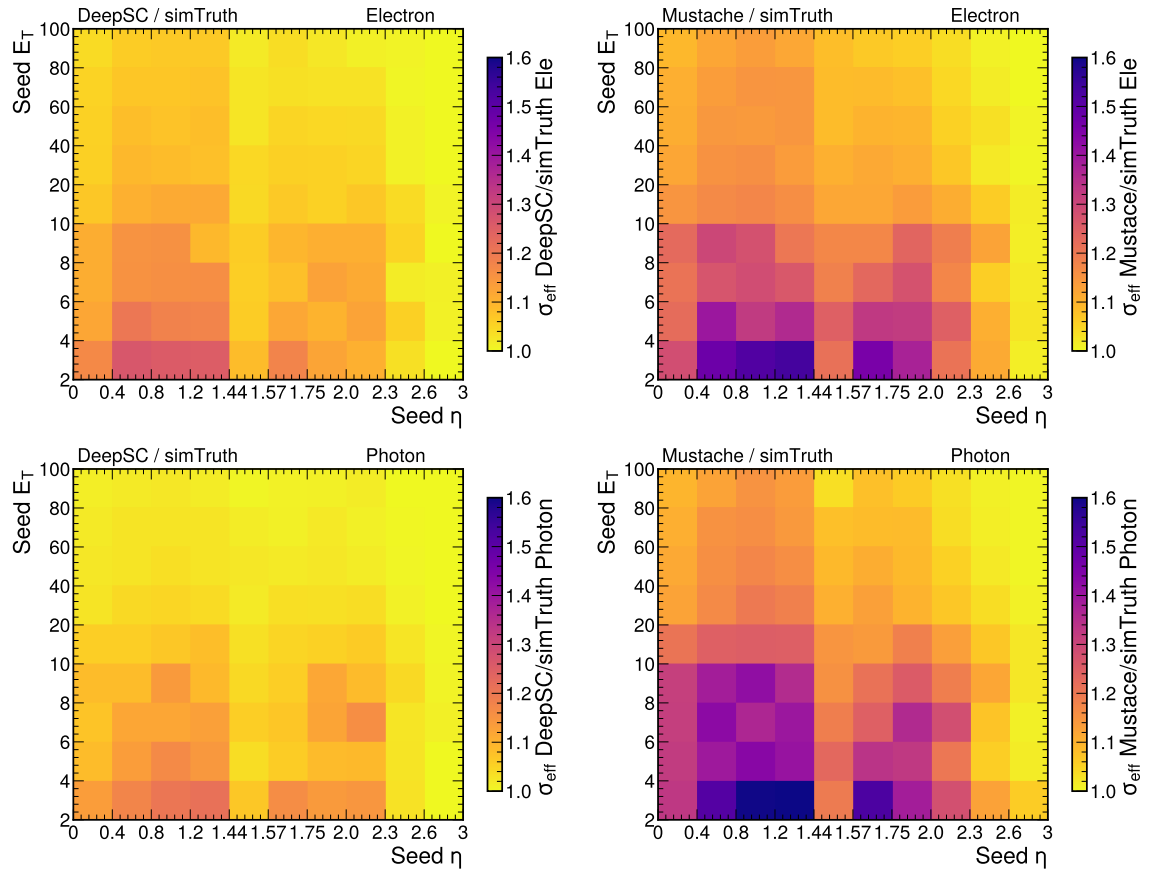


Figure 2.39: Comparison between the energy resolution of each algorithm and the ideal performance given by the truth matching. On the left the comparison with DeepSC, on the right with the Mustache; electrons on the first line and photons on the second.

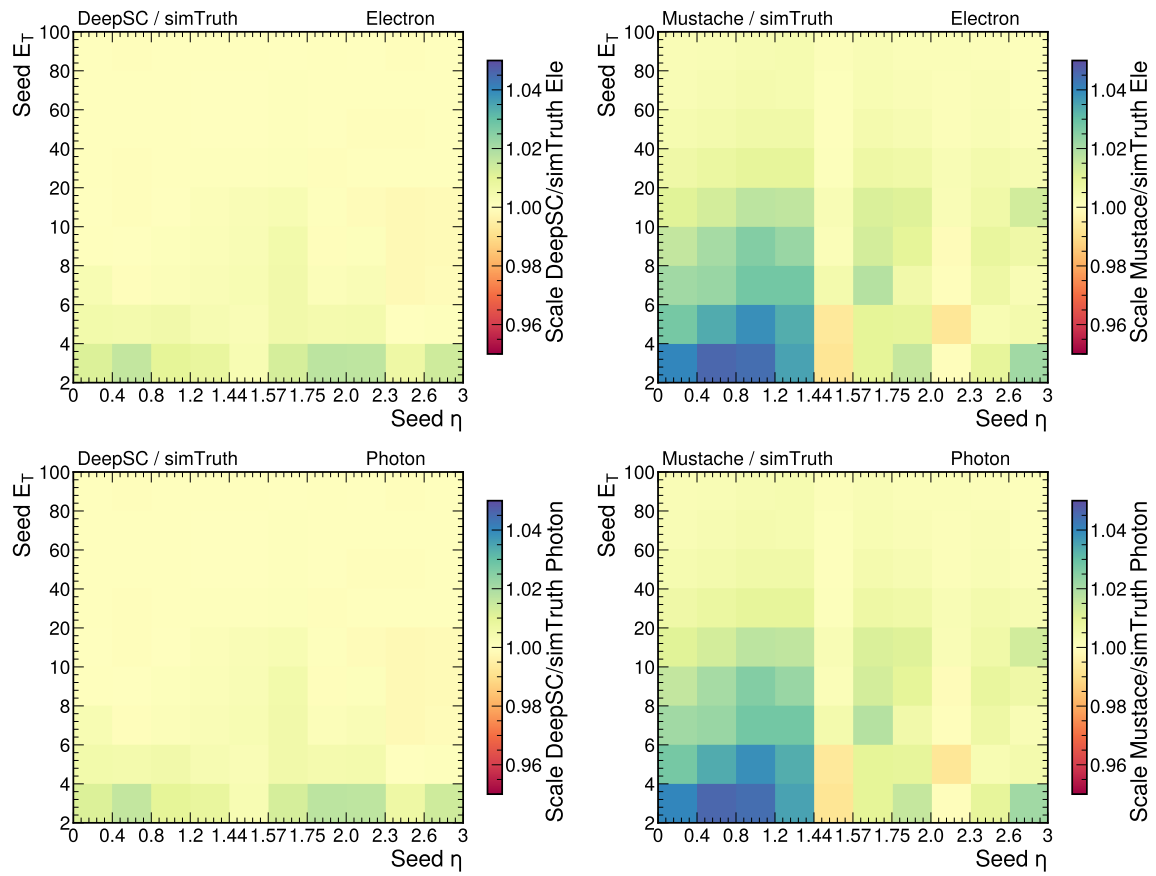


Figure 2.40: Comparison between the energy scale of each algorithm and the ideal performance given by the truth matching. On the left column the comparison with DeepSC, on the right column with the Mustache; electrons on the first line and photons on the second.

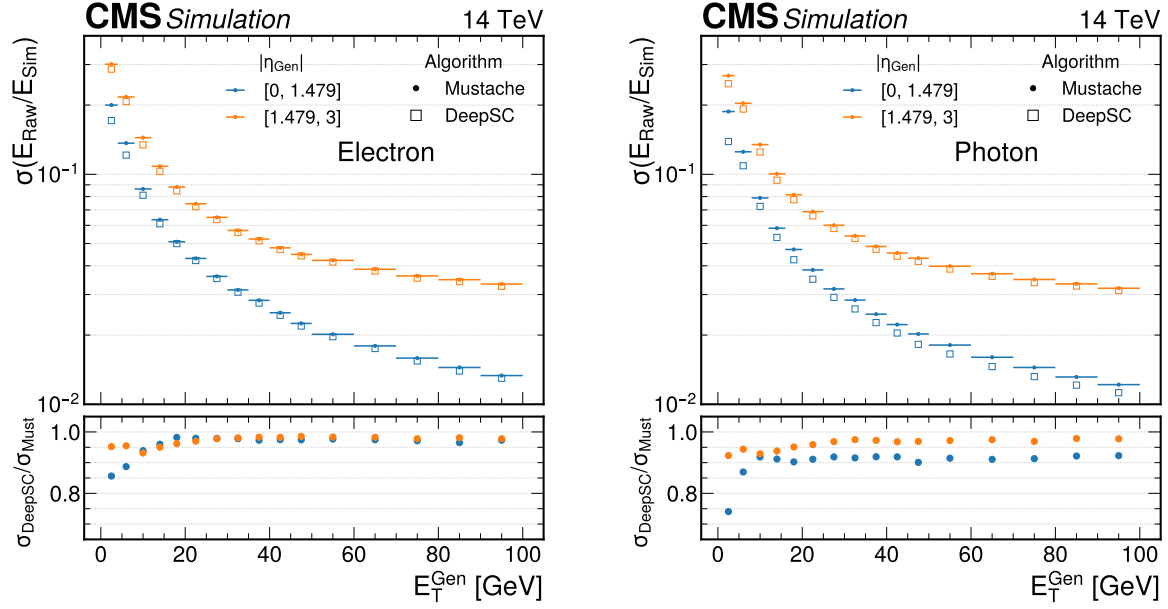


Figure 2.41: Direct comparison between the DeepSC and Mustache energy resolution in bins of  $E_T^{\text{Gen}}$  and  $\eta$ .

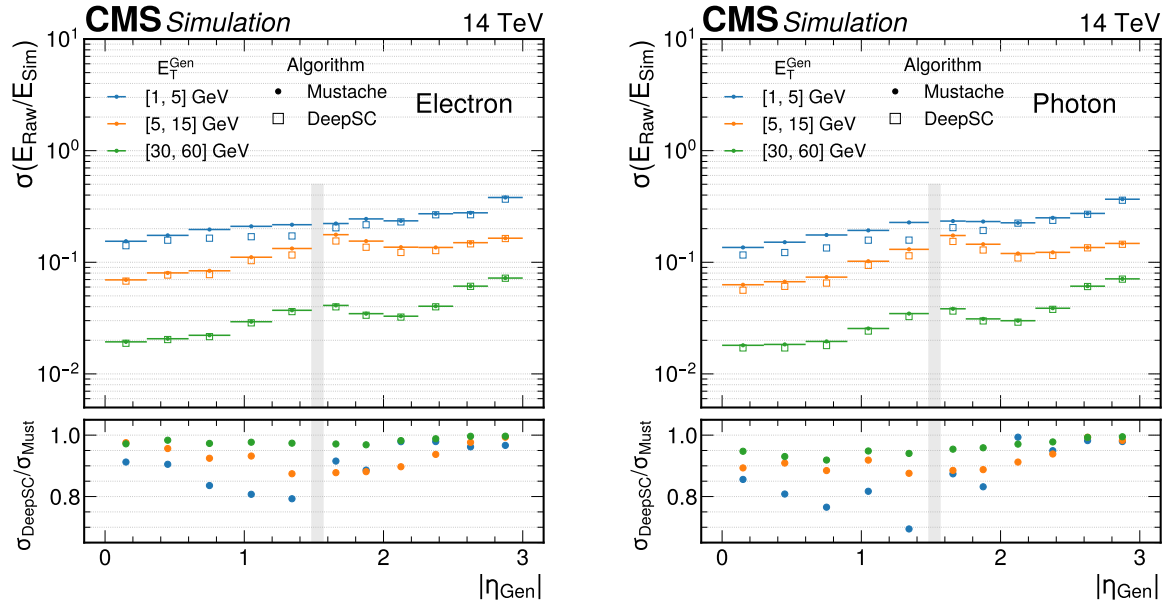


Figure 2.42: Direct comparison between the DeepSC and Mustache energy resolution in bins of seed  $|\eta|$  and  $E_T^{\text{Gen}}$ .



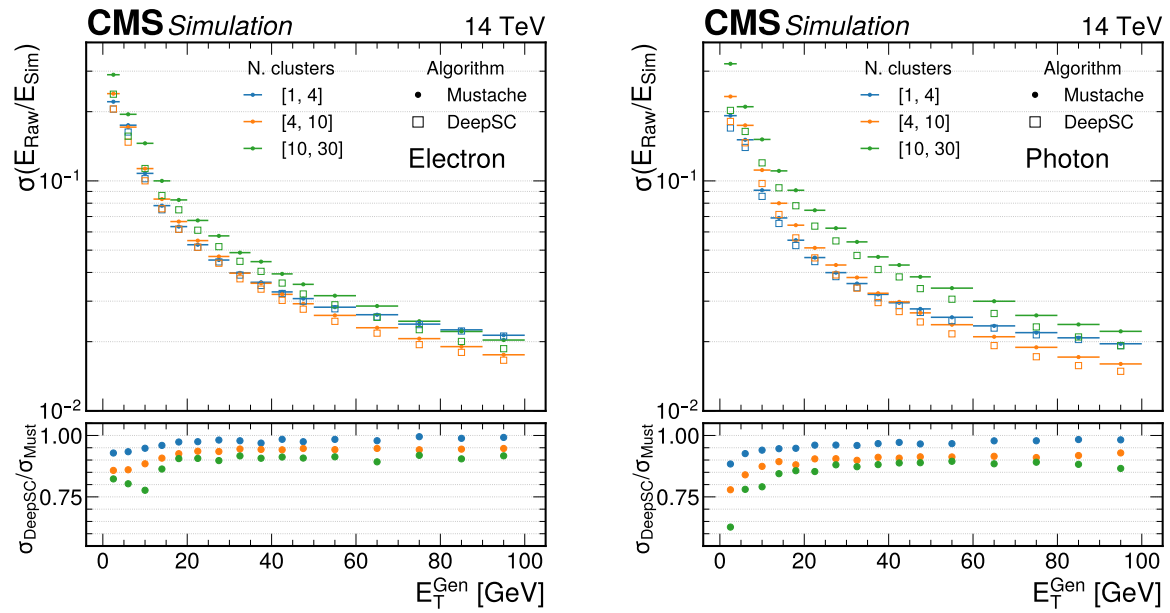


Figure 2.43: Direct comparison between the DeepSC and Mustache energy resolution in bins of  $E_T^{\text{Gen}}$  and total number of clusters in the window.

### 2.3.5.2 Difference between electrons and photons

The performance of SuperClustering for electrons and photons is intrinsically different: the fragmentation by photon conversion or bremsstrahlung emission gives different energy fractions to the seed and secondary clusters. The seed  $E_T$  and  $\eta$  distribution is shown in Fig 2.44: a uniform distribution for both electrons and photons is obtained along  $\eta$  from the simulation configuration. The seed energy distribution is instead different, but understood in the following terms. Fig. 2.45 shows the fraction of the total energy kept by the seed for electrons and photons, when a different number of secondary emission is considered. Electrons tend to have more secondary clusters, but with smaller energies; photons split less often but in larger energy chunks. The plot of the energy fraction of the most energetic secondary cluster is shown in Fig. 2.46 on the right and confirms this description.

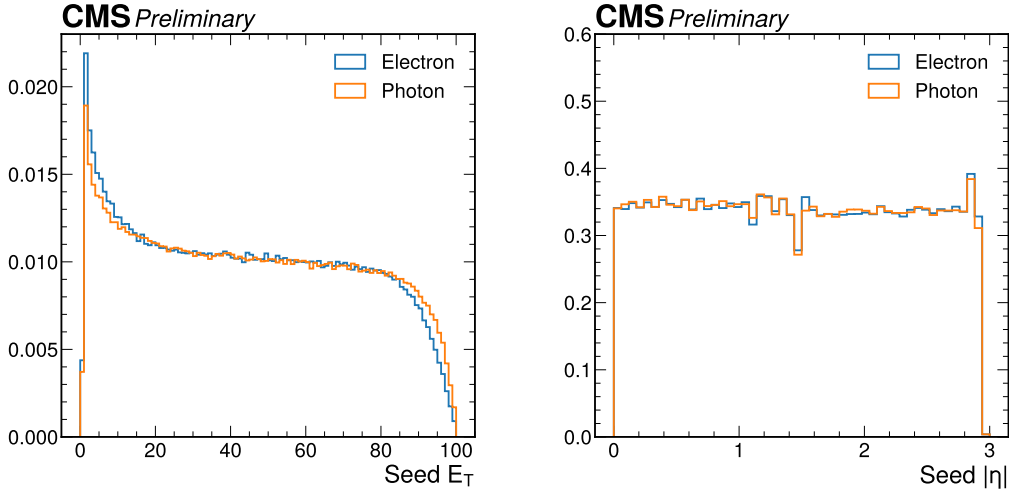


Figure 2.44: On the left: distribution of the seed  $E_T$  for electrons and photons. On the right: distribution of the seed  $|\eta|$ .

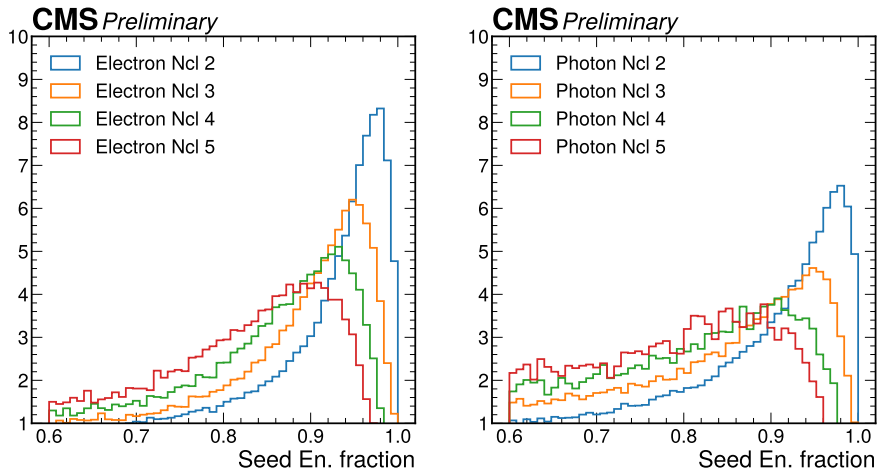


Figure 2.45: Distribution of the seed energy fraction for different number of secondary emissions for electrons (left) and photons (right)

The model is trained simultaneously on electrons and photons, balanced with weights applied to the loss to have the same distribution of the number of clusters and the seed  $E_T$ . Looking

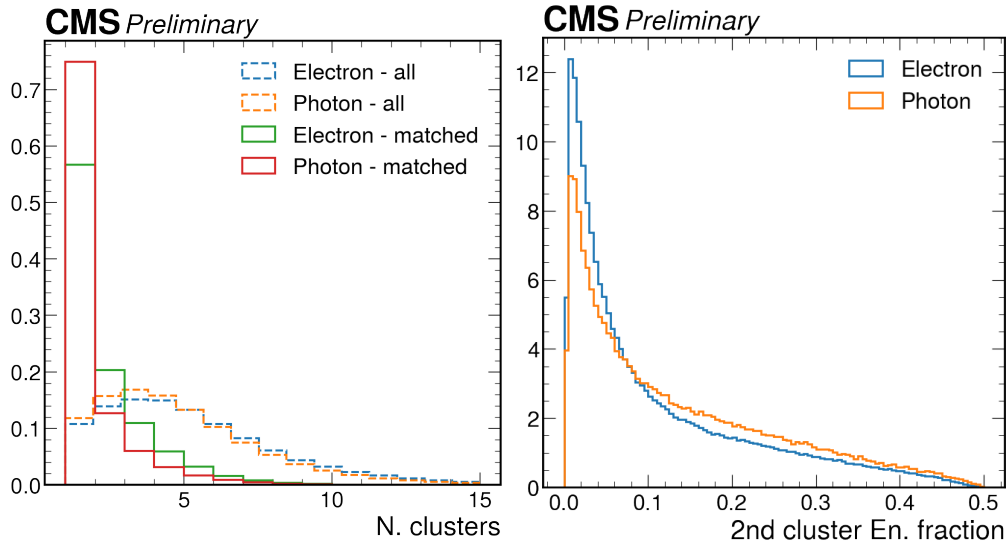


Figure 2.46: On the left: distribution of the total number of clusters and the truth-matched clusters in electron and photon windows. On the right: energy fraction of the most energetic secondary cluster for electrons and photons.

at the number of selected clusters in different energy bins, Fig. 2.47, the model is matching almost perfectly the cluster distribution for photons, while it tends to lose one cluster for the electrons. Since secondary emissions for electron have lower energy than the one from photon, the model has more difficulty in distinguishing them from PU and noise and tends to discard them. The different performance of DeepSC between electrons and photons is fully understood in these terms and a margin of improvement is still possible by tweaking the training to give more importance to electron secondary emissions.

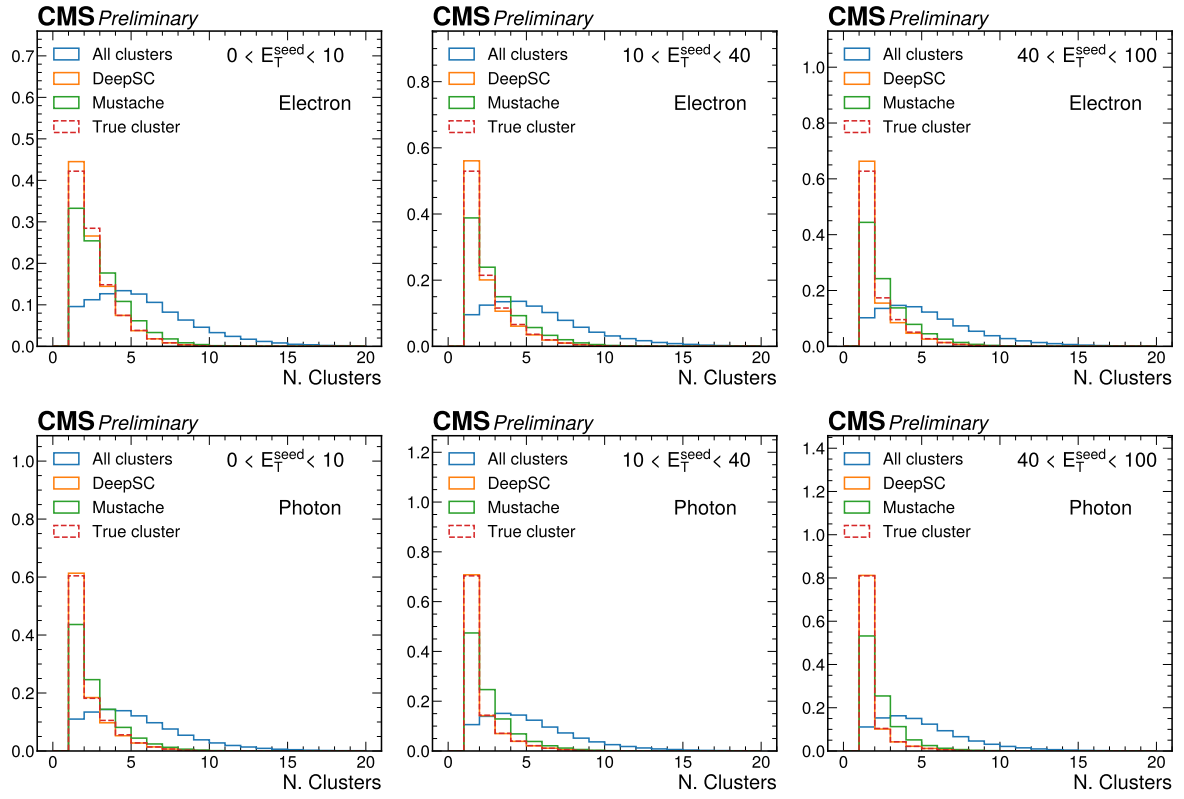


Figure 2.47: Distribution of selected number of clusters by DeepSC and Mustache compared with the truth-matched ones in different seed  $E_T$  bins. Electrons on the first line, photons on the second.

### 2.3.5.3 Recall and purity of cluster selection

For classification tasks, the terms true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) compare the results of the classifier under test with truth label. The terms positive and negative refer to the classifier's prediction, and the terms true and false refer to whether that prediction corresponds to the truth label. The recall, or efficiency, is defined as the true positives over the total true  $TP/(TP + FN)$ : translated in the SuperClustering words, it is the average ratio between the correctly selected energy (or clusters) over the truth-matched energy (or clusters). The precision, or purity, instead is defined the true positives over the total positives  $TP/(TP + FP)$ , corresponding to the average ratio between the correctly selected energy (or clusters) over all the selected energy (or clusters). Recall and precision gives a clear picture of the performance of a classifier.

The DeepSC and Mustache algorithm are compared in terms of recall and precision in the selection of energy and number of clusters in bins of seed  $E_T$  and  $\eta$ . Only the plots for the electrons are shown for brevity: Fig. 2.48 displays the energy recall and precision, whereas Fig. 2.49 the performance for the selection of the clusters. In terms of energy collection, the DeepSC has a comparable, or slightly better in some regions, energy recall, but it improves the energy purity by removing PU and noise clusters, as expected. Looking at the number of clusters plots instead, it is clear that the DeepSC tends to select less clusters than the mustache, loosing also some truth-matched ones, but it has a much better cluster selection purity. The DeepSC selects less clusters, but in a much more precise way.

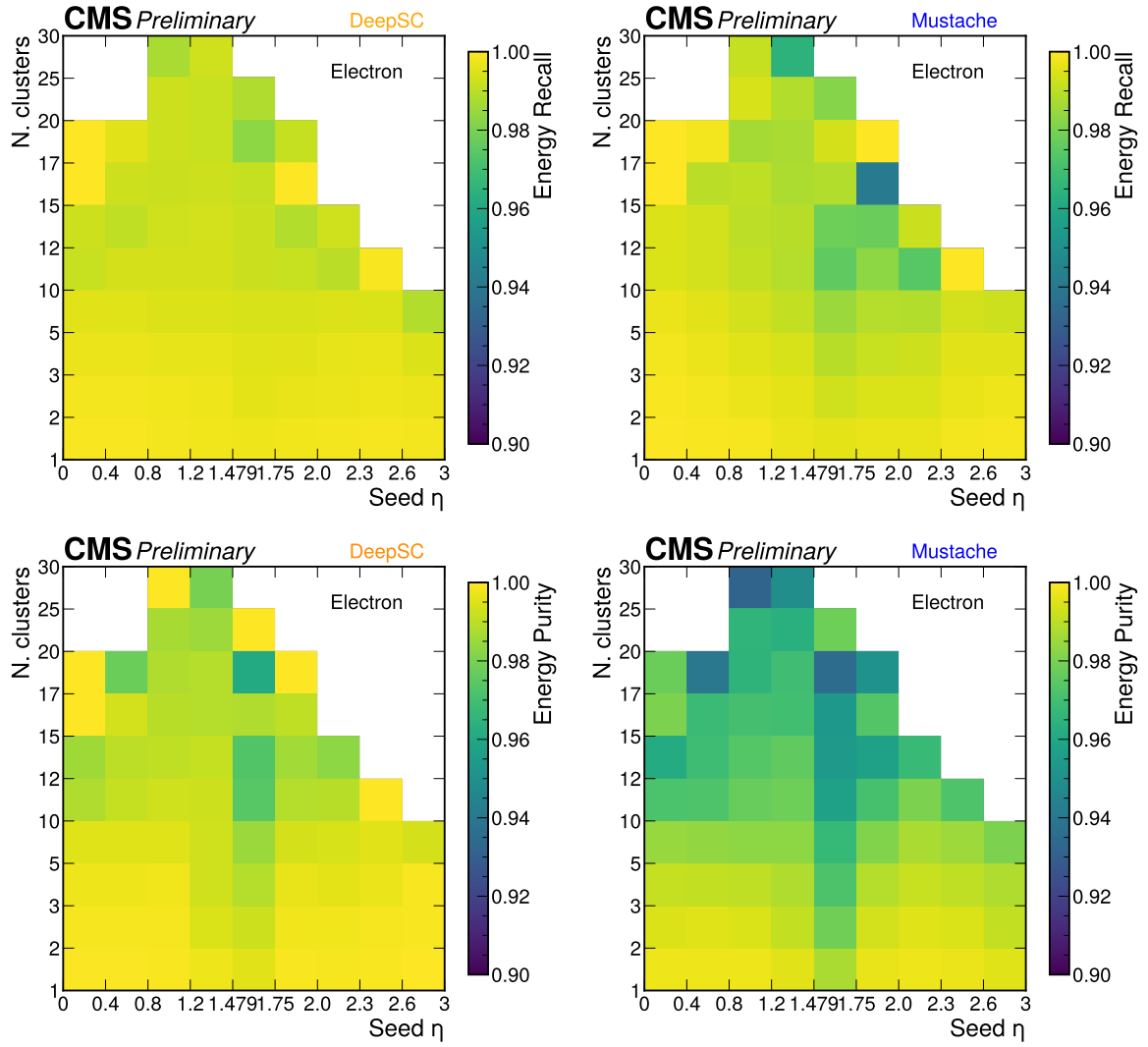


Figure 2.48: Energy recall (top)  $TP/(TP + FN)$  and precision (bottom)  $TP/(TP + FP)$  for DeepSC (left) and Mustache (right) for electrons.

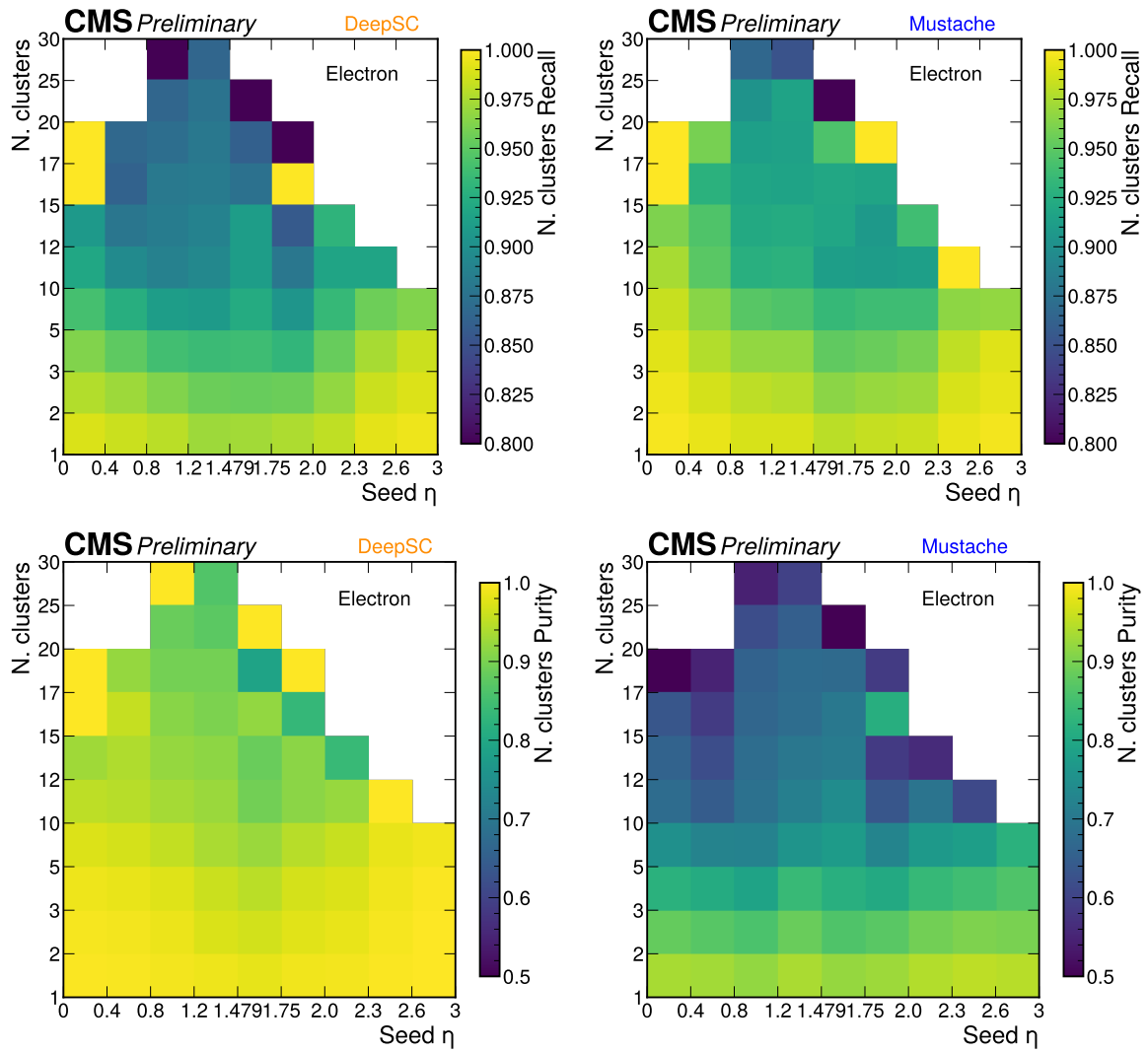


Figure 2.49: Number of cluster recall (top)  $TP/(TP + FN)$  and precision (bottom)  $TP/(TP + FP)$  for DeepSC (left) and Mustache (right) for electrons.

### 2.3.5.4 Dependence on number of vertices

Studying the dependence of the SuperClustering algorithm on different PU conditions is crucial to check the stability of its behavior during data taking. The DeepSC and Mustache energy resolution versus the number of simulated PU vertices for different  $E_T^{\text{Gen}}$  is analyzed in Fig. 2.50: for SuperCluster with  $E_T < 15$  GeV an higher PU is correlated with a loss of resolution. The DeepSC, being more robust again PU contamination specially at low energy, improves the energy resolution trends at high PU.

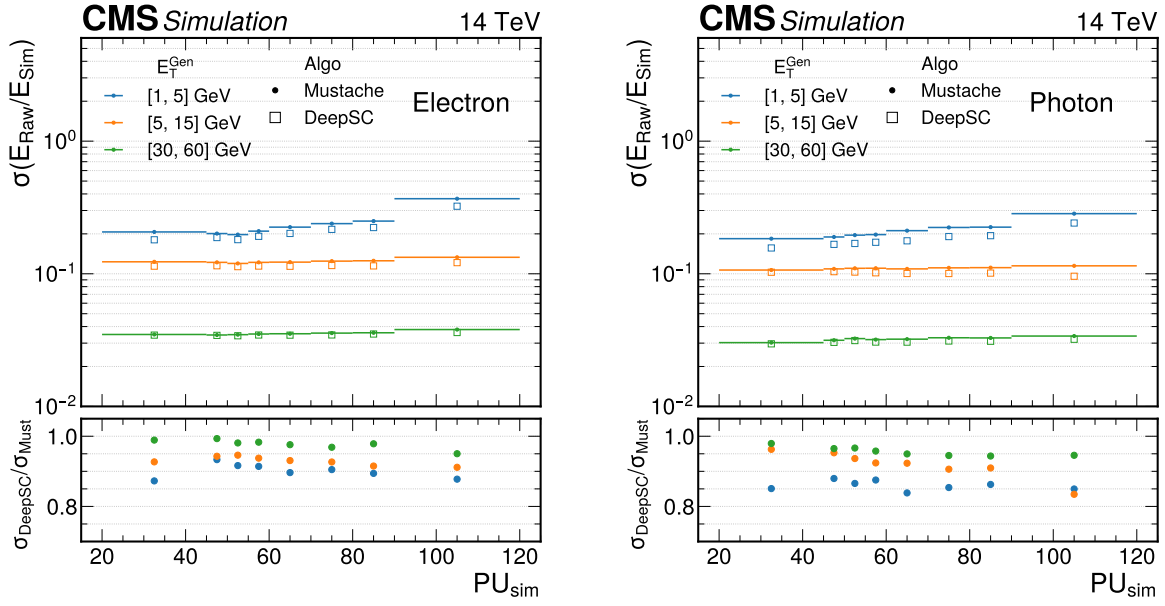


Figure 2.50: Energy resolution along the number of vertices. Top panel: DeepSC resolution for electrons (left) and photons (right). Bottom panel: comparison of DeepSC and Mustache energy resolution along the number of simulated PU vertices.

### 2.3.5.5 Window classification

The discrimination between electrons and photons is driven by the presence of tracks pointing to the ECAL SuperCluster, therefore it is actually performed during the Particle Flow reconstruction step. The DeepSC model, being a graph model applied on a portion of the detector, has enough information to attempt to discriminate electron and photons (and jets) using only ECAL local information. In fact, as shown in Sec. 2.3.5.2, SuperClusters generated by electrons and photons are intrinsically different in the distribution of energy between the secondary clusters and in the number of them. A 2D re-weighting of the loss function has been performed in order to make the training of the DeepSC unbiased with the respect of the seed energy and total number of clusters in the window.

The performance of the window classification, trained only for electron vs photon classes, is shown in Fig. 2.51 in terms of a ROC curve. The DeepSC reaches a discrete discriminator power at all energies. Looking at the ROC curve for different number of clusters in the window, it can be noticed that even with only the seed, the rechits information and shower shape variables have some discrimination power. With more than one cluster the relation between their energy is exploited by the graph model to identify the window.



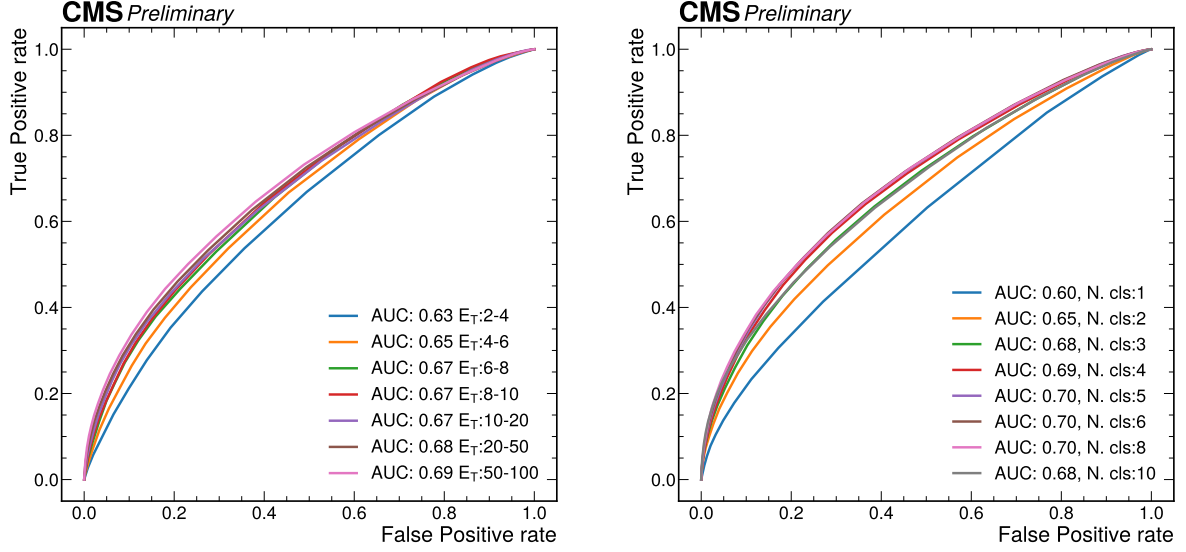


Figure 2.51: ROC curvee for window classes discriminator (electron vs photon) in differen bins of seed  $E_T$  (left) and number of clusters (right).

### 2.3.5.6 SuperCluster shape

The shape of the SuperCluster obtained with the Mustache and DeepSC algorithms is shown in Fig. 2.52 and Fig. 2.53. The image is built accumulating many electron and photon events and plotting for each cluster the  $\Delta\eta$  and  $\Delta\phi$  from the seed. The top row plots show the clusters matched to the electron or photon gen-particle: on the left the clusters selected by the algorithm, and on the right the not selected ones (loss of signal efficiency). The bottom row plots display clusters associated to PU or noise: on the left the selected clusters (contamination); on the right the correctly excluded ones. The seed position in the plot is  $|\eta| = [0.8, 1.4]$  and its  $E_T = [2, 30]$  GeV. The Mustache selects all the clusters in a parabolic region around the seed, both from signal and PU/noise; whereas the DeepSC has a higher purity and it is able to remove also spurious clusters close to the seed.

## 2.3.6 Conclusions

The DeepSC model is a promising ML based alternative to the Mustache algorithm [45]. It is robust again PU and it has been shown to outperform the Mustache in terms of energy purity and precision in the selection of the clusters. Until now the performance has been compared only at the uncalibrated energy level: further studies are ongoing to include the foreseen energy regression layer directly in the DeepSC model. Another planned improvement is the inclusion in the training of windows generated by jets interacting with ECAL: a third class can be added to the window classification to help the next reconstruction steps to discriminate between electrons photons and jets. Finally, an hyperparameter optimization has been planned to reach the ultimate performance of this model.

In terms of integration of the DeepSC model in the CMS reconstruction chain, software development is going in order to be able to evaluate the model on events from CMS real data and compare the performance with the MC simulation.

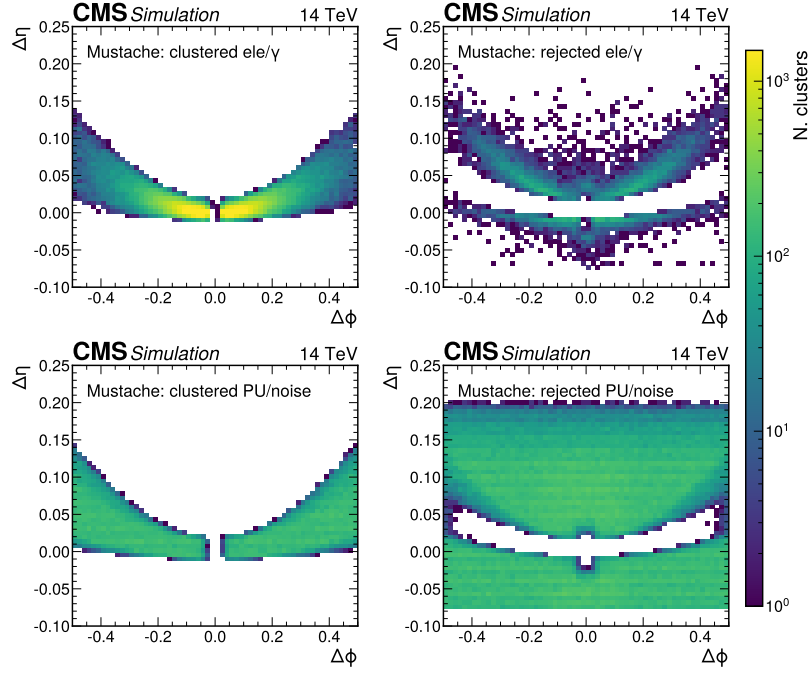


Figure 2.52: Shape of the SuperCluster obtained with the Mustache algorithm by accumulating many electron and photon events and plotting for each cluster the  $\Delta\eta$  and  $\Delta\phi$  from the seed.

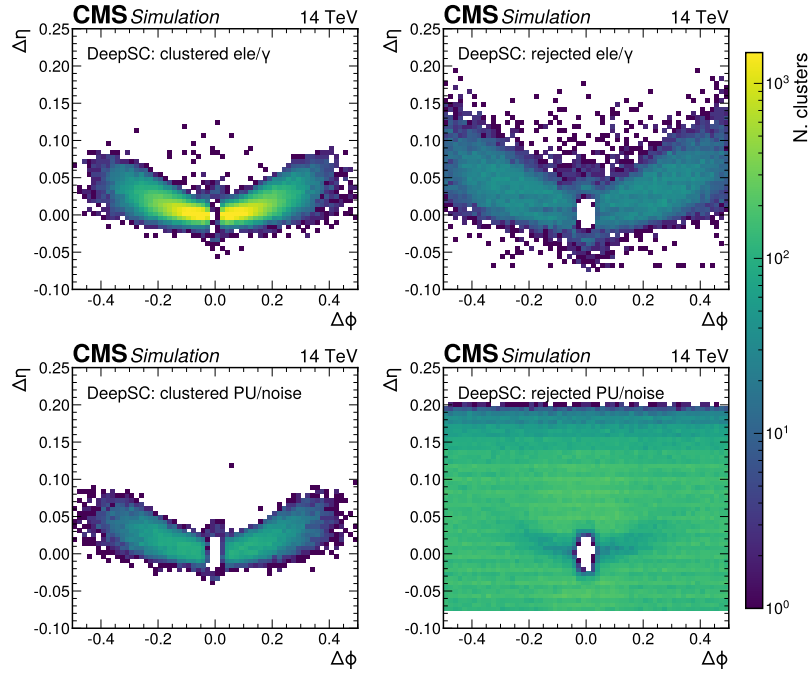


Figure 2.53: Shape of the SuperCluster obtained with the DeepSC algorithm by accumulating many electron and photon events and plotting for each cluster the  $\Delta\eta$  and  $\Delta\phi$  from the seed.



## Chapter 3

# Search for vector boson scattering in semileptonic decays with $WV \rightarrow l\nu qqjj$ final states

### 3.1 Theoretical introduction

#### 3.1.1 The Standard Model of particle physics

In this section the basic concepts of the model describing the elementary particles and their interactions, the standard model (SM), are introduced. The electroweak spontaneous symmetry breaking and the Higgs mechanism are then described as a way to explain the origin of particles masses. In order to describe the observed phenomena, four forces are believed to be fundamental: the electromagnetic, strong, weak and gravitational force. The SM explains how the particles interact, through the fundamental forces. Unfortunately, until now we are not able to include in the SM the gravitational force.

The background of the SM is relativistic quantum field theory. Basing on their spin, all particles are split into two groups: fermions and bosons. Fermions have half-integer spin, follow Fermi-Dirac statistics and are the constituents of matter. On the other hand, bosons have integer spin, follow Bose-Einstein statistics and can be seen as the intermediates of the interaction forces. Tables 3.1 and 3.2 show the fermions and bosons described by the SM and their main properties.

Table 3.1: Spin-1/2 fermions [46] masses and charges

|                | 1 <sup>st</sup> family             | 2 <sup>nd</sup> family                                        | 3 <sup>rd</sup> family                                            | Q           | Colour Charge      |
|----------------|------------------------------------|---------------------------------------------------------------|-------------------------------------------------------------------|-------------|--------------------|
| <b>Leptons</b> | $\nu_e (\sim 0)$<br>e (511 keV)    | $\nu_\mu (\sim 0)$<br>$\mu$ (105.7 MeV)                       | $\nu_\tau (\sim 0)$<br>$\tau$ (1.777 GeV)                         | 0<br>-1     | 0<br>0             |
| <b>Quarks</b>  | u (1.7-3.1 MeV)<br>d (4.1-5.7 MeV) | c ( $1.29^{+0.05}_{-0.11}$ GeV)<br>s ( $100^{+30}_{-20}$ MeV) | t ( $172.9^{+1.1}_{-1.1}$ GeV)<br>b ( $4.19^{+0.18}_{-0.01}$ GeV) | 1/3<br>-2/3 | r, b, g<br>r, b, g |

Table 3.2: Spin-1 bosons [46] masses and charges

|                                     | Mass (GeV)         | Q       | Colour Charge |
|-------------------------------------|--------------------|---------|---------------|
| <b>Photon (<math>\gamma</math>)</b> | 0                  | 0       | 0             |
| <b>Gluon (g)</b>                    | 0                  | 0       | r, b, g       |
| <b>W</b>                            | $80.385 \pm 0.015$ | $\pm 1$ | 0             |
| <b>Z<sup>0</sup></b>                | $91.188 \pm 0.002$ | 0       | 0             |

Matter is composed of two types of spin-1/2 constituents: leptons and quarks. Besides the masses and the electromagnetic charges, quarks and leptons differ for the couplings, i.e. quarks have additional charges, the colors:  $r$ ,  $b$ ,  $g$ ,  $\bar{r}$ ,  $\bar{b}$  and  $\bar{g}$ . Because of the colour charges, quarks cannot be detected alone as affected by the asymptotic freedom, i.e. the coupling is asymptotically weaker as energy increases and distance decreases. Hence, only their bound states can be detected: mesons ( $q\bar{q}$ ) and baryons ( $qqq$ ).

A quantum field is associated to each particle, based on the group of symmetries  $SU_C(3) \times SU_L(2) \times U_Y(1)$ . The three local symmetry groups dictate the three interactions between the particles.  $SU_C(3)$  is related to the colour charges, while  $SU_L(2)$  is related to the chirality. Each quark can be seen as a triplet under  $SU_C(3)$  symmetry due to the colour charges, whereas each lepton is a colourless singlet. The chirality is associated to the properties of the bispinors, i.e. representation of the fermions as solution of the Dirac equation:

$$\mathcal{L} = \bar{\psi}(i\partial - m)\psi \longrightarrow (-i\partial + m)\psi = 0 \quad (3.1)$$

Indeed, a bispinor can be separated as:  $\psi = \psi_R + \psi_L$ , where  $\psi_R$  (right component) and  $\psi_L$  (left component) are the eigenstate of chirality,  $\gamma_5$ . In the Dirac representation:

$$\begin{aligned} \psi_L &= \frac{1 - \gamma_5}{2} \psi \\ \psi_R &= \frac{1 + \gamma_5}{2} \psi \end{aligned}$$

Right bispinors and left bispinors interact in different ways. In fact, given the  $SU(2)$  symmetry it is possible to identify the pairs of the left-handed fermions as isospin doublets, while right-handed fermions are isospin singlets. The last symmetry,  $U_Y(1)$  is related to the weak hypercharge. The isospin and the hypercharge are quantum numbers associated with the electroweak interactions between the fermions and the bosons. Three bosons are associated with the electroweak force:  $W^\pm$  (charged),  $Z^0$  and photon ( $A^\mu$  field).

The coupling between the fermions and the charged bosons (charged currents) is associated to the vectorial components of the isospin. On the other hand, the neutral currents are generated by the interactions between the fermions and a mixing of photon and Z fields [47], where the couplings are given by the third component of the isospin and the hypercharge. The mixing between photons and Z bosons is performed by means of the Weinberg angle ( $\sin^2 \theta_W^{M_Z} = 0.23102 \pm 0.00005$  [46]). The relation between the electric charge  $Q$ , the third component of the isospin ( $T_3$ ) and the hypercharge ( $Y$ ) is:

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

In order to find out the Lagrangian terms that describe the currents, the gauge invariance principle is used. Consider a Dirac field  $\psi$  and the local gauge transformation:

$$\psi \rightarrow e^{ig\alpha(x)}\psi$$

The Lagrangian of Equation 3.1 is invariant under this gauge transformation only with the introduction of a gauge boson field ( $A^\mu$ ) which allows the following gauge transformation:

$$\begin{aligned} A_\mu &\rightarrow A_\mu - \partial_\mu \alpha(x) \\ \partial_\mu \psi &\rightarrow \partial_\mu \psi + ig A_\mu \psi \end{aligned}$$

Hence, using a representation of the  $SU_L(2)$  symmetry and the gauge invariance, after the definition of:

$$L^\ell \equiv \begin{pmatrix} \psi^\nu \\ \psi^\ell \end{pmatrix}_L, \quad R^\ell \equiv \begin{pmatrix} \psi^\nu \\ \psi^\ell \end{pmatrix}_R$$

$$\tau^\pm = \frac{\sigma_1 \pm i\sigma_2}{2}$$

where  $\sigma_i$  are the Pauli matrices, generators of  $SU(2)$  symmetry, the Lagrangian terms for the charged currents becomes:

$$\mathcal{L}_{CC} = \frac{g}{\sqrt{2}} \left[ \bar{L}^\ell W^+ \tau^+ L^\ell + \bar{L}^\ell W^- \tau^- L^\ell \right]$$

Concerning the terms for the neutral currents, after the definition of:

$$\Psi^\ell = \begin{pmatrix} \psi_L^\nu \\ \psi_L^\ell \\ \psi_R^\nu \\ \psi_R^\ell \end{pmatrix}, \quad T_3 = \begin{pmatrix} 1/2 & & & \\ & -1/2 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} Y_{\nu_L} & & & \\ & Y_{\ell_L} & & \\ & & Y_{\nu_R} & \\ & & & Y_{\ell_R} \end{pmatrix}$$

they can be decoupled between electromagnetic and weak interaction:

$$\mathcal{L}_A = e Q \bar{\Psi}^\ell \not{A} \Psi^\ell$$

$$\mathcal{L}_Z = \frac{e}{2 \sin \theta_W \cos \theta_W} \bar{\Psi}^\ell \not{Z} (g_V - g_A \gamma^5) \Psi^\ell$$

where:

$$Q \equiv T_3 + \frac{Y}{2}$$

$$g_V \equiv T_3 - 2Q \sin^2 \theta_W$$

$$g_A \equiv T_3$$

It is possible to generalize to all the leptons and quarks families, using the values of isospin, hypercharge and electromagnetic charges of Table 3.3. Since  $\nu_R$  has all the quantum numbers equal to zero it doesn't interact.

Table 3.3: Left: leptons quantum numbers; Right: quarks quantum numbers

|               | T             | T <sub>3</sub> | Y  | Q  |              | T             | T <sub>3</sub> | Y              | Q              |
|---------------|---------------|----------------|----|----|--------------|---------------|----------------|----------------|----------------|
| $\nu_L^\ell$  | $\frac{1}{2}$ | $\frac{1}{2}$  | -1 | 0  | $q_L^{up}$   | $\frac{1}{2}$ | $\frac{1}{2}$  | $\frac{1}{3}$  | $\frac{2}{3}$  |
| $\ell_L^\ell$ | $\frac{1}{2}$ | $-\frac{1}{2}$ | -1 | -1 | $q_L^{down}$ | $\frac{1}{2}$ | $-\frac{1}{2}$ | $\frac{1}{3}$  | $-\frac{1}{3}$ |
| $\nu_R^\ell$  | 0             | 0              | 0  | 0  | $q_R^{up}$   | 0             | 0              | $\frac{2}{3}$  | $\frac{2}{3}$  |
| $\ell_R^\ell$ | 0             | 0              | -2 | -1 | $q_R^{down}$ | 0             | 0              | $-\frac{2}{3}$ | $-\frac{1}{3}$ |

The mass terms:

$$m_W^2 W^{+\mu} W_\mu^- + \frac{1}{2} m_Z^2 Z^\mu Z_\mu + \sum_f m_f (\bar{\psi}_R \psi_L + \bar{\psi}_L \psi_R)$$

cannot be put directly in the Lagrangian because they are not gauge invariant. Maintaining the gauge invariance, a way to give mass to these particles is the Higgs mechanism [48], which introduces the last particle in this theory: the Higgs boson, which is a massive, chargeless, spin-0 particle. The Higgs boson is an elementary particle which was proposed at the beginning of the 1960s in order to explain in a very simple way how all the massive particles gain the mass.

### 3.1.2 The Higgs mechanism

Consider a complex scalar field  $\phi$ , a vectorial field  $A_\mu$  and the Lagrangian:

$$\mathcal{L} = D^\mu \phi^* D_\mu \phi - \mu^2 |\phi|^2 - \lambda |\phi|^4 - \frac{1}{4} F^{\mu\nu} F_{\mu\nu} \quad (3.2)$$

with  $D_\mu \phi = [\partial_\mu + iqA_\mu]\phi$  and  $V(x) \equiv \mu^2 |\phi(x)|^2 + \lambda |\phi(x)|^4$ .

Because  $\phi$  is a complex field it can be decomposed in two fields  $\phi_1$  and  $\phi_2$ . Indeed, the potential  $V(x)$  has very different physical meanings depending on  $\mu^2$  and  $\lambda$  signs. In order to have a lower limit the potential must have  $\lambda$  positive. If  $\mu^2 > 0$  the potential has only one minimum in zero, it becomes the Klein-Gordon potential with an extra auto-interaction term and  $\mu^2$  can be interpreted as the square mass of the particle. On the other hand, if  $\mu^2 < 0$ ,  $V(x)$  has a maximum of zero and minima at  $\phi = \sqrt{\frac{-\mu^2}{2\lambda}} e^{i\theta}$ ; thus, it is not possible to use the perturbation theory in  $\phi = 0$ . Figure 3.1 shows the potential for  $\mu^2 < 0$ .

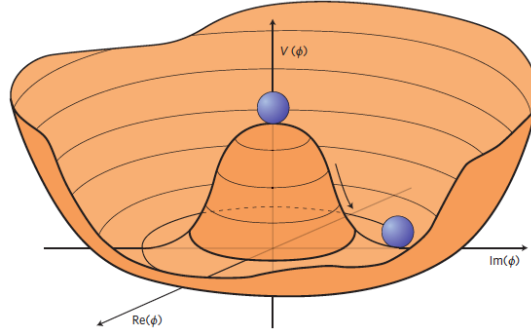


Figure 3.1: Higgs boson potential if  $\mu^2 < 0$

It is possible to expand the field on the minima-circle, choosing one specific minimum and breaking the  $\theta$ -symmetry of the potential (Spontaneous Symmetry Breaking). Fixing one minimum,  $v$ , expanding and inserting this expansion in the Lagrangian 3.2, it becomes:

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2}} [v + h(x) + i\eta(x)] \\ \mathcal{L} &= \frac{1}{2} \partial^\mu h \partial_\mu h(x) + \frac{1}{2} \partial^\mu \eta(x) \partial_\mu \eta(x) - \frac{1}{4} F^{\mu\nu} F_{\mu\nu} + \frac{1}{2} (qv)^2 A_\mu(x) A^\mu(x) \\ &\quad - \frac{1}{2} (2\lambda v^2) h^2(x) + vq A^\mu(x) \partial_\mu \eta(x) + \text{interaction terms} \end{aligned}$$

This Lagrangian is inconsistent since we begin with  $\phi$  and  $A^\mu$ , thus with 4 degrees of freedom, and after the spontaneous symmetry breaking, we arrive to 5 degrees of freedom: a massless real scalar field  $\eta(x)$  (Goldstone boson), a massive real scalar field  $h(x)$  and a massive vector

field  $A^\mu(x)$ . By means of invariance under  $U(1)$ , it is possible to perform a gauge transformation which eliminates the non physical field,  $\eta(x) = 0$ :

$$\mathcal{L} = \frac{1}{2}\partial^\mu h(x)\partial_\mu h(x) - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - \frac{1}{2}(2\lambda v^2)h^2(x) + \frac{1}{2}(qv)^2 A_\mu(x)A^\mu(x) + \text{interaction terms}$$

Now there is one massive vector field  $A^\mu(x)$  with mass  $qv$  and a massive scalar field  $h(x)$  with mass  $\sqrt{2\lambda v^2}$ , therefore 4 degrees of freedom.

In order to apply this mechanism to SM, it is necessary to introduce an Higgs boson doublet  $\Phi(x)$ , because of the  $SU(2)$  symmetry. With the Lagrangian of Equation 3.2 it is possible to give mass only to the bosons. In order to give mass to the fermions the additional Yukawa interaction term has to be added:

$$\mathcal{L}_{YK} = -g_\ell[\bar{L}^\ell\psi_R^\ell\Phi + \Phi^\dagger\bar{\psi}_R^\ell L^\ell]$$

Hence, the total Lagrangian to which applies the spontaneous symmetry breaking mechanism is:

$$\mathcal{L} = [D^\mu\Phi]^\dagger[D_\mu\Phi] - \mu^2\Phi^\dagger\Phi - \lambda[\Phi^\dagger\Phi]^2 \quad (3.3)$$

$$-g_\ell[\bar{L}^\ell\psi_R^\ell\Phi + \Phi^\dagger\bar{\psi}_R^\ell L^\ell] + i[\bar{L}^\ell\cancel{D}L^\ell + \bar{\psi}_R^\ell\cancel{D}\psi_R^\ell + \bar{\psi}_R^\nu\cancel{D}\psi_R^\nu] \quad (3.4)$$

Now, if  $\lambda > 0$  and  $\mu^2 < 0$ ,  $\langle 0|\Phi|0 \rangle \neq 0$ , it is possible to expand  $\Phi$  near the minimum  $v$  as:

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} \eta_1(x) + i\eta_2(x) \\ v + h(x) + i\eta_3(x) \end{pmatrix}$$

where  $\eta_1(x)$ ,  $\eta_2(x)$  and  $\eta_3(x)$  are Goldstone bosons. After the expansion, all the  $SU(2) \times U(1)$  symmetry but  $U_{EM}(1)$  is broken, since the photon is massless and cannot interact with the Higgs boson. As previously done, it is possible to cancel all the Goldstone bosons performing gauge transformation, taking to:

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}$$

Inserting this in the Lagrangian of Equation 3.3, interactions terms between Higgs boson, fermions and gauge bosons appear, as well as the desired mass terms, where:

$$\begin{aligned} m_\gamma^2 &= 0 & m_\ell^2 &= \frac{g_\ell v}{\sqrt{2}} \\ m_W^2 &= \frac{g^2 v^2}{4} & m_Z^2 &= \frac{v^2 e^2}{\sin^2 \theta_W \cos^2 \theta_W} \\ m_H^2 &= 2\lambda v^2 \end{aligned}$$

### 3.1.3 Vector Boson Scattering

The observation of a Higgs boson with a mass of about 125 GeV by the CMS and ATLAS collaborations [49–51] consolidated the validity of the Standard Model (SM) of elementary interactions: within the current uncertainties, the newly found particle is consistent with the SM



expectations. Nevertheless, several fundamental issues remain unsolved: gravitational interactions are not unified with other forces, no valid candidates exist for dark matter, no explanation is found for the relative abundance of matter over anti-matter in the Universe and the mass hierarchy in the SM particles is not naturally explained. In particular, new phenomena may be present in the electroweak sector, where a sensitive probe of new physics is naturally given by the study of the scattering at high energy of massive electroweak bosons, usually known as Vector Boson Scattering (VBS) [52].

The VBS process happens at the LHC when a quark from each of the incoming protons radiates a vector boson (see diagram 3.2). These two vector bosons interact with each other, and their decay products are detected in the LHC experiments. The two initial quarks, which carry a significant fraction of the proton energy, are deflected from their initial trajectory and can be used to tag the process looking for energetic jets in the forward regions of the detector.

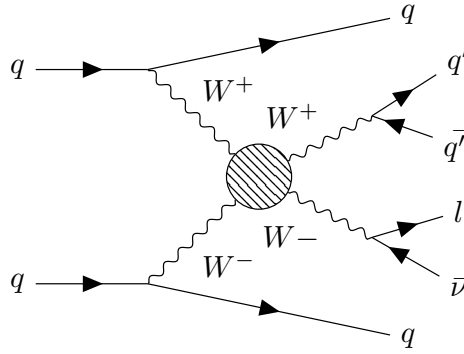


Figure 3.2: General form of Vector Boson Scattering Feynman diagrams

VBS plays a key role in the investigation of EW physics because it probes two key aspects of the SM together: gauge interactions on the one hand, being one of the few processes with tree-level sensitivity to the quadrilinear (or quartic) gauge couplings; the couplings between the Higgs and gauge bosons on the other hand, which are probed at energy scales which can sensibly differ from the Higgs mass [53]. In fact, in the absence of the SM Higgs boson, the amplitude of the longitudinal electroweak scattering ( $V_L V_L \rightarrow V_L V_L$ ) would increase as a function of  $\sqrt{\hat{s}}$  as  $\hat{s}/m_V^2$ , violating unitarity [54]. The inclusion of the SM Higgs boson in the t-channel precisely cancels this divergence. This poses a theoretical constraint on the Higgs boson mass from first principles of  $m_H < 800 \text{ GeV}$ , one of the reasons that convinced the scientific community to build the LHC accelerator. Moreover, if the  $H \rightarrow VV$  coupling, parametrized by the coupling modifier  $k_V$ , deviates from the SM predictions a distortion would show up in the di-boson spectrum, see Fig.3.3. This drives the quest for high precision measurements of VBS processes using the full Run II CMS dataset and the large amount of data that will be collected in Run II and HL-LHC.

### 3.1.4 Overview of the recent experimental results

Many recent experimental results have been published by ATLAS and CMS experiments on a large fraction of the VBS channels using the full Run II integrated luminosity. The cleanest ones, such as same-sign  $W^\pm W^\pm$  scattering, are moving from discovery to precision measurements such as Effective Field Theory (EFT) interpretation of the data. Moreover, the first searches looking for the polarized fraction of the  $WW$  scattering have been performed as a result of to the Run II integrated luminosity. A complete review describing both the theoretical and experimental status for VBS processes can be found in Ref. [53], and these are summarised

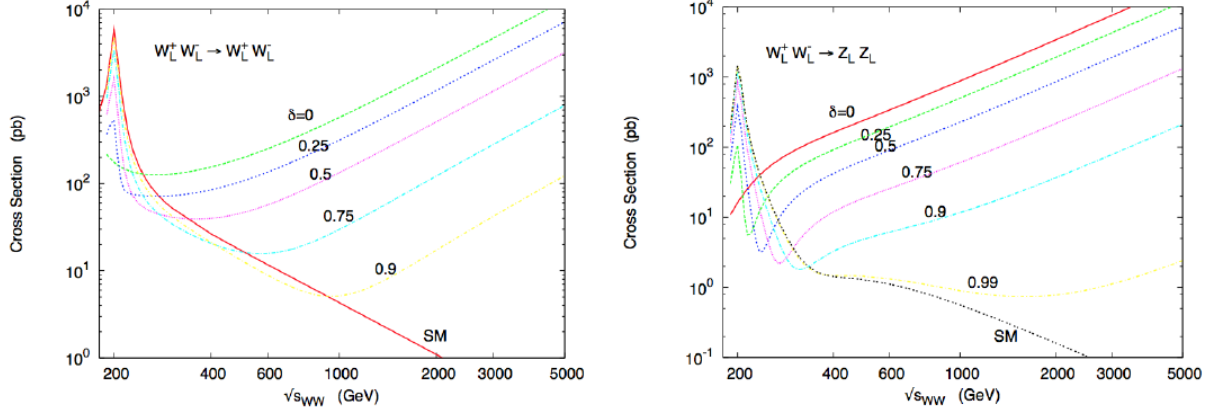


Figure 3.3: Scattering cross section for (left)  $W_L^+ W_L^- \rightarrow W_L^+ W_L^-$  and (right)  $W_L^+ W_L^- \rightarrow Z_L Z_L$  as a function of  $\sqrt{s_{WW}}$ . Different values of  $k_V$ , the  $H \rightarrow VV$  coupling modifier, with respect to unity are shown and a Higgs boson mass of GeV is assumed.

below:

**$W^\pm W^\pm jj \rightarrow l^\pm \nu l'^\pm \nu jj$**  The *golden* channel of VBS, where the EW cross-section is much larger than the QCD one. Both CMS and ATLAS observed this process using a subset of the Run II data, and has moved to a precision measurement of the cross-section. Both NLO QCD and EW calculations are available for this process [55]. CMS published a measurement of the differential cross-section performed on the full Run II dataset in combination with the  $W^\pm Z$  channel [56], while ATLAS published the search on a partial dataset [57]. All the results are in agreement with SM expectations, although the experimental uncertainties are of the order of 20% because of limited statistics.

**$W^\pm W^\pm jj$  polarized** In a separate analysis, CMS extracted from the same selected dataset a measurement of the polarization of W bosons in  $W^\pm W^\pm$  events [58]. A limit has been put on the longitudinal scattering  $W_L^\pm W_L^\pm < 1.17$  fb at 95% confidence-level (CL), compatible with the SM, and the signal significance for the observation of the  $W_X^\pm W_L^\pm$  scattering is  $2.3\sigma$ .

**$W^\pm Z jj \rightarrow l^\pm \nu l'^+ l'^- jj$**  This process has a larger cross-section than the same-sign WW one, but the Z branching ratio in leptons makes the number of final events comparable. However, it suffers from a larger background from  $W^\pm Z$  QCD production. Multivariate techniques have been implemented for signal extraction. ATLAS reported [59] differential measurements with the full Run II dataset, whereas CMS performed the analysis in combination with the same-sign W W channel [56].

**$ZZ jj \rightarrow l^+ l^- l'^+ l'^- jj$**  The ZZjj VBS process has the smallest cross-section among the final states containing heavy vector bosons, and is one of the rarest SM processes observed to date. However, its experimental signature is very clean and, being the only final state where full reconstruction of the 6 final-state fermions is possible, it could become important for polarization measurements with larger data sets in the future. ATLAS [60] reported an observation of the signal over 5 standard deviations, while CMS [61] a strong observed evidence ( $4\sigma$ ). CMS has also performed EFT interpretation of the data.

**$WV jj \rightarrow l^\pm \nu qq' jj$  and  $ZV jj \rightarrow l^+ l^- \nu qq' jj$**  The semileptonic VBS channels have a larger cross-section than the fully-leptonic ones, at the cost of a larger background from QCD production of single vector bosons plus jets. Semileptonic VBS involves one or two leptons and

four jets in the final state, therefore it is at the moment at the edge of the state of the art of MC generators and only calculations at leading-order accuracy are available. Both ATLAS and CMS have reported until now studies of semileptonic VBS final states with a partial subset of the 13 TeV LHC data. The ATLAS analysis [62] targeted the SM signal measurement combining all the semileptonic channels and using multivariate techniques. The search reached a significance for the rejection of the background-only hypothesis of  $2.7\sigma$  with  $36 \text{ fb}^{-1}$ . The CMS analysis [63], instead, was devoted to Beyond Standard Model (BSM) searches with a simple cut optimization, and obtained competitive limits on anomalous quartic gauge couplings.

**$VVjj \rightarrow qqjj$**  Fully hadronic VBS channels are currently under investigation by the experimental collaborations as they represent a huge technical and computational challenge both in terms of MC production and experimental signal extraction.

## 3.2 Analysis strategy

The target of this analysis is the extraction of the  $WV$  VBS production cross-section, signal strength, and its corresponding significance, in the semileptonic channel. The semileptonic VBS signal is characterized by the presence of a single isolated electron or muon and a moderate amount of missing transverse momentum  $p_T^{\text{miss}}$ , coming from the leptonic decay of one of the final state  $W$  vector bosons, and either three or four jets. One pair of jets, called VBS tag jets, originates from the scattered incoming partons, and it is required to have a large invariant mass  $m_{jj}^{\text{VBS}}$  and large pseudorapidity separation, the typical signature of VBS-like events. Among the remaining jets, the decay products of the hadronically-decaying final state vector boson  $W$  or  $Z$  must be identified to fully reconstruct the VBS topology. The signal efficiency is maximized targeting different energetic regimes of this hadronically-decayed vector boson. If it is boosted, its decay products are reconstructed with high probability as a single energetic large radius jet in the detector. If instead the  $p_T$  of the vector boson is smaller, it is reconstructed as a pair of smaller radius jets with invariant mass near to the  $W$  or  $Z$  boson mass.

The largest background contribution (see Sec.3.5) is given by the single  $W$  boson production process in association with jets, followed by the top quark and the QCD multi-jet backgrounds. The contamination from the single vector boson EW production in the VBF channel (VBF- $V$ ) is negligible in the resolved category, but more important in the boosted one. The signal region is bound inside a narrow interval of the invariant mass of the hadronic vector boson decays products around the  $W$  or  $Z$  mass peak. This choice permits to reduce the contamination from the  $W$ +jets background: a control region is defined in the invariant mass region outside the interval delimiting the signal one. The contamination from the top quark backgrounds is reduced drastically by implementing a b-jets veto using the CMS multivariate DeepCSV algorithm. Moreover, inverting the b-veto requirement a control region for the top quark backgrounds is obtained.

The signal observation significance is extracted including all the regions in a likelihood profiling, such that the fit jointly determines the background contribution importance in the signal region in the form of nuisances (see Sec. 3.9). The likelihood function is built on a signal model taken from simulation and corrected for all residual data-to-Monte-Carlo (MC) disagreements in particle reconstruction, selection efficiency, and on background models that are built based on estimates from control samples in data or corrected for data-to-simulation disagreements, as needed. In particular, the top quark background shape is determined with the MC simulation, while its normalization is left free to float in the fit. The  $W$ +jets background normalization is

also left free to float in the fit: this is performed in a coarse set of bins, allowing to correct differentially for residual discrepancies in its shape between data and Monte Carlo (see Sec. 3.8). Among the sub-leading backgrounds, the QCD multi-jet one, that may enter the signal region with fake charged leptons (also called nonprompt leptons), is determined in a data-driven way with jet-to-lepton fake rates, while the remaining ones are estimated with the simulation.

Because of the large background on the one hand, and of the sophisticated signal topology on the other, the most significant features that allow to separate them have been condensed in a single discriminator built with a Deep Neural Network (DNN) (see Sec. 3.7). One discriminant was trained for the resolved category and one for the boosted one, since the event topology and the particle kinematics significantly change between the two. In both cases, the resulting discriminators have been analyzed a-posteriori with the SHAP technique, a method borrowed from the field of the explainable artificial intelligence and built on the Shapley variables from game theory [64], that allowed for the isolation of the most significant variables among the DNN inputs. The resulting discriminators range from 0 – the background-like extreme – to 1 – the signal-like end and show a clear predominance of background events in their lower side, as expected. Therefore, the final analysis fit has been performed on the binned shape of the DNN discriminator, that has been transformed to obtain a flat yield for the signal.

### 3.2.1 Signal and background topology

A brief overview of the signal and main background processes is given in the next sections. Details about the configuration of the MC simulation of the different processes can be found Sec. 3.3.

#### 3.2.1.1 VBS signal

With VBS one refers to purely electroweak processes in which, at LO in perturbation theory, starting from two quarks in the initial state, a six fermion final state is produced involving only electroweak vertexes  $\mathcal{O}(\alpha_{em}^6)$ . Several Feynman diagrams, characterized by the same initial and final state, are involved:

- **Vector boson scattering diagrams** are divided into three independent logical steps: emission of a vector boson pair (W or Z) from the initial state partons, scattering of the emitted bosons and finally their decay into the final states particles (see Fig. 3.4). This category includes quartic gauge coupling (QGC) (diagram 3.5), double triple gauge (TGC) coupling in  $t$ - or  $s$ -channel (diagram 3.6),  $t$ -channel Higgs boson exchange and  $s$ -channel Higgs boson production (diagram 3.7) [53].

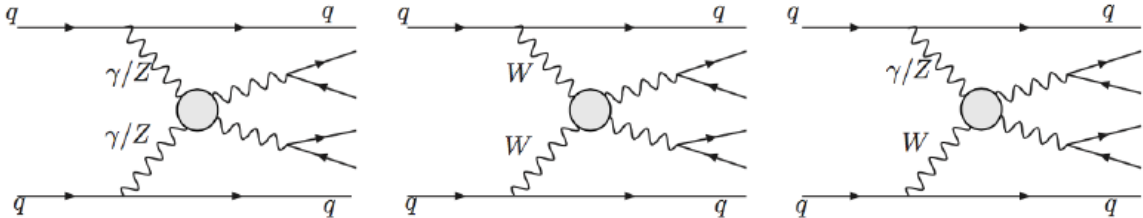


Figure 3.4: Generic form of the vector boson scattering diagrams. This structure is shared between double triple gauge coupling (TGC) in  $t$ - or  $s$ -channel, quartic gauge coupling (QGC),  $t$ -channel Higgs boson exchange and  $s$ -channel Higgs boson production diagrams.

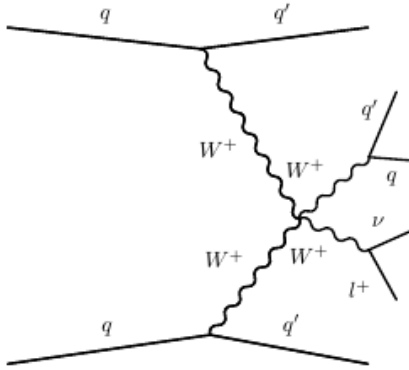


Figure 3.5: WW scattering with quartic gauge coupling

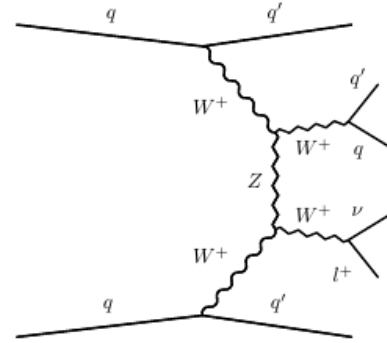


Figure 3.6: Double triple gauge coupling (TGC) in the t-channel. Also the s-channel is present

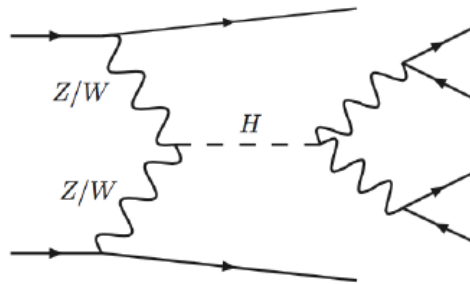


Figure 3.7: Higgs boson exchange in the s-channel, also the t-channel is present

- **Non scattering diagrams** strongly interfere with the scattering ones, making it not possible to disentangle scattering from non-scattering processes (diagram 3.8). The contribution of these processes is minimized by the analysis preselections.

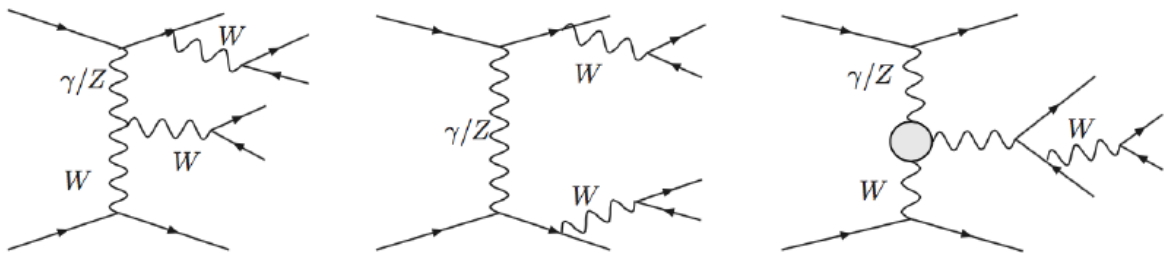


Figure 3.8: Example of non-scattering diagrams that are part of the pure EWK vector boson scattering process

- **Three vector boson production** includes the production of a vector boson or a photon in s-channel followed by triple gauge coupling (TGC), or the t-channel production of 3 vector bosons (diagrams in 3.9).

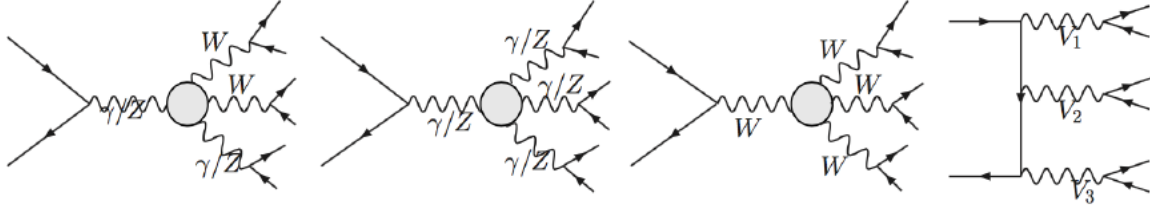


Figure 3.9: Schematic diagrams for three vector boson production in s- and t-channel that produces the same final state

In conclusion, several contributions must be considered when looking for a complete LO estimate of  $\alpha_{em}^6$  processes. The production times decay approximation is implemented in the VBS process simulation and its limitation must be taken into account [53]. It is impossible to separate only the scattering diagram contribution: it can be enhanced with the respect to other types of purely EW contributions with appropriate phase space cuts, for example requiring a high invariant mass of the pair of jets not coming from the vector boson (after identification).

### 3.2.1.2 QCD-WV background

The QCD production of two vector bosons is the main contributor to the irreducible background at order  $\mathcal{O}(\alpha_{em}^4 \alpha_S^2)$ . This process can be separated from the VBS signal only at LO in the matrix element expansion. At next-to-leading-order (NLO) the separation between the two processes is not well-defined without making assumptions[65]. In fact the EW and QCD corrections are mixed and interfere between each other, as shown schematically in Fig. 3.11. Also in this case appropriate phase space cuts are implemented in the analysis to enhance the VBS contribution over the QCD one.

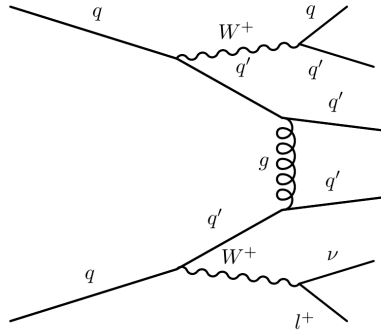


Figure 3.10: QCD di-boson production (QCD-WV) at order  $\alpha_{em}^4 \alpha_S^2$

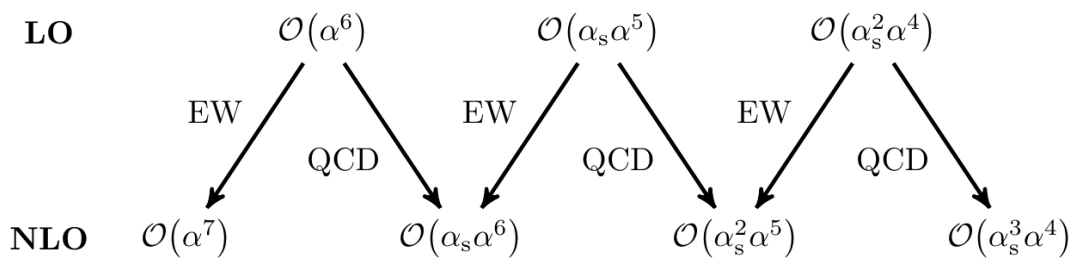


Figure 3.11: All contributing orders at both LO and NLO for the VBS processes at the LHC[65].

### 3.2.1.3 Single top and $t\bar{t}$ backgrounds

Another contribution at order  $\mathcal{O}(\alpha_{em}^4 \alpha_S^2)$  is the production of a single top quark or a  $t\bar{t}$  pair. The top quarks decay into a b quark and W boson and produce the same signature of VBS. The cross-section of this background is high at LHC, but it can be reduced as a result of b-jets identification algorithms.

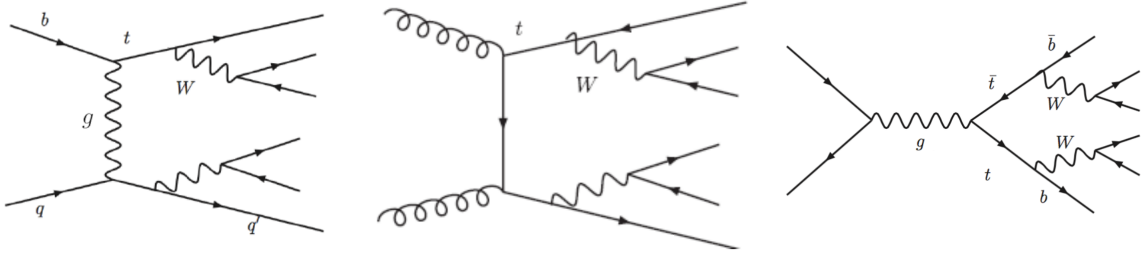


Figure 3.12: Single top quark and  $t\bar{t}$  production at order  $\alpha_{em}^4 \alpha_S^2$

### 3.2.1.4 W+jets background

The largest background contribution comes for the production of a W boson plus jets at order  $\alpha_{em}^2 \alpha_S^4$  (diagrams in 3.13). This is a reducible background since the final state is different from the signal (one vector boson is missing), but can be misidentified since vector bosons are not directly detected. Different multiplicities of jets enter the calculation of the cross-section and the generation of MC events for this background since the QCD radiation can add multiple jets in the final state. The larger the number of jets included in the computation of the hard-scatter cross-section, also called matrix-element, the higher the precision that can be reached in the theoretical simulation of this process. Additional softer emissions are modelled via the parton shower simulation with a more approximate computation, and matched with a special procedure with the matrix-element generated emissions [66]. More details about the MC generation of this background performed in bins of scalar transverse energy  $H^T$  can be found in Sec.3.5.

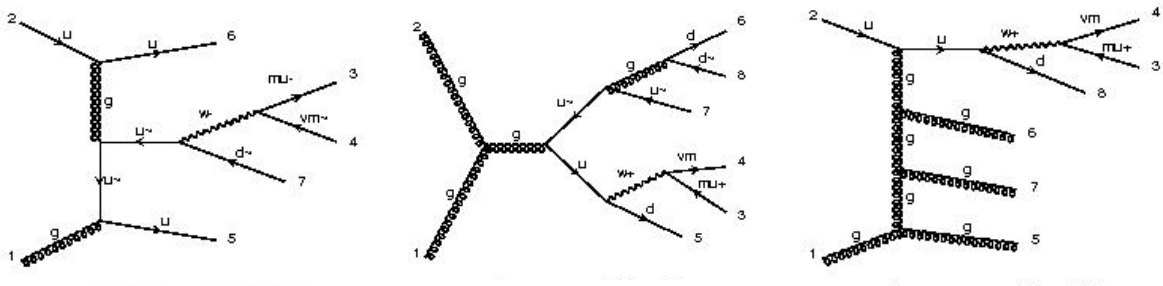


Figure 3.13: W+ 3 jets (left) process, W + 4 jets processes (center, right)

### 3.3 Data sets and MC simulations

#### 3.3.1 Data sets

Data events are selected using isolated single electron and single muon triggers. Only events certified with good data quality by the sub-detector experts are considered for further analysis. The  $p_T$  thresholds are different for each year following the detector evolution and the pileup conditions. Table 3.4 summarizes the list of the high level trigger (HLT) paths used in the analysis and their lepton  $p_T$  thresholds, while Tab. 3.5 lists the total luminosity available for the analysis for each year.

Table 3.4: Trigger paths used for the selection of events by year and CMS dataset.

| Year | Data set       | HLT trigger                  | $p_T$ threshold |
|------|----------------|------------------------------|-----------------|
| 2016 | SingleMuon     | HLT_IsoMu24                  | 24 GeV          |
|      |                | HLT_IsoTkMu24                |                 |
| 2017 | SingleMuon     | HLT_IsoMu27                  | 27 GeV          |
| 2018 | SingleMuon     | HLT_IsoMu24                  | 24 GeV          |
| 2016 | SingleElectron | HLT_Ele27_WPTight_Gsf        | 27 GeV          |
|      |                | HLT_Ele25_eta2p1_WPTight_Gsf |                 |
| 2017 | SingleElectron | HLT_Ele35_WPTight_Gsf        | 35 GeV          |
| 2018 | SingleElectron | HLT_Ele32_WPTight_Gsf        | 32 GeV          |

Table 3.5: Total integrated luminosity declared good for physics analysis for each year of data taking

| Year | Total integrated luminosity ( $\text{fb}^{-1}$ ) |
|------|--------------------------------------------------|
| 2016 | $36.3\text{fb}^{-1}$                             |
| 2017 | $41.5\text{fb}^{-1}$                             |
| 2018 | $59.7\text{fb}^{-1}$                             |

#### 3.3.2 Signal MC simulation samples

The VBS EW processes are simulated at LO with the MADGRAPH5\_aMC@NLO v2.6.5 generator requiring two vector bosons and two quarks in the final state: the intermediate-state vector boson pair is produced by implementing the narrow width approximation and then decayed by MadSpin [67] to partially account for finite-width effects and spin correlations. One vector boson decays leptonically to electrons, muons, and taus and the corresponding neutrino (in case  $V = W$ ); the second one decays into a pair of quarks.

The dipole recoil scheme is used in the parton shower PYTHIA simulation for the VBS signal MC sample, in order to get the best description of the additional jet emissions in the VBS topology [53, 55, 68]. A detailed comparison of the two simulation configurations is shown in Sec. 3.3.5.

The  $W^\pm W^\pm$ ,  $W^\pm W^\mp$ ,  $W^\pm Z$ , and  $ZZ$  processes are produced separately. The contribution from the VBS  $Z(\text{ll})V(\text{jj})$  production, where one of the leptons falls beyond the acceptance of the analysis, is considered as a background. Parton-level requirements are imposed to form a loosely defined fiducial region: all final state partons must have  $p_T > 10\text{ GeV}$  and one pair of final state partons is required to have an invariant mass greater than  $100\text{ GeV}$ . The complete list of the SM signal samples and the corresponding cross-sections can be found in Table 3.6. The fraction of events for each VBS channel in the resolved and boosted signal regions is shown in Fig. 3.14.



Table 3.6: EW standard model VBS samples.

| Process                       | cross-section (pb) |
|-------------------------------|--------------------|
| $W^+(\text{jj}) W^-(1\nu)$    | 0.9107             |
| $W^-(\text{jj}) W^+(1\nu)$    | 0.9114             |
| $W^-(\text{jj}) W^-(1\nu)$    | 0.0326             |
| $W^+(\text{jj}) W^+(1\nu)$    | 0.0879             |
| $W^-(1\nu) Z(\text{jj})$      | 0.1000             |
| $W^+(1\nu) Z(\text{jj})$      | 0.1825             |
| $Z(\text{ll}) Z(\text{jj})$   | 0.0159             |
| $W^+(\text{jj}) Z(\text{ll})$ | 0.0540             |
| $W^-(\text{jj}) Z(\text{ll})$ | 0.0298             |

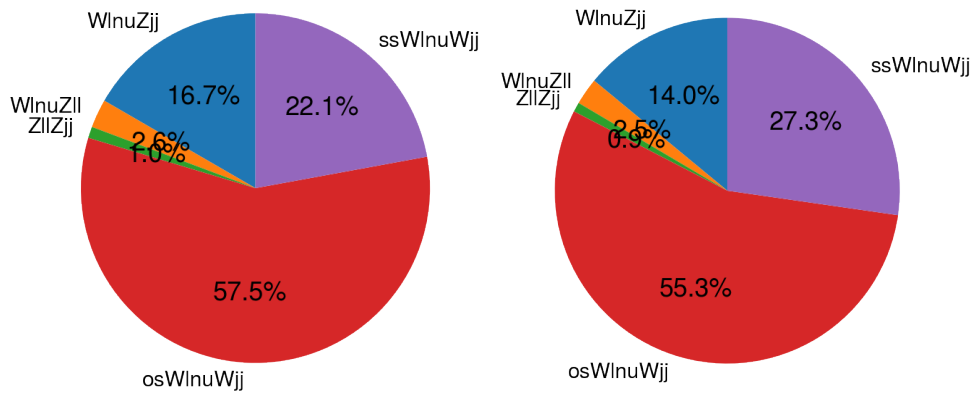


Figure 3.14: Composition of VBS events for each channel in the resolved (left) and boosted (right) signal regions.  $osWlnuWjj$  stands for the opposite sign WW process.  $ssWlnuWjj$  stands for the same sign WW process.

### 3.3.2.1 Interference between EW VBS and QCD-VV processes

The interference between the EW and QCD diagrams for the  $W^\pm W^\pm$ ,  $W^\pm W^\mp$ , and  $WZ$  processes is generated with MADGRAPH5\_aMC@NLO including the contributions of order  $\alpha_s \alpha^5$ . The small contributions of off-shell and nonresonant production [69] are not included in the simulated interference samples and the  $W$  and  $Z$  bosons are decayed with PYTHIA 8.230. The parton-level cuts applied are  $p_T > 15$  GeV and dijet mass greater larger than 200 GeV. The interference for the  $W^\pm W^\pm$  and  $WZ$  final states is positive for these fiducial requirements at the generation step with cross sections of 8.2 fb and 4.5 fb, respectively. The interference for the  $W^\pm W^\mp$  final state is negative with a cross section of -44.1 fb.

The interference samples have been processed in the same way as all the other samples and then compared to the signal in the analysis phase spaces. Figure 3.15 and 3.16 show the comparison in the resolved and boosted categories respectively. The total cross section of the interference evaluated on the 2018 MC sample in the resolved category is 0.28 fb, and in the boosted category is 0.09 fb. The interference is 2.6% of the signal in the resolved category and 3.8% in the boosted one: therefore it is neglected in the analysis.

### 3.3.3 Background MC simulation samples

The main sources of background contamination originate from the production of a single  $W$  boson accompanied by jets (called  $W$ +jets in the following),  $t\bar{t}$  production with one of the  $W$  bosons produced by the top quark decays hadronically, and the non-resonant QCD induced di-boson pair production. The background processes are modeled using Monte Carlo (MC) event generators (and corrected from data for  $t\bar{t}$  and  $W$ +jets): nonresonant quark- and gluon-induced  $W$  and  $Z$  boson pair production (QCD-WV, QCD-ZV);  $t\bar{t}$  and single top quark production (in  $s$ -,  $t$ -channel and  $tW$ ); Drell–Yan (DY) lepton pair production;  $W$  and  $Z$  boson production in association with a photon ( $W\gamma$  and  $Z\gamma$ ); single vector boson EW production in the VBF channel (VBF-V); and triboson production (VVV).

Most of the processes are simulated at next-to-leading order (NLO) in the strong coupling  $\alpha_s$  using POWHEG v2 [70–74], MADGRAPH5\_aMC@NLO v2.4.2 [66, 75], or MCFM v7.0 [76–79]. The  $W$ +jets, QCD-VV, VBF-V, and  $W\gamma$  events are generated with MADGRAPH5\_aMC@NLO v2.4.2 at leading order (LO) accuracy in perturbation quantum chromodynamics.

$W$ +jets and DY+jets backgrounds are modeled at leading order, with up to 4 jets simulated in the matrix-element, and generated in bins of total transverse scalar energy  $\sum |p_T^{\text{parton}}|$  (HT) to increase the number of MC events describing the high jets multiplicity phase space. A LO inclusive sample is used to cover the low HT phase space. The stitching of HT bins cross-sections is checked against LO sample and correction factors are applied for consistency.

Details about the scale factors applied to the simulated MC events to describe in the best way possible the data can be found in Sec. 3.3.4, while more information about the main backgrounds estimation ( $t\bar{t}$  and  $W$ +jets) are documented in Sec. 3.5. The list of sample names for different processes is shown in Tables 3.7 and 3.8.

### 3.3.4 MC simulation configuration

In all MC samples apart from VBF-V, parton shower, hadronization, and the simulation of the underlying event are provided by PYTHIA 8.226 (8.230) [74, 80]. The NNPDF 3.0 NLO [81] (NNPDF 3.1 NNLO [82]) parton distribution functions (PDFs) are used for simulating all 2016 (2017 and 2018) samples. The modeling of the underlying event is generated using the CUETP8M1 [83, 84] (CP5 [85]) tune for simulated samples corresponding to the 2016 (2017 and

Table 3.7: MC samples for background modeling.

| process              | year       | dataset name                                                                         |
|----------------------|------------|--------------------------------------------------------------------------------------|
| W+jets LO            | 2016       | WJetstoLNu_TuneCUEP8M1_13TeV-madgraphMLM-pythia8                                     |
|                      | 2017,2018  | WJetstoLNu_TuneCP5_13TeV-madgraphMLM-pythia8                                         |
|                      | 2016       | WJetstoLNu_HT*_TuneCUEP8M1_13TeV-madgraphMLM-pythia8                                 |
| W+jets LO HT bins    | 2017,2018  | WJetstoLNu_HT*_TuneCP5_13TeV-madgraphMLM-pythia8                                     |
|                      |            | * HT bins: 70, 100, 200, 400, 600, 800, 1200, 2500, + $\infty$ GeV                   |
|                      | 2016       | DYJetstoLL_M-50_TuneCUEP8M1_13TeV-amcatnloFXFX-pythia8                               |
| DY+jets              |            | DYJetstoLL_M-10to050_TuneCUEP8M1_13TeV-amcatnloFXFX-pythia8                          |
|                      |            | DYJetstoLL_M-5to050_HT*_TuneCUEP8M1_13TeV-madgraphMLM-pythia8                        |
|                      |            | HT bins: 70, 100, 200, 400, 600, + $\infty$ GeV                                      |
|                      |            | DYJetstoLL_M-50_HT*_TuneCUEP8M1_13TeV-madgraphMLM-pythia8                            |
|                      |            | HT bins: 70, 100, 200, 400, 600, 800, 1200, 2500, + $\infty$ GeV                     |
|                      |            | DYJetstoLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8                                   |
|                      |            | DYJetstoLL_M-10to050_TuneCP5_13TeV-madgraphMLM-pythia8                               |
|                      |            | DYJetstoLL_M-4to050_HT*_TuneCP5_13TeV-madgraphMLM-pythia8                            |
|                      |            | HT bins: 100, 200, 400, 600, + $\infty$ GeV                                          |
|                      |            | DYJetstoLL_M-50_HT*_TuneCP5_13TeV-madgraphMLM-pythia8                                |
| 2018                 |            | HT bins: 100, 200, 400, 600, 800, 1200, 2500, + $\infty$ GeV                         |
|                      |            | DYJetstoLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8                                   |
|                      |            | DYJetstoLL_M-10to050_TuneCP5_13TeV-madgraphMLM-pythia8                               |
|                      |            | DYJetstoLL_M-4to050_HT*_TuneCP5_13TeV-madgraphMLM-pythia8                            |
|                      |            | HT bins: 100, 200, 400, 600, + $\infty$ GeV                                          |
|                      |            | DYJetstoLL_M-50_HT*_TuneCP5_13TeV-madgraphMLM-pythia8                                |
|                      |            | HT bins: 70, 100, 200, 400, 600, 800, 1200, 2500, + $\infty$ GeV                     |
|                      |            | DYJetstoLL_M-50_TuneCUEP8M2_ttHtranche3_13TeV-powheg-pythia8                         |
|                      |            | TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8                                               |
|                      |            | TTToSemiLepton_TuneCUEP8M2_ttHtranche3_13TeV-powheg-pythia8                          |
| Semileptonic tt      | 2016       | TTToSemiLepton_TuneCUEP8M2_ttHtranche3_13TeV-powheg-pythia8                          |
|                      | 2017, 2018 | TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8                                               |
|                      | 2017, 2018 | TTToSemiLepton_TuneCP5_13TeV-powheg-pythia8                                          |
| Single top s-channel | 2016       | ST_s-channel_4f_leptonDecays_13TeV-amcatnlo-pythia8_TuneCUEP8M1                      |
|                      | 2017, 2018 | ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-amcatnlo-pythia8                          |
|                      | 2016       | ST_t-channel_(anti)top_4f_inclusiveDecays_13TeV-powhegV2-madspin-pythia8_TuneCUEP8M1 |
| Single top t-channel | 2016       | ST_t-channel_(anti)top_4f_inclusiveDecays_13TeV-powhegV2-madspin-pythia8             |
|                      | 2017       | ST_t-channel_(anti)top_4f_inclusiveDecays_TuneCP5_13TeV-powhegV2-madspin-pythia8     |
|                      | 2018       | ST_t-channel_(anti)top_4f_inclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8       |
| Single top tW        | 2016       | ST_tW_(anti)top_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUEP8M1                  |
|                      | 2017, 2018 | ST_tW_(anti)top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8                      |
|                      | 2016       | ttZJets_13TeV-madgraphMLM-pythia8                                                    |
| tt Z production      | 2017, 2018 | ttZJets_TuneCP5_13TeV-madgraphMLM-pythia8                                            |
|                      | 2016       | TTWJetstoLNu_TuneCUEP8M1_13TeV-amcatnloFXFX-madspin-pythia8                          |
|                      | 2017, 2018 | ttWJets_TuneCP5_13TeV-madgraphMLM-pythia8                                            |

Table 3.8: MC samples for background modeling.

| process  | year                | dataset name                                                               |
|----------|---------------------|----------------------------------------------------------------------------|
| QCD-VV   | 2016: TuneCUETP8M1  | WminusTo2JZTo2LJJ_QCD_LO_SM_*                                              |
|          |                     | WminusToLNuWminusTo2JJJ_QCD_LO_SM_*                                        |
|          |                     | WminusToLNuZTo2JJJ_QCD_LO_SM_*                                             |
|          |                     | WplusTo2JWminusToLNuJJ_QCD_LO_SM_*                                         |
|          | 2017: TuneCP5       | WplusTo2JZTo2LJJ_QCD_LO_SM_*                                               |
|          |                     | WplusToLNuWminusTo2JJJ_QCD_LO_SM_*                                         |
|          | 2018: TuneCP5       | WplusToLNuWplusTo2JJJ_QCD_LO_SM_*                                          |
|          |                     | WplusToLNuZTo2JJJ_QCD_LO_SM_*                                              |
|          |                     | ZTo2LZTo2JJJ_QCD_LO_SM_*                                                   |
|          |                     | * = MJJ100PTJ10_Tune*_13TeV-madgraph-pythia8                               |
| VVV      | 2016: TuneCUETP81M1 | WW_4F_Tune*_13TeV-amcatnlo-pythia8                                         |
|          | 2017: TuneCP5       | WWZ_Tune*_13TeV-amcatnlo-pythia8                                           |
|          | 2018: TuneCP5       | WZZ_Tune*_13TeV-amcatnlo-pythia8                                           |
|          |                     | ZZZ_Tune*_13TeV-amcatnlo-pythia8                                           |
| Vg + VgS | 2016                | WGToLNuG_TuneCUETP81M1_13TeV-madgraphMLM-pythia8                           |
|          |                     | ZGTo2LG_TuneCUETP81M1_13TeV-amcatnloFFX-pythia8                            |
|          | 2017, 2018          | WZTo3LNu_mllmin01_13TeV-powheg-pythia8                                     |
|          |                     | WGToLNuG_TuneCP5_13TeV-madgraphMLM-pythia8                                 |
| VBF-V    | 2016                | ZGToLLG_01J_5f_TuneCP5_13TeV-amcatnloFFX-pythia8                           |
|          |                     | WZTo3LNu_mllmin01_NNPDF31_TuneCP5_13TeV-powheg-pythia8                     |
|          | 2017, 2018          | EWK_LNuJJ_MJJ-120_13TeV-madgraph-herwigpp                                  |
|          |                     | EWK_LLJJ_MLL-50_MJJ-120_TuneEEC5_13TeV-madgraph-herwigpp_corrected         |
| ggWW     | 2017, 2018          | EWK_LNuJJ_MJJ-120 private production with same conf as EWK_LLJJ            |
|          | 2016, 2017, 2018    | EWK_LLJJ_MLL-50_MJJ-120_TuneCH3_PSweights_13TeV-madgraph-herwig7_corrected |
|          |                     | GluGluWWToLNuQQ_TuneCP5_13TeV-amCatNLO_pythia8                             |

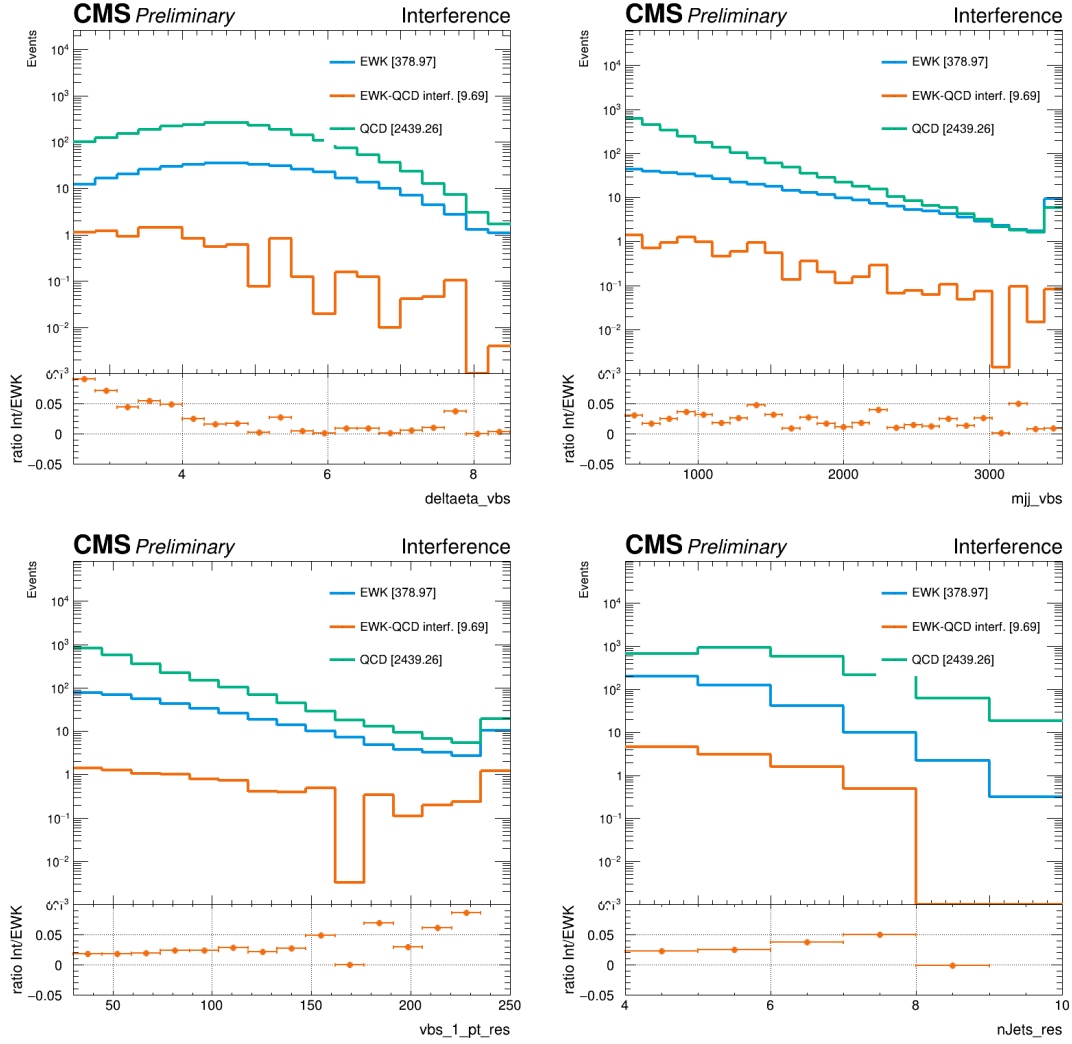


Figure 3.15: Resolved, muon channel (2018): comparison between EW signal, QCD-WV background and the EWK-QCD interference. The lower panel shows the ratio between the interference term and the EW signal one.

2018) data. The dipole recoil scheme is used in the parton shower PYTHIA simulation for the VBS signal MC sample, and the HERWIG 7.0 [86, 87] program is used for VBF-V background in order to improve the description of the additional jet emissions in the VBS topology [53, 55, 68].

For all processes, the detector response is simulated using a detailed description of CMS based on the GEANT4 package [88]. Additional interactions in the same or adjacent bunch crossings (pileup) based on minimum bias events simulated with PYTHIA are overlaid onto each event, with the number of interactions drawn from a distribution that is similar to the one observed in data. The average number of such interactions per event is  $\approx 23$  for the 2016 data, and  $\approx 32$  for the 2017 and 2018 data.

The samples are weighted event-by-event by a function dependent on the transverse momentum and pseudorapidity of the single lepton candidate to account for the HLT inefficiencies. The trigger efficiency is estimated using a tag and probe method on data. All MC events are required to have the highest  $p_T$  lepton matched to a gen-level one.

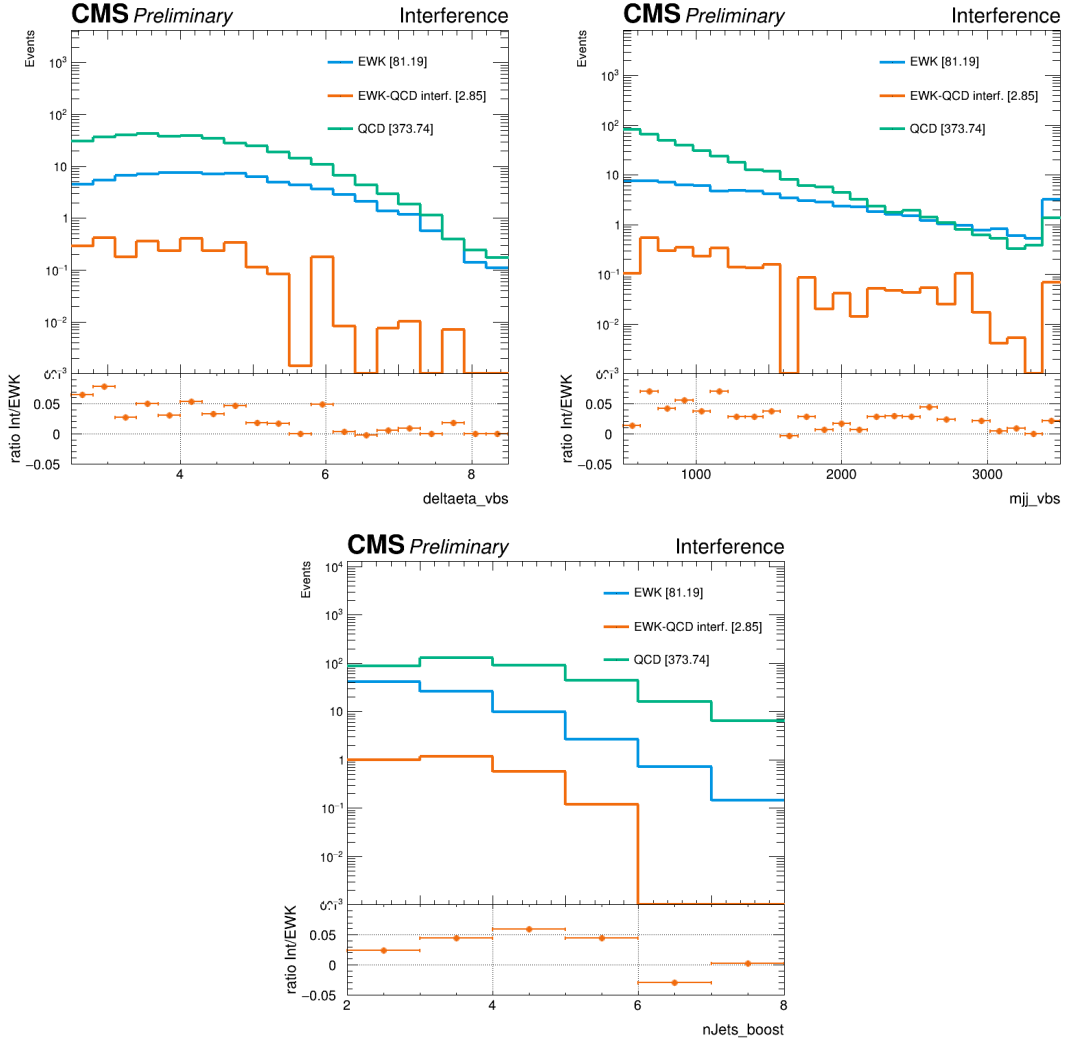


Figure 3.16: Boosted, muon channel (2018): comparison between EW signal, QCD-WV background and the EWK-QCD interference. The lower panel shows the ratio between the interference term and the EW signal one.

Standard MET filters are applied to all MC events and data events to remove events containing anomalous missing energy due to inefficiencies in specific detector regions.

Simulation samples are further modified with the following event-by-event weights to better approximate the collision data in terms of kinematics and object identification efficiency.

- **Pileup reweighting:** The ratio between the normalized distributions of the number of simultaneous inelastic collisions is computed for each year of data-taking and the MC pileup profile is used as an event weight. A proton inelastic cross section value of  $(69.2 \pm 4.6\%)$  mb is used to derive the distribution of the number of collisions per bunch crossing. The pileup reweighting factor is a function of the true number of generated inelastic collisions in the event.
- **Lepton reconstruction efficiency scale factor:** A data / MC scale factor for lepton reconstruction efficiency is applied as a function of the lepton  $p_T$  and  $\eta$ .
- **Lepton identification efficiency scale factor:** A data / MC scale factor of lepton

identification efficiencies is applied to the event as a function of the lepton  $p_T$  and  $\eta$ .

- **Heavy-flavor tagging efficiency scale factor:** For each jet in the event where the b-tagging discriminator is evaluated, a scale factor depending on the jet  $p_T$ ,  $\eta$ , flavor, and the b-tagging discriminator value is applied to the event.
- **Pileup jet ID scale factor:** Scale factors computed for the pileup jet identification at the loose working point are applied to the event. The total weight per event is a product of the scale factors for all jets with  $30 < p_T < 50$  GeV,  $|\eta| < 4.7$  and passing the tight Jet ID, in the event.
- **L1 prefiring correction:** This is only applied for MC simulations for the 2016 and 2017 data-taking periods. It is an additional trigger efficiency scale factor to account for the inefficiency due to the L1T prefiring effect.
- **DY  $p_T$  reweighting:** Drell-Yan events are reweighted to improve the agreement of the simulated transverse momentum ( $p_T$ ) distributions of the lepton-lepton system [89].
- **Top  $p_T$  reweighting:** The  $t\bar{t}$  component of the top quark background is reweighted to improve the agreement of the simulated transverse momentum  $p_T$  of the top quark. [90, 91]
- **$\tau_{21}$  boosted W tagging scale factor** Efficiency scale factors for AK8 boosted jets W\Z tagging based on the  $\tau_{21}$  selection are applied. Dedicated scale factors are applied to high purity (boosted V analysis category) and low purity (resolved analysis region) categories.
- **AK8 jet mass scale and resolution** The PUPPI softdrop mass of AK8 jets used in the analysis is rescaled (both in data and MC) and smeared (only in MC) to calibrate it and make the MC resolution equal to the data one.
- **Quark Gluon Likelihood calibration** The Quark Gluon Likelihood (QGL) discriminator is used in the analysis as one of the inputs of the DNN classifier to reduce the contamination of the W+jets background, where jets are produced more often by gluons instead of by quarks as in the VBS signal. The QGL variables need to be calibrated correctly for the phase space under study: the technique developed and used in this work is reported in Sec. 3.6.

### 3.3.5 Dipole shower simulation comparison

#### 3.3.5.1 VBS signal MC sample

The Pythia8 dipole shower simulation is known to be more correct than the default one for the simulation of the hadronic shower in VBS events [55]. Two MC samples have been produced for the VBS signal and have been compared for the observables used in the analysis. Figure 3.17 shows some distributions for the resolved category, whereas Figure 3.18 is for the boosted one. The largest difference is observed in the number of jets with  $p_T > 30$  GeV as expected from this kind of simulation configuration, and is shown in Fig 3.19. The effect of adopting the dipole recoil simulation on the DNN distribution, used in the final fit, is to increase the signal in the background depleted region since the DNN depends on the number of jets (for the full details about the input variables used for the DNN see Sec.3.7).

The VBS signal simulation has been also compared at parton-level with a MC sample generated with Madgraph+HERWIG. The effect on the number of jets distribution is shown in Fig. 3.21: the HERWIG distribution goes in an unexpected direction.

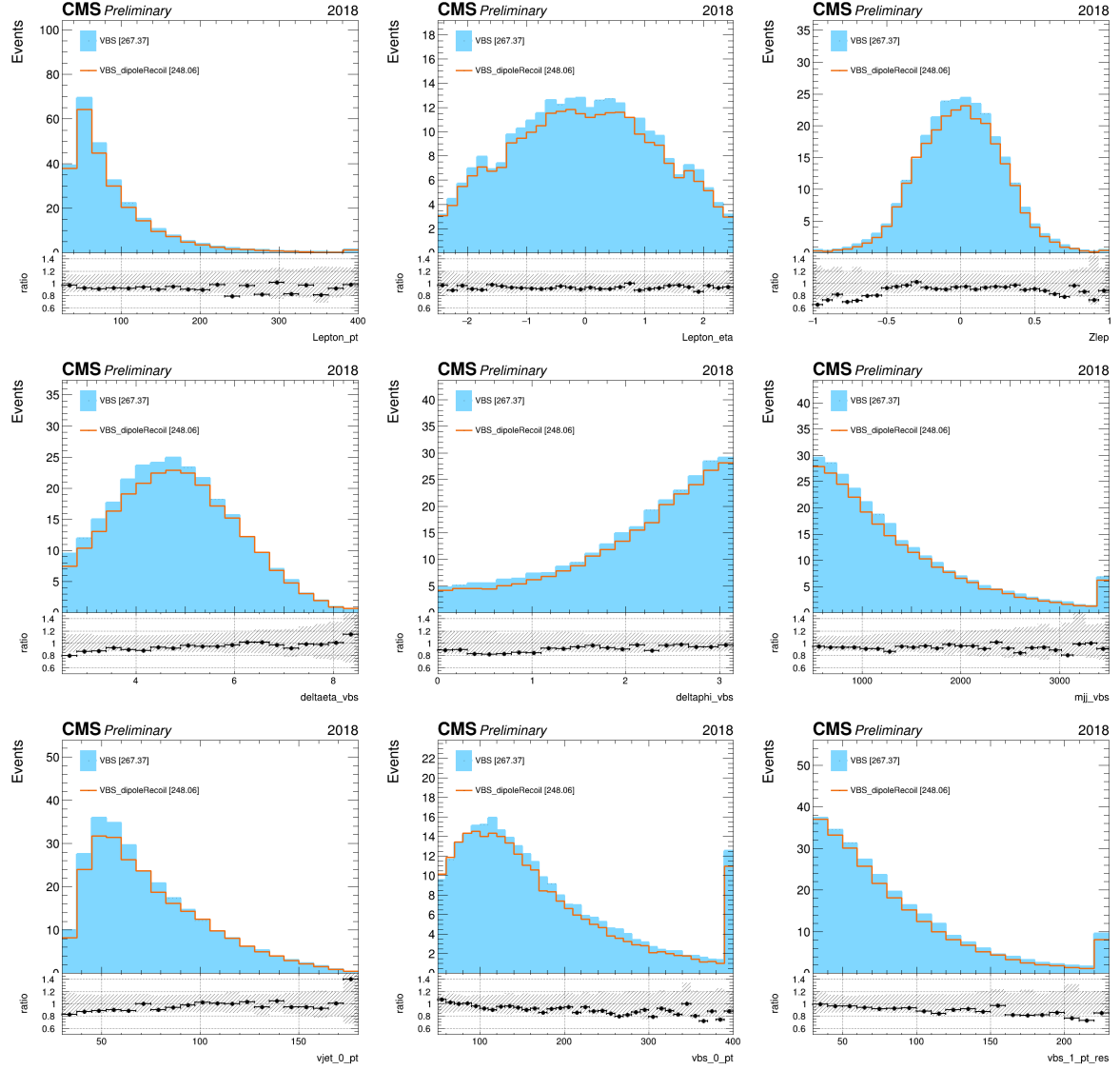


Figure 3.17: Comparison between VBS Madgraph+Pythia8 with and without dipole recoil shower option at RECO level. The 2018 sample is used, and the resolved electron phase space is shown. The lower panel shows the ratio between the dipole recoil simulation distribution and the nominal VBS one.



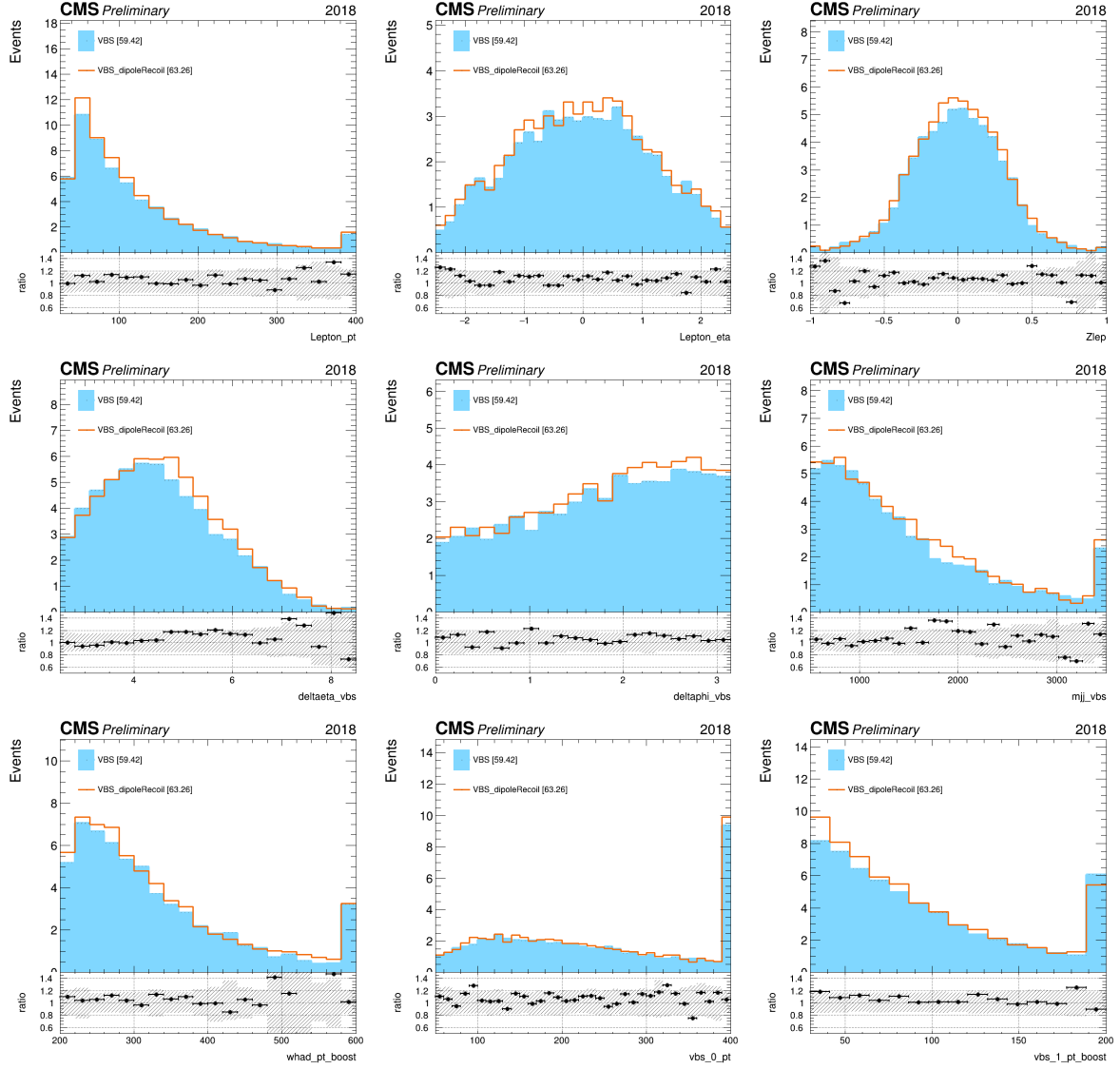


Figure 3.18: Comparison between VBS Madgraph+Pythia8 with and without dipole recoil shower option at reconstruction level. The 2018 sample is used, and the boosted electron phase space is shown. The lower panel shows the ratio between the dipole recoil simulation distribution and the nominal VBS one.

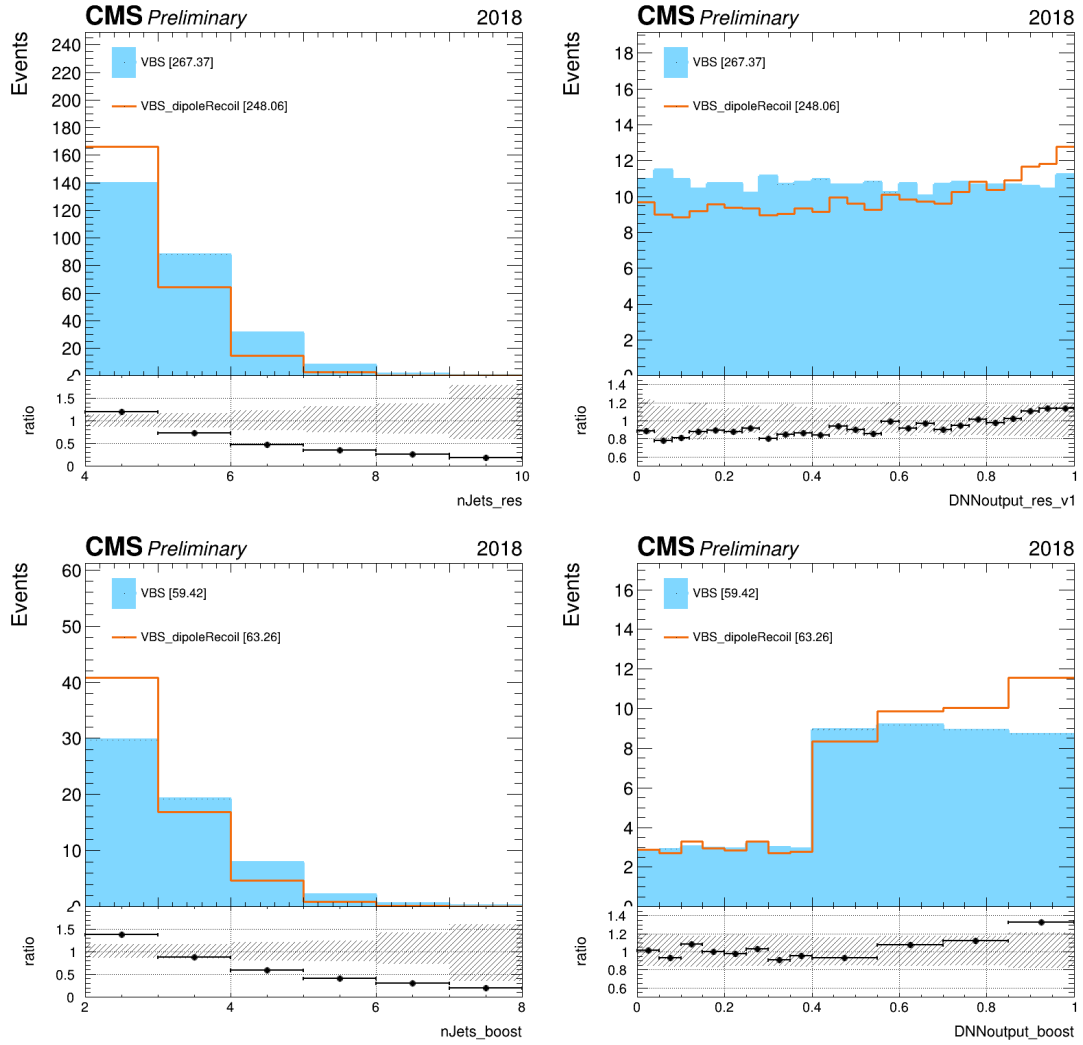


Figure 3.19: Comparison between VBS Madgraph+Pythia8 with and without dipole recoil shower option at reconstruction level. The 2018 sample is used, and the resolved (top) and boosted (bottom) electron phase space are shown. The lower panel shows the ratio between the dipole recoil simulation distribution and the nominal VBS one.

### 3.3.5.2 VBF-V MC sample

The VBF-V MC simulation is compared between MADGRAPH +Pythia8 and MADGRAPH +HERWIG configurations using 2016 MC samples. The dipole shower simulation, available in HERWIG, has been shown to describe the data in a better way than the default Pythia8 simulation [92]. The distribution of the  $p_T$  and  $\eta$  observables are compared in an inclusive single lepton phase space in Figure 3.20. The HERWIG sample has softer, less central jets, as expected for the dipole shower simulation. The effect on the number of jets with  $p_T > 30$  GeV is large. Moving into the analysis phase space, where 4 jets are required for the resolved category and 3 for the boosted category (1 Ak8, 2 Ak4 jets), the VBF-V contribution is smaller using the HERWIG sample. Figure 3.22 shows the number of jets distribution in the resolved and boosted category.

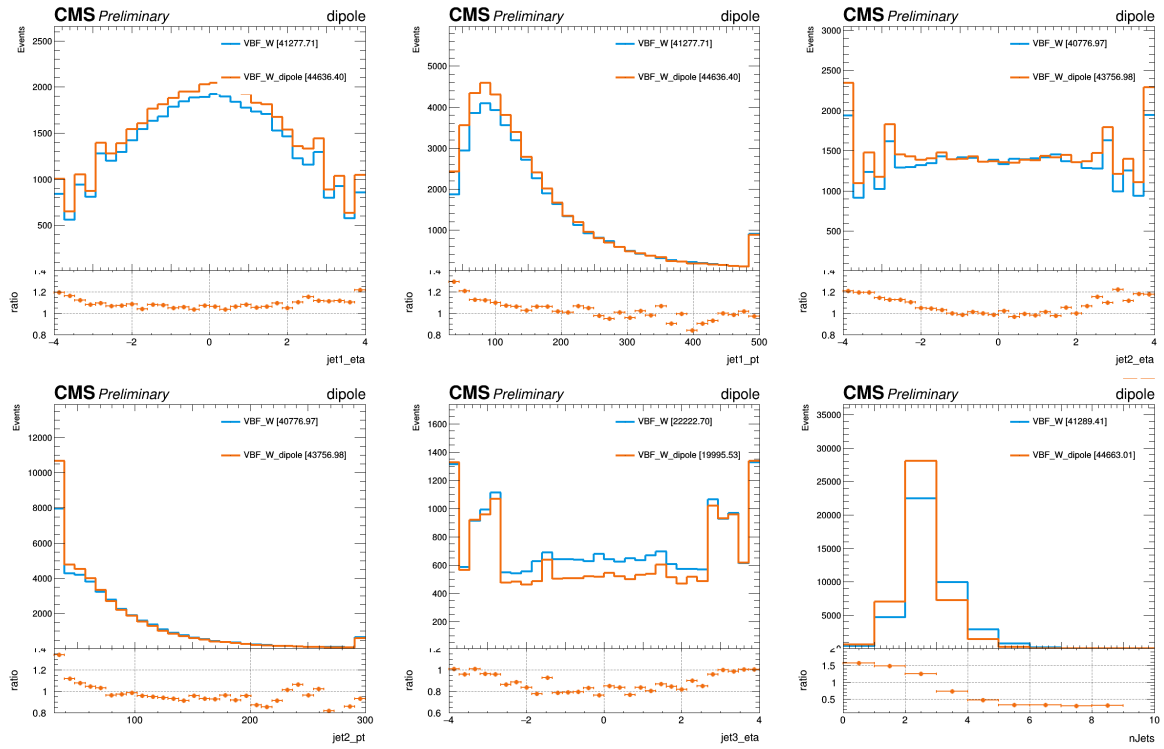


Figure 3.20: Comparison between MADGRAPH +Pythia8 and MADGRAPH +HERWIG for VBF-V simulation (2016 samples) at reconstruction level. Inclusive single lepton phase space. The lower panel shows the ratio between the HERWIG simulation distribution and the nominal VBF-V one.

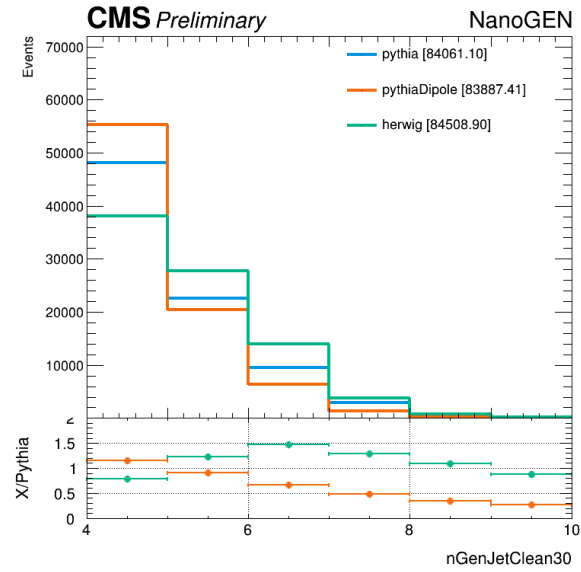


Figure 3.21: Comparison between VBS Madgraph+Pythia8 with and without dipole recoil shower option and with HERWIG simulation at GEN level. The 2018 sample is used, and the inclusive region is shown. The lower panel shows the ratio between the dipole recoil or the HERWIG simulation distribution and the nominal VBS one.

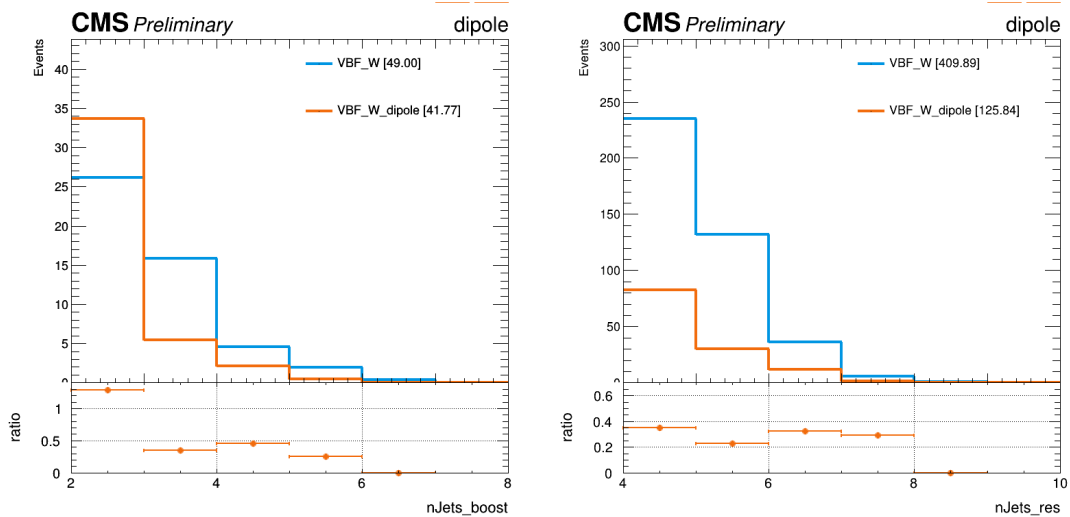


Figure 3.22: Comparison between MADGRAPH +Pythia8 and MADGRAPH +HERWIG for VBF-V simulation (2016 samples) at reconstruction level. Boosted (left) and resolved (right) electron analysis phasespaces. The lower panel shows the ratio between the Herwig simulation distribution and the nominal VBF-V one.

## 3.4 Object definition and event selection

### 3.4.1 Object definition

Events are selected for further analysis using isolated single lepton triggers with  $p_T$  thresholds of 27, 32, 35 GeV for electrons and of 24, 24, 27 GeV for muons, respectively for the 2016, 2017, 2018 data taking periods. The leptons are also required to have a reconstructed  $p_T$  of at least 35 GeV (30 GeV) for electron (muon) candidates, and a pseudorapidity of  $|\eta| < 2.5$  (2.4) for electrons (muons).

For each event, hadronic jets are clustered from reconstructed particles using the infrared- and collinear-safe anti- $k_T$  algorithm [8, 9] with a distance parameter of 0.4 (0.8), labeled in the following as AK4 (AK8) jets. Additional proton-proton interactions within the same or nearby bunch crossings (pileup) can contribute additional tracks and calorimetric energy depositions to the jet momentum. To mitigate this effect, charged particles identified as originating from pileup vertices are discarded and an offset correction is applied to correct for remaining contributions.

In an event, AK4 and AK8 jets are considered in the analysis if they have a  $p_T > 30$  GeV and  $|\eta| < 4.7$  or  $p_T > 200$  GeV and  $|\eta| < 2.4$ , respectively. Jets that overlap with loosely identified electrons or muons by a cone of  $\Delta R < 0.3$ , and those with  $p_T < 50$  GeV which fail the loose working point of the pileup-jet discriminator, are disregarded. Additionally, for the 2017 dataset, jets with  $2.65 < |\eta| < 3.139$  and  $p_T < 50$  GeV are required to pass a tighter threshold of the pileup-jet discriminator in order to mitigate the bias in the observed jet multiplicity coming from detector noise in this pseudorapidity range.

The pileup-per-particle identification algorithm (PUPPI) [10, 11] is applied to AK8 jet constituents to remove pileup tracks at the reconstructed particle level. Moreover, a grooming algorithm, known as “soft drop” (SD)[93–95], is applied to the constituents of AK8 jets reclustered using the Cambridge–Aachen algorithm [96, 97]. The SD algorithm, which has as an angular exponent  $\beta = 0$ , soft cutoff threshold  $z_{\text{cut}} < 0.1$ , and characteristic radius  $R_0 = 0.8$  [95], removes soft, wide-angle radiation from the large radius jet, improving the modeling of the jet mass observable. The parameters of the SD algorithm are calibrated in a top quark-antiquark sample enriched in hadronically decaying W bosons [98]. The AK8 jets are identified as hadronic decays of Lorentz-boosted W/Z bosons using the ratio between 2- and 1-subjettiness [99] variables denoted as  $\tau_{21} = \tau_2/\tau_1 < 0.45$  and with a groomed AK8 jet mass between 40 and 250 GeV.

### 3.4.2 Events selection and categories

Candidate events are required to contain exactly one tightly identified and isolated lepton [7, 100] associated with the W boson leptonic decay. Events containing a second loosely identified lepton with  $p_T > 10$  GeV are vetoed. Events with missing transverse momentum  $p_T^{\text{miss}} < 30$  GeV are discarded.

Two main categories are defined depending on the reconstruction regime of the hadronically decaying vector boson. An event is assigned to the boosted category if it contains only one AK8 jet, with  $p_T > 200$  GeV and  $|\eta| < 2.4$ , that passes the selection criteria as a hadronically decaying vector boson  $V_{\text{had}}$ , together with at least two AK4 jets. Otherwise, if no AK8 jet V boson candidate is found and instead at least four AK4 jets are reconstructed with  $p_T > 30$  GeV, the event is assigned to the resolved category. In both resolved and boosted categories, the pair of AK4 jets with the largest invariant mass are identified as the VBS tag jets. In the resolved category, considering the remaining jets after the VBS tag jets selection, the jet pair

with invariant mass closest to 85 GeV (the average between the W and Z boson masses) is then chosen as the decay product of  $V_{\text{had}}$ .

The fraction of VBS events in the sample is enhanced requiring a large invariant mass  $m_{jj}^{\text{VBS}} > 500$  GeV and large pseudorapidity interval  $\Delta\eta_{jj}^{\text{VBS}} = |\eta_{j1}^{\text{VBS}} - \eta_{j2}^{\text{VBS}}|$  for the tag jets. The leading VBS tag jet is required to have  $p_T > 50$  GeV and the transverse mass of the leptonically decaying W is required to be  $M_W^T < 185$  GeV, defined as

$$M_W^T = \sqrt{2 p_T(\ell) p_T^{\text{miss}} [1 - \cos(\Delta\phi(\ell, \vec{p}_T^{\text{miss}}))]}, \quad (3.5)$$

where  $p_T(\ell)$  is the  $p_T$  of the lepton and  $\Delta\phi(\ell, \vec{p}_T^{\text{miss}})$  is the azimuthal distance between the lepton and the missing transverse momentum  $\vec{p}_T^{\text{miss}}$ .

After these preselections, signal and control regions are defined for the main backgrounds, top quark and W+jets backgrounds, both in the resolved and boosted regions in a similar way, as illustrated in Fig. 3.23. The signal region is built by requiring that no b-jet candidates are found in the event according to the loose working point of the *DeepCSV* tagger [101], a ML-based b-tagging algorithm, corresponding to a b-tagging efficiency of  $\sim 85\%$  and mis-tag probability of  $\sim 20\%$ . The hadronically decaying vector boson invariant mass  $m_V$  should also be between 65-105 GeV (70-115 GeV) for the resolved (boosted) category, in order to be consistent with an on-shell W or Z boson decaying hadronically. Events falling in the same  $m_V$  interval as the signal but containing at least one b-jet are classified in the top quark control region. Finally, if no b-jets are found and  $m_V$  is not within the W or Z resonance range,  $m_V > 40$  GeV and  $m_V \notin (65, 105)$  GeV and  $m_V \notin (70, 115)$  GeV for the resolved and boosted cases, events are classified as part of the W+jets control region. All of the signal and control regions are split according to the flavor of the selected lepton (electron or muon).

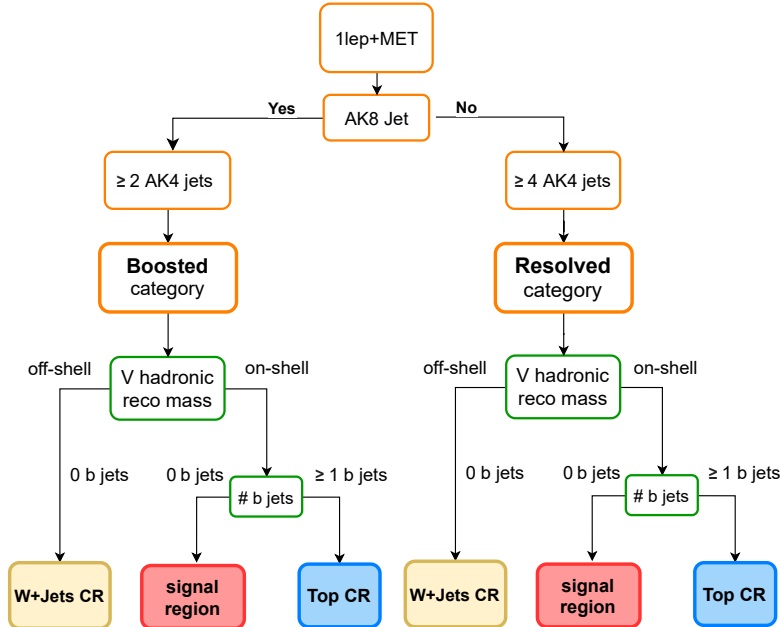


Figure 3.23: Analysis workflow: object and category selections followed by control regions (CR) and signal region definitions.

### 3.4.3 Jet tagging efficiency

In the resolved category, jets can be tagged as VBS jets or as hadronic decay products of the V boson with different algorithms. Moreover the presence of more than 4 jets introduces ambiguity in the tagging process. For the VBS signal simulation at parton level, the partons coming from the V boson decay are easily determined looking at their invariant mass and the partons coming from the initial interaction are simply the remaining one. The efficiency of the tagging algorithms at reco level can be obtained analyzing the VBS signal MC and matching the partons with the reco jets. This test is performed in a well defined phase space where all the 4 partons are associated withing a radius of 0.4 with 4 different jets (this requirement has an efficiency of 42% after baseline jets preselections). Moreover the efficiency is computed both in an inclusive region (requiring 4 jets with  $p_T > 30$  GeV, one lepton with  $p_T > 25$  GeV and vetoing b-jets) and in the resolved signal region.

Three matching efficiencies are computed: on VBS jets, on V jets and on the 4 jets together. Four different tagging algorithms are analyzed:

**maxMjj massV** First tag VBS jets looking at the maximum invariant mass pair, then tag the V-jets using the vector boson V mass (average between W and Z mass)

**massV maxMjj** First tag the V-jets using the V mass, then tag VBS jets looking at the maximum invariant mass jets pair

**massV maxPt** First tag the V-jets using the V mass, then tag VBS jets looking at the maximum  $p_T$  jets pair

**maxPt massV** First tag VBS jets looking at the maximum  $p_T$  jets pair, then tag the V-jets using the V mass

Table 3.9 shows the results in the inclusive region, while Table 3.10 shows the ones for the signal region. The analysis strategy (maxMjj massV) is among the most efficient in the signal region. It has been selected because tagging the VBS jets first avoids shaping the invariant mass spectrum in the W+jets background.

Table 3.9: Inclusive region, matching efficiencies

| Strategy     | VBS-jets match | V-jets match | Total match |
|--------------|----------------|--------------|-------------|
| maxMjj massV | 56.5 %         | 55.9 %       | 52 %        |
| massV maxMjj | 61.3 %         | 61.6 %       | 57 %        |
| massV maxPt  | 61.1 %         | 61.6 %       | 59%         |
| maxPt massV  | 41.6 %         | 37.6 %       | 36 %        |

Table 3.10: Resolved signal region, matching efficiencies

| Strategy      | VBS-jets match | V-jets match | Total match |
|---------------|----------------|--------------|-------------|
| maxMjj massWZ | 77.8 %         | 78.3 %       | 72 %        |
| massWZ maxMjj | 74.9 %         | 75.7 %       | 69 %        |
| massWZ maxPt  | 76.9 %         | 75.7 %       | 72%         |
| maxPt massWZ  | 52.7 %         | 49.5 %       | 48 %        |

## 3.5 Backgrounds

### 3.5.1 Top quark background

Control plots for several kinematic variables are shown in Figures 3.24, 3.25, and 3.26 for the resolved category and in Figures 3.27, 3.28, and 3.29 for the boosted category. Only the muon channels is shown for brevity, but the electron one shares the same the data/MC trends. All the distributions are plotted with nominal values of the nuisance parameters, as defined before the fit (pre-fit) and before the application of data-driven corrections to the W+jets background and normalization corrections for the top quark samples.

In general, the MC does not show noticeable discrepancies with respect to the data. The most important variables for the VBS signal discrimination,  $m_{jj}^{\text{VBS}}$ ,  $Z_{\text{lep}}$  (defined in Eq. 3.8), the number of jets and  $\Delta\eta_{jj}^{\text{VBS}}$ , are well described by the MC in this top enriched phase-space. Since  $t\bar{t}$  is an irreducible background with the same final state as the VBS signal under study, it is considered a proxy for how well the MC describes the signal.

Since the shape of the distributions is well described by MC, only the top quark background normalization will be left floating in the final fit and the top control region will be used to measure this normalization.



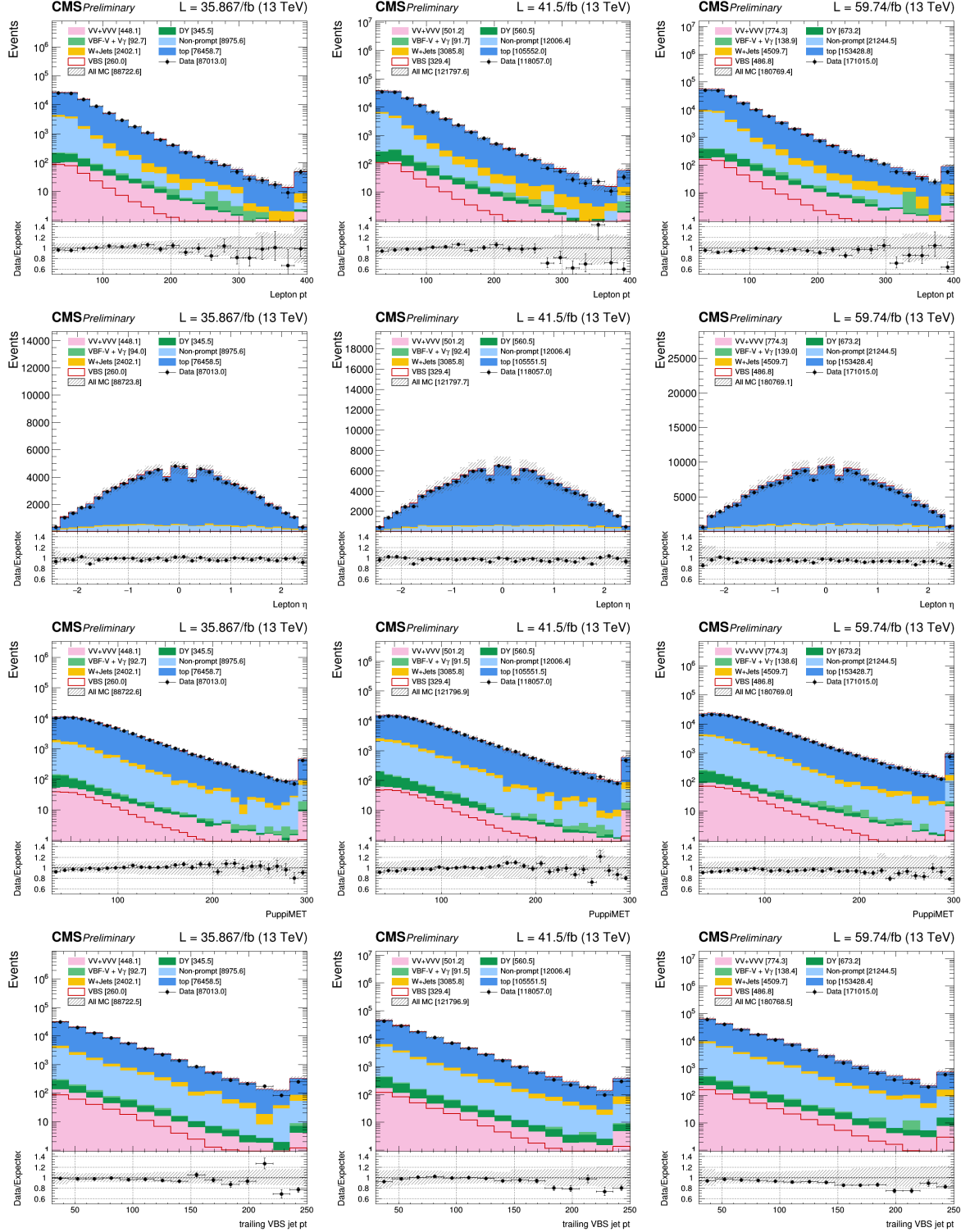


Figure 3.24: Resolved, top, muon control region. From left to right: 2016, 2017, 2018. From top to bottom: Lepton  $p_T$ ,  $\eta$ , PuppMET, and Pt of the trailing VBS-jet

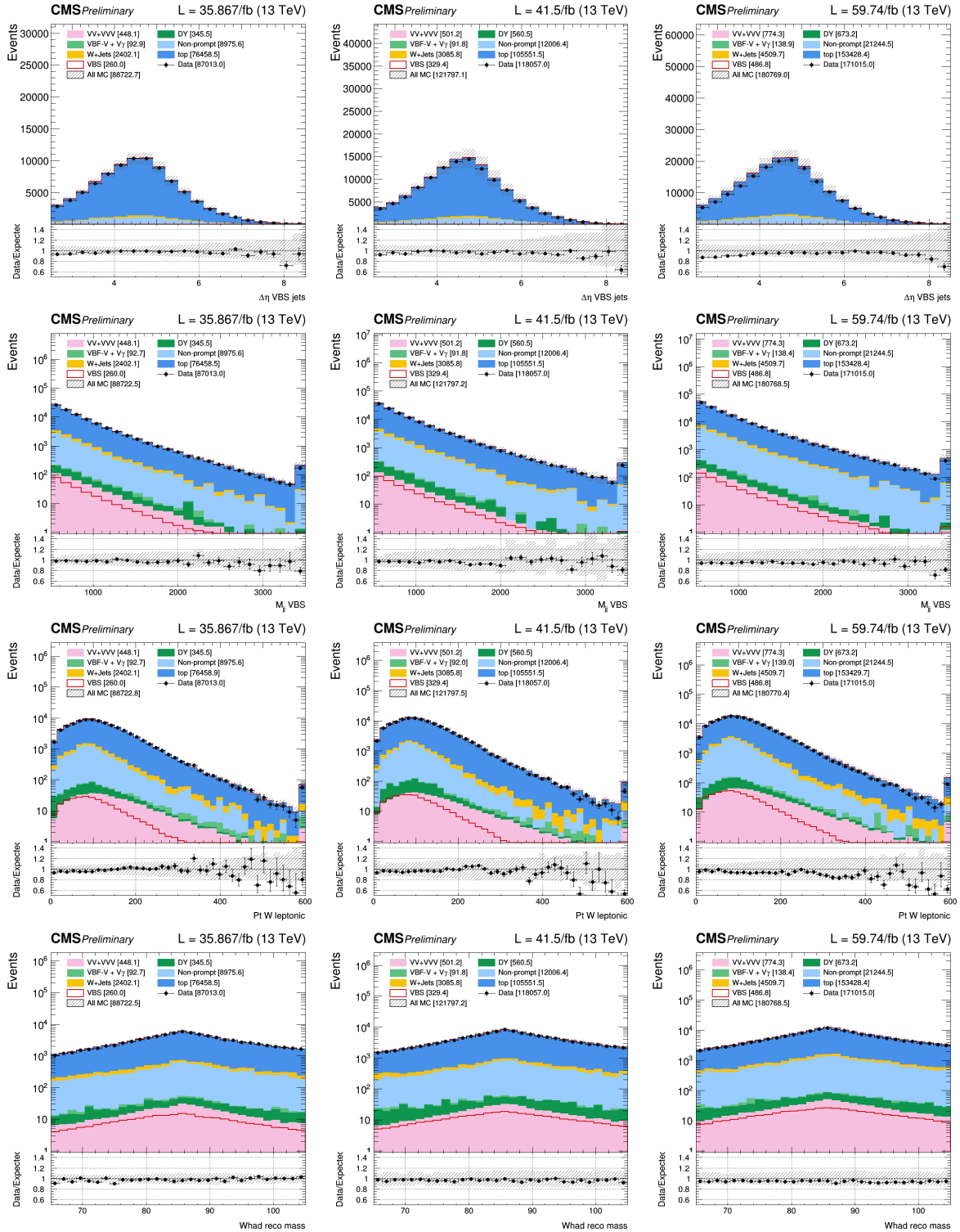


Figure 3.25: Resolved, top, muon control region. From left to right: 2016, 2017, 2018. From top to bottom:  $\Delta\eta_{\text{jj}}^{\text{VBS}}$ , and invariant mass of VBS-jets  $m_{\text{jj}}^{\text{VBS}}$ ,  $p_{\text{T}}^{\text{W},\ell}$ , and invariant mass of the jets associated to the hadronic V decay.

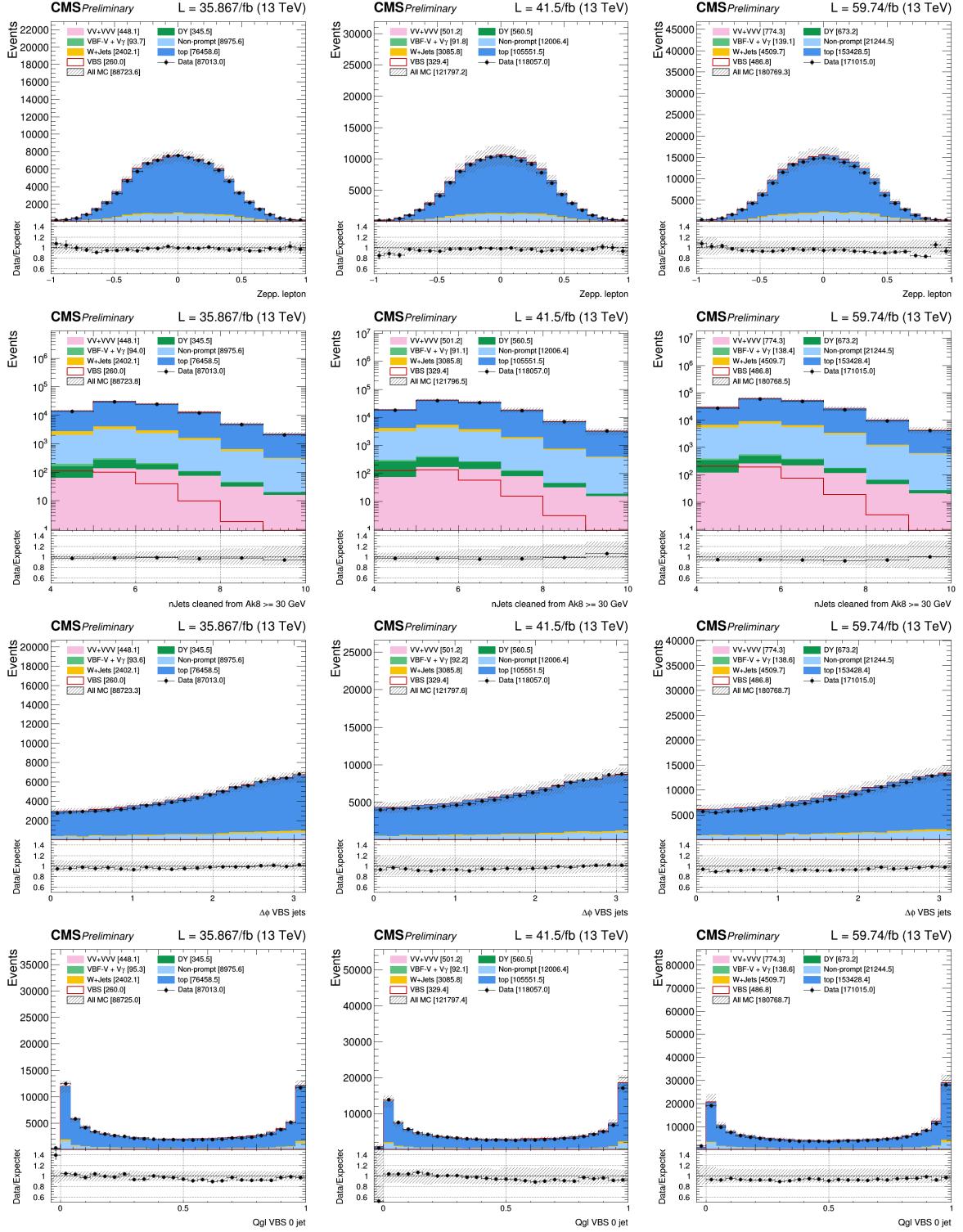


Figure 3.26: Resolved, top, muon control region. From left to right: 2016, 2017, 2018. From top to bottom: Lepton Zeppenfeld variable, number of jets with  $p_T > 30$  GeV, azimuthal distance between VBS jets  $\Delta\phi_{\text{VBS}}$ , and quark-gluon likelihood of the leading VBS jet.

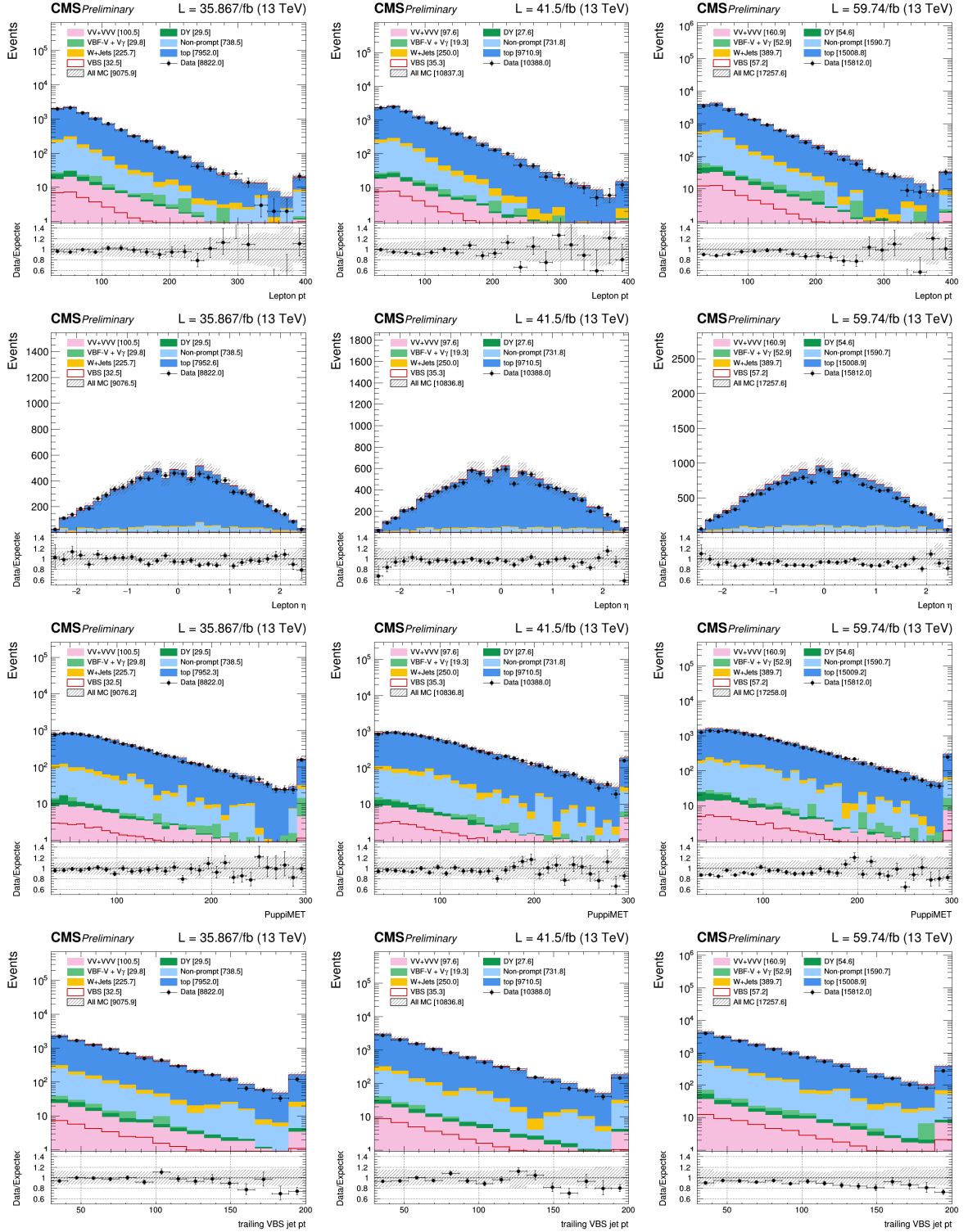


Figure 3.27: Boosted, top, muon control region. From left to right: 2016, 2017, 2018. From top to bottom: Lepton  $p_T$ ,  $\eta$ , PuppMET, and  $p_T$  of the trailing VBS-jet

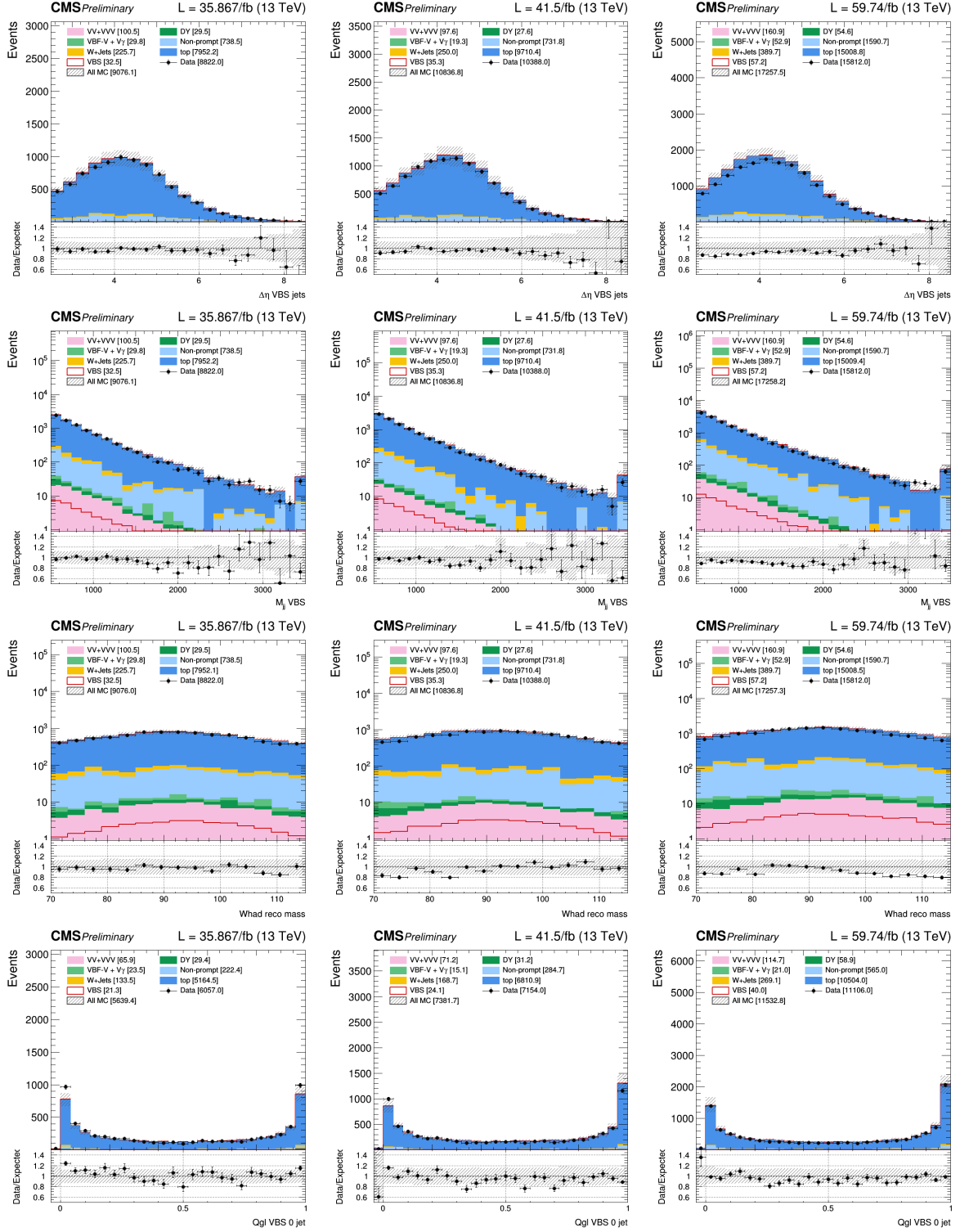


Figure 3.28: Boosted, top, muon control region. From left to right: 2016, 2017, 2018. From top to bottom:  $\Delta\eta_{\text{VBS jets}}$ , invariant mass of VBS-jets  $m_{jj}^{\text{VBS}}$ , invariant mass of the boosted AK8 jet, quark-gluon likelihood of the leading VBS jet.

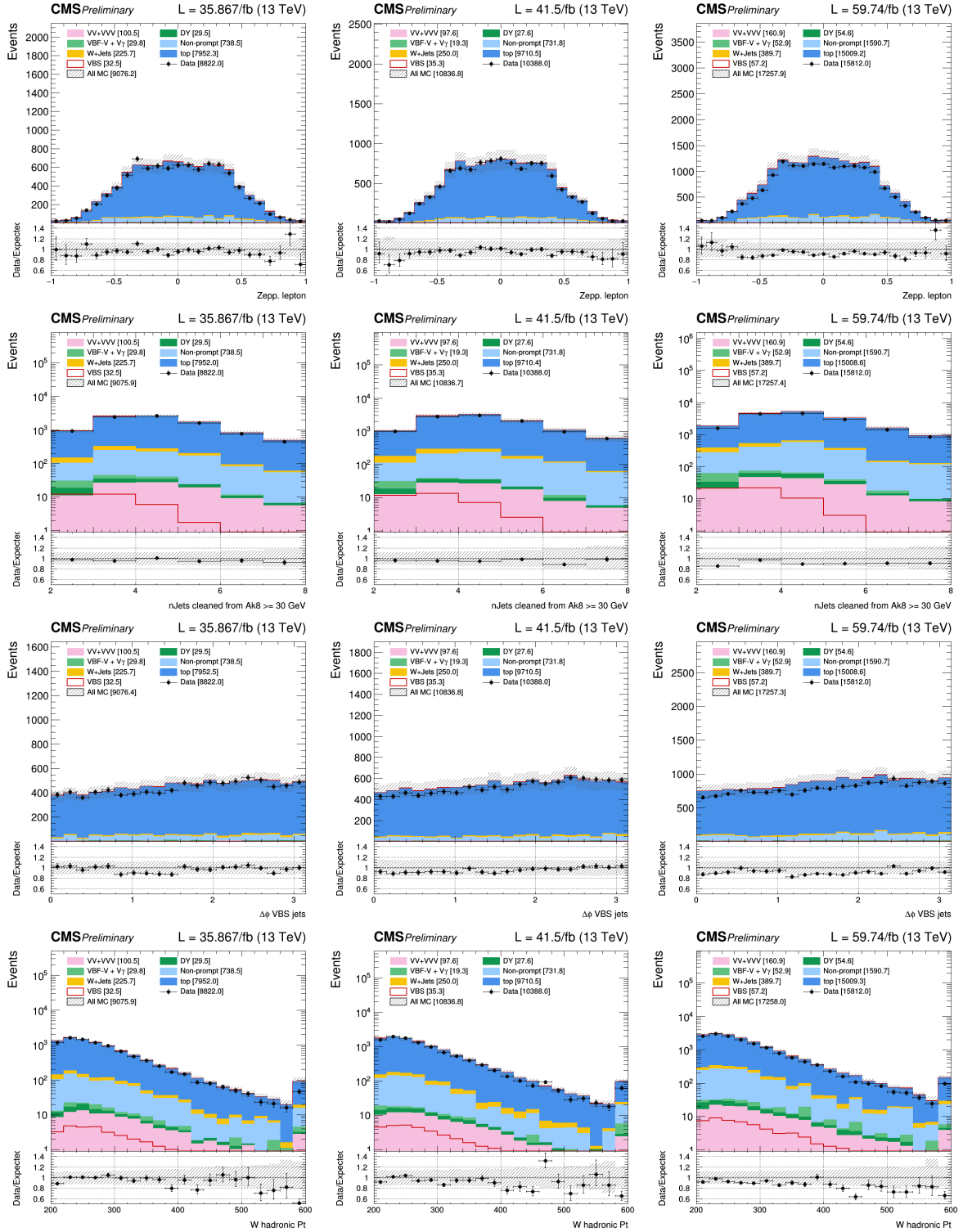


Figure 3.29: Boosted, top, muon control region. From left to right: 2016, 2017, 2018. From top to bottom: Lepton Zeppenfeld variable, number of jets with  $p_T > 30$  GeV,  $\Delta\phi_{VBS}$ , and V hadronic  $p_T$

### 3.5.2 W+jets background

A control region for the W+jets background is defined requiring the absence of any loosely b-tagged AK4 jet with  $p_T > 20$  GeV and inverting the requirement on the hadronically decaying V boson with respect to the signal region:  $40 \leq M_V \leq 65$  or  $M_V \geq 105$  GeV is imposed both in resolved and  $40 \leq M_V \leq 70$  or  $M_V \geq 115$  boosted categories.

The W+jets background MC is simulated at LO and with separated HT energy bins. This MC shows disagreements with the data, in particular in the kinematics of the selected jets in the resolved region. Control plots for several kinematic variables are plotted in the resolved category for the muon channel only, for brevity, in Figures 3.30, 3.31, and 3.32. The corresponding plots for the boosted muon category can be found in Figures 3.33, 3.34, and 3.35.

The  $p_T$  distributions of the VBS-jets and V-jets are affected by data/MC discrepancies, as well as the  $p_T$  of the leptonic W, with the same trend. Also the  $\Delta\eta_{jj}^{VBS}$  variable is not perfectly described at low values. Since the discrepancies show the same trend in different regions and different years, they are attributed to the imprecise matrix element calculation, showering and hadronization simulation in the LO W+jets MC. To cope with this problem, a data-driven technique to estimate this background from the control region has been developed and described in detail in Section 3.8. The most precise generation existing for the analysis final state of one W boson and four partons has been used. Nevertheless, a sample generated at NLO QCD but with a lower multiplicity of final state partons has been tested as well, as shown in Sec. 3.8.5, and does not bring improvements for this analysis.



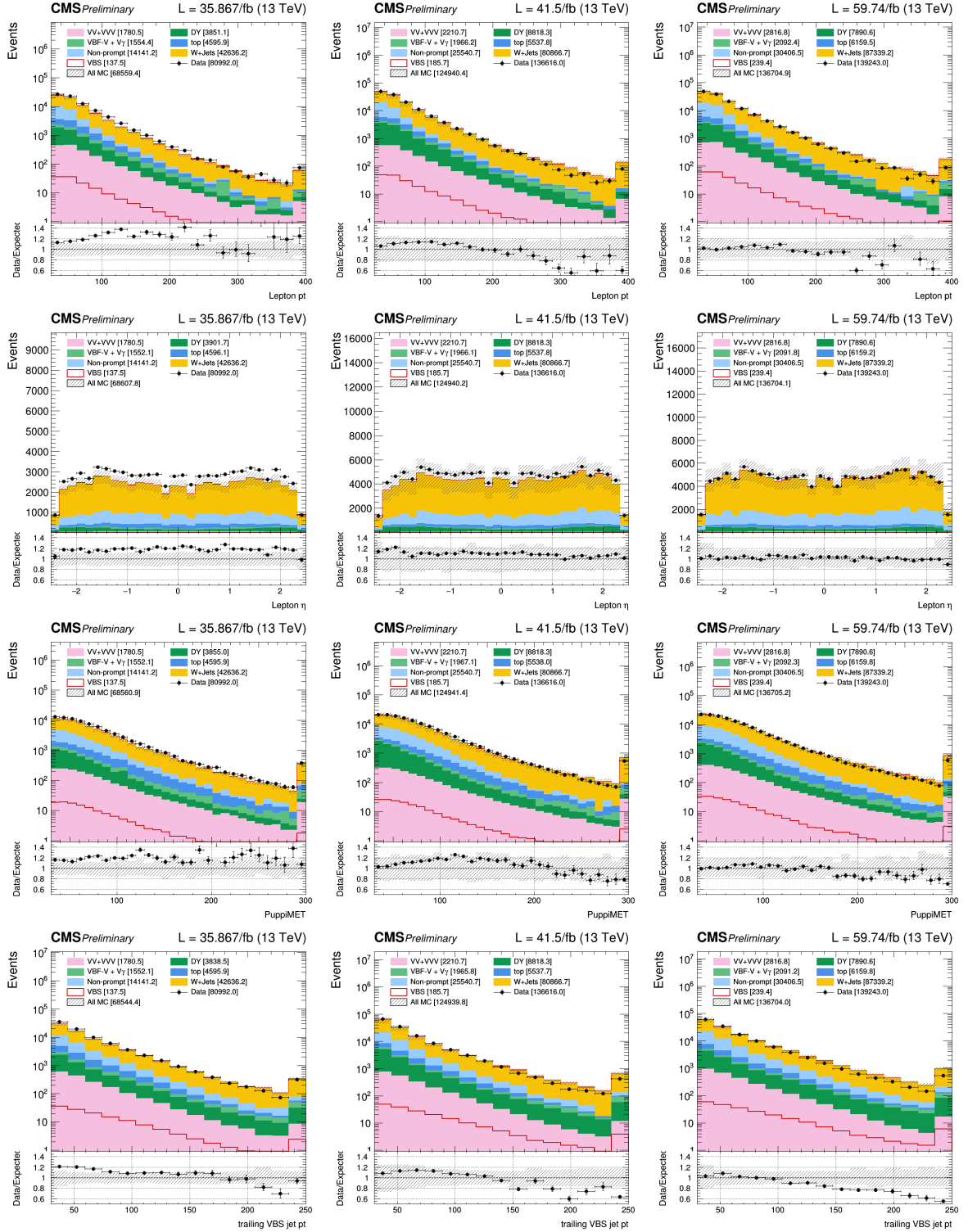


Figure 3.30: Resolved, W+jets, muon control region. From left to right: 2016, 2017, 2018. From top to bottom: Lepton  $p_T$ ,  $\eta$ , PuppMET, and Pt of the trailing VBS-jet



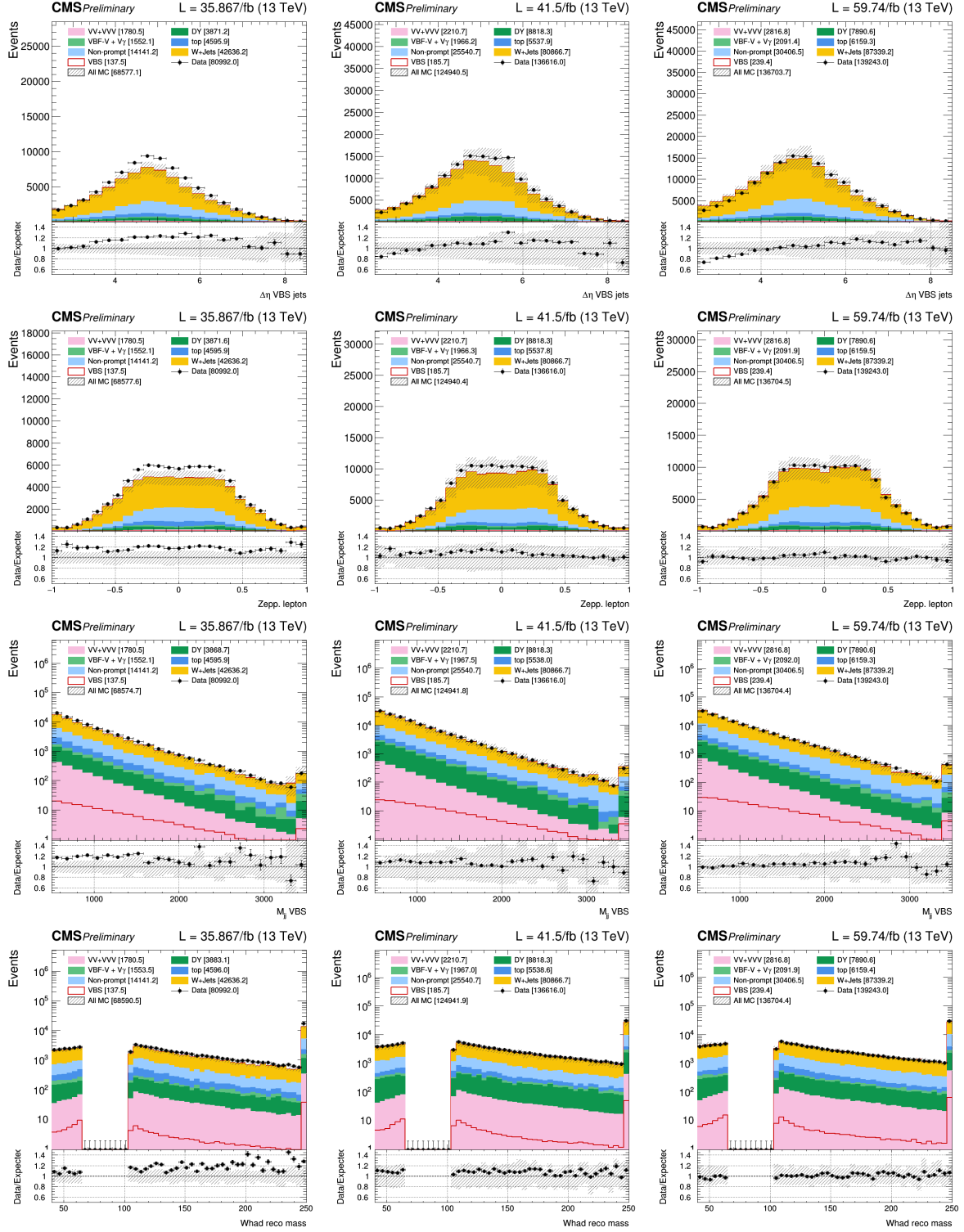


Figure 3.31: Resolved, W+jets, muon control region. From left to right: 2016, 2017, 2018. From top to bottom:  $\Delta\eta_{\text{VBS-jets}}$ , Zeppenfeld variable of the lepton  $Z_{\text{lep}}$ , invariant mass of VBS-jets  $m_{jj}^{\text{VBS}}$ , invariant mass of the V hadronic jets

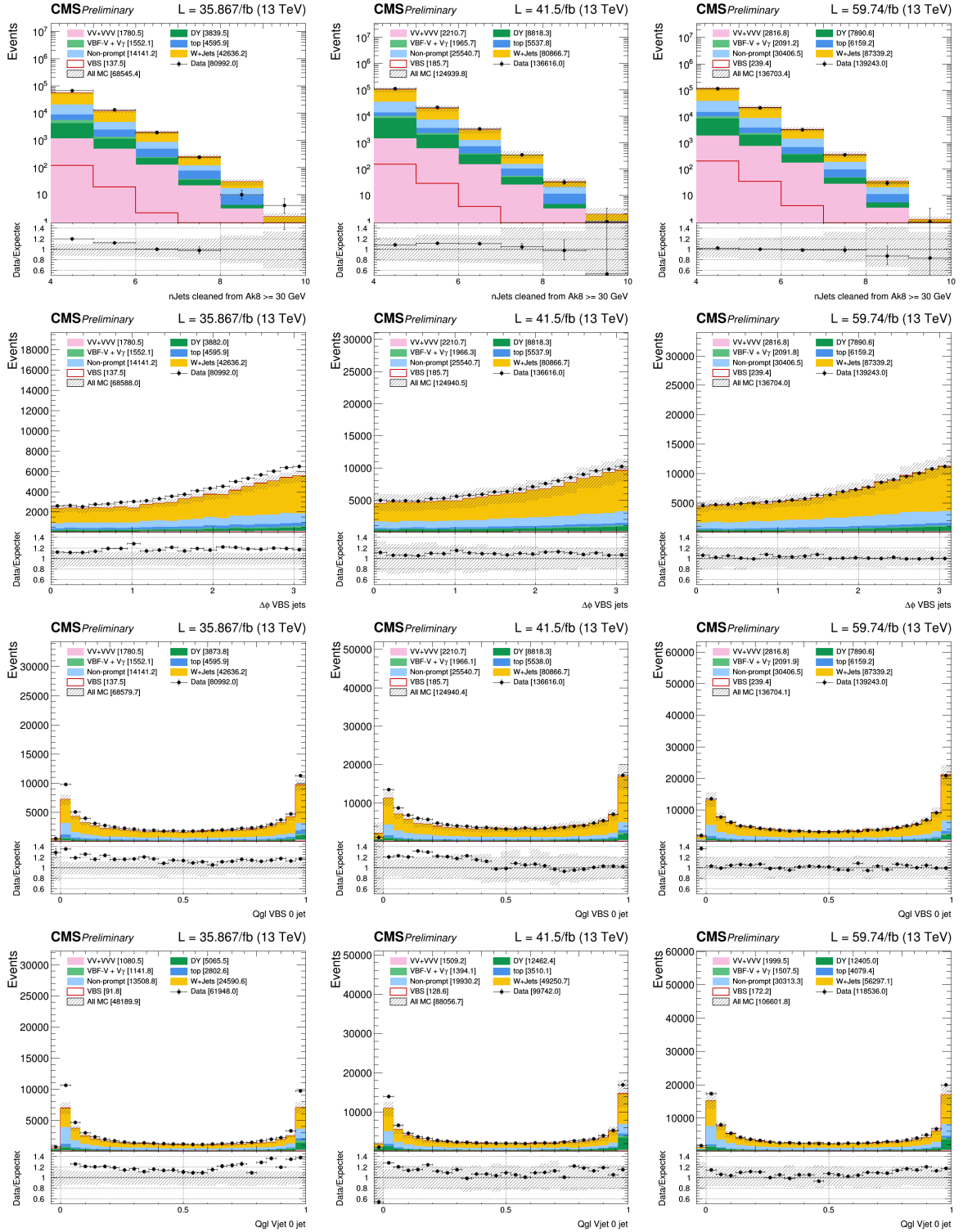


Figure 3.32: Resolved, W+jets, muon control region. From left to right: 2016, 2017, 2018. From top to bottom: number of jets with  $p_T > 30 \text{ GeV}$ ,  $\Delta\phi_{\text{VBS}}$ , quark-gluon likelihood of the leading VBS jet and leading V hadronic jet.

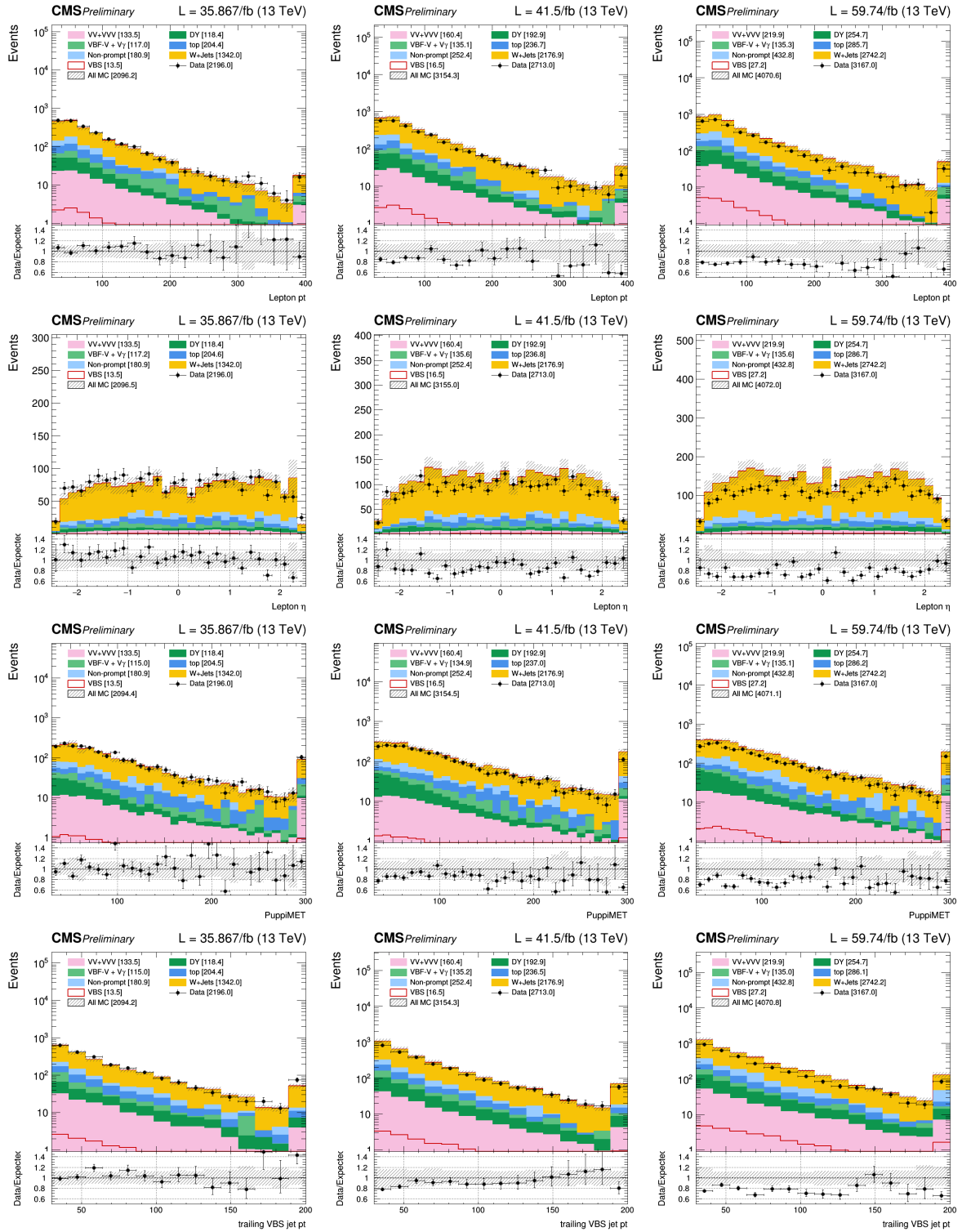


Figure 3.33: Boosted, W+jets, muon control region. From left to right: 2016, 2017, 2018. From top to bottom: Lepton  $p_T$ ,  $\eta$ , PuppMET, and Pt of the trailing VBS-jet

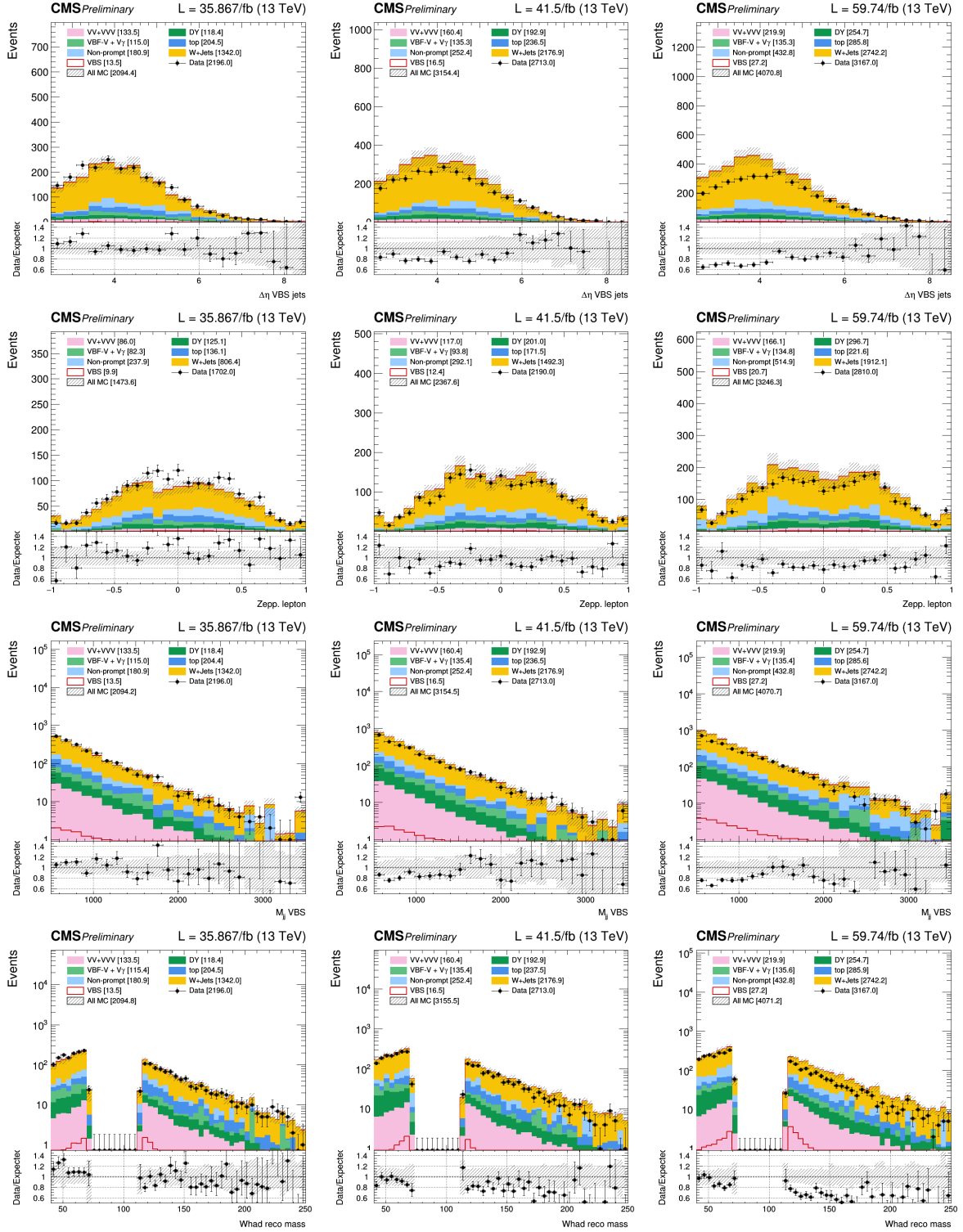


Figure 3.34: Boosted, W+jets, muon control region. From left to right: 2016, 2017, 2018. From left to right: 2016, 2017, 2018. From top to bottom:  $\Delta\eta_{jj}^{\text{VBS}}$ , Zeppenfeld variable of the lepton  $Z_{\text{lep}}$ , invariant mass of VBS-jets  $m_{jj}^{\text{VBS}}$ , mass of the boosted V

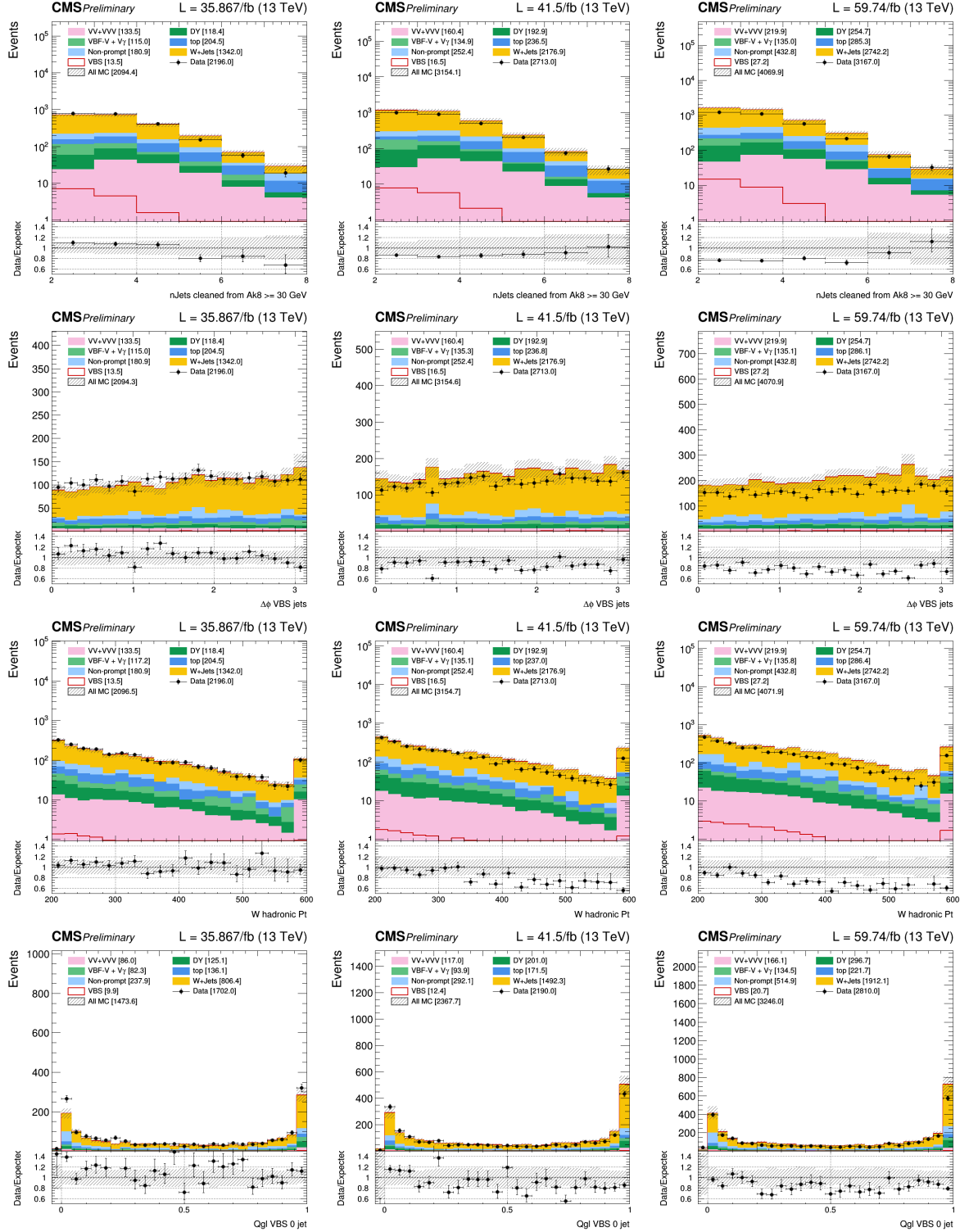


Figure 3.35: Boosted, W+jets, muon control region. From left to right: 2016, 2017, 2018. From top to bottom: number of jets with  $p_T > 30$  GeV,  $\Delta\phi_{\text{VBS}}$ , and V hadronic  $p_T$ , and quark-gluon likelihood of the leading VBS jet.

### 3.5.3 Non-prompt background

The QCD multijet background, which may enter the signal region as nonprompt leptons, is estimated from data by measuring the probability for a loosely defined reconstructed lepton originating from a jet to be misidentified as a tightly reconstructed lepton in a phase space region outside the analysis region. The QCD-enriched region is defined by the presence of at least one lepton with the same  $p_T$  requirement as for the rest of the analysis,  $p_T^{\text{miss}} < 20 \text{ GeV}$ ,  $M_W^T < 20 \text{ GeV}$ , at least one AK4 jet in the event with  $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} > 1$  from the lepton. The isolated jet is considered in the nonprompt estimation as the jet that recoils against the jet that is wrongly identified as a lepton: therefore the  $p_T$  of this recoil jet is a proxy for the  $p_T$  of the mis-identified jet.

The contribution from EW processes with a real lepton is subtracted from this QCD enriched phase space region by means of W+jets and DY MC events. The top quark background contribution in this phase space is found to be negligible. The method is data-driven since the control sample is selected in data, and the extrapolation factor is measured with data as well. The fake rate (see [102] for a complete definition of the formulation ) is then computed in bins of  $p_T$  and  $\eta$  of the lepton:

- $\eta$  binning  $-2.5, -2.15, -1.8, -1.479, -1.0, -0.5, 0., 0.5, 1.0, 1.479, 1.8, 2.15, 2.5$
- $p_T$  binning for electrons  $25, 30, 35, 40, 45, 50, 60, 100 \text{ GeV}$
- $p_T$  binning for muons  $25, 30, 35, 40, 45, 100 \text{ GeV}$

The fake rates are also computed varying the recoil jet  $p_T$  in order to build systematic variations: in fact, varying the recoil jet  $p_T$  requirement, different flavour compositions of the jet faked as a lepton are selected. The rates extracted with recoil jet  $p_T = 35 \text{ GeV}$  are considered as the nominal ones and are shown for the different Run II periods in Fig. 3.36.

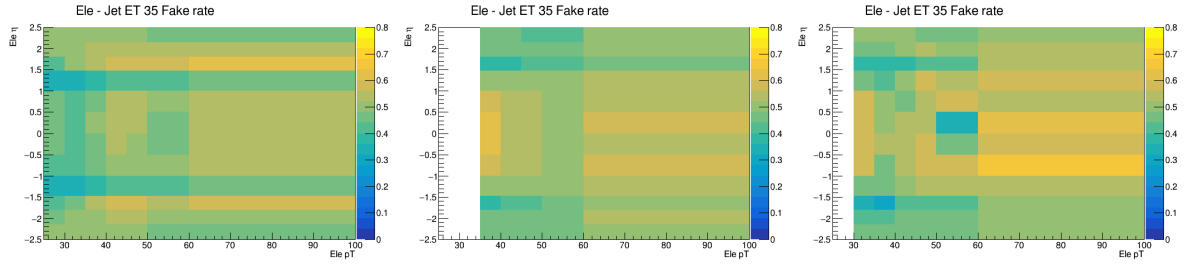


Figure 3.36: Fake rates for electrons with recoil jet  $p_T = 35 \text{ GeV}$  as a function of  $p_T$  (x) and  $\eta$  (y). From left to right: 2016, 2017, 2018

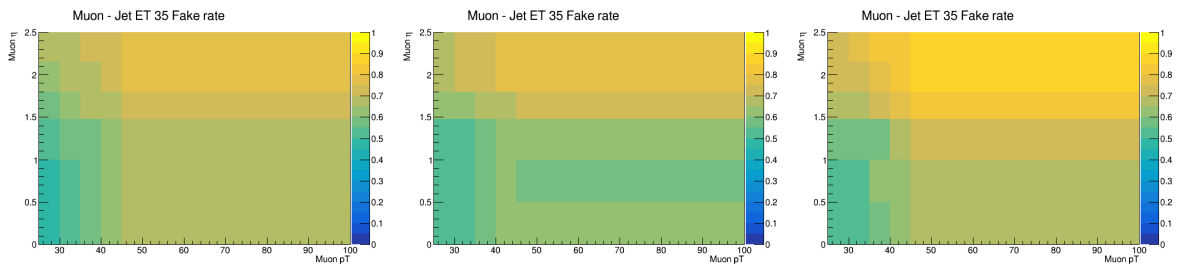


Figure 3.37: Fake rates for muon with recoil jet  $p_T = 35 \text{ GeV}$  as a function of  $p_T$  (x) and  $\eta$  (y). From left to right: 2016, 2017, 2018

### 3.5.3.1 Non-prompt closure test

The non-prompt estimate is checked relaxing the  $M_W^T$  cut but keeping a PUPPI  $\vec{p}_T^{\text{miss}} < 30$  GeV cut to select events orthogonally to the analysis regions. Moreover 4 jets with a least 30 GeV are requested to test it in a region nearby the signal. As Fig. 3.38 and Fig. 3.39 show the agreement between data and MC is satisfactory. The uncertainty drawn on the plot corresponds to the statistical and systematic uncertainty on the fake rate.



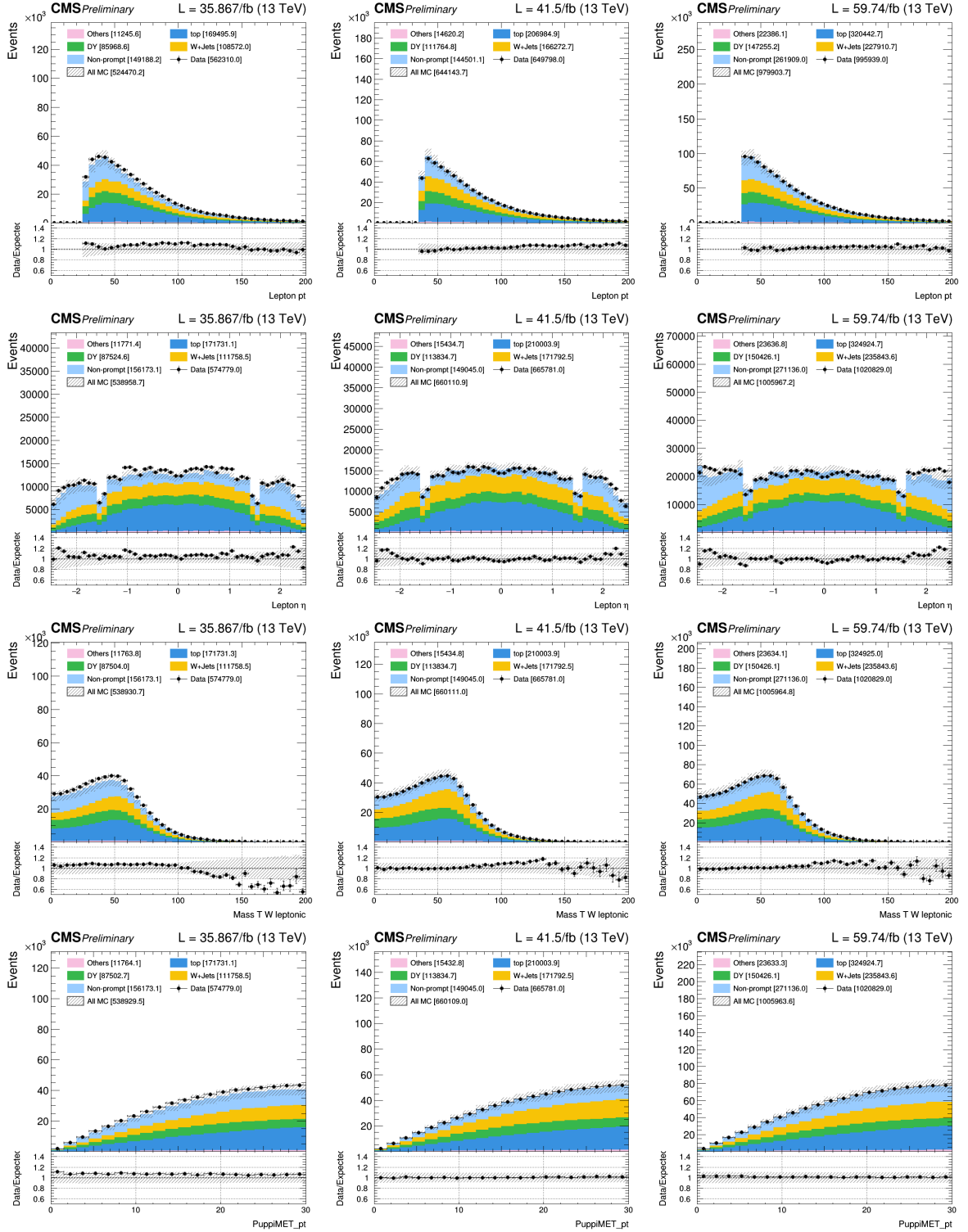


Figure 3.38: Electron channel:  $PuppiMET < 30\text{ GeV}$  and at least 4 jets with  $30\text{ GeV}$  control region.



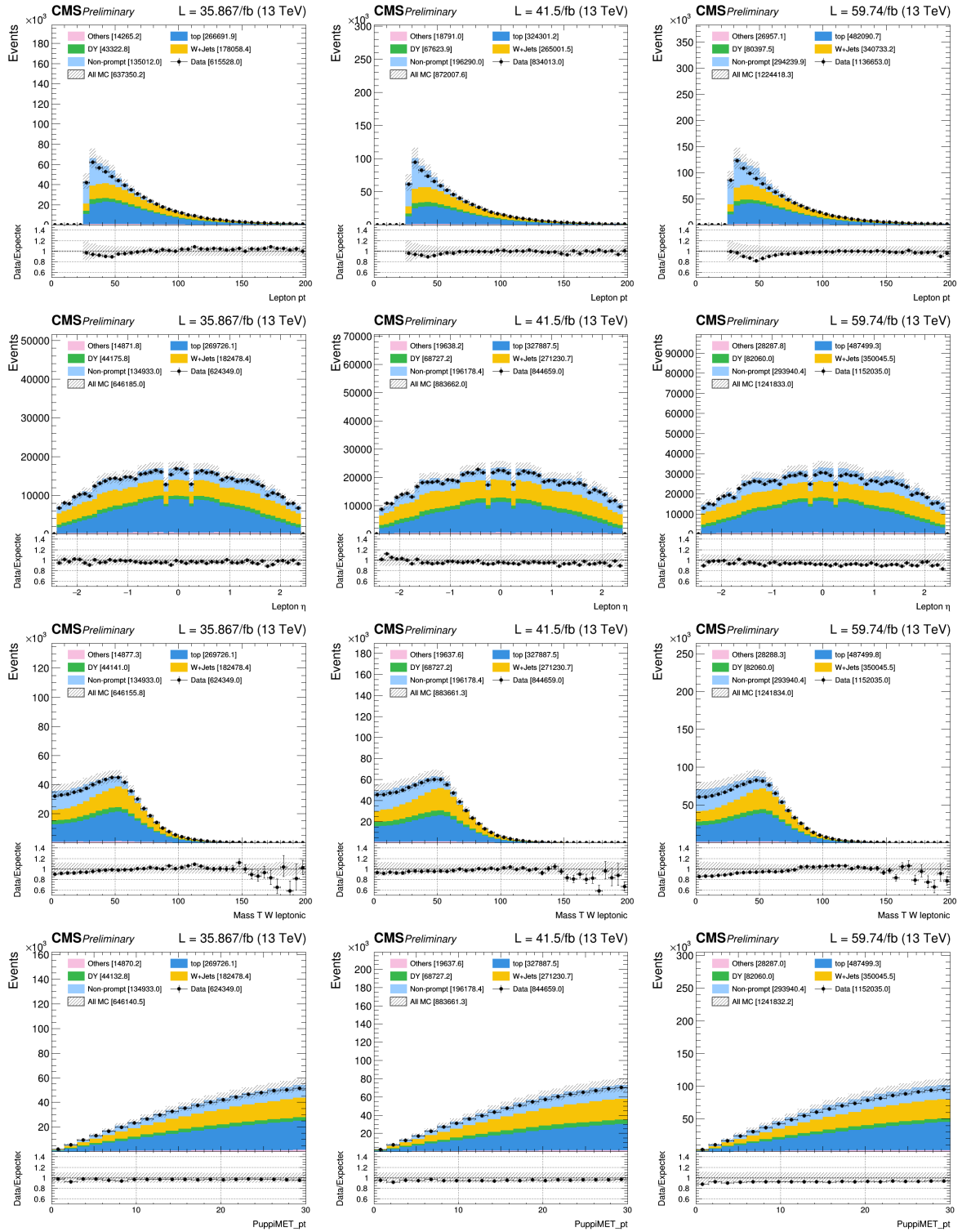


Figure 3.39: Muon channel:  $PuppiMET < 30$  GeV and at least 4 jets with 30 GeV control region.

### 3.6 Quark Gluon Likelihood calibration

The Quark Gluon Likelihood (QGL) is an observable built to discriminate between jets originating from light quarks and by a gluon. The tagger is a likelihood discriminant constructed out of three variables, calculated using the particle flow candidates withing the jet.

**ptD** jet energy variable, defined as  $\frac{\sqrt{(\sum_i p_{T,i}^2)}}{\sum_i p_{T,i}^2}$  with the sum over all particle flow candidates

**multiplicity** the total number of particle flow candidates reconstructed within the jet

**axis2** minor axis RMS in the  $\eta - \phi$  plane of the particle flow candidates

For the computation of all variables, charged hadron particle flow candidates are neglected unless they are found to be compatible with a high purity track and the event primary vertex, while neutral hadron particle flow candidates are restricted to those exceeding 1 GeV. At least three particles are needed to calculate the variables and the likelihood. In order to correct for dependencies on the transverse momentum and pseudorapidity of the jet, the likelihood p.d.f. sets are binned in categories of  $p_T$  (between 20 and 6500 GeV) and  $\eta$  (up to 4.7). The forward region is additionally binned in PU transverse momentum density,  $\rho$ , to correct for pile-up dependencies. The gluon and quark components are not perfectly described by the Pythia8 parton shower and therefore need a correction to account for data/MC discrepancies.

In this analysis, the QGL variable of VBS tag jets and  $V_{\text{had}}$  jets are used, therefore the correction needs to be applied on multiple jets. The QGL variable is powerful for the VBS signal extraction because in the VBS signal all the jets originate from quarks while in the  $W$ +jets background there is a large contribution from gluon jets.

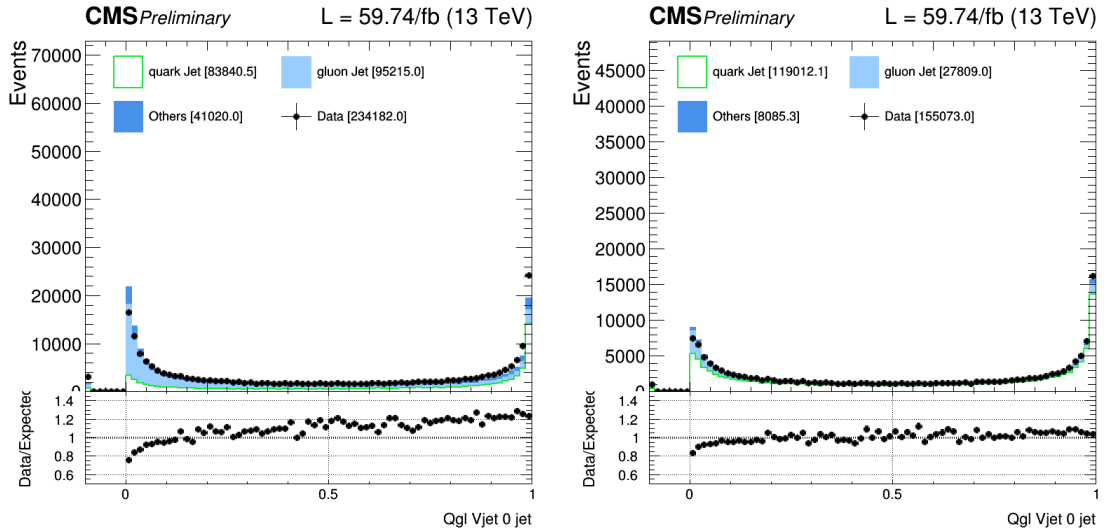


Figure 3.40: QGL variable for 1st V jet in resolved  $W$ +jets (left) and top (right) regions.

Fig.3.40 shows the QGL variable for the 1st  $V_{\text{had}}$  jet in resolved  $W$ +jets (left) and top quark (right) regions. The MC are split in gluon and quark components looking at the parton flavour of the parton gen matched to the 1st  $V_{\text{had}}$  jet. The non-prompt events are categorized as “Others” since they are data-driven and the parton flavour is unknown. The top quark control region is purer in quarks, whereas the  $W$ +jets one has a larger gluon component. The shape of

the QGL variable has larger discrepancies in the gluon purer region. The shape of the discrepancy is similar if the components are analyzed using another jet parton flavour for the splitting. The discrepancy is present in all the range of jet  $\eta$  as shown in Fig. 3.41.

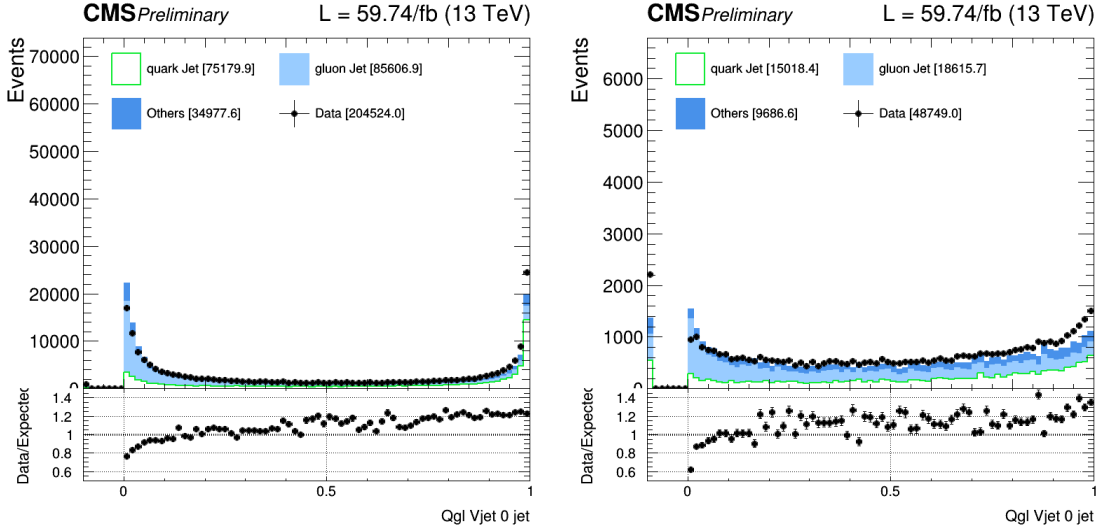


Figure 3.41: QGL variable for 1st V jet in resolved W+jets regions, at  $\eta < 2$  (left) and  $\eta > 2$  (right).

### 3.6.1 Correction strategy

In order to be able to correct the QGL shape without changing the normalization a quantile morphing procedure has been implemented, a strategy similar to the one used by CMS for the electron shower shape corrections.

In general, given two distributions  $f$  and  $g$  and their cumulative distribution functions  $F, G$ , an optimal mapping  $m$  from the first distribution to the second is defined as:

$$m(x) = G^{-1}(F(x)) \quad (3.6)$$

The goal is to compute different mappings for the gluon component and the quark component in order to fix the overall QGL distribution. The target distribution is the one needed to match the normalized data minus the other components.

To compute the mapping we need a gluon-like pure region and a quark pure one. Moreover we have at least 4 jets in our events, each with a different  $p_T$  spectrum, and we can choose between them to split the MC sample in gluon and quark components. The final correction will be then applied on all the jets. Different corrections are computed for 2016 and 2017-2018 given the different Pythia8 tune used. Finally since we want to model correctly the QGL for different  $p_T$  ranges and  $\eta$  regions, 4 corrections will be derived for each component: at  $\eta < 3$  and  $\eta > 3$  and for  $p_T < 75$  GeV and  $p_T > 75$  GeV. Figure 3.43 and Fig. 3.44 show the purity of the gluon and quark components for the first 4 jets in the 4  $p_T$  and  $\eta$  bins.

The following procedure has been carried out:

- Extract the morphing function in the purest regions available:
  - for gluons at  $\eta < 3$  using 4th jet splitting (in WJets CR)

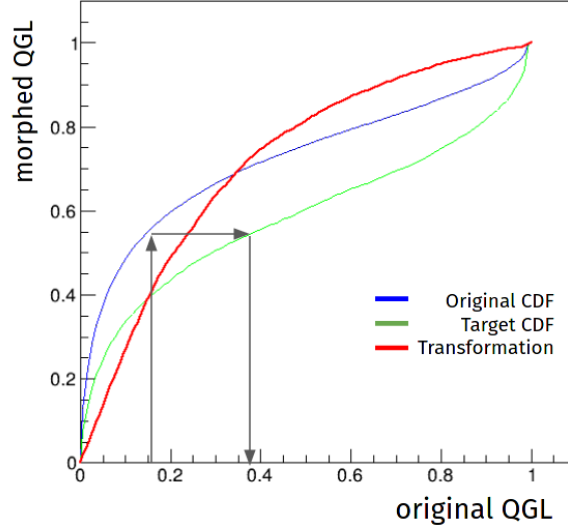


Figure 3.42: Quantile morphing method.

- for quarks at  $\eta > 3$  using 2nd jet splitting (in Top CR)
- Apply the derived morphing function on all the jets both in the W+jets and the top CR but only in the  $\eta$  region where the calculation has been extracted. Correct gluons at low eta and quarks at high eta
- Compute residual corrections on the other jet types:
  - for gluons at  $\eta > 3$  using 2st jet splitting (in WJets CR)
  - for quarks at  $\eta < 3$  using 2nd jet splitting (in Top CR)
- Apply all the corrections and check the residual trasformations that would be needed in each category
- Extract the uncertainty looking at the residual discrepancy

This method is always using the top control region to correct the quark-jets and the W+Jets region to correct the gluon jets. As an example in Fig. 3.45 the procedure for gluon in  $\eta < 3$ ,  $p_T < 75$  GeV is shown. In Fig.3.46 the same is shown for the quark component. The residual trasformations are close to unity and are used to extract a systematic uncertainty for the morphing function.

### 3.6.2 Application of the QGL morphing

The QGL morphing for gluons and quarks in the 4  $\eta$  and  $p_T$  bins are then applied in the W+Jets and top control region and the final shape is checked. Fig.3.47 shows the QGL variable in 2018 of the leading  $V_{\text{had}}$  jet in the top control region, the trailing  $V_{\text{had}}$  jet in W+jets region and leading VBS tag jet in the W+jet control region. Fig.3.48 shows the same for 2016 where the applied procedure has been the same but the correction factors are different because of the different Pythia8 tune. The shape is improved in both periods.

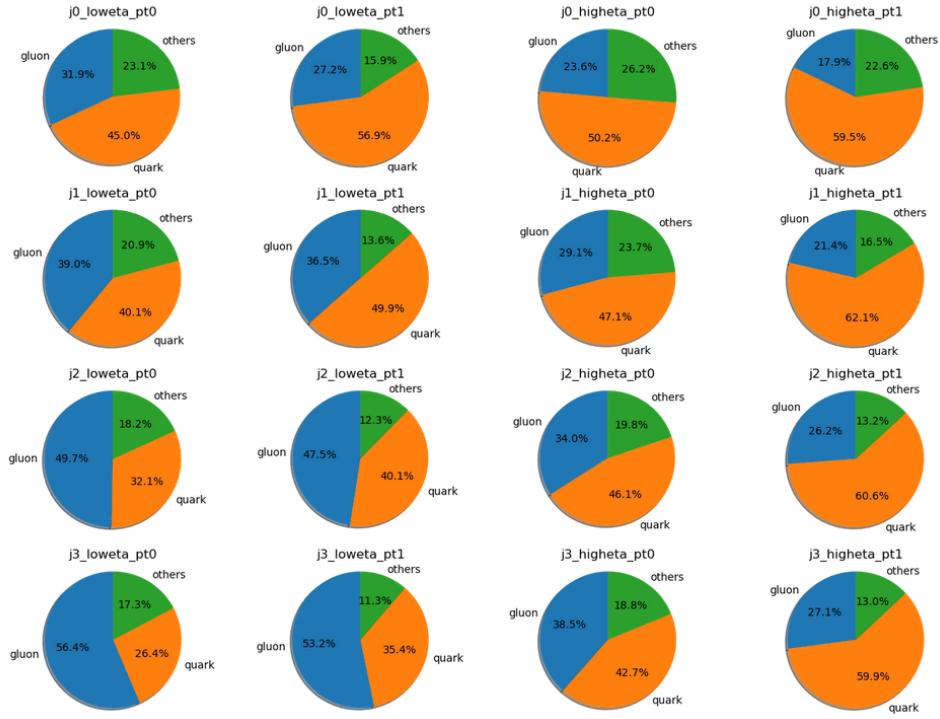


Figure 3.43: Jet flavour composition for the first 4 jets in top CR at  $\eta < 3$  and  $\eta > 3$  (columns) and for  $p_T < 75$  GeV (pt0) and  $p_T > 75$  GeV (pt1)

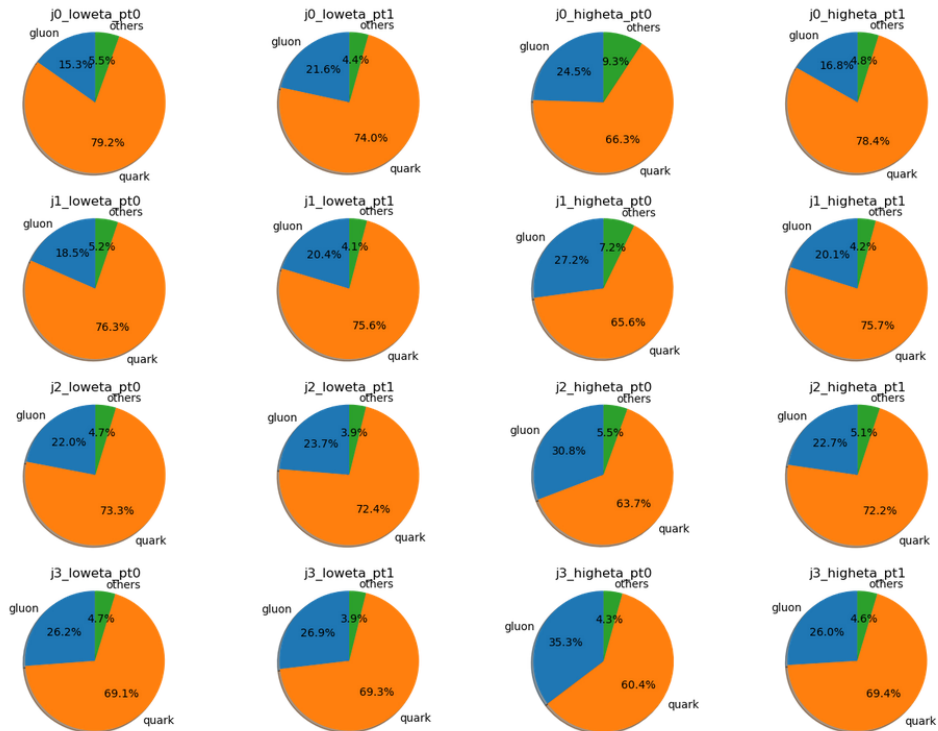


Figure 3.44: Jet flavour composition for the first 4 jets in W+jets CR at  $\eta < 3$  and  $\eta > 3$  (columns) and for  $p_T < 75$  GeV (pt0) and  $p_T > 75$  GeV (pt1)

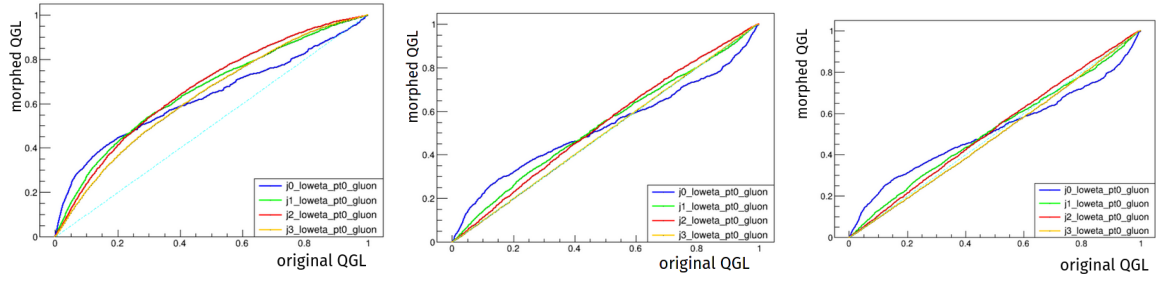


Figure 3.45: Correction to gluon jets in W+jets CR in the  $\eta < 3$ ,  $p_T < 75$  GeV bin. From left to right: initial gluon trasformations, trasformation needed after application of the gluon correction, finally the residual trasformation after quark correction.

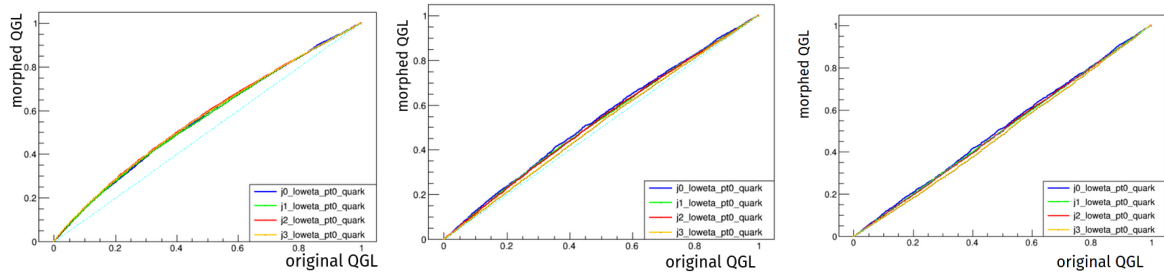


Figure 3.46: Correction on quark jets in Top CR in the  $\eta < 3$ ,  $p_T < 75$  GeV bin. From left to right: initial quark trasformations, trasformations needed after application of the gluon correction, finally the residual trasformation after quark correction.

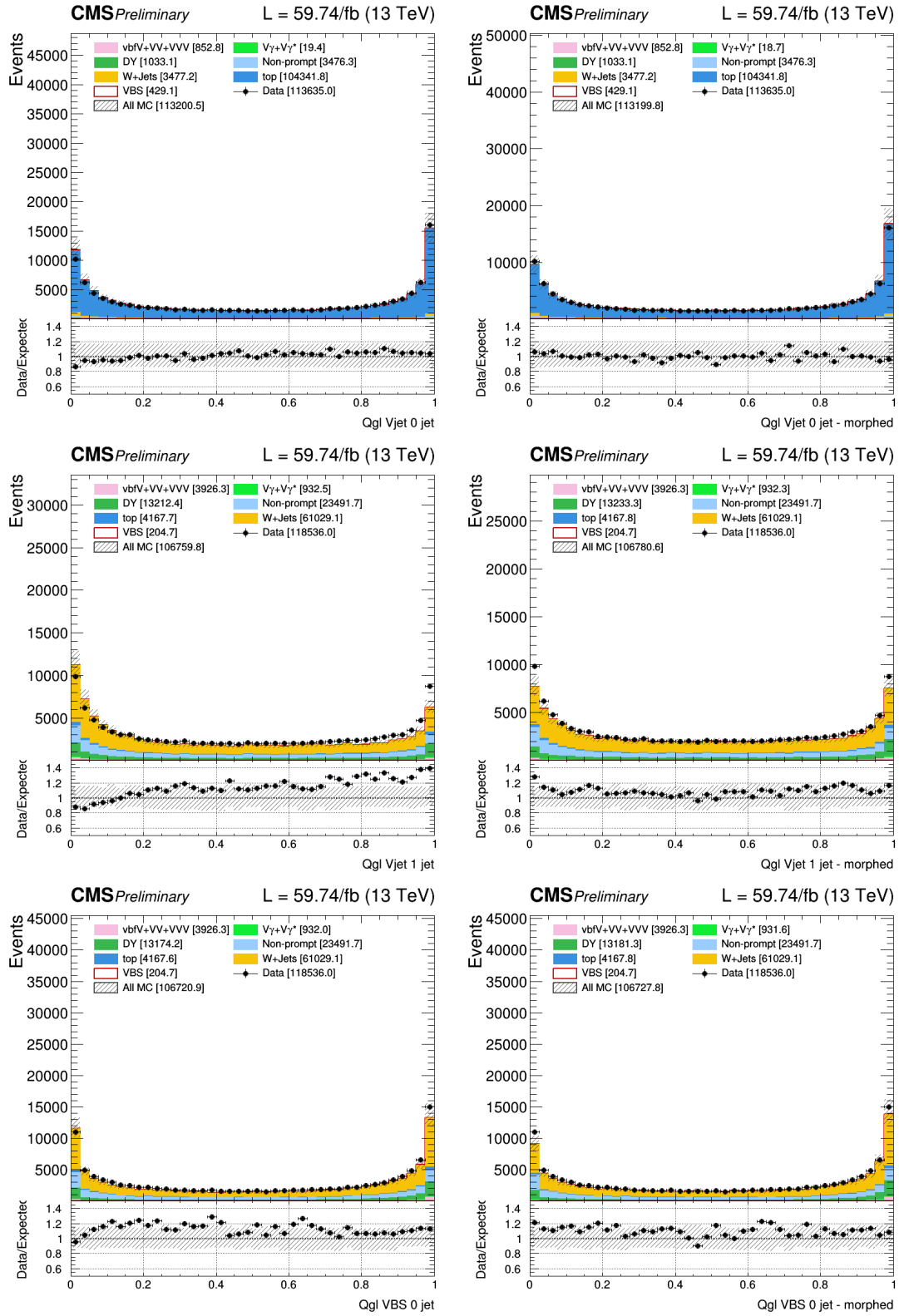


Figure 3.47: On the left: QGL variables without correction, on the right after morphing in 2018. From top to bottom: leading  $V_{had}$  jet in the top quark region, the trailing  $V_{had}$  jet in W+jets region and the leading VBS jet in W+jet region

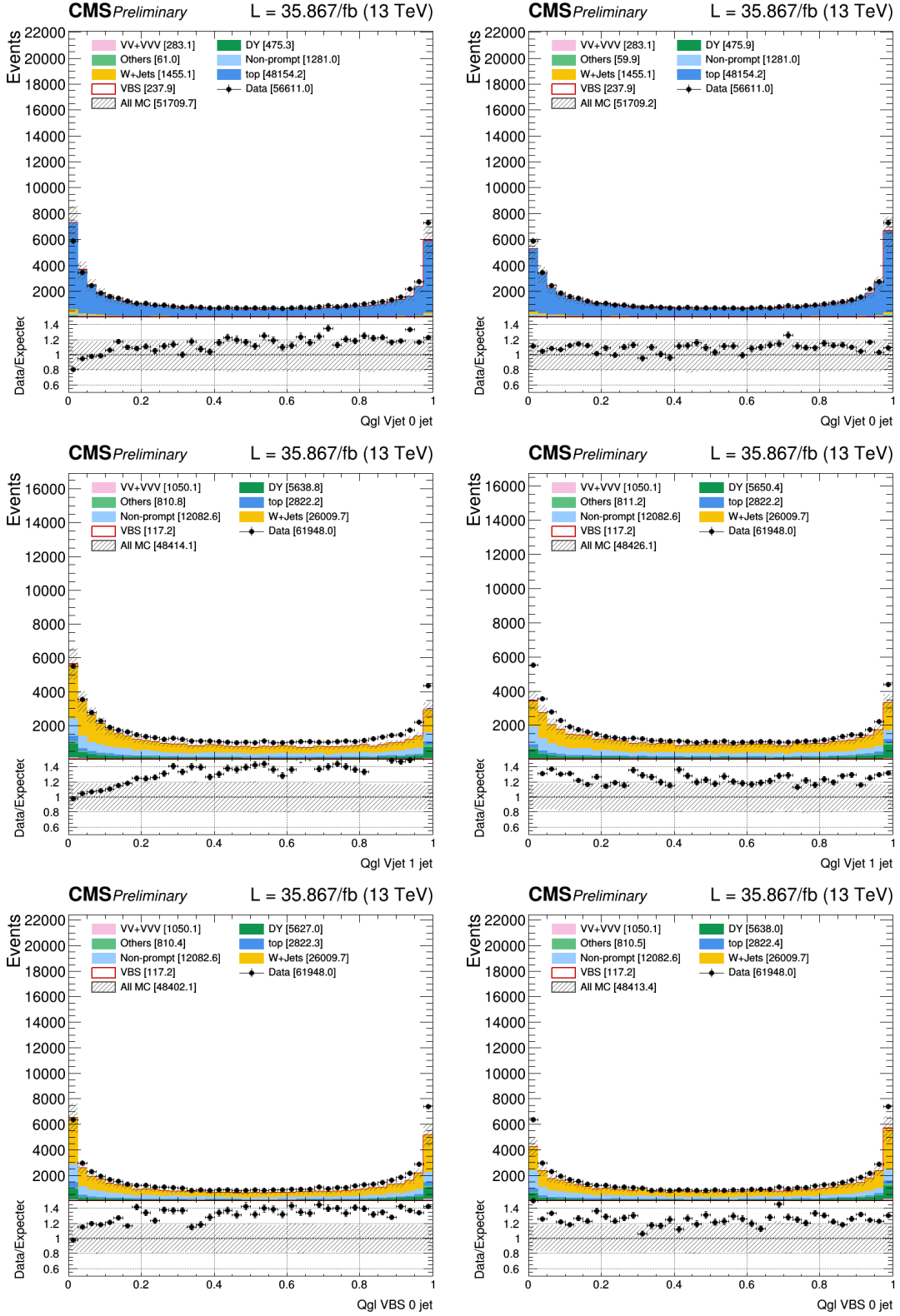


Figure 3.48: On the left: QGL variables without correction, on the right after morphing in 2016. From top to bottom: leading  $V_{\text{had}}$  jet Top region, the trailing  $V_{\text{had}}$  jet in W+jets region and the leading VBS jet in W+jet region



### 3.6.3 QGL morphing uncertainties

The uncertainty for the morphing functions is extracted from the residual transformations needed after the first round of morphing. Since we are using multiple jets for the calculation as shown above, the envelope of the residual transformations is taken to build the *morphUp* and *morphDown* shapes. This is a quite conservative approach resulting in a large a-priori uncertainty. Since their morphing functions are different, the gluon and quark uncertainties at high  $\eta$  and low  $\eta$  are uncorrelated in the fit. The high and low  $p_T$  morphing functions are separated but kept correlated in the fit. The final morphing function are drawn in Fig. 3.49 along with their uncertainty envelope.

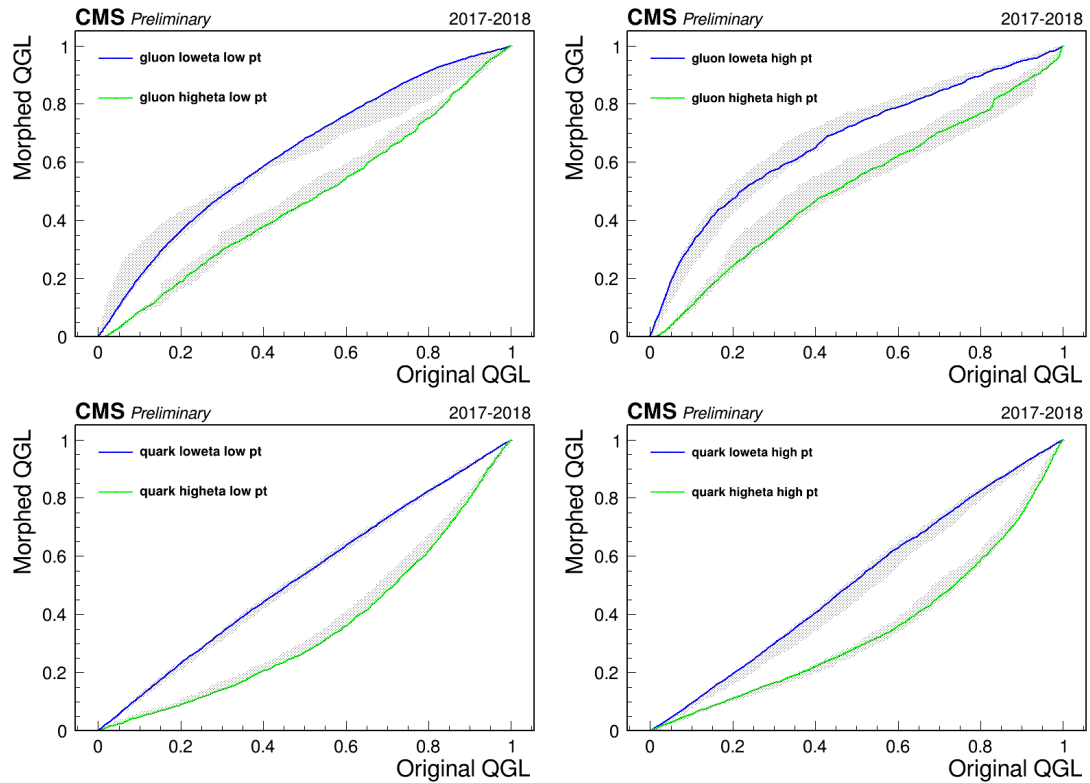


Figure 3.49: QGL morphing functions and their uncertainties for gluons (first line) and quarks (second line) for the 2017-2018 period.

## 3.7 Signal discrimination

### 3.7.1 Strategy

Some of the variables described show differences between the signal and the backgrounds, but none of them alone can be used as a powerful discriminant; therefore a multi-variate analysis technique is needed. Figures 3.50 and 3.51 show the normalized distribution of the signal and background respectively in the resolved and boosted category for some of the most discriminating kinematic observables:  $m_{jj}^{\text{VBS}}$ ,  $\Delta\eta_{jj}^{\text{VBS}}$ ,  $Z_{\text{lep}}$ , and the trailing VBS jet  $p_T$ . In order to achieve the best performance possible, a feed-forward deep neural network (DNN) has been chosen as the discriminator, built in the Keras/TensorFlow framework. Its output is a floating point number in the range  $[0, 1]$  and is referred to as *DNN output*.

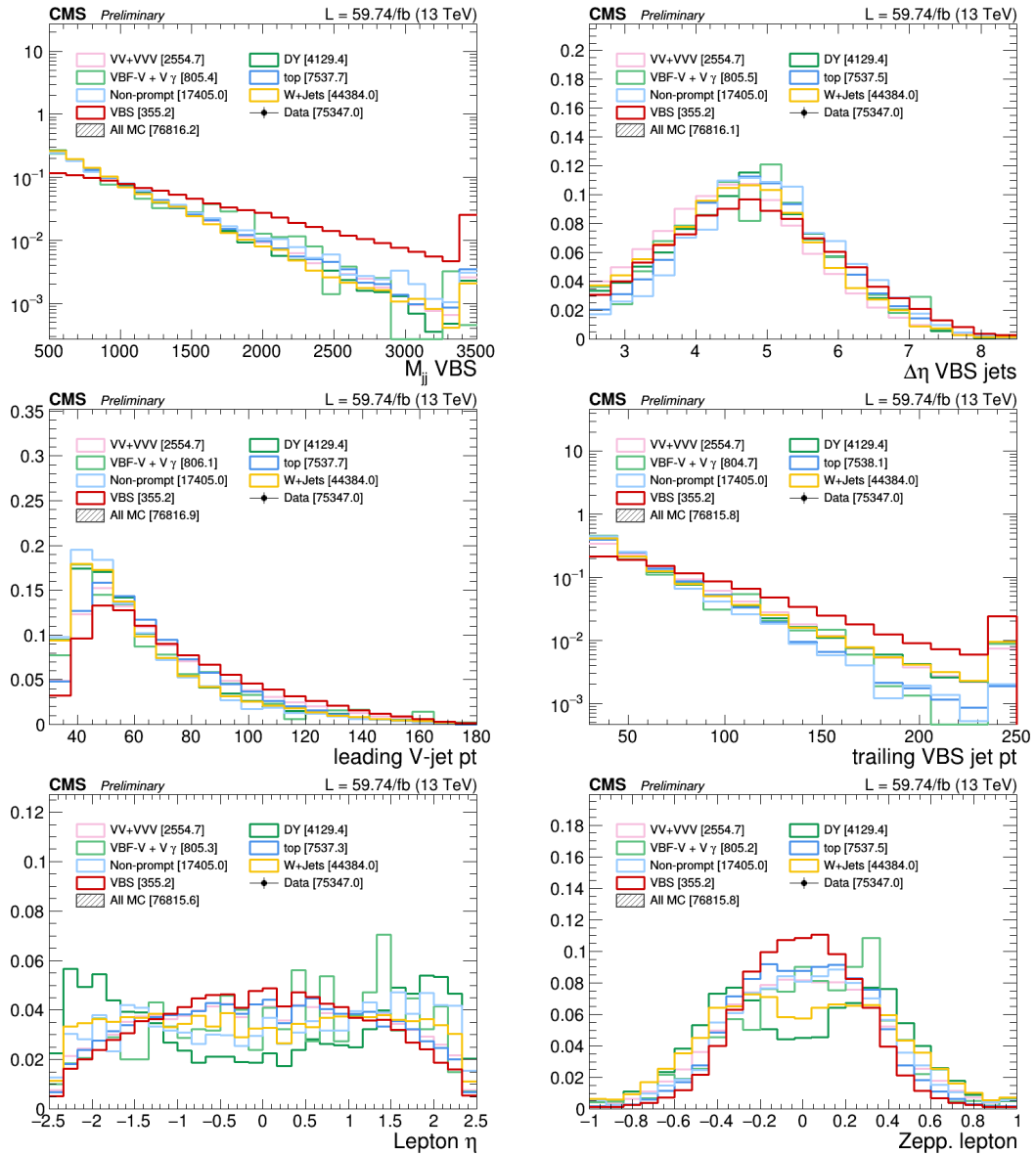


Figure 3.50: Shape comparison of background and signal processes in 2018 for muon resolved channel. Lepton selections and b-jet veto are applied. Plots in the signal region.

A single DNN is trained to discriminate the signal against all the backgrounds concurrently.

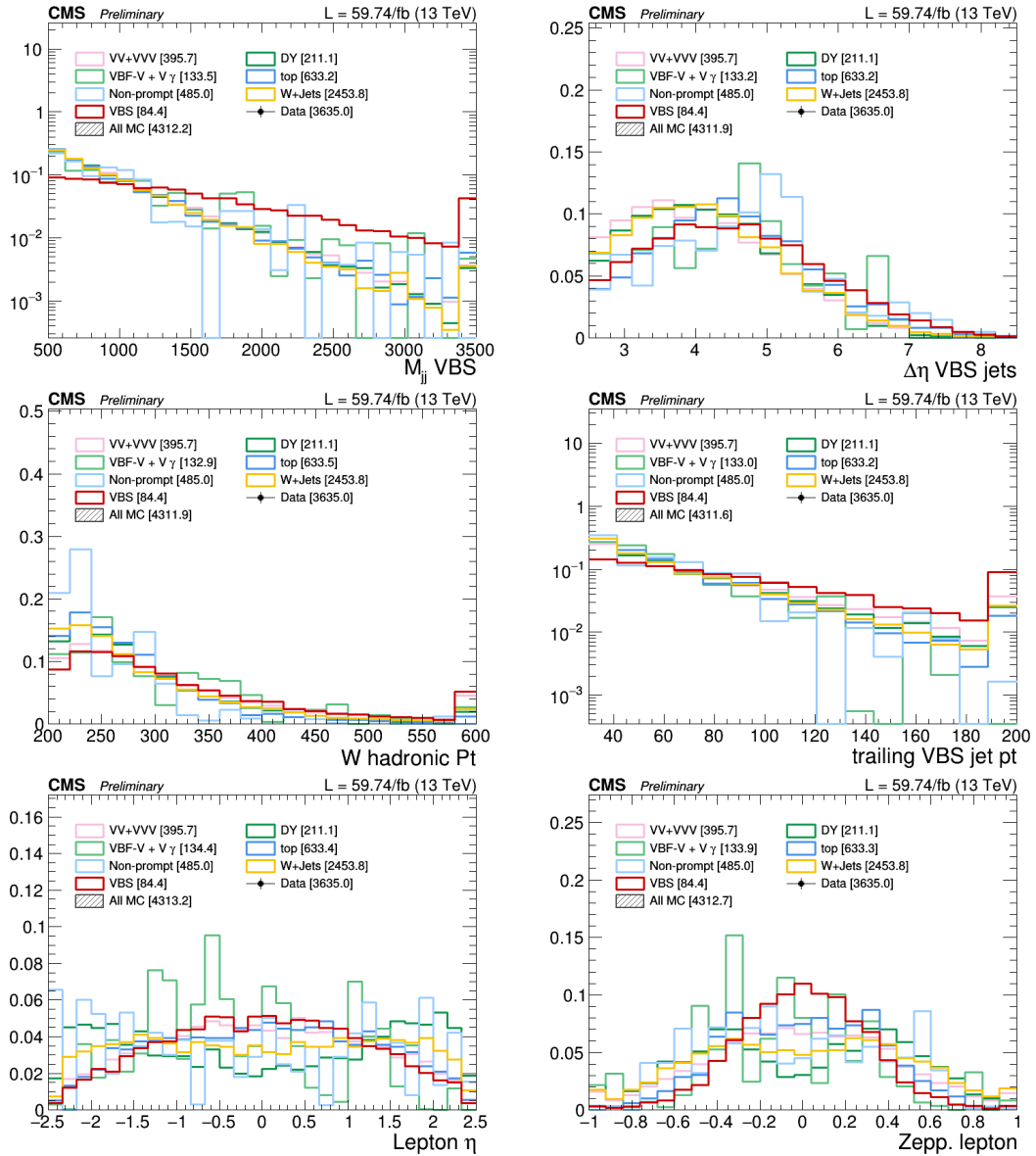


Figure 3.51: Shape comparison of background and signal processes in 2018 for muon boosted channel. Lepton selections and b-jet veto are applied. Plots in the signal region.

This is achieved creating a single sample containing events from all background processes, with the correct cross-section, selection and scale factor coefficients applied so that the relative weight between the backgrounds is corresponding to the expectation. This allows the DNN not to learn to discriminate the signal against a minor background, while losing discrimination power against the main backgrounds.

DNNs require to have balanced datasets as inputs for supervised learning, otherwise biases in the performance may be introduced. This means that the total number of weighted events of the signal dataset should be the same as that of the background datasets combined. This is achieved increasing the weight of the signal samples in the training.

The datasets are split into training (80 %) and validation (20 %) parts: the validation dataset is used to monitor the training for overtraining and compare the performance of different models

on an independent dataset. In order to be able to use all the events in the final fit, a Kolmogorov compatibility test is performed over the distribution of the DNN variable on the validation and training datasets. Models showing overtraining or failing the Kolmogorov test are discarded.

### 3.7.2 Architecture and training process

Two different fully connected DNNs have been trained for the resolved and boosted category. Only one model is trained for each category, merging the datasets of all the years for the training. This allows to have more samples for the training, and to avoid overtraining while being able to train larger regularized models. Since the task is binary classification, the network is optimized using the stochastic gradient descent algorithm to minimize the binary cross entropy loss on signal and background samples. The ReLU activation function is used for hidden layers, while the sigmoid is used for the output layer. The batch normalization [103] technique, applied on the input of all the layers, regularizes the network and increases the generalization power of the model. Moreover, L2 regularization [104] applied on the weights of the first two layers of the DNN has shown to be helpful in regularizing the network and avoid over-training. Both networks use the *Adam* gradient-descent optimizer [44], with a different learning rate for each weight and adapting it as the learning progresses keeping a memory of the previous steps. The learning rate is decreased during the learning process to improve the convergence. The training proceeds until the  $\Delta$  of the loss computed on the validation sample after 10 epochs is less than 0.0001. This early stopping prevents overtraining.

During the training, events are weighted as a function of  $m_{jj}^{\text{VBS}}$  as in Eq. 3.7 to enhance the discrimination power in the signal rich region. This weighting is applied to both signal and backgrounds as an optimization during the training and it is not applied in the valuation of the DNN.

$$w_{\text{train}} = 1 * (m_{jj}^{\text{VBS}} < 2000) + 2 * (m_{jj}^{\text{VBS}} > 2000) \frac{m_{jj}^{\text{VBS}}}{2000} \quad (3.7)$$

The size of samples available for the DNN training directly affects how complex the architecture can be. Having more training events allows for a larger network size with a reduced risk of overtraining. Table 3.11 shows the number of events available for each signal and background for the resolved and boosted category.

The architecture of the neural networks for the boosted and resolved category are summarized

| Sample | Resolved category |                    | Boosted category |                    |
|--------|-------------------|--------------------|------------------|--------------------|
|        | N. events         | N. weighted events | N. events        | N. weighted events |
| W+jets | 773089            | 184640.22          | 266456           | 10346.89           |
| top    | 268952            | 32902.33           | 25291            | 2911.63            |
| DY     | 292782            | 27222.56           | 63429            | 1092.98            |
| VV     | 473627            | 10193.86           | 739007           | 1609.37            |
| VBF-V  | 13041             | 4582.90            | 1184             | 426.16             |
| VBS    | 233531            | 1502.52            | 51967            | 329.24             |
| Vg     | 5618              | 2838.00            | 275              | 203.12             |
| VgS    | 4322              | 141.47             | 230              | 5.82               |
| VVV    | 4238              | 88.61              | 1077             | 24.95              |

Table 3.11: Total number of events available for the training in the resolved and boosted categories. The number of events weighted by the cross-section and efficiencies is shown in the rightmost column.

in Table 3.12 and 3.13 respectively. The latter benefits from a larger dataset, which allows for the use of a larger network and more input variables without causing overtraining.

Table 3.12: Architecture of the DNN model for the boosted category.

| Layer N. | Nodes | Regulatization                              | activation |
|----------|-------|---------------------------------------------|------------|
| Inputs   | 13    | scaled to $\bar{x} = 0$ and $\sigma(x) = 1$ |            |
| 1        | 64    | L2 + Batch normalization                    | ReLU       |
| 2        | 32    | L2 + Batch normalization                    | ReLU       |
| 3        | 32    | L2 + Batch normalization                    | ReLU       |
| 4        | 32    | L2 + Batch normalization                    | ReLU       |
| Output   | 1     |                                             | Sigmoid    |

Table 3.13: Architecture of the DNN model for the resolved category.

| Layer N. | Nodes | Regulatization                              | activation |
|----------|-------|---------------------------------------------|------------|
| Inputs   | 16    | scaled to $\bar{x} = 0$ and $\sigma(x) = 1$ |            |
| 1        | 64    | L2 + Batch normalization                    | ReLU       |
| 2        | 64    | L2 + Batch normalization                    | ReLU       |
| 3        | 64    | L2 + Batch normalization                    | ReLU       |
| 4        | 64    | L2 + Batch normalization                    | ReLU       |
| Output   | 1     |                                             | Sigmoid    |

### 3.7.3 DNN input variables

It has been observed that when the network dimension or the number of input variables increase, the DNN training is susceptible to overtraining. Nonetheless, greater DNN size and more input features increase the discrimination power of the network. An optimum value between network dimension and choice of input variables has to be reached.

The chosen approach involves fixing the network structure and training it with as many input variables as possible. If it overtrains, then reduce the number of input variables, discarding the less important ones. Repeat until the DNN does not overtrain. If the resulting performance is not satisfactory, increase the network size and repeat the process. The non-trivial task is the sorting of input variables based on their importance, accomplished using the *SHAP* (SHapley Additive exPlanations) routine, described in detail in [105], which describes a set of algorithms to compute or approximate Shapley values and provide global interpretation methods based on aggregations of Shapley values.

The Shapley value, coined 1953, is a method of cooperative game theory and was designed to assign value to players depending on their contribution to the total payoff of the game. Players cooperate in a coalition and receive a certain profit from this cooperation. This game theory approach adapts well to prediction algorithms: the *game* is the prediction task for a single instance of the dataset, the *gain* is the actual prediction for this instance minus the average prediction for all instances, the *players* are the feature values of the instance that collaborate to receive the gain. The Shapley value of an input variable is the average contribution of a feature value to the prediction in different coalitions of features [106]. It has to be noted that the Shapley value of a feature is not the difference in prediction when we would remove the feature from the inputs of a model. The feature contribution score can be computed on single events to interpret the prediction of a ML model. One can visualize feature attributions such as Shapley values as *forces* as shown in Fig. 3.52: each feature Shapley value is a force that either increases

or decreases the model output prediction from a baseline given by the average prediction on the overall dataset.

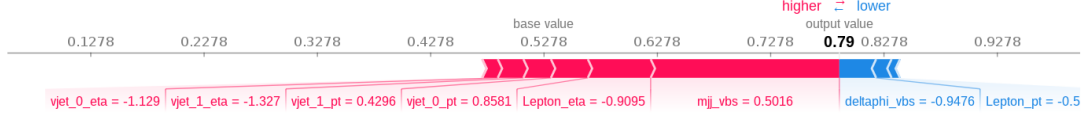


Figure 3.52: Representation of the Shapley values of each input feature of the DNN model selected for the resolved category on a single event

Table 3.14 contains the list of the variables used for the resolved and boosted models. The Zeppenfeld variable is defined as follows:

$$Z_x = \frac{\eta^x - \bar{\eta}}{\Delta\eta^{vbs}}, \quad (3.8)$$

where  $\bar{\eta}$  is the average  $\eta$  of the tag jets. It can be computed for the lepton or for V hadronic jets for example.

The centrality variable [107, 108] for the VW system is defined as follows:

$$\begin{aligned} \Delta\eta_+ &= \max(\eta^{vbs}) - \max(\eta^V, \eta^W), \\ \Delta\eta_- &= \min(\eta^V, \eta^W) - \min(\eta^{vbs}), \\ C_{VW} &= \min(\Delta\eta_-, \Delta\eta_+), \end{aligned} \quad (3.9)$$

where W stands for the leptonic W and V for the hadronic vector boson.

### 3.7.3.1 Correlation matrices

In Fig. 3.53 the correlation matrix for the input variables for signal in the resolved and boosted categories is shown. It is also interesting to see the difference between the correlation matrices of the backgrounds and the signal one (plotted in Fig. 3.54 for the resolved category and in Fig. 3.55 for the boosted one). For example the correlation between the  $m_{jj}^{VBS}$  and  $\Delta\eta_{jj}^{VBS}$  variables for the signal is quite different with respect to the correlation for the W+jets background. The same comment also applies to the correlation between  $m_{jj}^{VBS}$  and the  $p_T$  of the trailing VBS jet. Looking at the difference between the signal correlation matrix and the top, VV, and DY ones, it is milder with the respect to the W+jets ones, but they are all distinct between each other. The DNN discriminator will exploit these peculiarities for the signal discrimination against all the different backgrounds.

Table 3.14: Variables used as input to the DNN for the resolved and boosted models. They are ranked by their contribution to the signal discrimination power of the DNN model using the SHAP[64, 109] technique and their rank is shown in the table for the resolved and boosted categories models.

| Variable                                                                | Resolved | Boosted | SHAP ranking |         |
|-------------------------------------------------------------------------|----------|---------|--------------|---------|
|                                                                         |          |         | Resolved     | Boosted |
| Lepton pseudorapidity                                                   | ✓        | ✓       | 13           | 12      |
| Lepton transverse momentum                                              | ✓        | ✓       | 16           | 10      |
| Zeppenfeld variable for the lepton $Z_{\text{lep}}$                     | ✓        | ✓       | 2            | 2       |
| Number of jets with $p_T > 30$ GeV                                      | ✓        | ✓       | 7            | 3       |
| VBS leading tag jet $p_T$                                               | -        | ✓       | -            | 11      |
| VBS trailing tag jet $p_T$                                              | ✓        | ✓       | 7            | 6       |
| Pseudorapidity interval $\Delta\eta_{jj}^{\text{VBS}}$ between tag jets | ✓        | ✓       | 4            | 4       |
| Quark/gluon discriminator of leading tag jet                            | ✓        | ✓       | 9            | 7       |
| Azimuthal angle distance between VBS tag jets                           | ✓        | -       | 10           | -       |
| Invariant mass of the VBS tag jets pair                                 | ✓        | ✓       | 1            | 1       |
| $p_T$ of the leading jet from $V_{\text{had}}$                          | ✓        | -       | 14           | -       |
| $p_T$ of the trailing jet from $V_{\text{had}}$                         | ✓        | -       | 12           | -       |
| Pseudorapidity difference between $V_{\text{had}}$ jets                 | ✓        | -       | 8            | -       |
| Quark/gluon discriminator of the leading $V_{\text{had}}$ jet           | ✓        | -       | 3            | -       |
| Quark/gluon discriminator of the trailing $V_{\text{had}}$ jet          | ✓        | -       | 5            | -       |
| $p_T$ of the AK8 $V_{\text{had}}$ jet candidate                         | -        | ✓       | -            | 8       |
| Invariant mass of the $V_{\text{had}}$                                  | ✓        | ✓       | 11           | 5       |
| Zeppenfeld variable for the $V_{\text{had}}$                            | -        | ✓       | -            | 9       |
| Centrality                                                              | -        | ✓       | 15           | 13      |

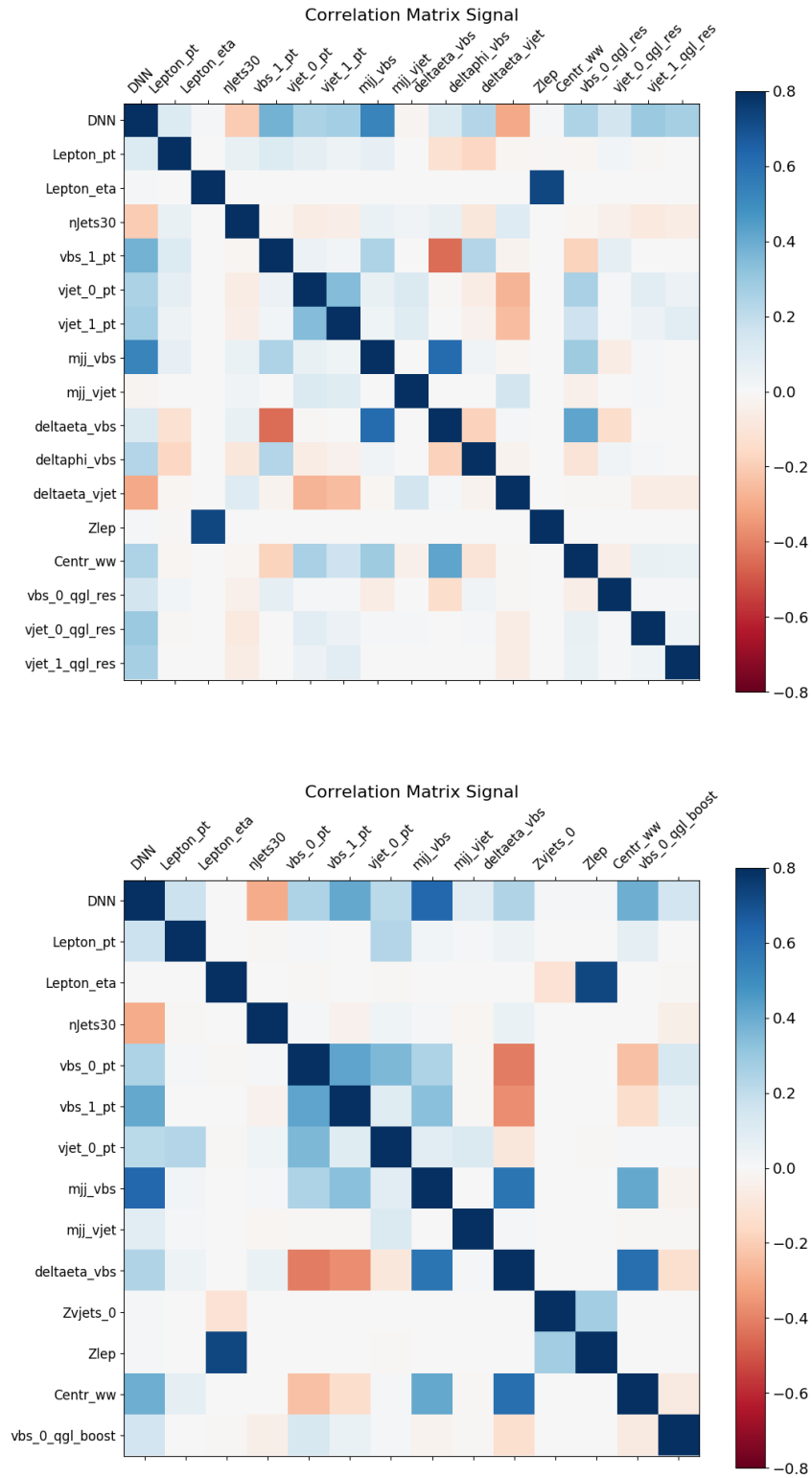


Figure 3.53: Correlation matrix of input variables for signal in resolved (top) and boosted (bottom) categories.



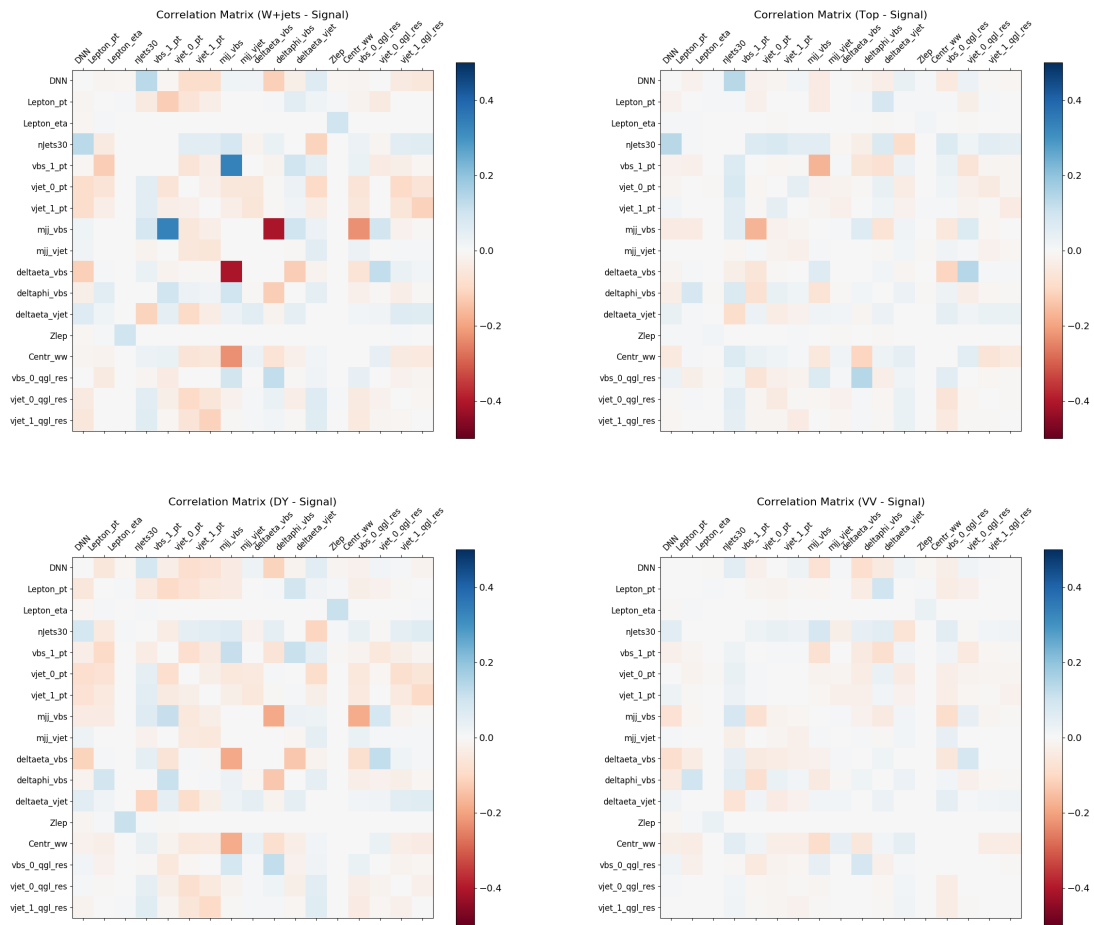


Figure 3.54: Difference between the input variables correlation matrices of backgrounds and signal in the resolved category. From top left to bottom right: W+jets, top, DY, VV.

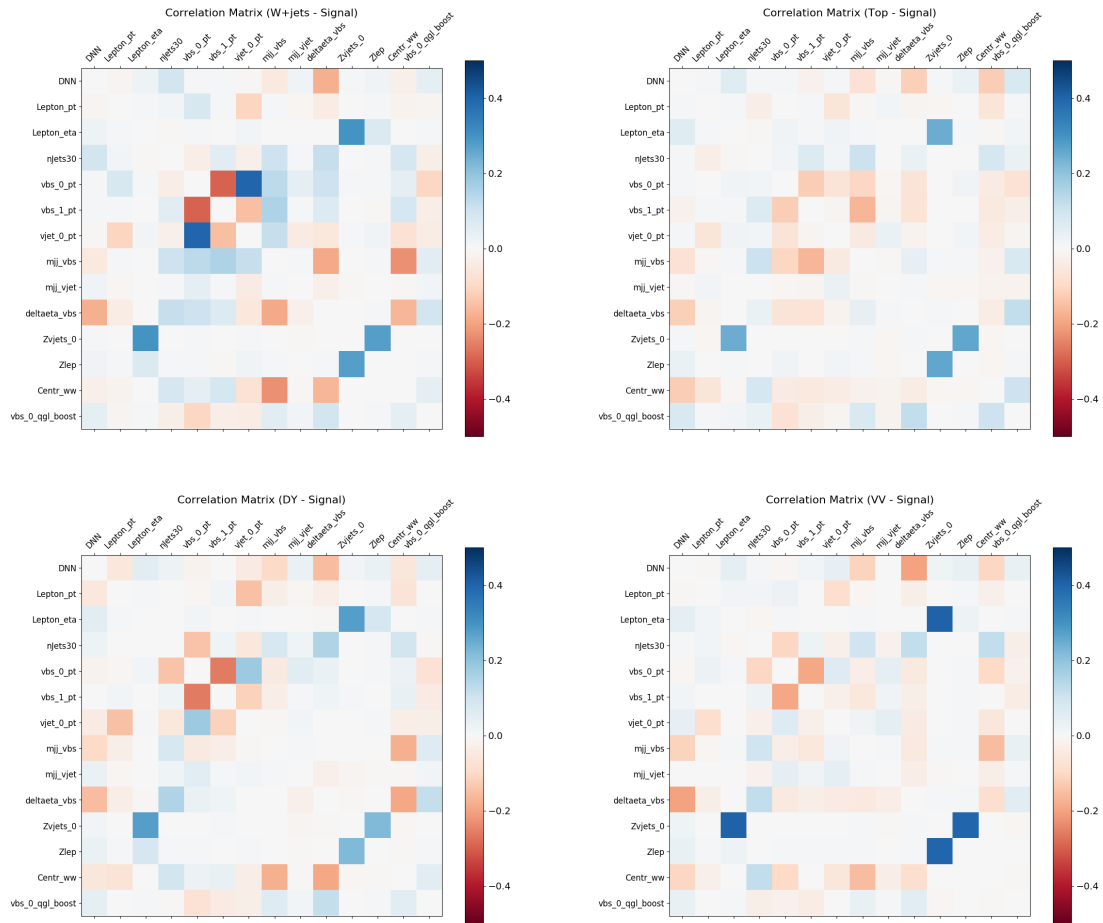


Figure 3.55: Difference between the input variables correlation matrices of backgrounds and signal in the boosted category. From top left to bottom right: W+jets, top, DY, VV.

### 3.7.3.2 SHAP values analysis

The plot of the SHAP values of each feature for the DNN architecture and input feature set chosen for the resolved and for the boosted case are reported in Fig. 3.56. Between the most important variables for the model discrimination power there are:  $m_{jj}^{\text{VBS}}$ ,  $Z_{\text{lep}}$ , the number of jets and the quark/gluon likelihood.

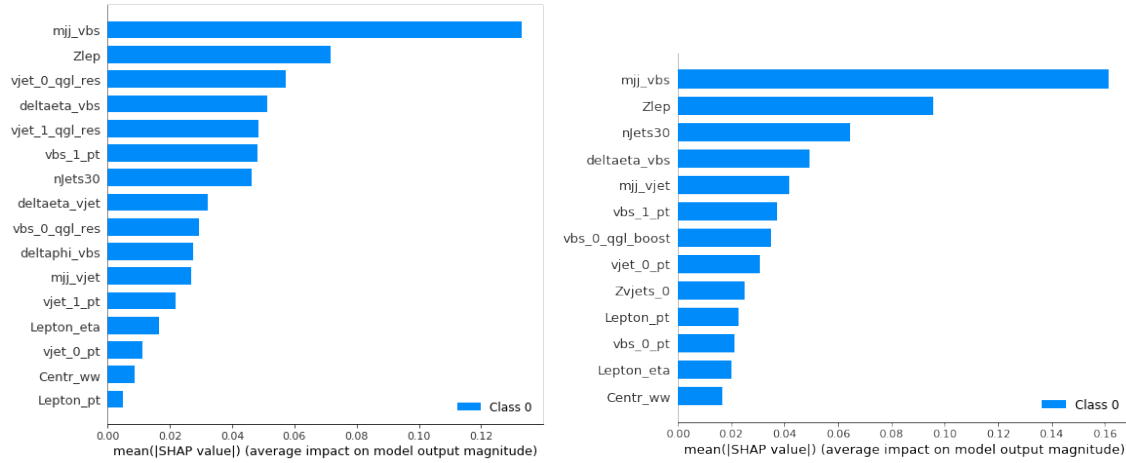


Figure 3.56: Average absolute value of the Shapley values of each feature at the last iteration for the desired DNN architecture, for resolved and boosted category

Another plot related to the Shapley value is shown in Fig. 3.57 and it is useful to understand the influence of each input variable on the DNN output. For each variable the distribution of the Shapley value for each event is plotted. The X-axis of the plot is the Shapley value, while the color variation shows the variation along the variables values from low (blue) to high (violet). The plot is built event by event: for each event the input variables obtain different Shapley values. A positive Shapley value indicates that the value of the variable is pushing toward the signal, a negative one towards the background. The Shapley values are accumulated for many events to build the distribution shown along X-axis. The plot is built using both signal and background events. The plot shows that high values of  $m_{jj}^{\text{VBS}}$  are correlated with signal-like events, whereas high or low values of the  $Z_{\text{lep}}$  variables are background-like. Moreover, high values of  $\Delta\eta_{jj}^{\text{VBS}}$  are considered strongly background like, and low values are slightly signal-like. This plot is useful because it is showing the effect of the variables on the DNN output after the training and it offers an interpretation that is matching the expected physical one.

The Shapley values associated to an input variable can be also understood in terms of the correlation of that variable with others. For example high values of the  $\Delta\eta_{jj}^{\text{VBS}}$  variable are given negative Shapley values, they are identified as background like. By looking at the distribution of the Shapley values for a second input variable like the trailing VBS jet  $p_T$ , one can see, Fig. 3.58 top, that large  $\Delta\eta_{jj}^{\text{VBS}}$  values are background-like since they are correlated with small trailing VBS jet  $p_T$ . On the contrary by looking at the Shapley values for the  $m_{jj}^{\text{VBS}}$  variables with respect to the  $\Delta\eta_{jj}^{\text{VBS}}$  values, large  $m_{jj}^{\text{VBS}}$  values have larger Shapley importance if  $\Delta\eta_{jj}^{\text{VBS}}$  is also large.

### 3.7.4 Training monitoring

In order to monitor the neural network training and to reject models that are over-trained, the training sample is divided into three subsets: the first used for training and called *training*

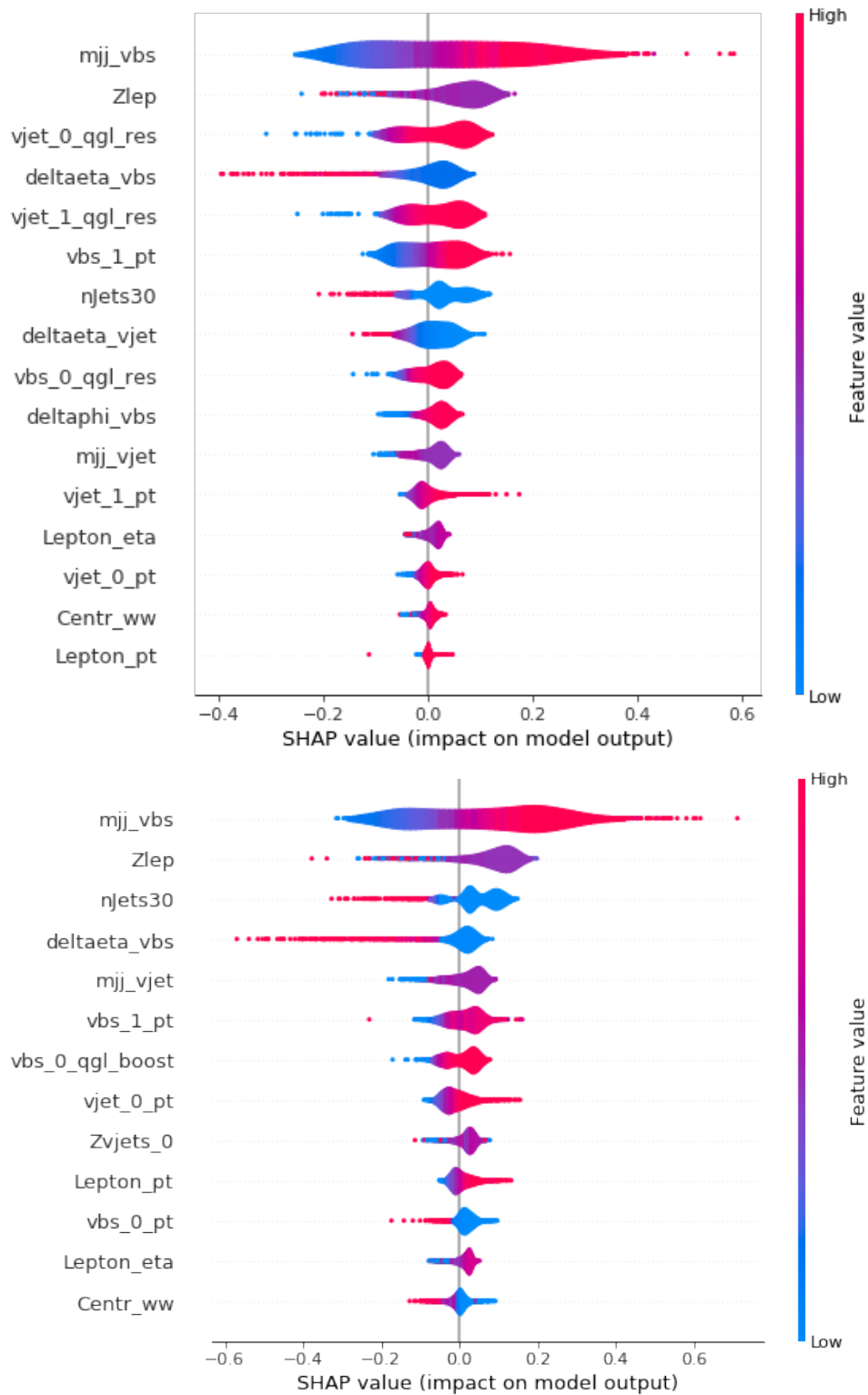


Figure 3.57: Shapley values distributions for each feature for the resolved and boosted models. The Shapley values are accumulated over many events. The thickness of the band represents the frequency of that Shapley value. The color of the band represents the value of the variable corresponding to each Shapley value.

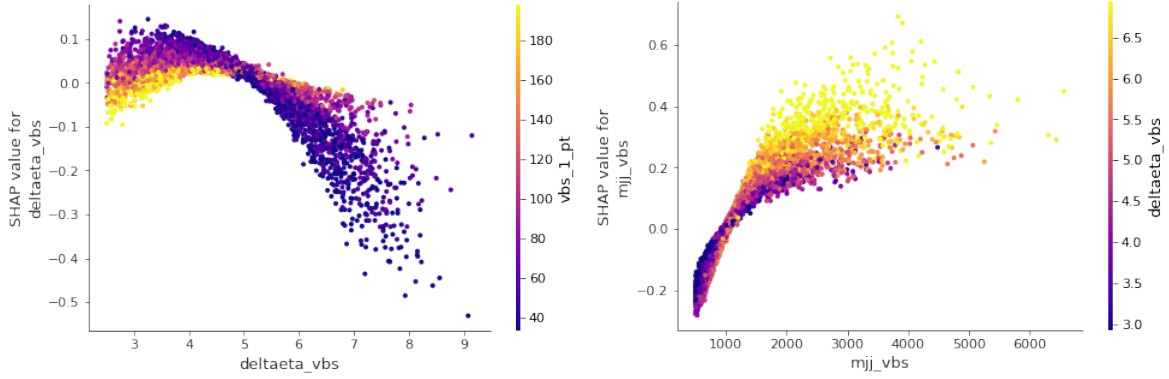


Figure 3.58: Shapley values of the  $\Delta\eta_{jj}^{\text{VBS}}$  (left) and  $m_{jj}^{\text{VBS}}$  (right) features plotted for many events. The color of the point represents the value of a second variable correlated with the observable under analysis: trailing VBS tag jet  $p_T$  (left) and  $\Delta\eta_{jj}^{\text{VBS}}$  (right).

*dataset*, the second called *validation dataset* for controlling whether the network is able to generalize to data that it has not been shown before, and the third one, called *test dataset*, is used to compared in an unbiased way the performance of different models when an optimization of the hyperparameters is performed. A set of metrics are chosen to compare how the network behaves on the training and validation dataset.

The first metric chosen is the loss function used to train the network itself. Overtraining is monitored during the training looking at the value of the loss function evaluated on the training and validation datasets: both should decrease and the loss on the validation dataset should never increase while the other decreases. Then, signal efficiency ( $R$ ) and purity ( $P_u$ ) (also known as recall and precision) computed with a cutoff of  $DNN > 0.85$  are monitored to check the performance of the training. Calling  $TP$  the true positives,  $FP$  the false positives,  $T$  the total number of signal events.

$$\begin{aligned} \text{Signal efficiency} = R &= \frac{TP}{T} \\ \text{Signal purity} = P_u &= \frac{TP}{FP + TP} \end{aligned} \quad (3.10)$$

An overall metric called the F1-score is used as a measure of the discriminator accuracy since it is computed as the harmonic mean of precision and recall.

$$F^1 = 2 \frac{R \cdot P_u}{R + P_u} \quad (3.11)$$

The metrics are computed after each training epoch, a complete round over the training dataset, to check their trends during the training. It may happen that the validation loss function and metrics values do not clearly show overtraining, but the network produces a different shape of the signal and background in the DNN output variable between the training and validation dataset. In order to check that this is not occurring, a Kolmogorov-Smirnov test (KS test) is performed between signal training and validation datasets, and background training and validation datasets. If KS-test p-value drops significantly, than the model is rejected even if the loss and accuracy look fine. This can happen when the network is not sufficiently regularized or the number of training samples is not large enough.

In Fig.3.59 the evolution of the metrics along the training for the resolved and boosted models are plotted as a function of the epoch. Both the models reach the convergence and do not show evidence of overtraining. Moreover in Fig. 3.60 the training and validation DNN distributions are shown for the resolved model along with their ROC curve.

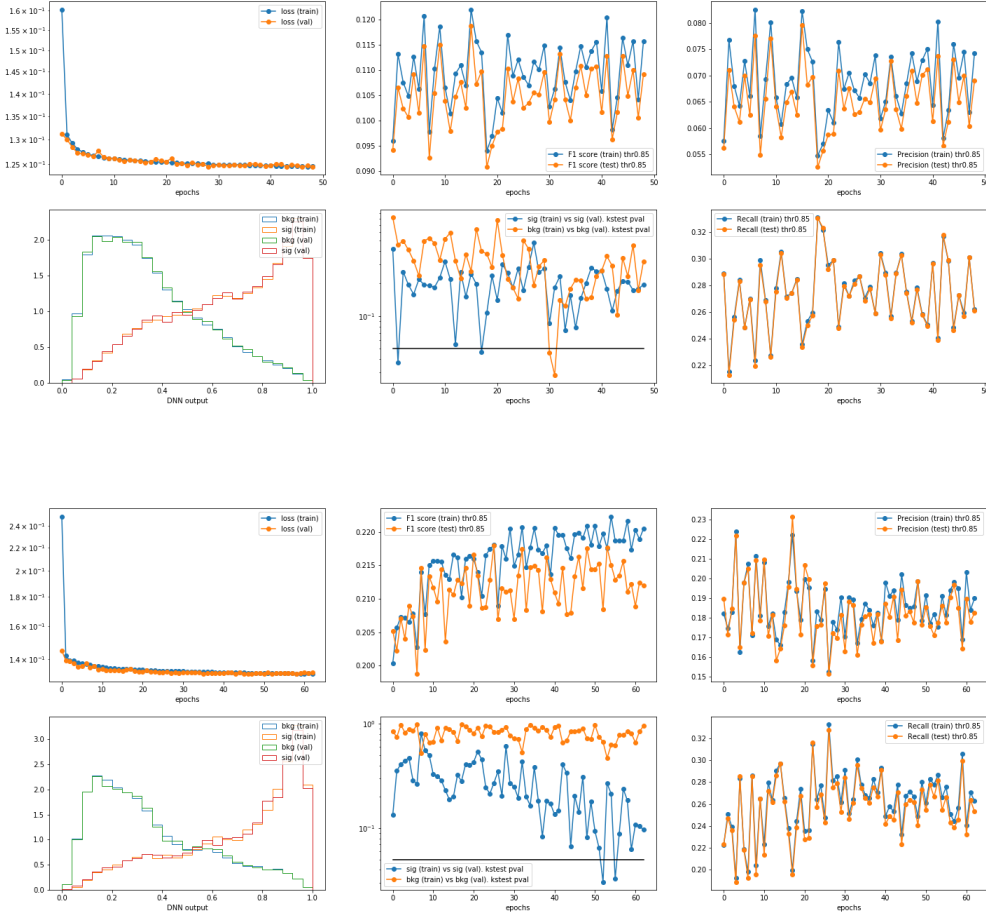


Figure 3.59: Training monitor of the two selected DNN models. The plots are the evolution during the training of the metrics defined in this section. Top: resolved model, bottom: boosted model.

### 3.7.5 Performance

The output of the DNN is rescaled in order to make the distribution for the signal flat, allowing for a uniform definition of the binning for the final fit. The transformation is calculated from the cumulative function of the signal distribution separately for each year. Figure 3.61 shows the DNN normalized distributions, taken from simulation, for the signal and backgrounds.

Figure 3.62 show the DNN output distribution for the 2018 dataset in the resolved and boosted categories for the electron and muon channel in the signal region and in the top and W+jets control regions. The DNN shape in data is well described by MC in all the control regions, although these are pre-fit distributions where the data-driven corrections from the control regions are not applied yet. Since the discriminator has been trained in the signal region, when

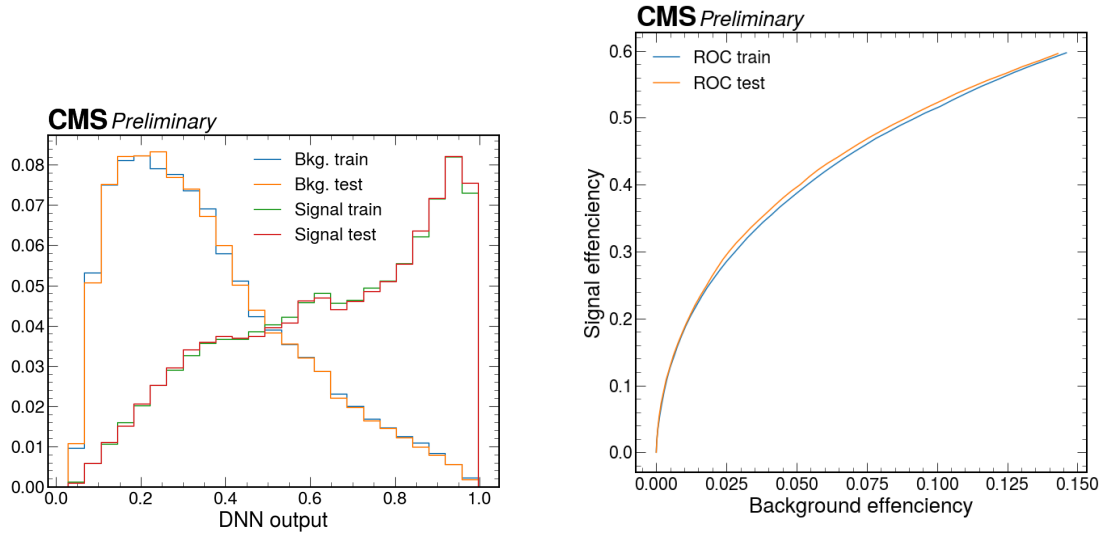


Figure 3.60: On the left: training and testing distributions for resolved DNN. On the right: ROC curve for resolved model.

evaluating the DNN on the W+jets control region the input corresponding to the invariant mass of the V boson is set to its mean value for the W+jets process in the signal region. During the optimization of the analysis samples in the high DNN region ( $> 0.7$ ) of the signal region were blinded to avoid any biases.

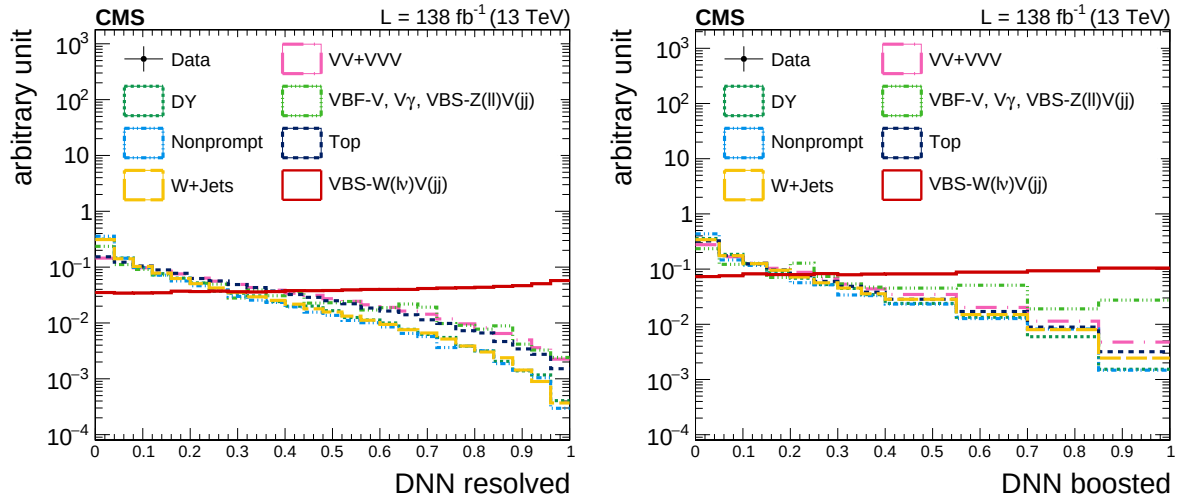


Figure 3.61: The DNN discriminator distribution, taken from simulation, for VBS signal and backgrounds in the resolved (left) and boosted (right) signal regions normalized to unity.

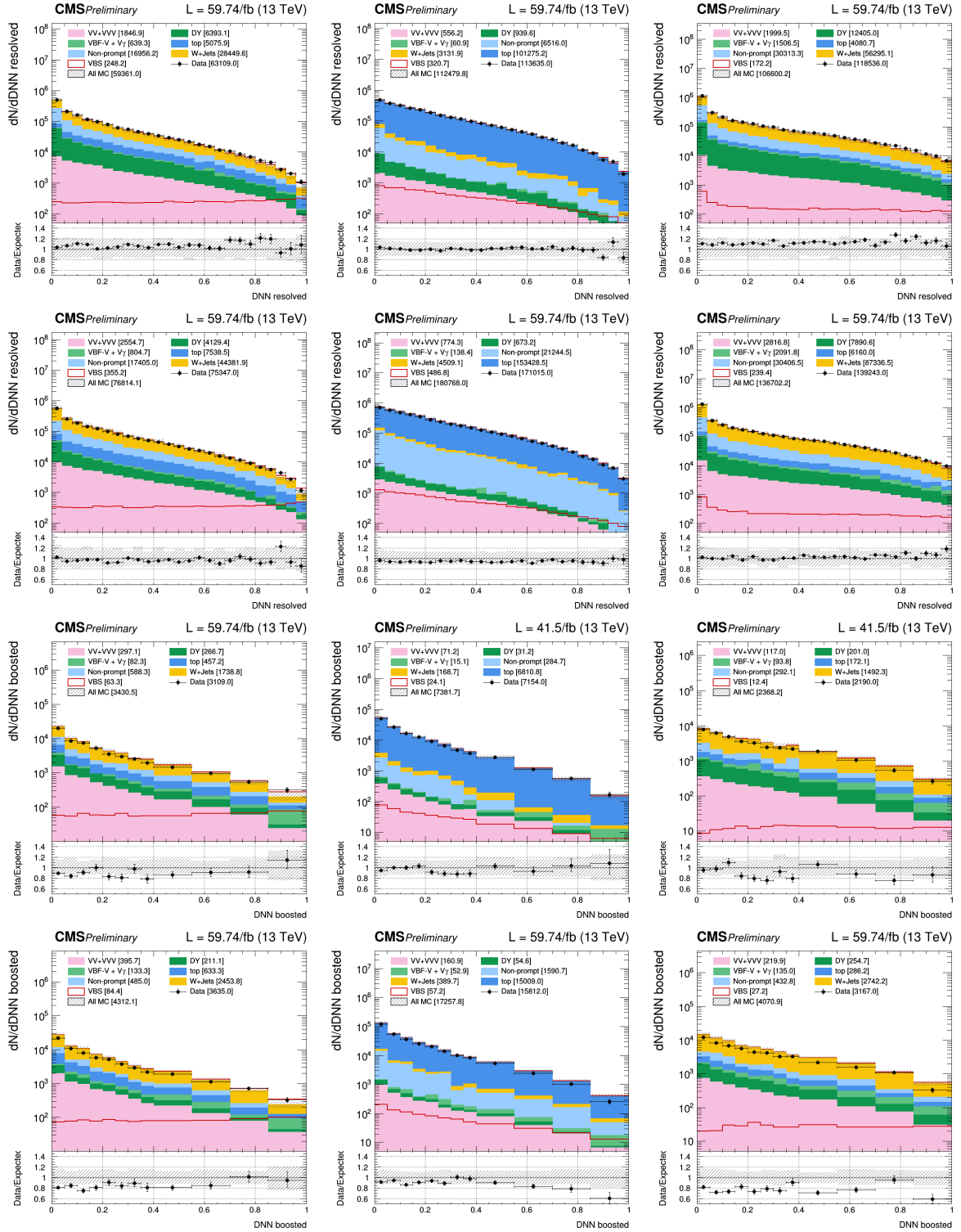


Figure 3.62: DNN output in signal region (left), top quark control region (center), and W-jets control region (right). Pre-fit distributions on 2018 dataset. Categories by line: resolved electron category (1st line), resolved muon (2nd line), boosted electron (3rd line), boosted muon (4th line)



### 3.7.6 DNN bins kinematic regions

In Fig. 3.63 and Fig. 3.64 the most interesting kinematic distributions are shown for subsets of events belonging to different DNN bins. The contribution of signal and background MC samples is summed. This check confirms that the events selected as signal-like in the  $DNN > 0.8$  region correspond to the expectation for the VBS process: high  $m_{jj}^{VBS}$  events, central leptons ( $Z_{lep} \rightarrow 0$ ), low additional number of jets. Signal-like events have on average larger  $\Delta\eta_{jj}^{VBS}$  than background like ones, but at high  $\Delta\eta_{jj}^{VBS}$  the background contamination is still consistent and the DNN relies on other observables, like the  $m_{jj}^{VBS}$ , for the signal discrimination.

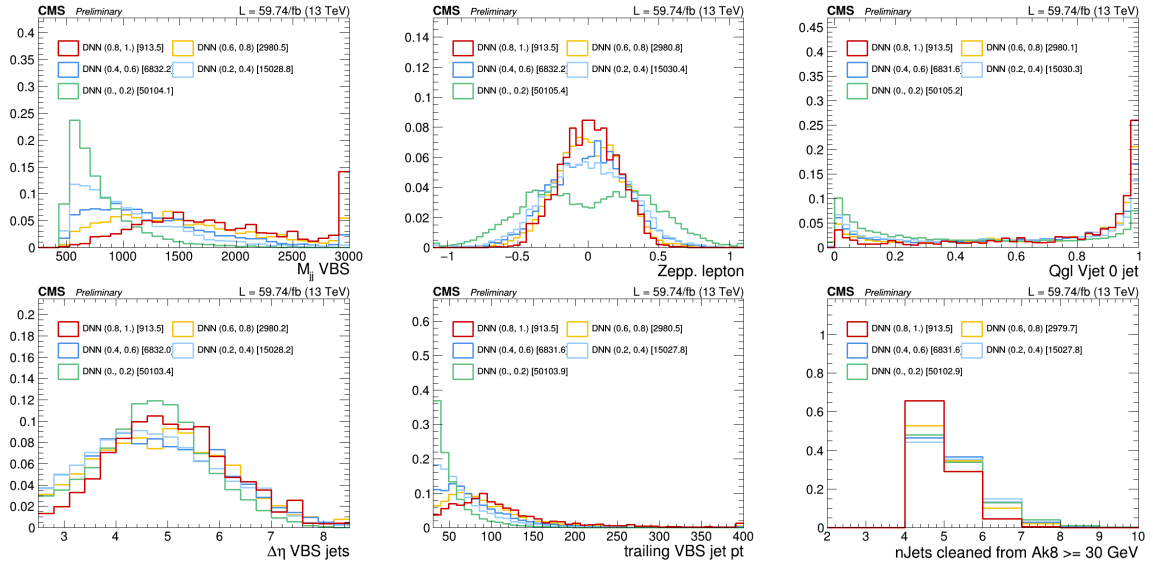


Figure 3.63: DNN bins kinematic distribution. Resolved signal region.

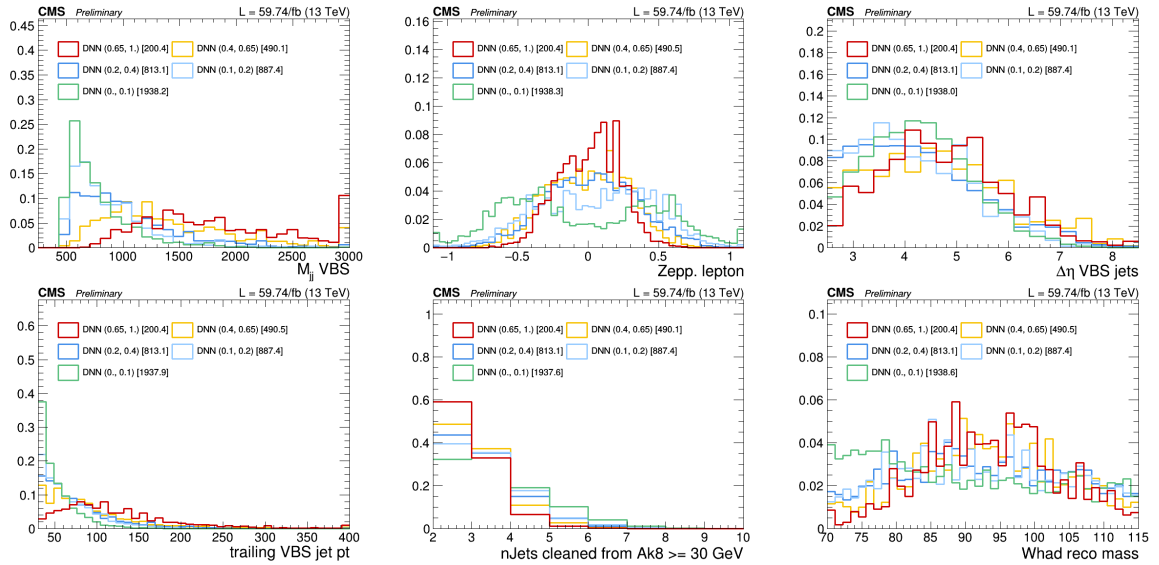


Figure 3.64: DNN bins kinematic distribution. Boosted signal region.

## 3.8 W+jets background data-driven estimation

### 3.8.1 Data/MC discrepancies

The W+jets background is the main one in the signal region and it is crucial to guarantee that it is well modeled. This analysis requires at least 4 jets to be present in the resolved category, which is a regime that is not well described by the showering and hadronization Monte Carlo simulation performed on top of the LO matrix element calculation of these samples. Moreover, it has been shown by other analyses in CMS that the  $p_T$  of the leptonic W is not well described by the W+jets LO simulation used for the analysis.

In both the resolved and boosted regions discrepancies are evident in jets variables and show similar trends for all the years. In Figure 3.65 the  $p_T$  of the leptonic W  $p_T^{W,\ell}$ ,  $\Delta\eta_{jj}^{VBS}$ , and the  $p_T$  of the trailing VBS tag jets are shown for the resolved category, only in the muon channel for brevity. Figure 3.66 displays instead distributions in the boosted category.

A data-driven method has been designed and implemented to estimate the contribution of the main background in the signal region. The  $p_T$  of the leptonic W shows an evident trend in the data/MC discrepancy that is correlated with the discrepancy at high  $p_T$  of the jets tagged in the event. Hence, the W+jets background is split into subcategories using bins of the  $p_T^{W,\ell}$  and trailing VBS jet  $p_T$  observables, and their normalization is measured from data in the fit thanks to the W+jets control regions. Thanks to this procedure, the correction on the W+jets MC normalization (for each  $p_T^{W,\ell}$ ) can be propagated to the signal region taking into account its uncertainty directly in the fit.

A closure test, reported in the Sec 3.8.3, is performed by splitting the W+jets in two regions, as in Sec. 3.8.4 the normalization correction is applied on prefit plots in the signal and W+jets regions to check their effect on the analysis distributions.

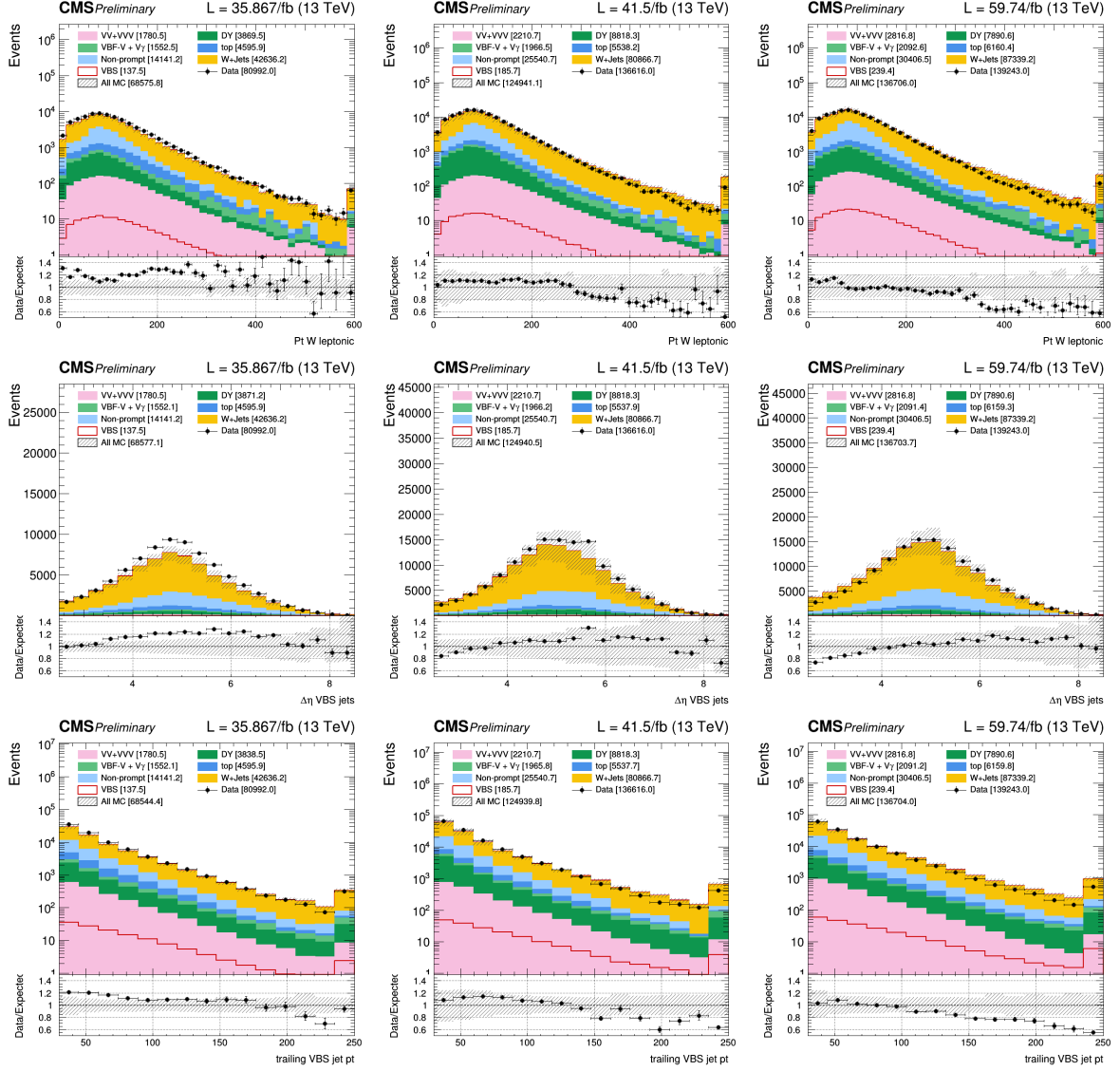


Figure 3.65: Resolved, W+jets, muon control region. Pre-control distributions. From left to right: 2016, 2017, 2018. From top to bottom:  $p_T^{W, \ell}$ ,  $\Delta\eta_{VBS}^{jets}$ , trailing VBS tag jet  $p_T$

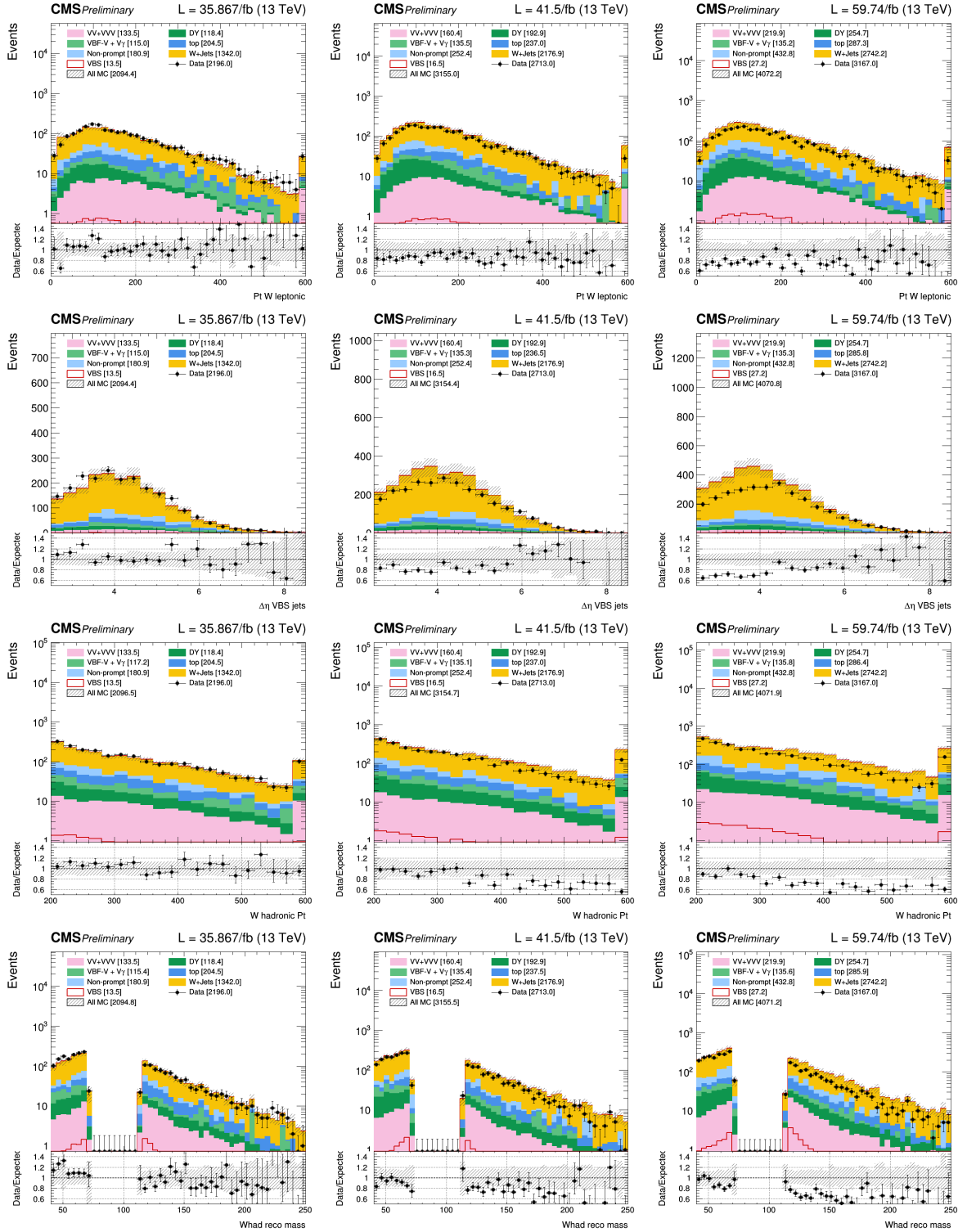


Figure 3.66: Boosted, W+jets, muon control region. Pre-fit distributions. From left to right: 2016, 2017, 2018. From top to bottom:  $p_T^{W,\ell}$ ,  $\Delta\eta^{VBS}$ , boosted V  $p_T$  and mass.

### 3.8.2 W+jets subcategories

The W+jets sample splitting is performed separately for electron and muon final states in both the boosted and resolved region. Table 3.15 shows the binning used to define the categories in the resolved region, whereas Tab.3.16 the one for the boosted category.

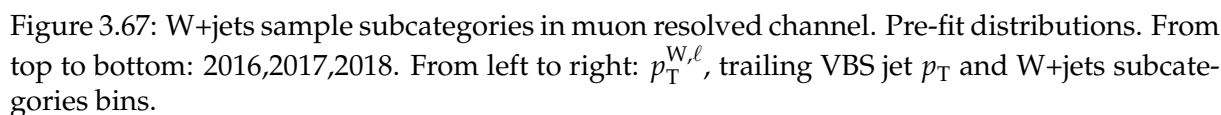
The value of the W+jet background normalization in every sub-category is left free to float in the fit, so that it can match the observed data in the W+jet control region defined with the same sub-category selection. The same normalization correction factors are applied on the W+jets sample in the signal region, to keep the correction correlated. This approach changes both the W+jets shapes and normalization. The rescaling is left independent for each year and lepton flavour. In order for this method to work the discrepancies in  $p_T^{W,\ell}$  and  $p_T^{\text{VBS 2nd}}$  must have a compatible shape in the W+jets control region and in the signal region. A closure test has been implemented to check this method and it is described in Sec 3.8.3.

Table 3.15: W+Jets data-driven method binning in the resolved category.

| Bin | $p_T^{W,\ell}$ bin                        | $p_T^{\text{VBS 2nd}}$ bin                        |
|-----|-------------------------------------------|---------------------------------------------------|
| 1   | $p_T^{W,\ell} < 100 \text{ GeV}$          | $p_T^{\text{VBS 2nd}} < 55 \text{ GeV}$           |
| 2   |                                           | $55 \leq p_T^{\text{VBS 2nd}} < 75 \text{ GeV}$   |
| 3   |                                           | $75 \leq p_T^{\text{VBS 2nd}} < 100 \text{ GeV}$  |
| 4   |                                           | $100 \leq p_T^{\text{VBS 2nd}} < 135 \text{ GeV}$ |
| 5   |                                           | $135 \leq p_T^{\text{VBS 2nd}} < 170 \text{ GeV}$ |
| 6   |                                           | $p_T^{\text{VBS 2nd}} \geq 170 \text{ GeV}$       |
| 7   | $100 \leq p_T^{W,\ell} < 200 \text{ GeV}$ | $p_T^{\text{VBS 2nd}} < 55 \text{ GeV}$           |
| 8   |                                           | $55 \leq p_T^{\text{VBS 2nd}} < 75 \text{ GeV}$   |
| 9   |                                           | $75 \leq p_T^{\text{VBS 2nd}} < 100 \text{ GeV}$  |
| 10  |                                           | $100 \leq p_T^{\text{VBS 2nd}} < 135 \text{ GeV}$ |
| 11  |                                           | $135 \leq p_T^{\text{VBS 2nd}} < 170 \text{ GeV}$ |
| 12  |                                           | $p_T^{\text{VBS 2nd}} \geq 170 \text{ GeV}$       |
| 13  | $200 \leq p_T^{W,\ell} < 300 \text{ GeV}$ | $p_T^{\text{VBS 2nd}} < 90 \text{ GeV}$           |
| 14  |                                           | $90 \leq p_T^{\text{VBS 2nd}} < 125 \text{ GeV}$  |
| 15  |                                           | $125 \leq p_T^{\text{VBS 2nd}} < 160 \text{ GeV}$ |
| 16  |                                           | $p_T^{\text{VBS 2nd}} \geq 160 \text{ GeV}$       |
| 17  | $300 \leq p_T^{W,\ell} < 400 \text{ GeV}$ | $p_T^{\text{VBS 2nd}} < 90 \text{ GeV}$           |
| 18  |                                           | $p_T^{\text{VBS 2nd}} \geq 90 \text{ GeV}$        |
| 19  | $400 \leq p_T^{W,\ell} < 500 \text{ GeV}$ | $p_T^{\text{VBS 2nd}} < 85 \text{ GeV}$           |
| 20  |                                           | $p_T^{\text{VBS 2nd}} \geq 85 \text{ GeV}$        |
| 21  | $p_T^{W,\ell} \geq 500 \text{ GeV}$       |                                                   |

Figure 3.67 shows the  $p_T^{W,\ell}$  and  $p_T^{\text{VBS 2nd}}$  in the W+jets resolved control region (muon channel) for all the years. Figure 3.68 shows the same distributions for the boosted region. In 2016 the normalization of the W+jets sample is different from the other two years due to the different tune, PDF and Madgraph version used for the 2016 MC. The plot on the right shows the normalization in each of the subcategories: these are the variables used in the final fit.

| Bin | $p_{\text{T}}^{W,\ell}$ bin                |
|-----|--------------------------------------------|
| 1   | $p_{\text{T}}^{W,\ell} < 50$ GeV           |
| 2   | $50 \leq p_{\text{T}}^{W,\ell} < 100$ GeV  |
| 3   | $100 \leq p_{\text{T}}^{W,\ell} < 150$ GeV |
| 4   | $150 \leq p_{\text{T}}^{W,\ell} < 200$ GeV |
| 5   | $200 \leq p_{\text{T}}^{W,\ell} < 300$ GeV |
| 6   | $300 \leq p_{\text{T}}^{W,\ell} < 400$ GeV |
| 7   | $p_{\text{T}}^{W,\ell} \geq 400$ GeV       |



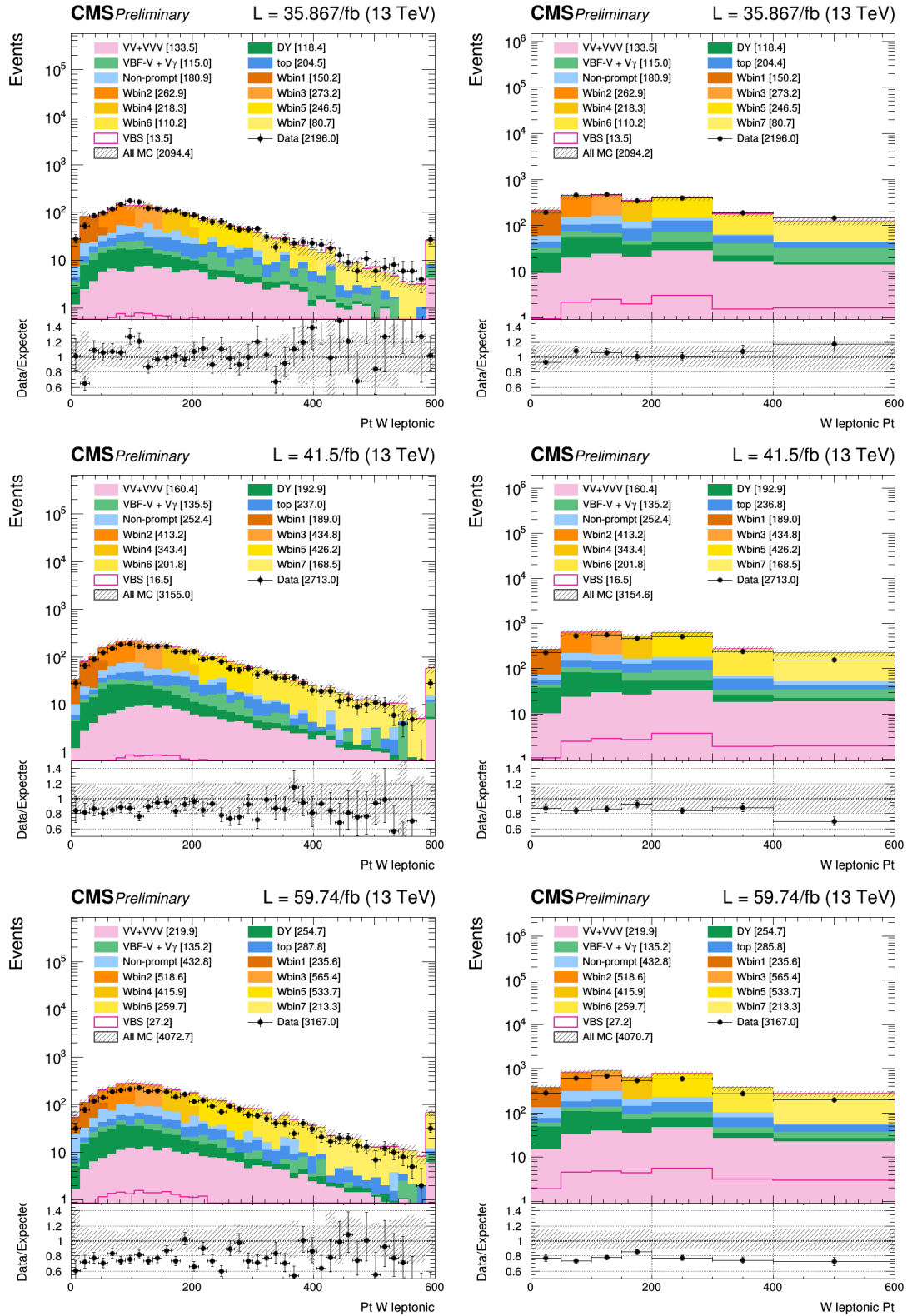


Figure 3.68: W+jets sample subcategories in muon boosted channel. Pre-fit distributions. From top to bottom: 2016,2017,2018. From left to right:  $P_T^{W, \ell}$  and W+jets subcategories bins.

### 3.8.3 W+jets closure test

The W+jets control region is split into two regions: one closest to the signal W mass window  $M_W \in [50, 65] \cup [105, 150]$  GeV ( $M_W \in [55, 70] \cup [115, 150]$  for boosted), and one farther from it  $M_W < 50 \cup M_W > 150$  GeV ( $M_W < 55 \cup M_W \in [150, 250]$  GeV for boosted). These regions are depicted in Figure 3.69.

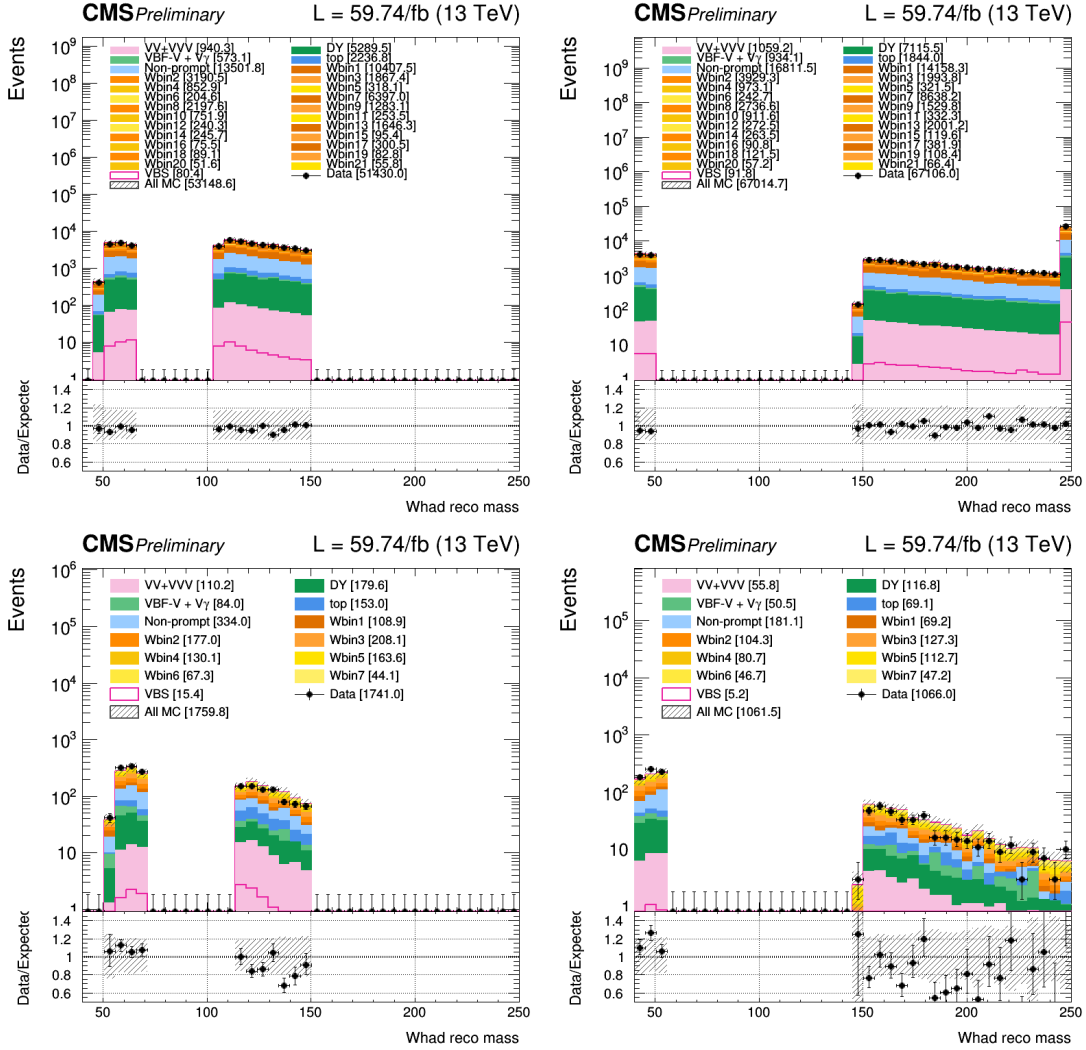


Figure 3.69: W+jets control region division for closure tests in the resolved (top) and boosted (bottom) region: on the left the *internal* region, nearer to the signal mass window, on the right the *external* region.

The correction factors to be applied on W+jets subcategories normalization to fix the normalization to the data are calculated in the external region and propagated in the internal region to check their effect on the kinematic variables. Figure 3.70 shows the W+jets bins in the internal region with the correction applied. The correction factors computed in the external and internal regions are also compared directly in the Fig. 3.71 for the resolved category and Fig. 3.72 for the boosted category: they are always compatible within statistical and systematic error.

This shows that the correction can be extrapolated to the signal region. Since the normalization of the W+jets subcategories is determined directly in the combined fit, the uncertainty of this method is correctly propagated to the signal region and correlated in the fit procedure.



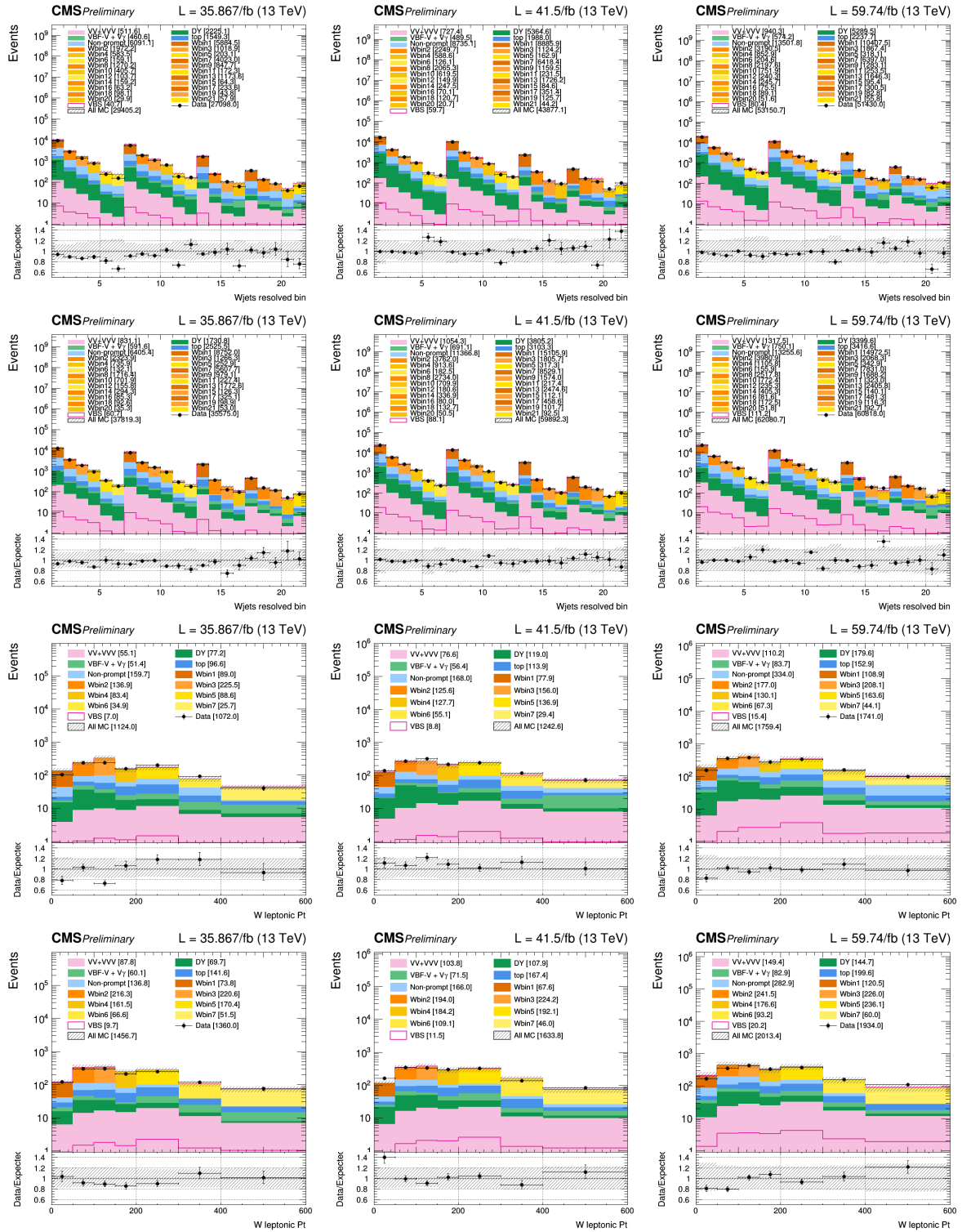


Figure 3.70: W+jets subcategories in the internal W+jets control region after the application of correction factors derived in the external region. From left to right: 2016,2017,2018. From top to bottom: resolved ele and mu, boosted ele and mu categories.

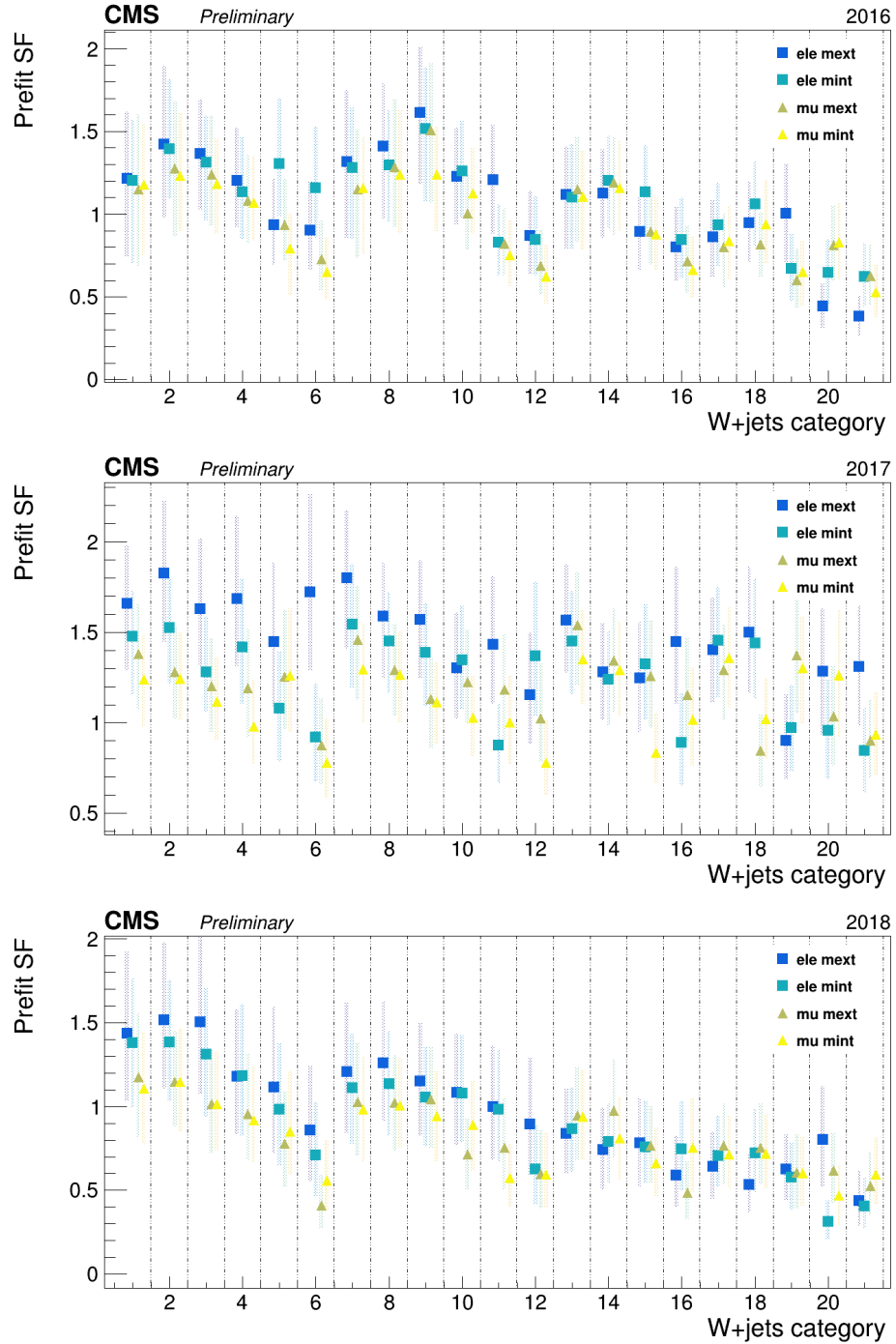


Figure 3.71: Resolved category: comparison of W+jets scale factors derived on internal and external regions. The uncertainty band is stat+syst.

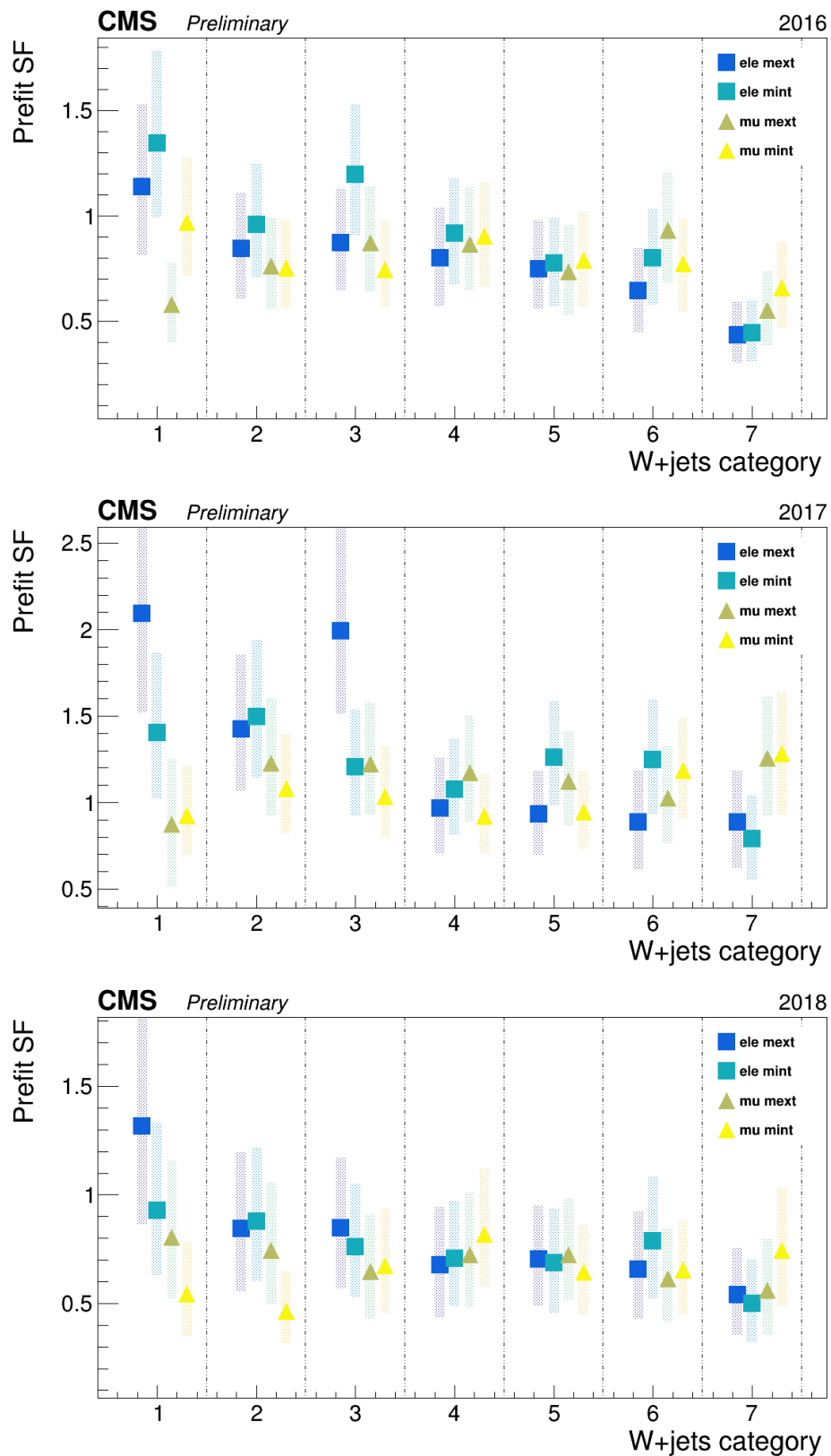


Figure 3.72: Boosted category: comparison of W+jets scale factors derived on internal and external regions. The uncertainty band is stat+syst.

### 3.8.4 W+jets rescaling effect

We extract the correction factors to the W+jets subcategories normalization from the prefit yields in the W+jets control regions and apply them to visualize the effect of the W+jets subcategories rescaling on the kinematic distributions of the analysis and on the DNN discriminant shape in both the control regions and signal region. The normalization and shape of all the other samples is kept fixed.

In Figure 3.73 the factors are applied in W+jets resolved and boosted control regions, where they are extracted, on the  $p_T^{W,\ell}$  variable while in Figure 3.74 the  $\Delta\eta_{jj}^{\text{VBS}}$  and DNN shapes are plotted. The first line shows the prefit distribution, the second line the distribution after the correction. The DNN does not depend strongly on the  $p_T^{W,\ell}$  and this adds stability to the fit.

The same test is performed on the boosted W+jets control region to check the effect of the rescaling. In Figure 3.76 the  $\Delta\eta_{jj}^{\text{VBS}}$  and DNN in the boosted region are plotted: the W+jets correction is improving the description of these variables.

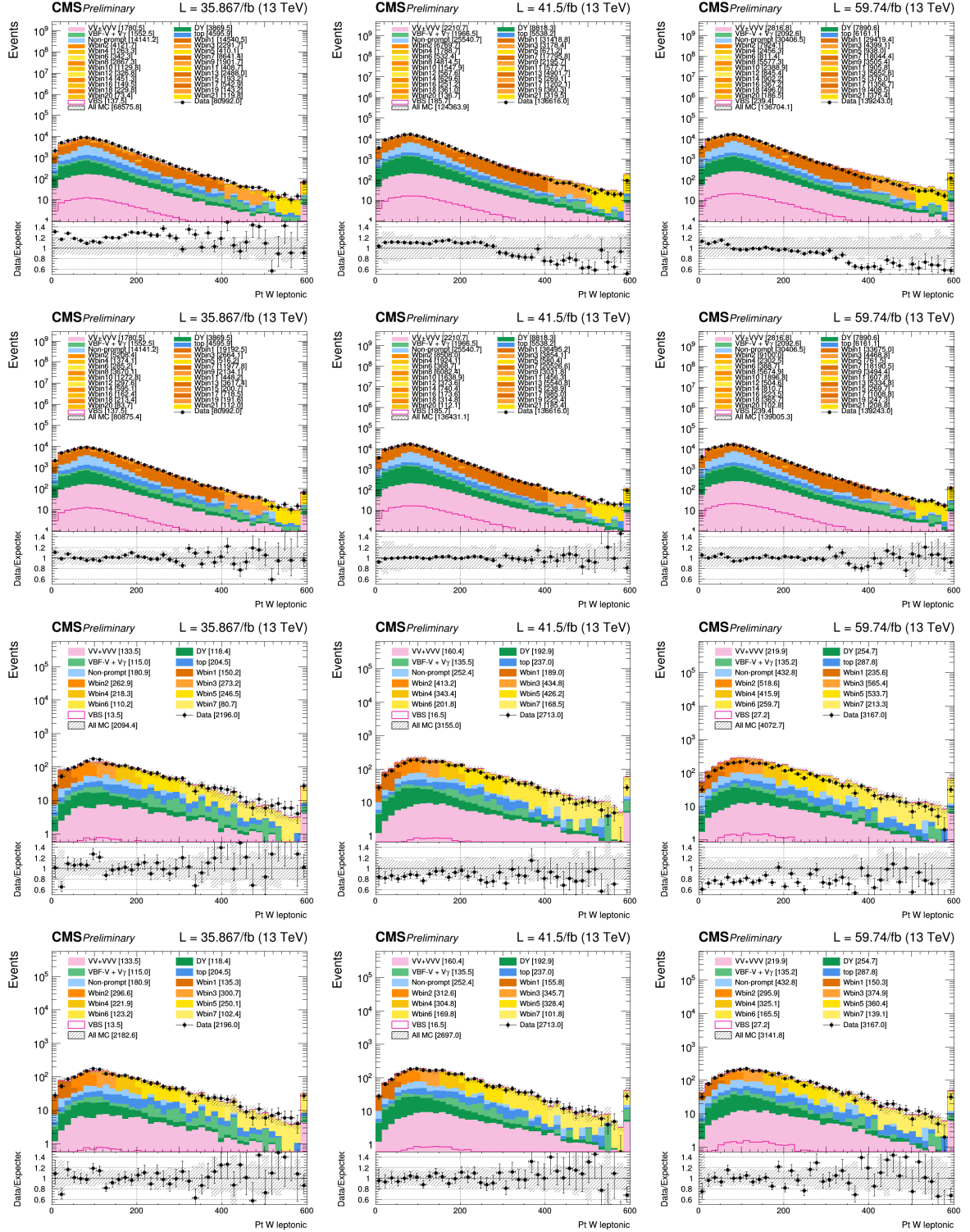


Figure 3.73: W+jets, muon resolved (top lines) and boosted (bottom lines) control regions.  $p_T^{W,\ell}$  before and after applying prefit correction factor to W+jets subcategories normalization. From left to right: 2016, 2017, 2018.

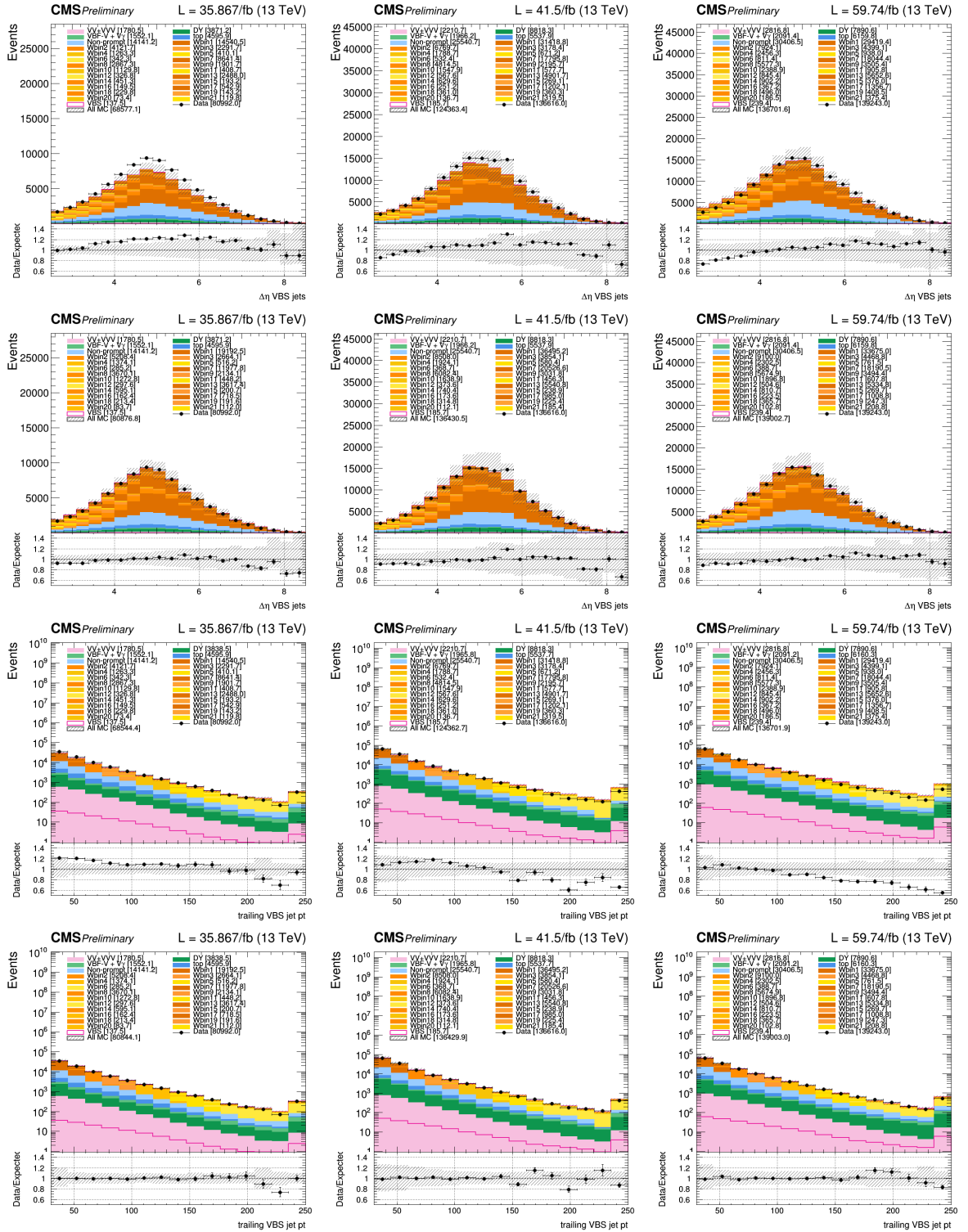
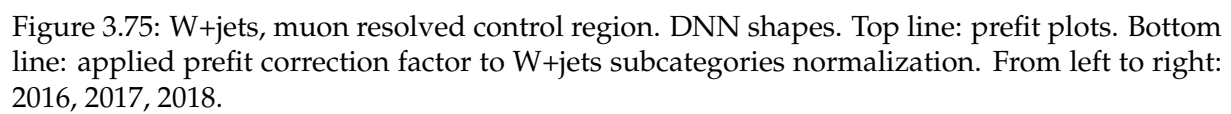


Figure 3.74: W+jets, muon resolved control region.  $\Delta\eta_{\text{VBS}}$  and trailing VBS jet  $p_T$ . Top line: prefit plots. Bottom line: applied prefit correction factor to W+jets subcategories normalization. From left to right: 2016, 2017, 2018.





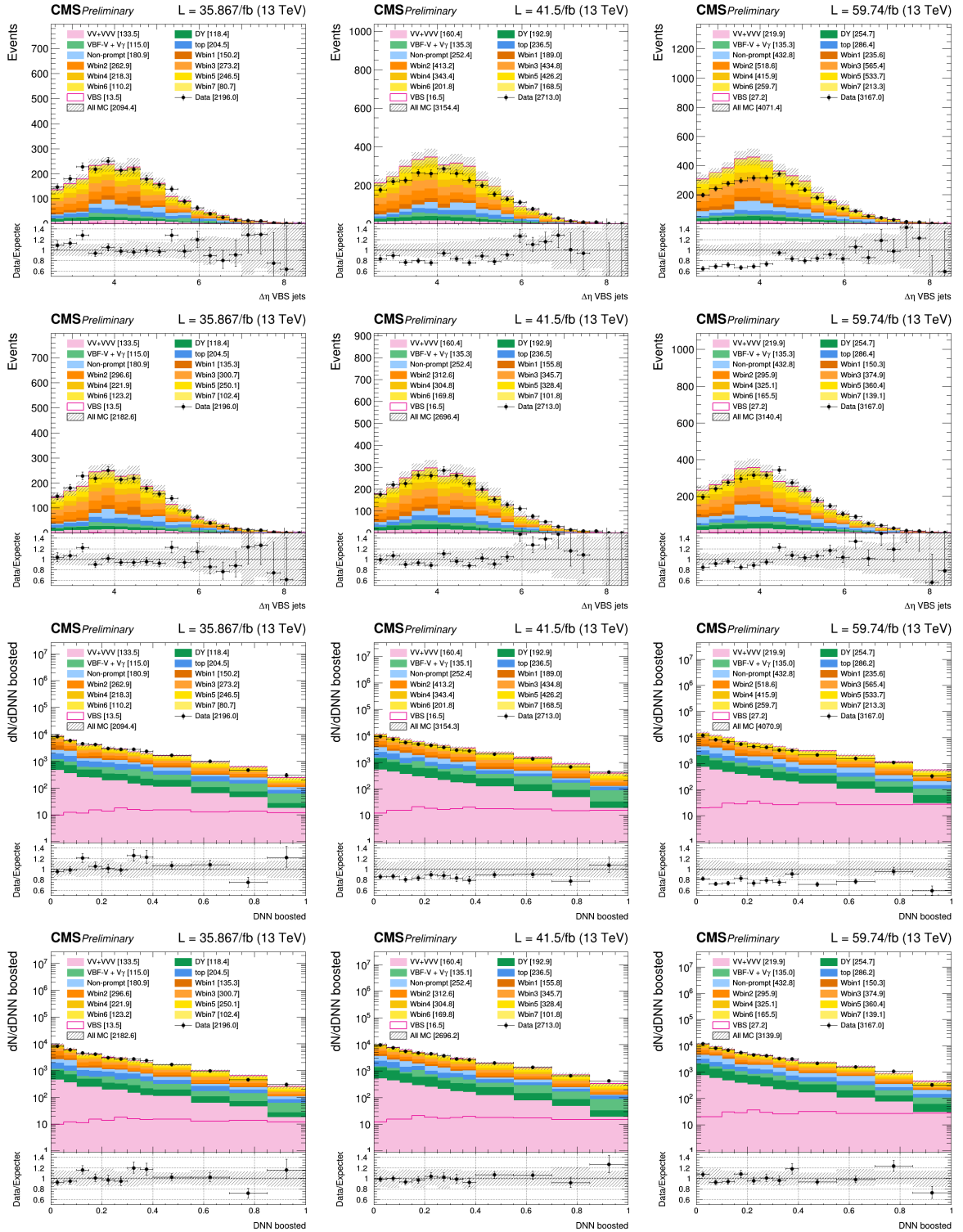


Figure 3.76: W+jets, muon boosted control region.  $\Delta\eta_{\text{VBS}}$  and DNN shape. Top line: profit plots. Bottom line: applied profit correction factor to W+jets subcategories normalization. From left to right: 2016, 2017, 2018.



### 3.8.5 W+jets NLO vs LO samples

The W+jets background MC used in the analysis is simulated with the MADGRAPH generator at LO in HT bins. Up to four jets have been included at LO in the matrix element calculation. A NLO calculation with Madgraph is available and generated in bins of jet multiplicity up to 3 jets, but including only up to 2 jets at NLO in the matrix element calculation.

A comparison of the two samples in the 2018 resolved W+jets control region is shown in Fig. 3.77, and for the boosted region in Fig. 3.78. Although the  $\Delta\eta_{jj}^{\text{VBS}}$  shape difference goes in the correct direction to improve the data/MC ratio, the NLO sample has not enough events in the tails for the resolved categories, and also in the bulk for the boosted one. For these reason, after a detailed comparison, the W+Jets HT binned samples have been used in the analysis and corrected with the data driven method described above.

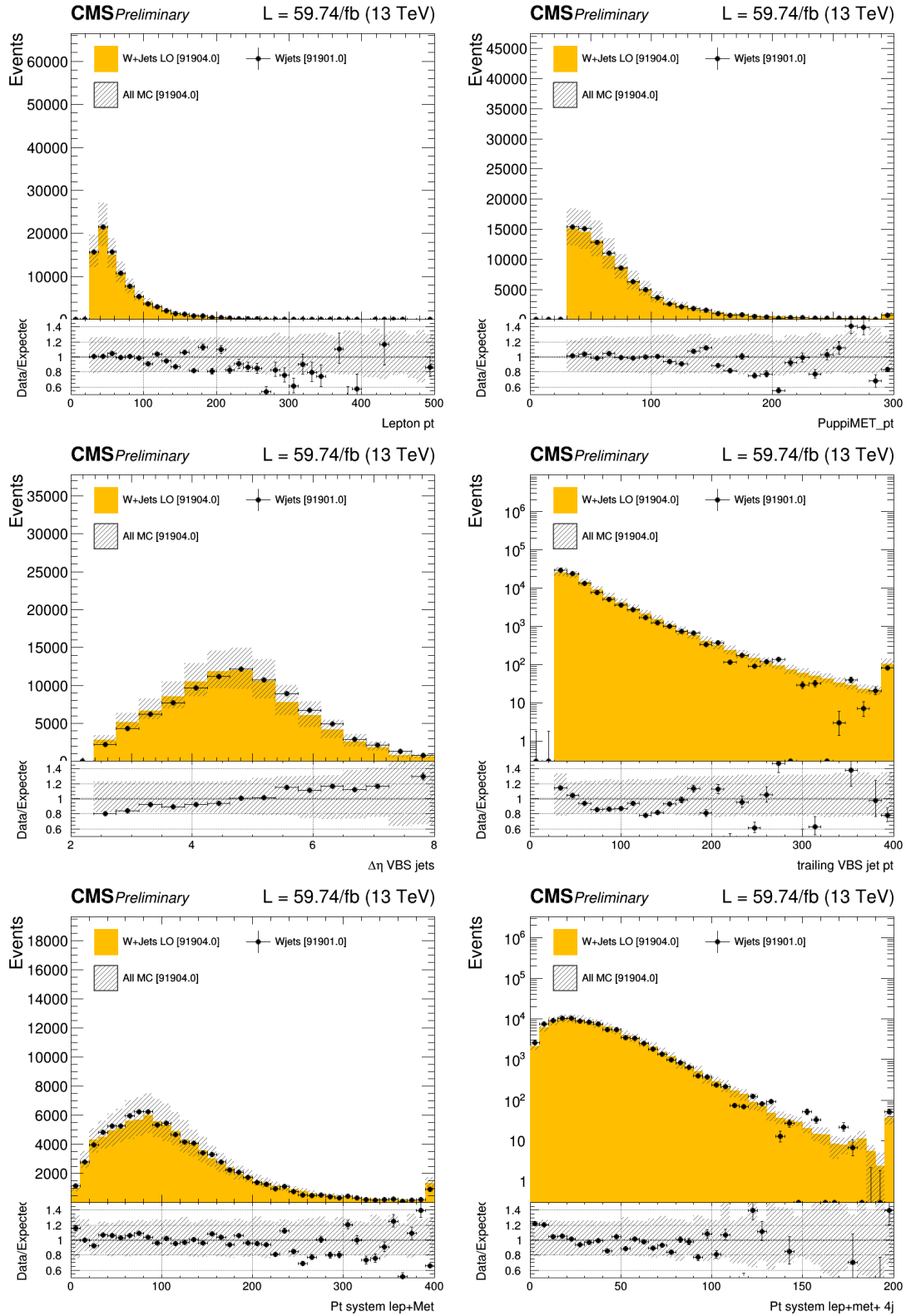


Figure 3.77: W+jets NLO sample (dots) compared with W+jets LO (fill) in resolved W+jets muon region in 2018. The error bars associated to the black markers reflects the MC statistical uncertainty only.

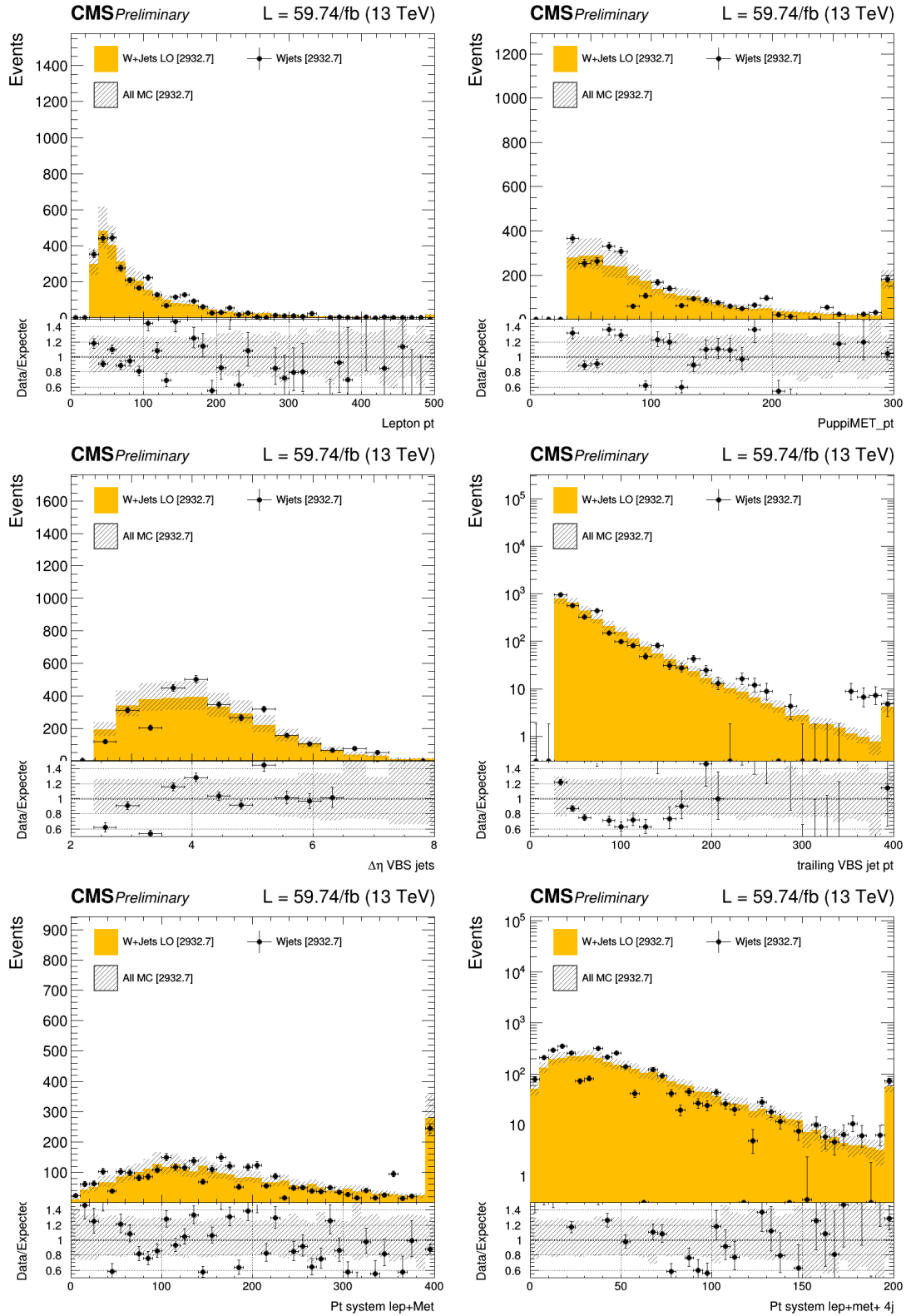


Figure 3.78: W+jets NLO sample (dots) compared with W+jets LO (fill) in boosted W+jets muon region in 2018. The error bars associated to the black markers reflects the MC statistical uncertainty only.

## 3.9 Systematic uncertainties

In the signal extraction fit, each uncertainty is represented by a nuisance parameter that morphs the template shape or scales the normalization of the templates. In the itemized list below, uncertainties indicated as correlated between three years are represented by a single common nuisance parameter, while ones indicated as uncorrelated are given three independent parameters. The correlation scheme of the nuisance parameters is also summarized in Table 3.17.

### 3.9.1 Experimental systematic uncertainties

#### 3.9.1.1 Uncertainties specific to individual background processes

- **Non-prompt background:** Non-prompt background is modeled by applying weights to a set of data events identified by leptons passing the loose selection criteria. Systematic uncertainties on the non-prompt background prediction arise from the limited size of the samples used to measure the fake rates and the difference in the flavor composition of the jets faking the leptons between the measurement sample and the signal region. The latter is assessed by varying the  $p_T$  of the jet that recoils against the lepton object in the dijet control sample used to measure the fake rate. Assuming that the control sample is dominated by events with two back-to-back jets, varying the  $p_T$  of the “away-side” jet for a fixed lepton object  $p_T$  corresponds to varying the fraction of the lepton object  $p_T$  to the total  $p_T$  of the jet that contains it, which in turn has an effect of probing different jet flavor compositions. The statistical uncertainty of the control samples affects the non-prompt background prediction by at most 10% and 5% for the samples electron and muon fakes, respectively. The effect of jet  $p_T$  variation is at most 8% and 7%, also from electron and muon fake rate measurements, respectively.

The jet  $p_T$  variation is, however, expected to not fully cover the uncertainty due to flavor variation. Therefore, a conservative 30% normalization uncertainty is additionally assigned to the fake background prediction.

Uncertainties on the non-prompt background from the statistical uncertainty in the measurement and the jet  $p_T$  variation are uncorrelated among three years. The additional 30% normalization uncertainties are correlated within each observable.

- **Top quark background:** Since the mixture of  $t\bar{t}$  and single top production samples are treated as one background, the internal ratio of the two processes is given an uncertainty of about 8%. The kinematic shapes of this background process are predicted by simulation and normalized to their theoretical predictions. Uncertainty on the shape and normalization is assessed by applying event-by-event weights corresponding to the up and down variations of the renormalization and factorization scales by a factor of two, correlated between the three years. The top quark background overall normalization is left floating in the fit independently for each year for each category and lepton flavour, and measured in its dedicated control regions. Renormalization and factorization scale uncertainties are evaluated and cast as a shape uncertainty that applies in all the regions. More details about the treatment of normalization uncertainties are in Sec.3.9.3.
- **W+jets background** A dedicated approach to correct the W+jets background estimation in a data-driven way is described in detail in Sec. 3.8. Renormalization and factorization scale uncertainties are evaluated and cast as a shape uncertainty that applies in all the regions. More details about the treatment of normalization uncertainties are in Sec.3.9.3.

### 3.9.1.2 Uncertainties affecting all MC-based background and signal modeling

This category mostly consists of uncertainties in MC modeling of reality. All uncertainties under this category are uncorrelated between the three years unless otherwise stated.

- **Integrated luminosity:** The uncertainty determined by the CMS luminosity monitoring is 1.2%, 2.3%, and 2.5% for 2016, 2017, and 2018 data sets, respectively. The total Run II (2016–2018) integrated luminosity has an uncertainty of 1.6%. The improvement in precision reflects the (uncorrelated) time evolution of some systematic effects [110–112].
- **Trigger efficiency:** Trigger efficiency uncertainties are less than 1%. These uncertainties are computed by varying the tag selection as well as the Z window in the tag and probe method used to compute the corresponding scale factors.
- **Prefiring correction:** Present only in 2016 and 2017 with an effect smaller of 0.2%: the variation is uncorrelated between years.
- **Lepton reconstruction and identification efficiency:** The lepton reconstruction and identification efficiencies are measured with the tag and probe method in data. To correct for the difference in the lepton identification efficiencies between data and MC, data/MC scale factors dependent on  $p_T$  and  $\eta$  are applied to the MC. The resulting uncertainty in the signal region is  $\sim 1\%$  for electrons and  $\sim 2\%$  for muon.
- **Muon momentum and electron momentum scale:** These uncertainties arise due to different detector effects and are  $p_T$  and  $\eta$ -dependent. Uncertainties on both the scale and resolution individually amount to  $\sim 0.6\text{--}1\%$  for electrons and  $\sim 0.2\%$  for muons. The effect on the signal region is about 1% from the muons and  $\sim 1\text{--}3\%$  from electrons.
- **Jet energy scale (JES) uncertainties:** We estimate the effect of this uncertainty by shifting the jet energy scale, and thereby the type-1 corrected  $p_T^{\text{miss}}$ , by  $\pm 1\sigma$  variations. The fit templates are modified in normalization as a result of this variation since the effect of this variation is found to be flat. JES uncertainty affects the rates in the signal region at the level of up to  $\sim 8\%$ . Some JES uncertainty components are uncorrelated for each year, but most of them are correlated across all the years. The JES uncertainties are also correlated between all the backgrounds and signal.
- **Jet energy resolution (JER) uncertainties:** We estimate the effect of this uncertainty by smearing the jet energy scale, by  $\pm 1\sigma$  variations. The total JER uncertainty is included correlated between all the years.
- **$p_T^{\text{miss}}$  unclustered component:** Uncertainties on  $p_T^{\text{miss}}$  that are not due to lepton and jet momentum scales are assessed by varying the momentum of each PF candidate that is not associated to a lepton or a jet according to the reconstructed particle type. The effect on the rates in the signal region is of the order of 1% to 10% depending on the process.
- **b-tagging scale factor:** Uncertainties on the b-tagging scale factor are evaluated by shifting the per-jet scale factors. There are ten distinct variations, corresponding to different sources of scale factor measurement uncertainty. Out of the ten, four variations accounting for the statistical uncertainties of the measurement samples are uncorrelated among the three years. The remaining six correlated variations are related to unknown jet flavor compositions of the samples.
- **Pileup reweighting:** Uncertainty in the pileup reweighting applied to the MC is taken into account and correlated between all categories and processes. The nui-

sance parameter is uncorrelated between years. Since we want to model only the kinematical variations due to the PU reweighting uncertainty, the normalization effect per sample of the up/down variations is computed after the inclusive single lepton preselections and removed from the templates.

- **Jet PU ID SF:** Uncertainty on the Jet PU ID scale factor is applied on all the jets of the analysis ( $p_T < 50$  GeV). The variation is uncorrelated between the years.
- **QGL morphing uncertainty** Quark Gluon Likelihood variables have been corrected using a custom quantile morphing procedure described in Sec 3.6. A shape uncertainty has been extracted from the closure of the method and included in the fit (see Sec. 3.6.3). Since the morphing functions have been derived separately for gluons and quarks at low  $\eta$  and high  $\eta$ , the uncertainty is split in four uncorrelated parameters. Finally the 2017 and 2018 morphing uncertainties are correlated since the same functions are applied. For 2016, given the different tune, a separate set of morphing functions has been computed.

### 3.9.1.3 Uncertainties concerning the use of boosted FatJets

- **$\tau_{21}$  boosted W tagging scale factor** Efficiency scale factors for AK8 boosted jets W/Z tagging based on the  $\tau_{21}$  selection are applied. The uncertainty on the efficiency scale factors is applied both on the high purity (boosted V analysis category) and low purity (events with fatjets failing the  $\tau_{21}$  selection fall in the resolved analysis region) categories in an anti-correlated way. The uncertainty is uncorrelated between the years and is up to 2% effect in the resolved category and  $\sim 6\%$  in the boosted category.
- **$\tau_{21}$   $p_T$  regime extrapolation uncertainty:** the effect of extrapolating the  $\tau_{21}$  selection to higher momenta of the W/Z boosted particles than the  $p_T$  200-300 GeV in the reference  $t\bar{t}$  sample, is taken into account. It has a negligible effect in the resolved category,  $\sim 3\%$  in the boosted one.
- **AK8 Jet energy scale (fatjetJES) and resolution (fatjetJER)** The fatjet jet energy is varied and smeared in the analysis in the same way as the AK4 jets. This uncertainty follows the same correlation schema as the AK4 JES and JER. The uncertainty is applied only on the boosted category since in the resolved category events containing one fatjet failing the  $\tau_{21}$  selection but with  $p_T > 180$  GeV are vetoed to avoid migration of events between the two categories.
- **AK8 Jet Mass Scale (fatjetJMS)** The fatjet soft-drop Puppi mass is scaled by  $\pm 1\sigma$  variation and the effect is propagated in the analysis. The uncertainty is applied only in the boosted category and uncorrelated between the years. The effect is always less than 1%.
- **AK8 Jet Mass Resolution (fatjetJMR)** The fatjet soft-drop Puppi mass is smeared by  $\pm 1\sigma$  variation and the effect is propagated in the analysis. The uncertainty is applied only in the boosted category and uncorrelated between the years. The effect is always less than 1%.

### 3.9.2 Theoretical uncertainties

All nuisance parameters associated to uncertainties in this category are correlated between three years unless stated otherwise.

- **Parton shower modeling:** Uncertainty in parton shower modeling primarily affects the jet multiplicity.

For the 2018 analyses, event reweighting with weights calculated by Pythia 8 is used to evaluate the uncertainty. Reweighting factors correspond to per-event cross section variations with ISR and FSR scales varied up and down by a factor of two. The envelope of the variation distributions are considered as up and down variations. The parton shower uncertainties computed in 2018 are applied also on 2016 and 2017 samples where the weights are not available.

Because we are interested only in modeling the kinematics variations due to the parton shower uncertainty, the variations per sample are renormalized to give the overall event yield prediction identical to the nominal prediction in an inclusive phase space after the single lepton preselections.

- **Renormalization and factorization scale:** Missing higher-order terms in the perturbation series of the cross section calculations will also affect the shapes and the overall cross-sections. For the background processes, MC simulation events are reweighted to alternative conditions where renormalization and factorization scales are varied up and down by a factor of two, and the envelopes of the varied distributions are taken as one standard-deviation variations. Uncertainties on VBS semileptonic signal process are split into overall cross section and event acceptance in the inclusive phase space after single lepton preselection cuts are applied. The former is absorbed in the signal strength parameter, the latter is calculated by means of event reweighting. Moreover, the acceptance effect on the VBS signal is not included in the measurement of the cross-section, while it is included in the uncertainty of the signal strength parameter.
- **Parton distribution function:** The PDF uncertainty for the background processes are evaluated by reweighting the MC simulation events with different PDF sets based on parton-level kinematics, and then adding the resulting weights over the events in the signal and control regions. The PDF affects the distribution shapes as well as the overall cross-sections of the processes. The PDF uncertainty is ignored for backgrounds measured directly in control regions, such as  $W$ +jets and top. Uncertainties on VBS semileptonic signal process are split into overall cross section and event acceptance in the inclusive phase space after single lepton preselection. The former is absorbed in the signal strength parameter, the latter is calculated by means of event reweighting.
- **Underlying event modeling:** Uncertainty on underlying event (UE) is evaluated by varying with their uncertainties the underlying-event UE tunes (CUETP8M1 for the 2016 analysis, CP5 for 2017 and 2018) of Pythia8. From previous studies the effect on the number of jets distribution is small: therefore only a 1.5% effect on the overall normalization is applied for all MC except  $W$ +jets and top.

### 3.9.3 Special treatment for normalization uncertainty for Top and $W$ +jets

The normalization of top and  $W$ +jets processes are measured in dedicated control regions, and the correction is propagated in the fit with a single parameter in the signal region yields. A distinct rate parameter is defined for each year, category (resolved, boosted) and lepton flavour. Since the QCD scale and parton shower uncertainties have a large normalization effect, their variation for each sample (top quark and  $W$ +jets categories) is re-normalized to give the overall event yield prediction identical to the nominal prediction summing all channels in which the rate parameter of the sample is included. For example the overall up/down yields for QCD scale variation of the top sample in the electron resolved category is normalized looking both at the signal channel and at top and  $W$ +jets control regions. In this way both the shape ef-

fects and the migration effect between signal and control regions are modeled, but the overall normalization is removed.

### 3.9.4 Impact plots

The impact of each nuisance parameter on the best-fit measurement of the signal strength is evaluated following this procedure:

1. Take the best-fit value of the nuisance parameter  $\hat{\theta}$  and its 68% confidence level interval  $\sigma_{\theta}^{\pm}$  (estimated via the Hessian matrix computed at the minimum of the likelihood).
2. Set  $\theta$  at the values  $\theta_{\pm} = \hat{\theta} \pm \sigma_{\theta}^{\pm}$
3. Perform a new maximum likelihood fit keeping the  $\theta_{\pm}$  value fixed and profiling all the other parameters.
4. Extract the impact on the parameter of interest (the signal strength) by taking the distance of the new best-fit value from the global minimum one ( $\hat{\mu}$ ).  $I_{\mu}^{\pm} = \hat{\mu}_{\theta_{\pm}} - \hat{\mu}$ .

The best-fit and impact values of the nuisance parameters are summarized in the Fig. 3.79. The parameters are ordered by their impact and only the most important ones are shown. The impacts of the normalization parameters used to correct the W+jets and top quark backgrounds are shown separately.



| Uncertainty                                 | Process                   | Type          | Correlation                                             |
|---------------------------------------------|---------------------------|---------------|---------------------------------------------------------|
| Integrated luminosity                       | all MC except W-jets, top | normalization | partially correlated                                    |
| Fake rate (statistical and away-jet $p_T$ ) | non-prompt                | shape         | uncorrelated                                            |
| Fake rate (30% for jet composition)         | non-prompt                | normalization | uncorrelated by lepton flavour                          |
| B-tag SF                                    | all MC                    | shape         | partially correlated                                    |
| Trigger efficiency                          | all MC                    | shape         | uncorrelated                                            |
| Prefiring weight                            | all MC                    | shape         | uncorrelated, only for 2016 and 2017                    |
| Lepton ID efficiency                        | all MC                    | shape         | uncorrelated                                            |
| Lepton $p_T$ scale                          | all MC                    | shape         | uncorrelated                                            |
| Jet energy scale (AK4)                      | all MC                    | normalization | partially correlated                                    |
| Jet energy resolution (AK4)                 | all MC                    | normalization | uncorrelated                                            |
| Fatjet energy scale (AK8)                   | all MC                    | normalization | partially correlated                                    |
| Fatjet energy resolution (AK8)              | all MC                    | normalization | uncorrelated                                            |
| Fatjet mass scale (AK8)                     | all MC                    | normalization | uncorrelated                                            |
| Fatjet mass resolution (AK8)                | all MC                    | normalization | uncorrelated                                            |
| $\tau_{21}$ SF                              | all MC                    | shape         | uncorrelated                                            |
| $\tau_{21}$ $p_T$ extrapolation             | all MC                    | shape         | uncorrelated                                            |
| Jet PU ID SF                                | all MC                    | shape         | uncorrelated                                            |
| Unclustered MET                             | all MC                    | shape         | uncorrelated                                            |
| Pileup reweighting                          | all MC                    | normalization | uncorrelated                                            |
| Single top tt composition                   | all MC                    | normalization | uncorrelated                                            |
| Top $p_T$ reweighting                       | top                       | shape         | correlated                                              |
| QGL morphing                                | all MC                    | shape         | correlated                                              |
| PS FSR and ISR                              | all MC                    | shape         | uncorrelated 2016 and 2017-18, split in four components |
| QCD renormalization and factorization scale | all MC                    | shape         | correlated for each background                          |
| PDF                                         | all MC except W-jets, top | shape         | uncorrelated for 2016 and 2017-2018 periods             |
| UE                                          | all MC except W-jets, top | shape         | uncorrelated for 2016 and 2017-2018 periods             |

Table 3.17: Summary of uncertainty applied in the template fit and correlation between years.

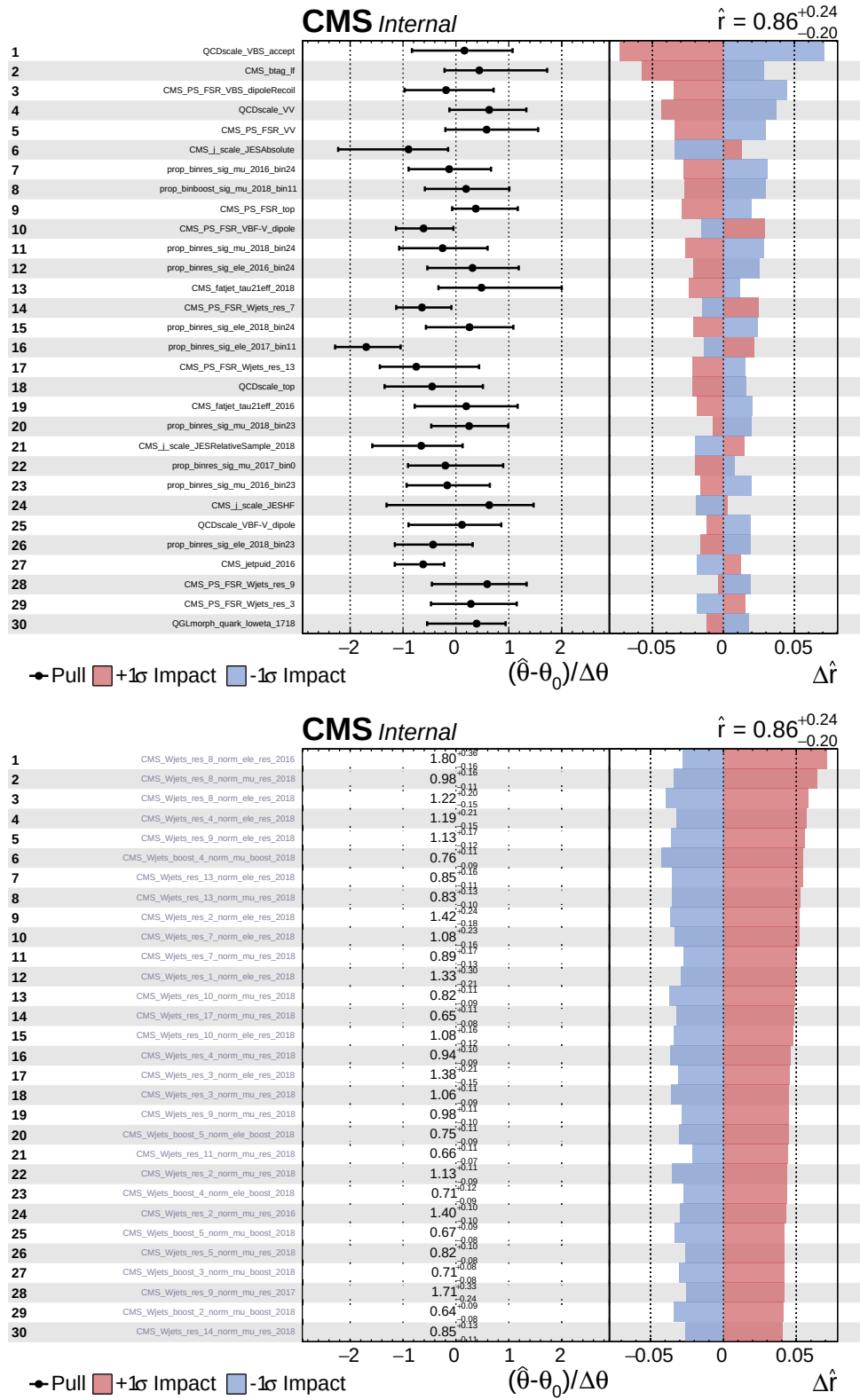


Figure 3.79: Best-fit value and impact of most important nuisance and normalization parameters extract from the full Run II fit on data.

## 3.10 Results

### 3.10.1 Statistical setup

#### 3.10.1.1 Signal significance extraction

Given an observed data sample, claiming the discovery of a new signal requires determining that the sample is sufficiently *inconsistent* with the hypothesis that only background is present in the data. A test statistic  $t$  can be used to measure the inconsistency of the observation in the hypothesis of the presence of background only, typically assumed as a null hypothesis  $H_0$ . The probability that the test statistic under  $H_0$  assumed a greater value of the observed one (assuming implicitly that larger values of  $t$  correspond to a more signal-like sample) is called the *p-value*:

$$p = \int_{t_{obs}}^{+\infty} g(t|H_0) dt \quad (3.12)$$

where  $g(t|H_0)$  is the PDF of the test statistic  $t$  under the null hypothesis. Usually the p-value is reported as a *significance* level using the equivalent number of standard deviations that correspond to an area equal to the p-value under the right-most tail of a normal distribution.

#### 3.10.1.2 Likelihood ratio test statistic

Typically, two hypotheses are considered:  $H_1$  represents the presence of both signal and background, i.e. the expected event yields is  $\nu = \mu s + b$ , while  $H_0$  represents the presence of only background events in the sample,  $\nu = b$ . The multiplier  $\mu$  is called *signal strength* and regulates the amount of expected signal yield from the theory  $s$  is present. A value of  $\nu = 0$  corresponds to signal only,  $\nu = 1$  corresponds to the theory prediction [113].

The best test statistic is given by the Neyman-Pearson lemma as the likelihood ratio:

$$\lambda(\mu, \vec{\theta}) = \frac{\mathcal{L}_{s+b}(\vec{x}_1, \dots, \vec{x}_n | \mu, \vec{\theta})}{\mathcal{L}_b(\vec{x}_1, \dots, \vec{x}_n | \vec{\theta})} \quad (3.13)$$

where  $\vec{\theta}$  are the parameters of the model that can be divided in parameters of interest and nuisance parameters.

A minimum of  $-2 \log \lambda(\mu)$  at  $\mu = \hat{\mu}$  indicates the possible presence of a signal having a signal strength equal to  $\hat{\mu}$ . To extract the significance of this signal the distribution of the test statistic under the background only hypothesis ( $\mu = 0$ ) has to be calculated: a possible way is generating a large number of Monte Carlo pseudo-experiments assuming no signal. The other possibility is to use Wilk's theorem approximation.

#### 3.10.1.3 Profile likelihood

Assume that two hypotheses  $H_0$  and  $H_1$  can be defined in terms of a set of parameters  $\vec{\theta} = (\theta_1, \dots, \theta_m)$  that appear in the definition of the likelihood function. The condition that  $H_1$  is true can be expressed as  $\vec{\theta} \in \Theta_1$ , while the condition that  $H_0$  is true can be expressed as  $\vec{\theta} \in \Theta_0$ . Let's assume that the  $H_0$  and  $H_1$  are nested hypothesis,  $\Theta_0 \subseteq \Theta_1$ .

Given a data sample made of independent measurements  $(\vec{x}_1, \dots, \vec{x}_N)$ , Wilk's theorem [114] ensures that the quantity:

$$\chi^2 = -2 \log \frac{\sup_{\vec{\theta} \in \Theta_0} \prod_{i=1}^N \mathcal{L}(\vec{x}_i; \vec{\theta})}{\sup_{\vec{\theta} \in \Theta_1} \prod_{i=1}^N \mathcal{L}(\vec{x}_i; \vec{\theta})} \quad (3.14)$$

has a distribution that can be approximated, for  $N \rightarrow \infty$ , and if  $H_0$  is true, with a  $\chi^2$  distribution having a number of degrees of freedom equal to the difference between the dimensionality of the set  $\Theta_1$  and  $\Theta_0$ .

As a specific example, let's assume that  $\vec{\theta} = (\mu, \vec{\theta})$  where  $\mu$  is the signal strength and  $\vec{\theta}$  are nuisance parameters. The  $H_0$  hypothesis corresponding to fixed  $\mu = \mu_0$  is a nested hypothesis. So the equation becomes:

$$\chi^2(\mu_0) = -2 \log \frac{\sup_{\vec{\theta}} \prod_{i=1}^N \mathcal{L}(\vec{x}_i; \mu_0, \vec{\theta})}{\sup_{\mu, \vec{\theta}} \prod_{i=1}^N \mathcal{L}(\vec{x}_i; \mu, \vec{\theta})} \quad (3.15)$$

The denominator is the likelihood function evaluated at the parameters  $\mu = \hat{\mu}$  and  $\vec{\theta} = \hat{\vec{\theta}}$  that maximise it. In the numerator only the nuisance parameters are fitted and  $\mu$  is fixed to the constant value  $\mu = \mu_0$ .  $\hat{\vec{\theta}}$  is the value of  $\vec{\theta}$  that maximize the likelihood at  $\mu_0$  fixed.

The final test statistic is known as the *profile likelihood*:

$$\chi^2(\mu_0) = -2 \log \frac{\mathcal{L}(\vec{x}|\mu_0, \hat{\vec{\theta}})}{\mathcal{L}(\vec{x}|\hat{\mu}, \hat{\vec{\theta}})} \quad (3.16)$$

Going back to the likelihood ratio test statistic in equation 3.13 the significance can be extracted using the *profile likelihood*: according to Wilk's theorem the distribution of  $2 \log \lambda(\hat{\mu})$  can be approximated by a  $\chi^2$  distribution with the number of degrees of freedom equal to the number of nuisance parameters. The significance from the test statistic is calculated thanks to approximation formulae [113] as

$$Z_0 \simeq \sqrt{2 \log \lambda(\hat{\mu})}$$

#### 3.10.1.4 Nuisance parameters

Uncertainties on the prediction of signal and background yields are handled introducing nuisance parameters  $\vec{\theta}$ . The systematic errors are not exactly known but they are assigned a p.d.f.  $\rho(\theta|\vec{\theta})$ , where  $\vec{\theta}$  is the default values of the nuisance parameter. This p.d.f. can be re-interpreted as a posterior probability from some real measurements of  $\vec{\theta}$ :

$$\rho(\theta|\vec{\theta}) \sim p(\vec{\theta}|\theta) \cdot \pi(\theta) \quad (3.17)$$

where  $\pi(\theta)$  is a prior for this "measurement", usually considered flat. Such shift in the point of views allows one to repress all the systematic errors in a frequentist context and insert  $p(\vec{\theta}|\theta)$  in the likelihood [115].

### 3.10.2 Inputs to the likelihood fits

The final analysis maximum-likelihood fits are performed combining the information from the signal regions and control regions. For each year the DNN score variable in the signal region, evaluated separately for electron and muon channels, is used in the fit for the likelihood ratio test. W+jets and Top regions are also used in the fit to control the contribution of the backgrounds. In particular only the number of events in the Top control regions is inserted in the fit to be able to constrain the normalization of the top background, whereas in the W+jets control

region the normalization of the different W+jets subcategories is used in the fit to correct for the data/MC discrepancies (as described in details in the Sec 3.8).

The likelihood is then defined as [115]:

$$\mathcal{L}(\text{data}|\mu, \vec{\theta}) = \text{Poisson}(\text{data}|\mu s(\vec{\theta}) + b(\vec{\theta})) \cdot p(\vec{\theta}) \quad (3.18)$$

The  $\text{Poisson}(\text{data}|\mu s(\vec{\theta}) + b(\vec{\theta}))$  stands for the product of Poisson probabilities to observe  $n_i$  events in bins  $i$ :

$$\mathcal{L}(n_1, \dots, n_m; \mu, \vec{\theta}) = p(\vec{\theta}) \cdot \prod_i^m \frac{(\mu s_i(\vec{\theta}) + b_i(\vec{\theta}))^{n_i}}{n_i!} e^{-\mu s_i(\vec{\theta}) - b_i(\vec{\theta})} \quad (3.19)$$

where  $s_i(\vec{\theta})$  is the expected signal yield for each bin and  $b_i(\vec{\theta})$  the one of the background.

Fig. 3.80 shows the shapes fit in the signal region. Fig. 3.81 shows the subcategories normalizations fit in the W+jets resolved control region, whereas Fig. 3.82 shows the prefit normalization of the top process.

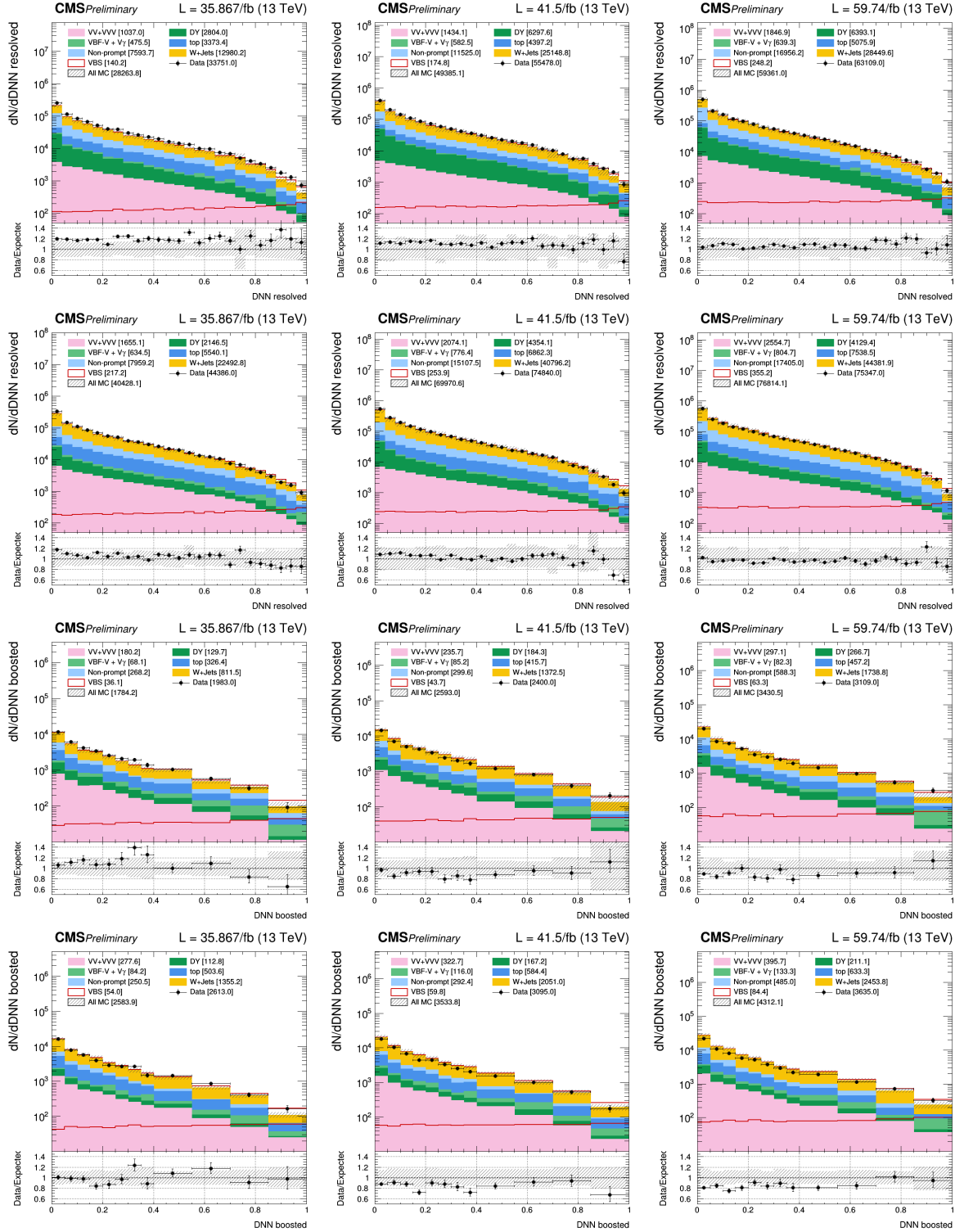


Figure 3.80: DNN discriminator used in the fit. From top to bottom: Resolved electron, muon; Boosted electron, muon channels. From left to right: 2016, 2017, 2018.

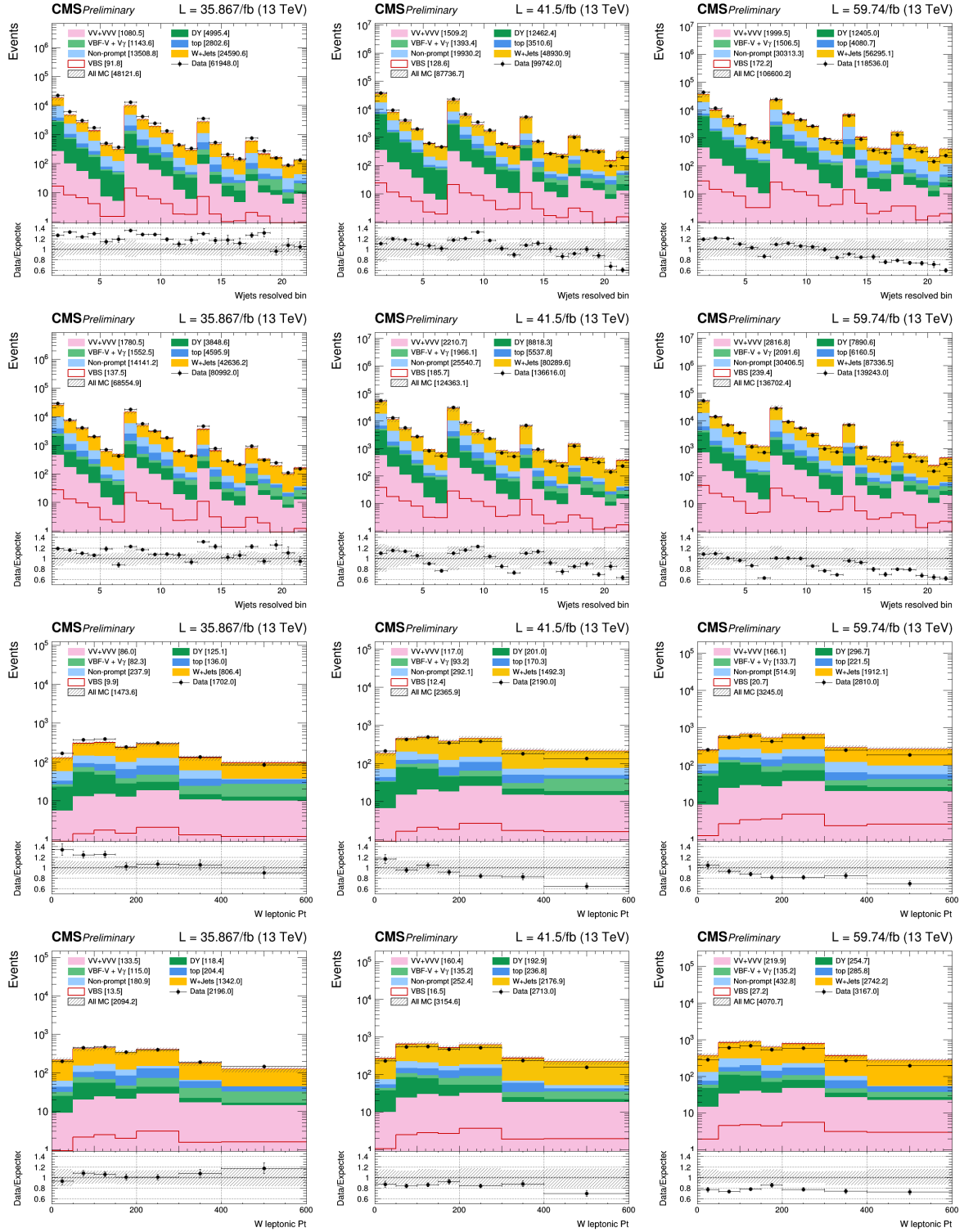


Figure 3.81: W+jets subcategories normalizations in the W+jet control regions. This variable is used in the fit to constrain the W+jets background normalization and shape. From top to bottom: Resolved electron, muon. Boosted electron, muon. From left to right: 2016, 2017, 2018.

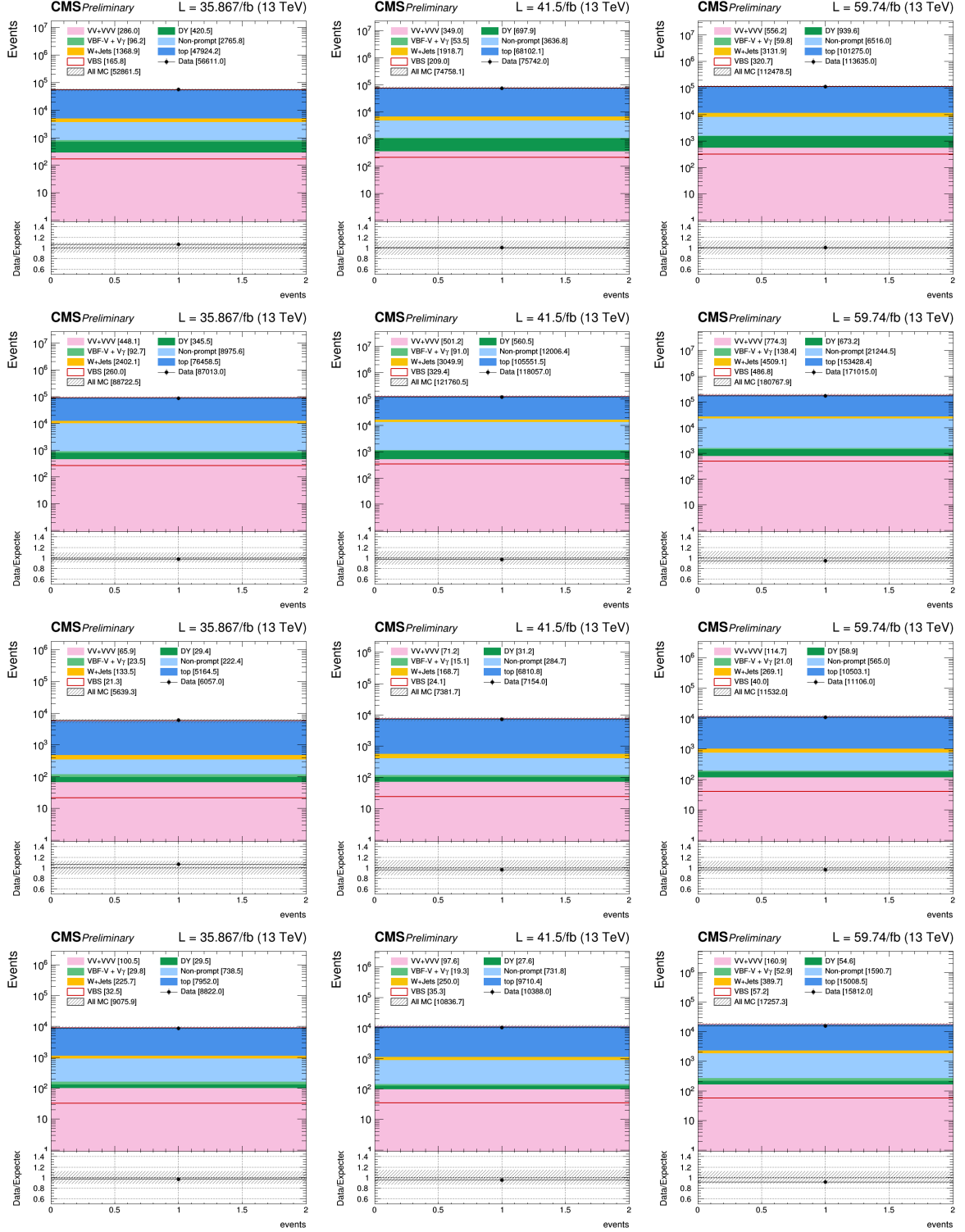


Figure 3.82: Top normalizations in the top control regions. From top to bottom: Resolved, Boosted and electron,muon. From left to right: 2016, 2017, 2018.



### 3.10.3 Measurements

Three separated maximum likelihood fits are performed: the measurement of the purely EW signal strength  $\mu_{\text{EW}}$  keeping the QCD WV production contribution fixed to the SM prediction  $\mu_{\text{QCD}} = 1$ ; the measurement of the signal strength considering as signal the EW and QCD WV processes together; a two-dimensional simultaneous measurement of the signal strengths  $\mu_{\text{EW}}$  and  $\mu_{\text{QCD}}$ .

#### 3.10.3.1 EWK signal strength

Figure 3.83 shows the post-fit DNN distribution for the resolved (left) and the boosted (right) signal phase space in the EW-only fit. The background-subtracted plot, where the evidence for the signal is clearly visible, is also shown.

A fiducial phase space region is defined at parton level requiring all partons to have  $p_T > 10 \text{ GeV}$  and at least one pair of outgoing partons with invariant mass  $m_{qq} > 100 \text{ GeV}$ . The SM prediction for the EW WV production cross section in this fiducial region is  $2.23^{+0.08}_{-0.11} (\text{scale}) \pm 0.05 (\text{PDF}) \text{ pb}$ . The measured EW WV production cross section is  $1.90^{+0.53}_{-0.46} \text{ pb}$ , corresponding to an observed EW-only signal strength of:

$$\mu_{\text{EW}} = \sigma^{\text{obs}} / \sigma^{\text{SM}} = 0.85 \pm 0.12 (\text{stat})^{+0.19}_{-0.17} (\text{syst}) = 0.85^{+0.23}_{-0.21}, \quad (3.20)$$

with  $1^{+0.24}_{-0.22}$  expected, where  $\sigma^{\text{obs}}$  and  $\sigma^{\text{SM}}$  are the observed and predicted cross section, respectively. The observed significance for the SM EW WV signal is 4.4 standard deviations with 5.1 expected. The contribution of each channel to the total expected and observed signal significance is shown in Table 3.18. The likelihood scan for the EWK signal strength is also shown in Fig. 3.84.

Table 3.18: Expected and observed significance for the different years and categories.

|          | Category | 2016 | 2017 | 2018 | Combined |
|----------|----------|------|------|------|----------|
| Expected | Boosted  | 2.3  | 2.0  | 2.7  | 3.7      |
|          | Resolved | 2.4  | 1.8  | 3.2  | 3.6      |
|          | Combined | 3.1  | 2.6  | 4.1  | 5.3      |
| Observed | Boosted  | 1.4  | 1.5  | 3.8  | 3.8      |
|          | Resolved | 2.4  | 0.0  | 3.7  | 3.6      |
|          | Combined | 2.2  | 0.7  | 4.9  | 4.4      |

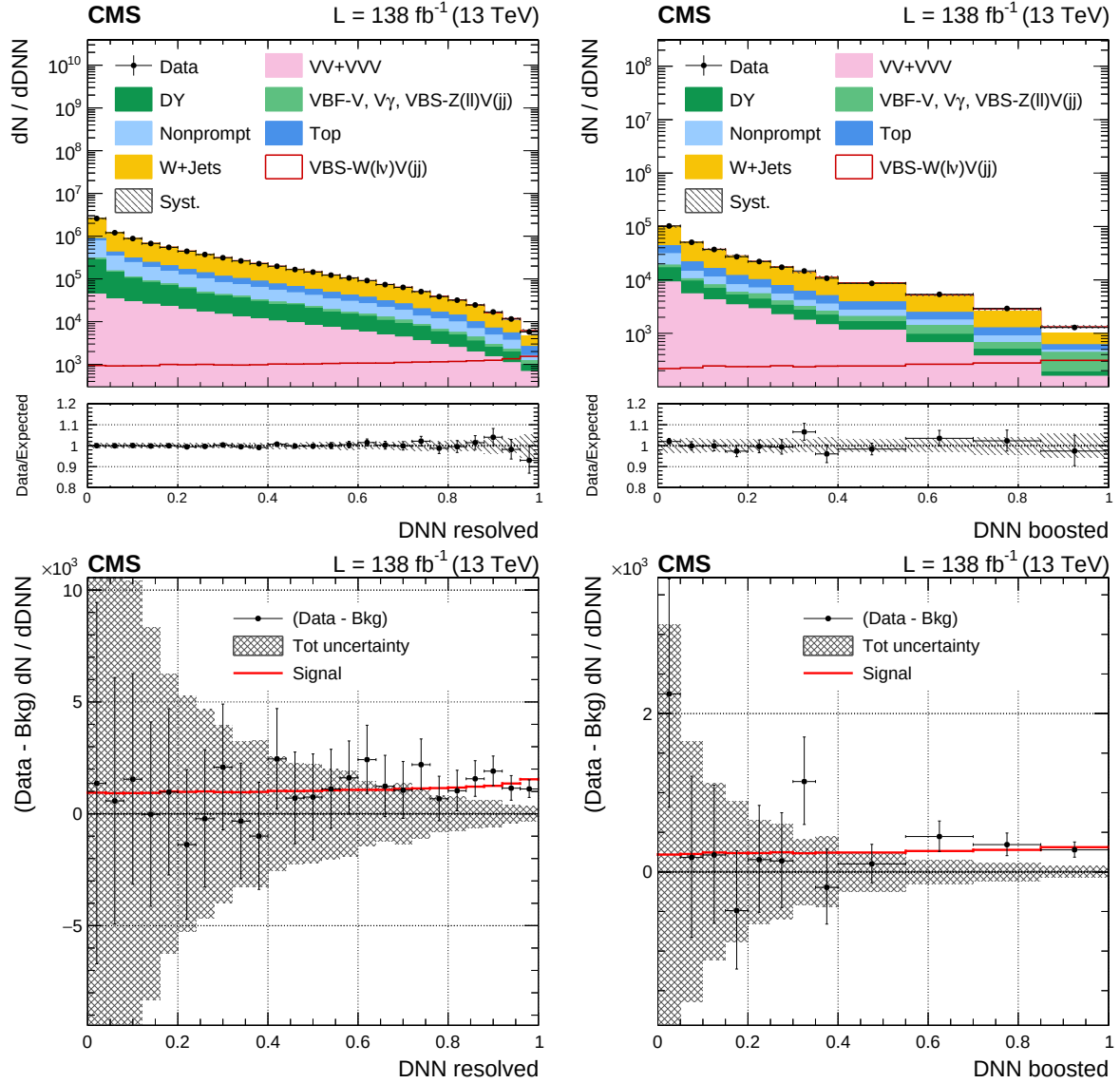


Figure 3.83: Results for the EW-signal-only fit, keeping the QCD WV contribution fixed to the SM prediction. Upper plots: post-fit DNN discriminator distributions for the resolved (left) and the boosted (right) signal regions. The signal contribution is plotted both stacked on top of the background processes and also overlaid to show the signal postfit distribution. The expected yield is the sum of signal and background. Lower plots: background-subtracted DNN discriminator distribution for the resolved (left) and the boosted (right) categories. Post-fit background yields in each bin are subtracted from data and compared with the signal post-fit distribution, plotted as a red line. The post-fit uncertainty band is also shown in the plots with all systematic uncertainties included.

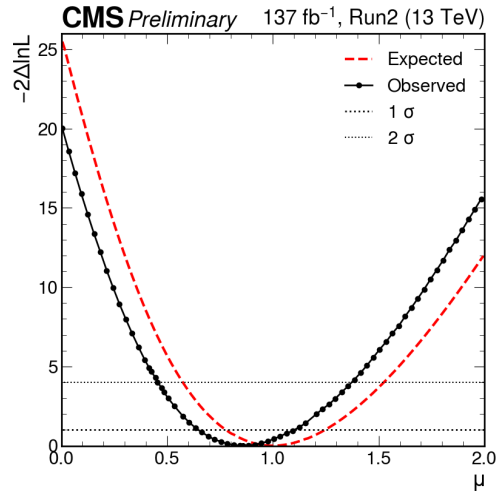


Figure 3.84: Likelihood scan versus signal strength from Run II data.

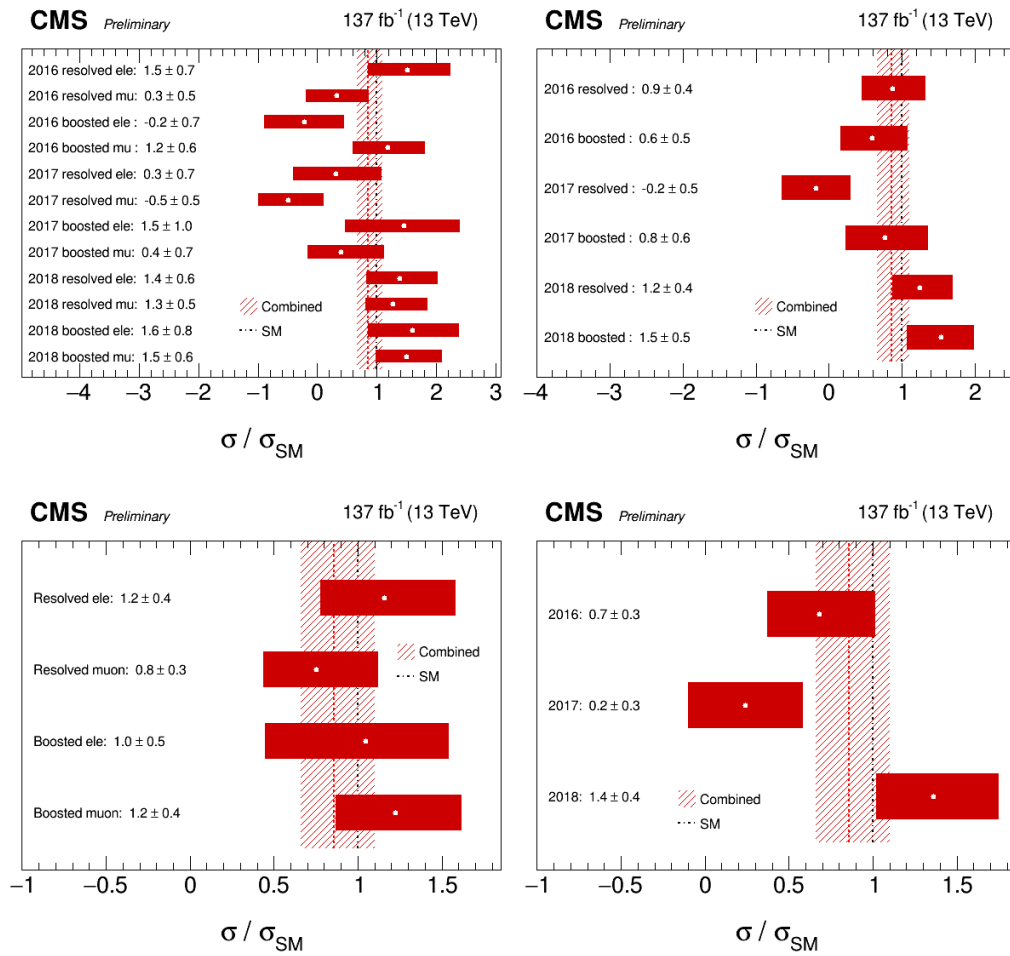


Figure 3.85: Signal strength measured in data in each channel (top left), combining the lepton flavour categories (top right), in each jet category (bottom left) and in each year (bottom right).

### 3.10.3.2 EWK and QCD WV production

Considering instead the signal as the overall EW and QCD WV production, the cross-section is measured to be  $16.4^{+3.5}_{-2.8}$  pb,  $16.9^{+2.9}_{-2.1}$  (scale)  $\pm 0.5$  (PDF) expected, in the same fiducial phase space region as the EW-only one. The overall signal strength  $\mu = \sigma^{\text{obs}} / \sigma^{\text{SM}}$ ,  $1^{+0.21}_{-0.20}$  expected, is measured as:

$$\mu_{\text{EW+QCD}} = \sigma^{\text{obs}} / \sigma^{\text{SM}} = 0.97 \pm 0.06 \text{ (stat)}^{+0.19}_{-0.21} \text{ (syst)} = 0.97^{+0.20}_{-0.22}, \quad (3.21)$$

The fit is performed also leaving as free independent parameters the signal strength of the EW and QCD WV production components ( $\mu_{\text{EW}}$  and  $\mu_{\text{QCD}}$ ). The result of the 2D fit is shown in Fig. 3.86, where the expected and observed minima are presented, together with the 68 and 95% confidence level contours built from the likelihood function. The measured signal strengths are in agreement with the SM predictions within the 68% confidence interval.

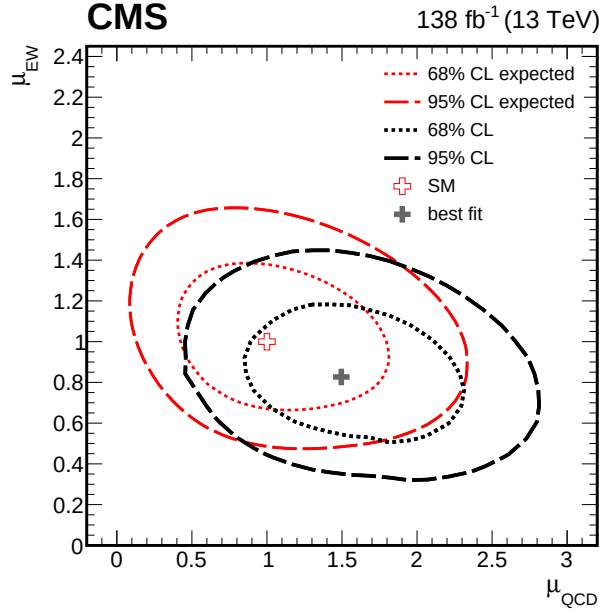


Figure 3.86: Simultaneous EW and QCD WV production fit: the expected and observed 68 and 95% confidence level (CL) contours on the signal strengths are shown in the plot. The best fit result is compatible with the SM prediction within the 68% confidence level area.

Overall, the different fit results are consistent between each other, and the two-dimensional one shows a small correlation between the EW and QCD component; in all cases, the measurement is in agreement with the SM predictions within the 68% confidence interval.

## 3.11 Fit checks

### 3.11.0.1 Goodness of fit test

A goodness of fit test is performed using the saturated likelihood model as prescribed by the CMS Statistical Committee [116]. Toys are generated after a preliminary fit on data keeping the signal strength fixed to 1, to adjust the initial values of the nuisance parameters. The distribution of the test statistic  $t$  estimated with toys is shown in Fig 3.87: the data is found to be compatible with the model.

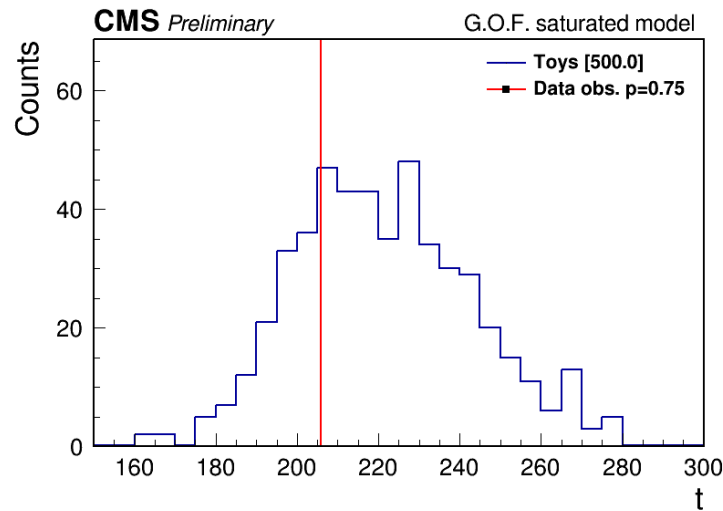


Figure 3.87: Goodness-of-fit test statistic estimated with toys and compared to the observed value for data.

### 3.11.1 W+jets post-fit normalizations summary

In Figure 3.88 and 3.89 the ratio between the post-fit and pre-fit W+jets subcategories normalizations is shown for each year and lepton flavour in the resolved and per=category phase spaces. Each bin represents a W+jets subcategory. The empty boxes represent the prefit correction factors and their (statistical and systematic) uncertainties computed to correct the data/MC ratio in the W+jets control regions (for each year and lepton flavour) as explained in the section 3.8.4. The marker and filled bars show the post-fit correction factors and the post-fit uncertainties. The postfit values are computed using a background-only fit on the data-asimov dataset in the control regions only, including both W+jets and top control regions.

The normalization factors obtained for 2017 and 2018 for each lepton flavour are compatible. The 2016 normalization factors follow a similar trend with a different initial normalization: the difference is given by the different tune and generator configuration of the W+jets MC sample between 2016 and 2017-2018. The normalization factors in the boosted category follow a similar trend as the resolved ones.

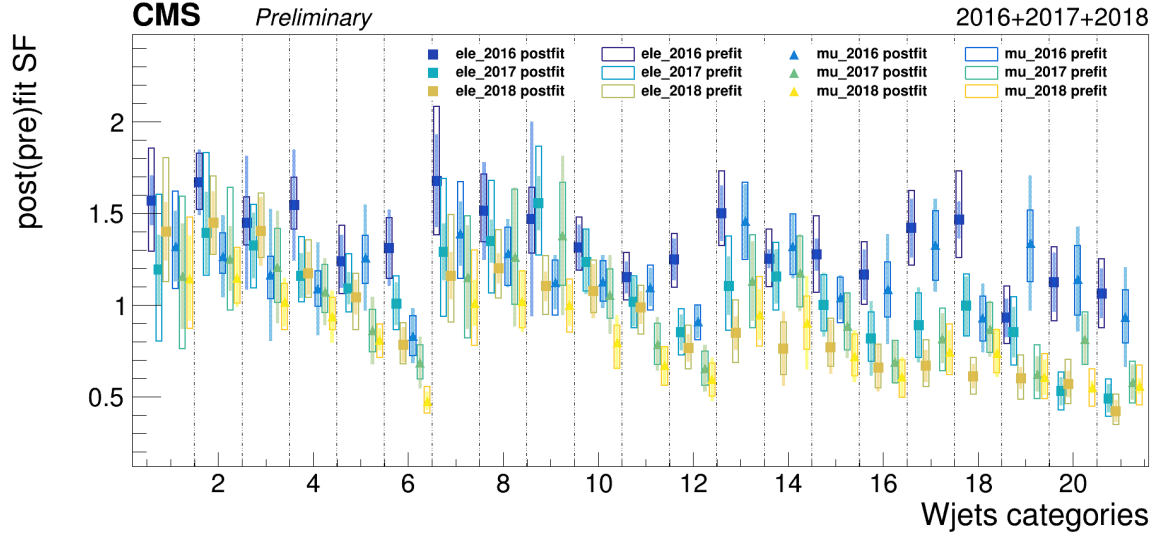


Figure 3.88: Plot of the pre-fit and post-fit W+jets subcategories normalization correction factors for each year and lepton flavour. Each bin represents a W+jets subcategory in the resolved phase space. The boxes represent the prefit correction factors and their uncertainties. The marker and filled bars show the post-fit correction factors and the post-fit uncertainties.

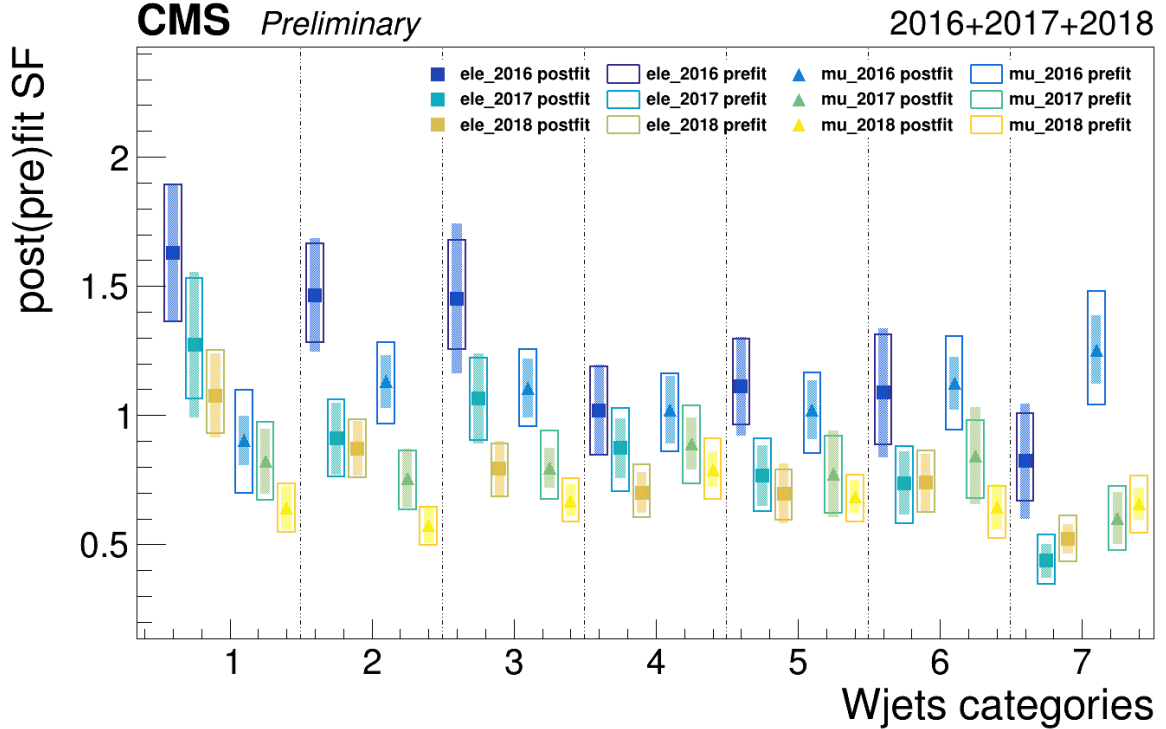


Figure 3.89: Plot of the pre-fit and post-fit W+jets subcategories normalization correction factors for each year and lepton flavour. Each bin represents a W+jets subcategory in the boosted phase space. The boxes represent the prefit correction factors and their uncertainties. The marker and filled bars show the post-fit correction factors and the post-fit uncertainties.

### 3.11.2 Postfit plots

The post-fit distribution of some of the most important DNN input variables are shown in the following figures combining all years and flavours in the signal regions. They are computed using the parameters extracted on the EW-only fit performed on the data. The postfit uncertainties on the shapes are computed exploiting the full fit parameters covariance matrix by means of toys and are included in the plots. The variables used to extract the W+jets data-driven corrections are also included: the correction constrained in the fit from the W+jets control regions is correctly fixing the W+jets contribution also in the signal regions.

Some discrepancy is still present in the post-fit plots in the  $\Delta\eta_{jj}^{\text{VBS}}$  and  $Z_{\text{lep}}$  variables, therefore dedicated bias tests have been conducted to assess the impact on the DNN distribution before performing the final unblinding of the analysis results. The bias tests are described in Sec 3.11.3.

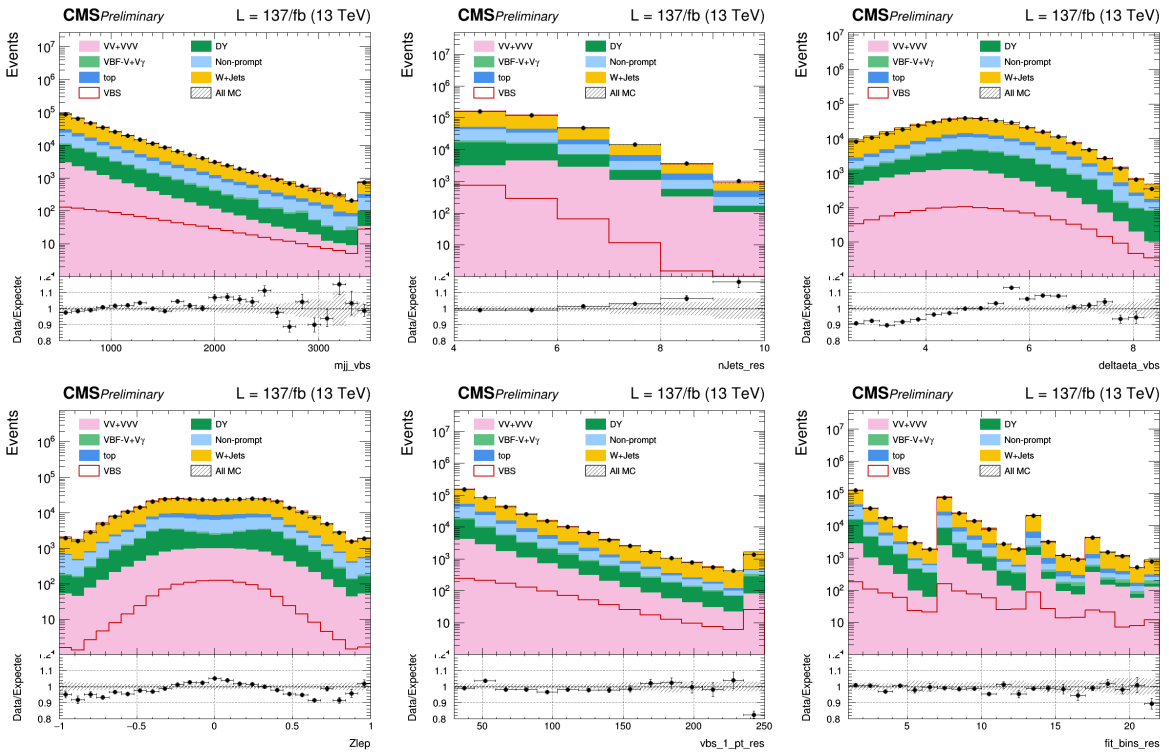


Figure 3.90: Signal, resolved category. Combined full Run II post-fit distribution.

### 3.11.3 Residual discrepancies bias test

The unblinded post-fit distributions in the signal region of  $\Delta\eta_{jj}^{\text{VBS}}$  and  $Z_{\text{lep}}$  (see Fig. 3.90) show a residual discrepancy after the W+jets and top background correction in the fit. We perform a simple bias test to assess the impact of the discrepancy on the DNN distribution before unblinding it: the prefit distribution of the input variable under test is reweighted to perfect agreement with data and the impact of this one-sided uncertainty on the DNN distribution is checked, both in the W+jets control region and in the signal region. The test is performed on the 2018 dataset, after the rescaling the W+jets subcategories normalization.

Figure 3.92 show the residual corrections extracted from prefit distributions. Figure 3.93 shows the effect on the prefit DNN compared with the postfit distribution and the total uncertainty in the W+jets control region and signal region for the  $\Delta\eta_{jj}^{\text{VBS}}$  discrepancy. The effect of the

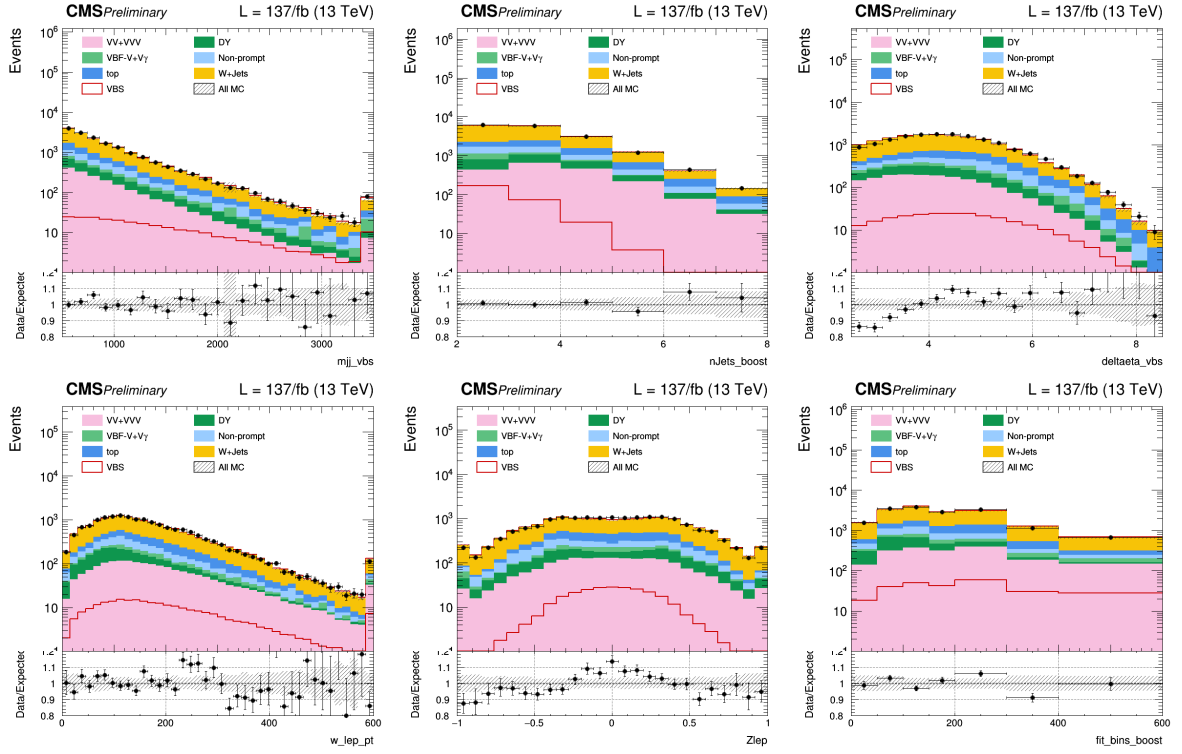


Figure 3.91: Signal, boosted category. Combined full Run II post-fit distribution.

reweighting is almost negligible on the DNN distribution. The same effect is shown in Fig. 3.94 of the  $Z_{lep}$  distribution. The conclusion of the test is that the effect of the  $\Delta\eta_{jj}^{VBS}$  and  $Z_{lep}$  mis-modeling is negligible with respect to the total postfit uncertainty. Plots are shown only in the muon channel for brevity.



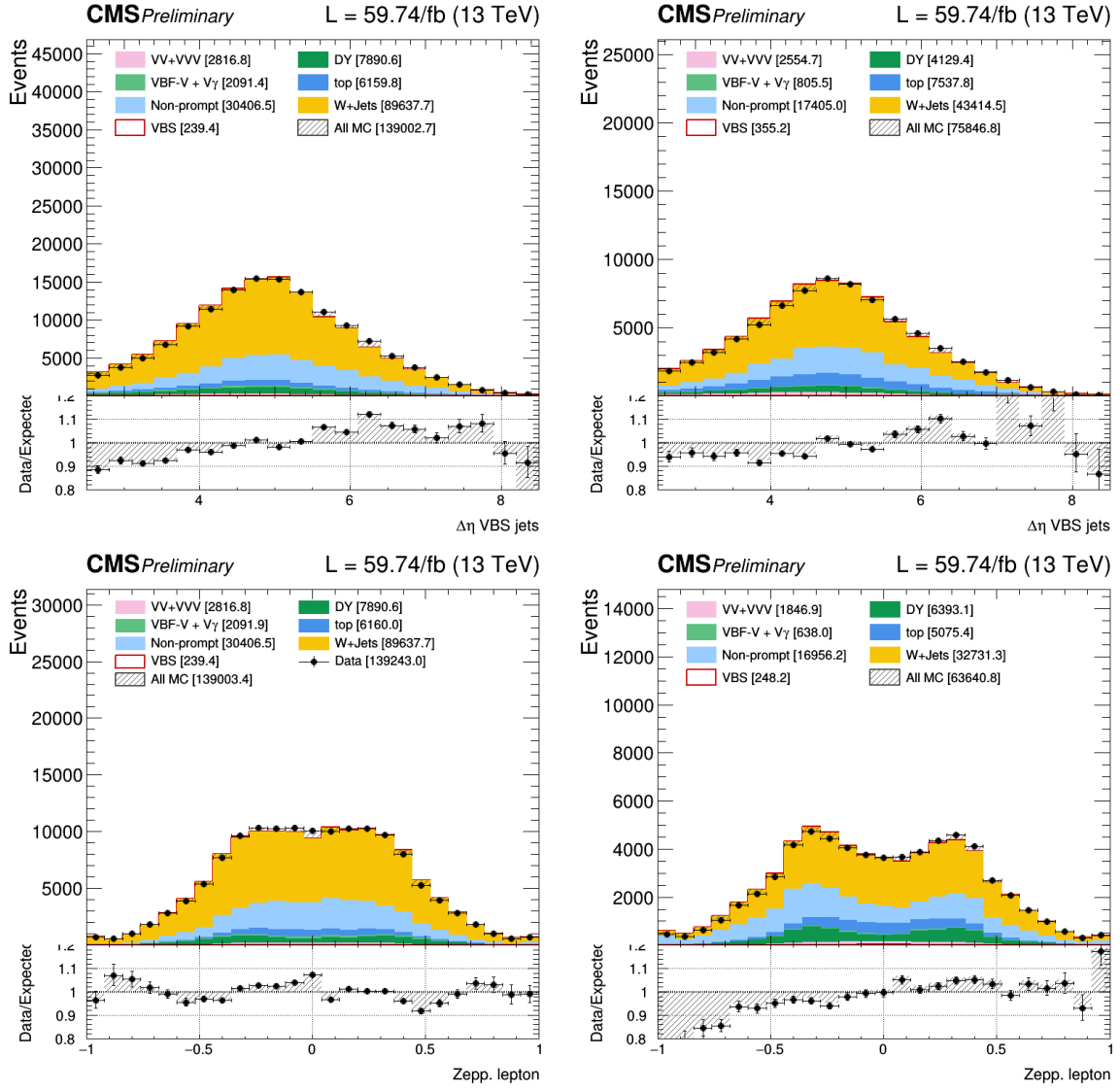


Figure 3.92: Residual discrepancy in post-fit  $\Delta\eta_{jj}^{\text{VBS}}$  distributions (first line) and  $Z_{\text{lepton}}$  (second line) in the resolved W+jets control region (left) and signal region (right). Muon channel only. The dashed band shows the one-sided uncertainty computed reweighting the observable to perfect agreement.

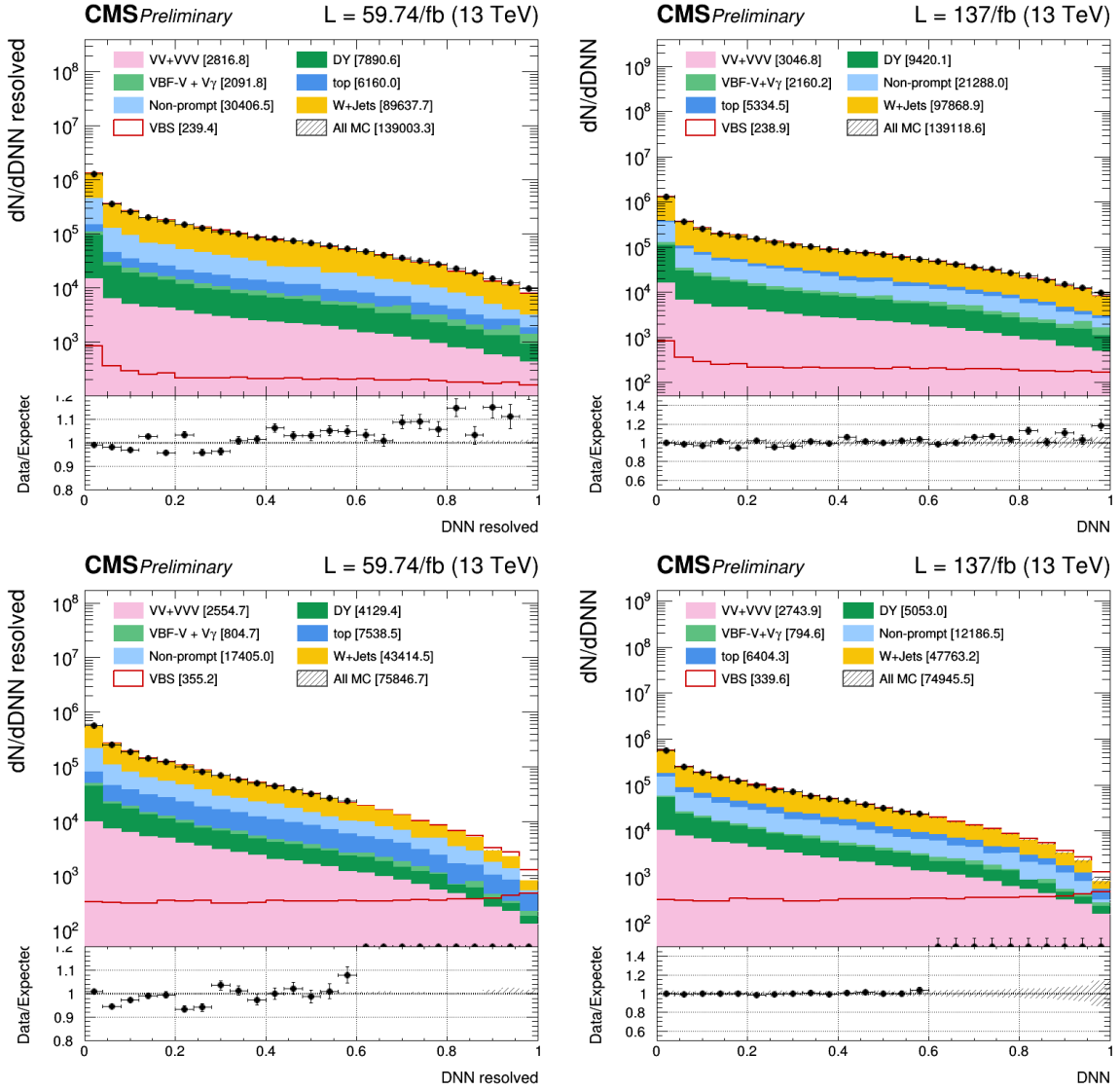


Figure 3.93: Muon, resolved control region. Comparison between the effect of reweighting the residual discrepancy on the prefit DNN (left) and the total postfit shape and uncertainty. W+jets control region distributions on the first line, the signal region in the second.

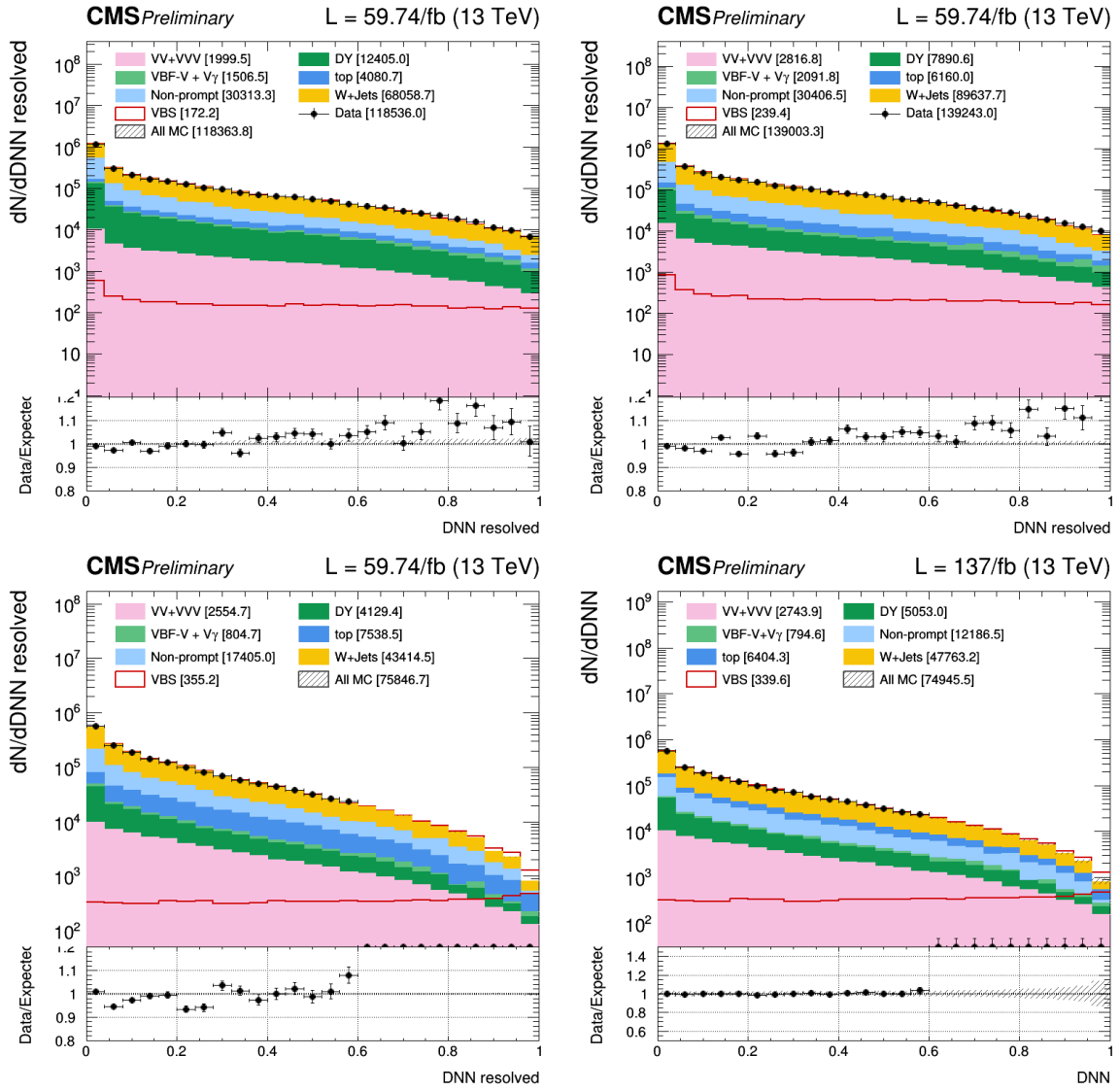


Figure 3.94: Muon, resolved control region. Comparison between the effect of reweighting the residual discrepancy on the prefit DNN (left) and the total postfit shape and uncertainty. W+jets control region distributions on the first line, the signal region in the second.

## 3.12 Future perspectives

The investigation of VBS semileptonic channels has just started: the first evidence of this process observed with the full Run II dataset in the analysis described in this thesis confirms the feasibility of the rare signal extraction thanks to machine learning models and careful modelling of the main backgrounds. Moreover, a complementary analysis involving the  $ZV$  channel is ongoing in CMS and it will be made public soon.

The next step will be the interpretation of the selected events in the framework of the Effective Field Theory (EFT) Lagrangian. The previous VBS semileptonic CMS analysis on a  $36 \text{ fb}^{-1}$  dataset [63] already shown the power of this channel in constraining the anomalous quartic gauge couplings. More recently, many phenomenological studies are being completed in order to assess the most influential operators in the EFT lagrangian of 6th order that can be probed in this VBS channel. Soon after the publication of this thesis, a fully-fledged CMS analysis on this topic will start.

Another important step will be the investigation of the polarized  $WW$  scattering. Feasibility studies for HL-LHC[117] show that the semileptonic channels can be competitive with the fully-leptonic one in terms of sensitivity in the search for longitudinal scattering.

In the future, the improvements in theory modelling and the large recorded dataset of experimental data at HL-LHC, will make possible even more precise analyses of VBS processes, in order to shed stronger and stronger light on the EW symmetry breaking sector of the SM.



# Conclusions

In this thesis two main contributions have been presented: innovative optimizations developed for the ECAL sub-detector for Run III, and a search for VBS in the semileptonic  $WVjj \rightarrow lvqqjj$  final state.

The increase in integrated luminosity and pileup expected for Run III will create a challenging environment for the ECAL sub-detector, whose performance is affected by radiation induced ageing effects. To cope with these conditions many important developments have been carried out in this thesis to help maintain or improve the performance of the ECAL trigger primitives and offline energy reconstruction.

On the ECAL trigger side two main contributions are reported. First, the parameters governing the real-time measurement of the energy deposits used by the trigger, have been re-optimized for the first time from the beginning of the data-taking to make them more robust against PU and improve the energy resolution up to 20% in the high  $\eta$  region of the ECAL endcaps. Secondly, a novel configuration of the ECAL frontend electronics, already foreseen by the designers of the firmware but never used during data-taking, has been re-discovered and characterized in order to reduce the contamination of anomalous signals or tag out-of-time energy in a new way. The work carried out in this thesis consists of the full development of this new trigger capability from hardware studies to performance optimization.

Also the algorithms performing the ECAL energy reconstruction will be affected by the worsening of pileup and noise conditions during Run III. Another original contribution described in this work is the exploration of modern machine learning architectures, like graph neural networks, to perform the electron and photon energy re-collection or SuperClustering. This is an unprecedented application of this kind of model in ECAL and initial performance studies show that an improvement of the energy resolution of the ECAL SuperCluster up to 10-15% is achievable.

The first evidence for the electroweak (EW) production of a  $WV$  ( $V = W$  or  $Z$ ) pair plus two jets in the semileptonic channel is reported in this thesis. The search uses a data set of proton-proton collisions at 13 TeV, corresponding to an integrated luminosity of  $137 \text{ fb}^{-1}$  collected with the CMS detector during 2016-2018 (Run II). Events are triggered requiring a single isolated lepton and then separated into two categories: either the hadronically decaying  $W$  or  $Z$  boson is reconstructed as one large-radius jet, or it is identified as a pair of jets with a dijet mass close to the boson mass. Multivariate machine learning discriminators are optimized to separate the signal from the background in each category and their outputs are exploited in the statistical analysis. The large background from single  $W$  boson production accompanied by jets is estimated on control samples in data to reduce the impact of Monte Carlo mismodeling in this multi-jet phase space region.

Three separate maximum likelihood fits are performed: the measurement of the purely EW signal strength  $\mu_{\text{EW}}$  keeping the QCD  $WV$  production contribution fixed to the SM prediction  $\mu_{\text{QCD}} = 1$ ; the measurement of the signal strength considering as signal the EW and QCD  $WV$

processes together; a two-dimensional simultaneous measurement of the signal strengths  $\mu_{EW}$  and  $\mu_{QCD}$ .

A fiducial phase space region is defined at parton level requiring all partons to have  $p_T^q > 10$  GeV and at least one pair of outgoing partons with invariant mass  $m_{qq} > 100$  GeV. The SM prediction for the EW WV production cross section in this fiducial region is  $2.23_{-0.11}^{+0.08}$  (scale)  $\pm 0.05$  (PDF) pb, where PDF is the uncertainty coming from the parton distribution function. The measured EW WV production cross section is  $1.90_{-0.46}^{+0.53}$  pb, corresponding to an observed EW-only signal strength of:  $\mu_{EW} = 0.85 \pm 0.12$  (stat)  $_{-0.17}^{+0.19}$  (syst) =  $0.85_{-0.21}^{+0.23}$ ,  $1_{-0.22}^{+0.24}$  expected, where  $\sigma^{obs}$  and  $\sigma^{SM}$  are the observed and predicted cross sections respectively. The observed significance for the SM EW WV signal is 4.4 standard deviations with 5.1 expected. Considering instead the signal as the overall EW and quantum chromodynamics (QCD) WV production, the cross-section, measured in the same fiducial phase space region as the EW-only one, is  $16.4_{-2.8}^{+3.5}$  pb,  $16.9_{-2.1}^{+2.9}$  (scale)  $\pm 0.5$  (PDF) expected. The overall signal strength  $\mu_{EW+QCD}$ , with an expectation of  $1_{-0.20}^{+0.21}$ , is measured as:  $\mu_{EW+QCD} = 0.97 \pm 0.06$  (stat)  $_{-0.21}^{+0.19}$  (syst) =  $0.97_{-0.22}^{+0.20}$ . Finally, a simultaneous two-dimensional fit of the EW and QCD WV production components is performed. Overall, both the WV EW-only measurement and the simultaneous EW and QCD WV measurements are in agreement with the SM predictions within the 68% confidence interval.

# References

---

- [1] L. Evans and P. Bryant, “LHC machine”, *Journal of Instrumentation* **3** (aug, 2008) S08001–S08001, doi:10.1088/1748-0221/3/08/S08001.
- [2] CMS Collaboration, “The CMS experiment at the CERN LHC”, *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.
- [3] CMS Collaboration, “Performance of the CMS Level-1 trigger in proton-proton collisions at  $\sqrt{s} = 13$  TeV”, *JINST* **15** (2020) P10017, doi:10.1088/1748-0221/15/10/P10017, arXiv:2006.10165.
- [4] CMS Collaboration, “The CMS trigger system”, *JINST* **12** (2017) P01020, doi:10.1088/1748-0221/12/01/P01020, arXiv:1609.02366.
- [5] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector”, *JINST* **12** (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.
- [6] CMS Collaboration, “Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at  $\sqrt{s} = 8$  TeV”, *JINST* **10** (2015) P08010, doi:10.1088/1748-0221/10/08/P08010, arXiv:1502.02702.
- [7] 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.
- [8] M. Cacciari, G. P. Salam, and G. Soyez, “The anti- $k_t$  jet clustering algorithm”, *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.
- [9] 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.
- [10] CMS Collaboration, “Pileup mitigation at CMS in 13 TeV data”, *JINST* **15** (2020), no. 09, P09018, doi:10.1088/1748-0221/15/09/P09018, arXiv:2003.00503.
- [11] D. Bertolini, P. Harris, M. Low, and N. Tran, “Pileup per particle identification”, *JHEP* **10** (2014) 059, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013.
- [12] 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.
- [13] CMS Collaboration, “Performance of missing transverse momentum reconstruction in proton-proton collisions at  $\sqrt{s} = 13$  TeV using the CMS detector”, *JINST* **14** (2019) P07004, doi:10.1088/1748-0221/14/07/P07004, arXiv:1903.06078.



- [14] CMS Collaboration, “CMS: The electromagnetic calorimeter”, Technical Report CERN-LHCC-97-33, CMS-TDR-4, 12, 1997.  
<https://cds.cern.ch/record/349375>.
- [15] CMS Collaboration, “The CMS experiment at the CERN LHC”, *J. Instrum.* **3** (aug, 2008) S08004–S08004, doi:10.1088/1748-0221/3/08/S08004.
- [16] M. Anfreville et al., “Laser monitoring system for the CMS lead tungstate crystal calorimeter”, *Nucl. Instrum. Meth. A* **594** (2008) 292–320,  
doi:10.1016/j.nima.2008.01.104.
- [17] CMS Collaboration, “CMS Physics: Technical Design Report Volume 1: Detector Performance and Software”. Technical design report. CMS. CERN, Geneva, 2006.
- [18] P. Adzic et al., “Reconstruction of the signal amplitude of the CMS electromagnetic calorimeter”, *Eur. Phys. J. C* **46S1** (2006) 23–35,  
doi:10.1140/epjcd/s2006-02-002-x.
- [19] P. Paganini and I. Van Vulpen, “Pulse amplitude reconstruction in the CMS ECAL using the weights method”, Technical Report CMS-NOTE-2004-025, 10, 2004.  
<https://cds.cern.ch/record/801816>.
- [20] D. A. Petyt, “Anomalous APD signals in the CMS Electromagnetic Calorimeter”, *Nucl. Instrum. Methods Phys. Res. A* **695** (2012) 293–295,  
doi:<https://doi.org/10.1016/j.nima.2011.10.025>. New Developments in Photodetection NDIP11.
- [21] CMS Collaboration, “Electromagnetic calorimeter commissioning and first results with 7 TeV data”, Technical Report CMS-NOTE-2010-012, CERN-CMS-NOTE-2010-012, 7, 2010. <http://cds.cern.ch/record/1278160>.
- [22] D. Valsecchi, “ECAL trigger performance in Run 2 and improvements for Run 3”, *JINST* **15** (2020), no. 06, C06037, doi:10.1088/1748-0221/15/06/C06037.
- [23] CMS Collaboration, “Level-1 E/Gamma trigger performance in Run II”, Technical Report CMS-DP-2019-020, Jul, 2019. <https://cds.cern.ch/record/2682202>.
- [24] CMS Collaboration Collaboration, “CMS ECAL Response to Laser Light”, Technical Report CERN-CMS-DP-2019-005, Mar, 2019.  
<https://cds.cern.ch/record/2668200>.
- [25] CMS Collaboration, “Performance of the CMS Level-1 trigger in proton-proton collisions at  $\sqrt{s} = 13$  TeV”, *JINST* **15** (2020), no. 10, P10017,  
doi:10.1088/1748-0221/15/10/P10017, arXiv:2006.10165.
- [26] A. Tishelman-Charny, “Optimizing the Performance of the CMS ECAL Trigger for Runs 2 and 3 of the CERN LHC”, in *Meeting of the Division of Particles and Fields of the American Physical Society*. 10, 2019. arXiv:1910.06232.
- [27] CMS Collaboration, “Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC”, *JINST* **16** (2021), no. 05, P05014,  
doi:10.1088/1748-0221/16/05/P05014, arXiv:2012.06888.

- [28] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, “Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC”,  
doi:10.1088/0954-3899/31/9/n01.
- [29] CMS Collaboration, “Description and performance of track and primary-vertex reconstruction with the CMS tracker”, *JINST* **9** (May, 2014) P10009. 80 p,  
doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569.
- [30] CMS Collaboration, “Reconstruction of signal amplitudes in the CMS electromagnetic calorimeter in the presence of overlapping proton-proton interactions”, *JINST* **15** (2020), no. 10, P10002, doi:10.1088/1748-0221/15/10/P10002,  
arXiv:2006.14359.
- [31] CMS Collaboration Collaboration, “ECAL 2016 refined calibration and Run2 summary plots”, Technical Report CERN-CMS-DP-2020-021, Apr, 2020.  
<https://cds.cern.ch/record/2717925>.
- [32] CMS Collaboration, “Measurements of Inclusive  $W$  and  $Z$  Cross Sections in  $pp$  Collisions at  $\sqrt{s} = 7$  TeV”, *JHEP* **01** (2011) 080, doi:10.1007/JHEP01(2011)080,  
arXiv:1012.2466.
- [33] CMS Collaboration Collaboration, “CMS ECAL trigger plots”, Technical Report CERN-CMS-DP-2019-031, Sep, 2019. <https://cds.cern.ch/record/2690933>.
- [34] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library”, arXiv:1912.01703.
- [35] J. Shlomi, P. Battaglia, and J.-R. Vlimant, “Graph Neural Networks in Particle Physics”,  
doi:10.1088/2632-2153/abbf9a, arXiv:2007.13681.
- [36] P. Battaglia et al., “Relational inductive biases, deep learning, and graph networks”,  
*arXiv* (2018) arXiv:<https://arxiv.org/pdf/1806.01261.pdf>.
- [37] P. W. Battaglia et al., “Interaction Networks for Learning about Objects, Relations and Physics”, arXiv:1612.00222.
- [38] A. Vaswani et al., “Attention is all you need”, arXiv:1706.03762.
- [39] V. Mikuni and F. Canelli, “Point cloud transformers applied to collider physics”, *Mach. Learn. Sci. Tech.* **2** (2021), no. 3, 035027, doi:10.1088/2632-2153/ac07f6,  
arXiv:2102.05073.
- [40] J. Pata et al., “MLPF: Efficient machine-learned particle-flow reconstruction using graph neural networks”, *Eur. Phys. J. C* **81** (2021), no. 5, 381,  
doi:10.1140/epjc/s10052-021-09158-w, arXiv:2101.08578.
- [41] X. Xin, A. Karatzoglou, I. Arapakis, and J. M. Jose, “Graph highway networks”,  
arXiv:2004.04635.
- [42] F. Wu et al., “Simplifying graph convolutional networks”, arXiv:1902.07153.
- [43] T. Developers, “Tensorflow”, technical report, October, 2021.  
doi:10.5281/zenodo.5593257.
- [44] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization”, technical report, 2014. arXiv:1412.6980.

- [45] CMS Collaboration Collaboration, “ECAL SuperClustering with Machine Learning”, Technical Report CERN-CMS-DP-2021-032, Nov, 2021.  
<https://cds.cern.ch/record/2792321>.
- [46] P. D. Group et al., “Review of Particle Physics”, *Progress of Theoretical and Experimental Physics* **2020** (08, 2020) doi:10.1093/ptep/ptaa104,  
<https://academic.oup.com/ptep/article-pdf/2020/8/083C01/34673722/ptaa104.pdf>  
083C01.
- [47] G. S. Franz Mandl, “Quantum Field Theory”. Wiley, 2nd edition, 2010.
- [48] P. W. Higgs, “Broken Symmetries and the Masses of Gauge Bosons”, *Phys. Rev. Lett.* **13** (1964) 508–509, doi:10.1103/PhysRevLett.13.508.
- [49] 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.
- [50] 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.
- [51] CMS Collaboration, “Observation of a New Boson with Mass Near 125 GeV in  $pp$  Collisions at  $\sqrt{s} = 7$  and 8 TeV”, *JHEP* **06** (2013) 081, doi:10.1007/JHEP06(2013)081, arXiv:1303.4571.
- [52] C. F. Anders et al., “Vector boson scattering: Recent experimental and theory developments”, *Rev. Phys.* **3** (2018) 44, doi:10.1016/j.revip.2018.11.001, arXiv:1801.04203.
- [53] R. Covarelli, M. Pellen, and M. Zaro, “Vector-boson scattering at the LHC: unraveling the electroweak sector”, *Int. J. Mod. Phys. A* **36** (2021) 2130009, doi:10.1142/S0217751X2130009X, arXiv:2102.10991.
- [54] B. W. Lee, C. Quigg, and H. B. Thacker, “The Strength of Weak Interactions at Very High-Energies and the Higgs Boson Mass”, *Phys. Rev. Lett.* **38** (1977) 883–885, doi:10.1103/PhysRevLett.38.883.
- [55] A. Ballestrero et al., “Precise predictions for same-sign W-boson scattering at the LHC”, *Eur. Phys. J. C* **78** (2018) 671, doi:10.1140/epjc/s10052-018-6136-y, arXiv:1803.07943.
- [56] CMS Collaboration, “Measurements of production cross sections of WZ and same-sign WW boson pairs in association with two jets in proton-proton collisions at  $\sqrt{s} = 13$  TeV”, *Phys. Lett. B* **809** (2020) 135710, doi:10.1016/j.physletb.2020.135710, arXiv:2005.01173.
- [57] ATLAS Collaboration, “Observation of electroweak production of a same-sign W boson pair in association with two jets in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector”, *Phys. Rev. Lett.* **123** (2019), no. 16, 161801, doi:10.1103/PhysRevLett.123.161801, arXiv:1906.03203.

- [58] CMS Collaboration, “Measurements of production cross sections of polarized same-sign W boson pairs in association with two jets in proton-proton collisions at  $\sqrt{s} = 13$  TeV”, *Phys. Lett. B* **812** (2021) 136018, doi:10.1016/j.physletb.2020.136018, arXiv:2009.09429.
- [59] ATLAS Collaboration, “Observation of electroweak  $W^\pm Z$  boson pair production in association with two jets in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector”, *Phys. Lett. B* **793** (2019) 469–492, doi:10.1016/j.physletb.2019.05.012, arXiv:1812.09740.
- [60] ATLAS Collaboration, “Observation of electroweak production of two jets and a Z-boson pair with the ATLAS detector at the LHC”, 2020. arXiv:2004.10612. Submitted to *NPHYS*.
- [61] CMS Collaboration, “Evidence for electroweak production of four charged leptons and two jets in proton-proton collisions at  $\sqrt{s} = 13$  TeV”, *Phys. Lett. B* **812** (2021) 135992, doi:10.1016/j.physletb.2020.135992, arXiv:2008.07013.
- [62] ATLAS Collaboration, “Search for the electroweak diboson production in association with a high-mass dijet system in semileptonic final states in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector”, *Phys. Rev. D* **100** (2019), no. 3, 032007, doi:10.1103/PhysRevD.100.032007, arXiv:1905.07714.
- [63] CMS Collaboration, “Search for anomalous electroweak production of vector boson pairs in association with two jets in proton-proton collisions at 13 TeV”, *Phys. Lett. B* **798** (2019) 134985, doi:10.1016/j.physletb.2019.134985, arXiv:1905.07445.
- [64] L. S. Shapley, “A value for n-person games.”, *Contributions to the Theory of Games* **2.28** (1953) 303–317.
- [65] B. Biedermann, A. Denner, and M. Pellen, “Complete NLO corrections to  $W^+W^+$  scattering and its irreducible background at the LHC”, *JHEP* **10** (2017) 124, doi:10.1007/JHEP10(2017)124, arXiv:1708.00268.
- [66] 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.
- [67] 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.
- [68] B. Jäger et al., “Parton-shower effects in Higgs production via Vector-Boson Fusion”, *Eur. Phys. J. C* **80** (2020), no. 8, 756, doi:10.1140/epjc/s10052-020-8326-7, arXiv:2003.12435.
- [69] A. Ballestrero, E. Maina, and G. Pelliccioli, “W boson polarization in vector boson scattering at the LHC”, *JHEP* **03** (2018) 170, doi:10.1007/JHEP03(2018)170, arXiv:1710.09339.
- [70] S. Frixione and B. R. Webber, “Matching NLO QCD computations and parton shower simulations”, *JHEP* **06** (2002) 029, doi:10.1088/1126-6708/2002/06/029, arXiv:hep-ph/0204244.

- [71] P. Nason, “A New method for combining NLO QCD with shower Monte Carlo algorithms”, *JHEP* **11** (2004) 040, doi:10.1088/1126-6708/2004/11/040, arXiv:hep-ph/0409146.
- [72] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with parton shower simulations: the POWHEG method”, *JHEP* **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.
- [73] S. Alioli, P. Nason, C. Oleari, and E. Re, “NLO vector-boson production matched with shower in POWHEG”, *JHEP* **07** (2008) 060, doi:10.1088/1126-6708/2008/07/060, arXiv:0805.4802.
- [74] S. Alioli, P. Nason, C. Oleari, and E. Re, “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”, *JHEP* **06** (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.
- [75] J. Alwall et al., “Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions”, *Eur. Phys. J. C* **53** (2008) 473, doi:10.1140/epjc/s10052-007-0490-5, arXiv:0706.2569.
- [76] J. Campbell and T. Neumann, “Precision phenomenology with MCFM”, *JHEP* **12** (2019) 034, doi:10.1007/JHEP12(2019)034, arXiv:1909.09117.
- [77] J. M. Campbell, R. K. Ellis, and W. T. Giele, “A multi-threaded version of MCFM”, *Eur. Phys. J. C* **75** (2015) 246, doi:10.1140/epjc/s10052-015-3461-2, arXiv:1503.06182.
- [78] J. M. Campbell, R. K. Ellis, and C. Williams, “Vector boson pair production at the LHC”, *JHEP* **07** (2011) 018, doi:10.1007/JHEP07(2011)018, arXiv:1105.0020.
- [79] J. M. Campbell and R. K. Ellis, “An update on vector boson pair production at hadron colliders”, *Phys. Rev. D* **60** (1999) 113006, doi:10.1103/PhysRevD.60.113006, arXiv:hep-ph/9905386.
- [80] T. Sjöstrand et al., “An Introduction to PYTHIA 8.2”, *Comput. Phys. Commun.* **191** (2015) 159, doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012.
- [81] NNPDF Collaboration, “Parton distributions for the LHC Run II”, *JHEP* **04** (2015) 040, doi:10.1007/JHEP04(2015)040, arXiv:1410.8849.
- [82] NNPDF Collaboration, “Parton distributions from high-precision collider data”, *Eur. Phys. J. C* **77** (2017) 663, doi:10.1140/epjc/s10052-017-5199-5, arXiv:1706.00428.
- [83] P. Skands, S. Carrazza, and J. Rojo, “Tuning PYTHIA 8.1: the Monash 2013 Tune”, *Eur. Phys. J. C* **74** (2014), no. 8, 3024, doi:10.1140/epjc/s10052-014-3024-y, arXiv:1404.5630.
- [84] CMS Collaboration, “Event generator tunes obtained from underlying event and multiparton scattering measurements”, *Eur. Phys. J. C* **76** (2016), no. 3, 155, doi:10.1140/epjc/s10052-016-3988-x, arXiv:1512.00815.
- [85] CMS Collaboration, “Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements”, *Eur. Phys. J. C* **80** (2020), no. 1, 4, doi:10.1140/epjc/s10052-019-7499-4, arXiv:1903.12179.

- [86] M. Bähr et al., “Herwig++ physics and manual”, *Eur. Phys. J. C* **58** (2008) 639, doi:10.1140/epjc/s10052-008-0798-9, arXiv:0803.0883.
- [87] J. Bellm et al., “Herwig 7.0/Herwig++ 3.0 release note”, *Eur. Phys. J. C* **76** (2016), no. 4, 196, doi:10.1140/epjc/s10052-016-4018-8, arXiv:1512.01178.
- [88] GEANT4 Collaboration, “GEANT4: a simulation toolkit”, *Nucl. Instrum. Meth. A* **506** (2003) 250, doi:10.1016/S0168-9002(03)01368-8.
- [89] CMS Collaboration, “Measurements of differential Z boson production cross sections in proton-proton collisions at  $\sqrt{s} = 13$  TeV”, *JHEP* **12** (2019) 061, doi:10.1007/JHEP12(2019)061, arXiv:1909.04133.
- [90] M. Czakon et al., “Top-pair production at the LHC through NNLO QCD and NLO EW”, *JHEP* **10** (2017) 186, doi:10.1007/JHEP10(2017)186, arXiv:1705.04105.
- [91] CMS Collaboration, “Measurement of differential cross sections for top quark pair production using the lepton+jets final state in proton-proton collisions at 13 TeV”, *Phys. Rev. D* **95** (2017), no. 9, 092001, doi:10.1103/PhysRevD.95.092001, arXiv:1610.04191.
- [92] CMS Collaboration, “VBF H parton shower modelling study”. [https://indico.cern.ch/event/1028137/contributions/4333291/attachments/2232979/3784093/cooperstein\\_VBFPSmodelling\\_26april2021.pdf](https://indico.cern.ch/event/1028137/contributions/4333291/attachments/2232979/3784093/cooperstein_VBFPSmodelling_26april2021.pdf).
- [93] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, “Towards an understanding of jet substructure”, *JHEP* **09** (2013) 029, doi:10.1007/JHEP09(2013)029, arXiv:1307.0007.
- [94] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, “Jet substructure as a new Higgs search channel at the LHC”, *Phys. Rev. Lett.* **100** (2008) 242001, doi:10.1103/PhysRevLett.100.242001, arXiv:0802.2470.
- [95] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, “Soft drop”, *JHEP* **05** (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.
- [96] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, “Better jet clustering algorithms”, *JHEP* **08** (1997) 001, doi:10.1088/1126-6708/1997/08/001, arXiv:hep-ph/9707323.
- [97] M. Wobisch and T. Wengler, “Hadronization corrections to jet cross-sections in deep inelastic scattering”, in *Proceedings of the Workshop on Monte Carlo Generators for HERA Physics, Hamburg, Germany*, p. 270. 1998. arXiv:hep-ph/9907280.
- [98] CMS Collaboration, “Identification techniques for highly boosted W bosons that decay into hadrons”, *JHEP* **12** (2014) 017, doi:10.1007/JHEP12(2014)017, arXiv:1410.4227.
- [99] J. Thaler and K. Van Tilburg, “Identifying boosted objects with  $N$ -subjettiness”, *JHEP* **03** (2011) 015, doi:10.1007/JHEP03(2011)015, arXiv:1011.2268.
- [100] 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.

- [101] CMS Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”, *JINST* **13** (2018) P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.
- [102] J. Erdmann et al., “Reformulation of a likelihood approach to fake-lepton estimation in the framework of Bayesian inference”, arXiv:2106.13628.
- [103] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, arXiv:1502.03167.
- [104] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning”. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [105] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions”, arXiv:arXiv:1705.07874.
- [106] C. Molnar, “Interpretable machine learning, a guide for making black box models explainable”, technical report, 2020. <https://christophm.github.io/interpretable-ml-book/>.
- [107] ATLAS Collaboration, “Measurement of  $W^\pm W^\pm$  vector-boson scattering and limits on anomalous quartic gauge couplings with the ATLAS detector”, *Phys. Rev. D* **96** (2017) 012007, doi:10.1103/PhysRevD.96.012007, arXiv:1611.02428.
- [108] ATLAS Collaboration, “Observation of electroweak  $W^\pm Z$  boson pair production in association with two jets in  $pp$  collisions at  $\sqrt{s} = 13$  TeV with the ATLAS detector”, *Phys. Lett. B* **793** (2019) 469–492, doi:10.1016/j.physletb.2019.05.012, arXiv:1812.09740.
- [109] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions”, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, p. 4768. Curran Associates Inc., Red Hook, NY, USA, 2017. arXiv:arXiv:1705.07874.
- [110] CMS Collaboration, “Precision luminosity measurement in proton-proton collisions at  $\sqrt{s} = 13$  TeV in 2015 and 2016 at CMS”, *Eur. Phys. J. C* **81** (2021), no. 9, 800, doi:10.1140/epjc/s10052-021-09538-2, arXiv:2104.01927.
- [111] CMS Collaboration, “CMS luminosity measurement for the 2017 data-taking period at  $\sqrt{s} = 13$  TeV”, Technical Report CMS-PAS-LUM-17-004, CERN, Geneva, 2018. <https://cds.cern.ch/record/2621960>.
- [112] CMS Collaboration, “CMS luminosity measurement for the 2018 data-taking period at  $\sqrt{s} = 13$  TeV”, Technical Report CMS-PAS-LUM-18-002, CERN, Geneva, 2019. <https://cds.cern.ch/record/2676164>.
- [113] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics”, *Eur. Phys. J. C* **71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727. [Erratum: *Eur.Phys.J.C* **73**, 2501 (2013)].
- [114] S. S. Wilks, “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses”, *Ann. Math. Stat.* **9** (mar, 1938) 60–62, doi:10.1214/aoms/1177732360.

- 
- [115] ATLAS, CMS, LHC Higgs Combination Group Collaboration, “Procedure for the LHC Higgs boson search combination in summer 2011”, Technical Report ATL-PHYS-PUB-2011-011, CMS-NOTE-2011-005, 8, 2011.  
<http://cds.cern.ch/record/1379837>.
- [116] R. D. Cousins, “Generalization of chisquare goodness-of-fit test for binned data using saturated models, with application to histograms”, 2013. [http://www.physics.ucla.edu/~Ecousins/stats/cousins\\_saturated.pdf](http://www.physics.ucla.edu/~Ecousins/stats/cousins_saturated.pdf).
- [117] J. Roloff, V. Cavaliere, M.-A. Pleier, and L. Xu, “Sensitivity to longitudinal vector boson scattering in semi-leptonic final states at the HL-LHC”, [arXiv:2108.00324](https://arxiv.org/abs/2108.00324).