

2 **Opportunities and Challenges of Standard Model**
3 **Production Cross Section Measurements in**
4 **Proton–Proton Collisions at $\sqrt{s}=8$ TeV using CMS Open**
5 **Data**

6 **Aram Apyan^a William Cuozzo^b Markus Klute^b Yoshihiro Saito^b Matthias Schott^{1,b,c} Bereket**
7 **Sintayehu^b**

8 ^a*Fermilab, USA*

9 ^b*Massachusetts Institute of Technology, Cambridge, USA*

10 ^c*Johannes Gutenberg-University, Mainz, Germany*

11 *E-mail: matthias.schott@cern.ch*

12 **ABSTRACT:** The CMS Open Data project offers new opportunities to measure cross sections of
13 standard model (SM) processes which have not been probed so far. We evaluate the challenges
14 and the opportunities of the CMS Open Data project in the view of cross section measurements.
15 In particular, we reevaluate the SM cross sections of the production of W bosons, Z bosons,
16 top-quark pairs and WZ dibosons in several decay channels at a center of mass energy of 8 TeV
17 with an integrated luminosity of 1.8 fb^{-1} . These cross sections were previously measured by the
18 ATLAS and CMS Collaborations and are used to validate our analysis and calibration strategy. The
19 results indicate the achievable level of precision for future measurements using the CMS Open Data
20 performed by scientists who are not members of the LHC Collaborations and hence lack detailed
21 knowledge of experimental and detector related effects and their handling.

22 **KEYWORDS:** Analysis and statistical methods

¹corresponding author

23	Contents	
24	1 Introduction	1
25	2 The CMS Detector and CMS Open Data	2
26	2.1 The CMS Detector and Reconstructed Objects	2
27	2.2 Software and Infrastructure	3
28	2.3 Selected Open Data	4
29	2.4 MC Simulated Samples	4
30	3 Calibration	6
31	3.1 Muon Performance	6
32	3.2 Electron Performance	6
33	3.3 Jet Energy Scale and Resolution	7
34	3.4 Tagging of b-Quarks	8
35	3.5 Missing Transverse Energy	8
36	4 Standard Model Cross Section Measurements	10
37	4.1 Standard Processes and Signal Selection	10
38	4.2 Background Estimations	12
39	4.3 Systematic Uncertainties	13
40	4.4 Results and Comparisons	15
41	5 Opportunities and Challenges of the CMS Open Data Initiative	17

42 1 Introduction

43 Precision measurements of standard model (SM) processes at the Large Hadron Collider (LHC)
44 made tremendous progress in recent years. The differential measurement of the production cross
45 sections of W and Z bosons, as well as top-quark pairs, reached a precision of a few percent (e.g.
46 [1–3]), sometimes a few per-mil (e.g. [4, 5]). These form the basis for testing and improving
47 modern Monte Carlo (MC) event generators that aim to describe those processes in high-energy
48 hadron collisions. Numerous of these high-precision measurements are at the core of the research
49 program of the ATLAS and CMS Collaborations since the beginning of the LHC.

50 The CMS Collaboration has published significant amounts of recorded and simulated proton-
51 proton collisions at a center-of-mass energy of 8 TeV within the context of the Open Data initiative
52 [6]. In principle, the availability of these data sets allows physicists who are not member of the
53 LHC Collaborations to perform measurements. With this work, we systematically evaluate the
54 physics potential of the available CMS Open Data for cross section measurements and broaden

55 the perspective of previous studies using CMS Open Data [7, 8]. Special focus is drawn on the
56 limitations of such measurements and possible future improvements.

57 As a starting point, we estimate and derive several physics object calibration constants, either
58 from previous publications, or using the data itself. In a second step, we measure several SM cross
59 sections at a center-of-mass energy of 8 TeV with an integrated luminosity of 1.8 fb^{-1} and compare
60 them to the results published by the CMS and ATLAS Collaborations. The agreement with the
61 published results, as well as the assigned uncertainties on our measurements, indicate to which
62 precision new cross section measurements can also be performed using CMS Open Data.

63 The paper is structured as follows: The CMS detector, its physics objects, and the data-sets
64 used in this analysis are summarized in Section 2. The calibration of the physics objects, such as
65 electrons, muons or particle jets, is discussed in Section 3. The cross section measurements of SM
66 processes are discussed in Section 4, leading to a discussion of the opportunities and challenges of
67 cross section measurements within the CMS Open Data Initiative in Section 5.

68 2 The CMS Detector and CMS Open Data

69 2.1 The CMS Detector and Reconstructed Objects

70 The data used in this analysis has been recorded with the CMS detector at the LHC in the year
71 2012. CMS is a typical high-energy physics experiment, using a superconducting solenoid of 6 m
72 internal diameter with a magnetic field of 3.8 T [9]. The inner detector (ID) of CMS can reconstruct
73 trajectories of charged particles using silicon pixel and strip trackers. Electrons and photons are
74 identified and measured in a crystal electromagnetic calorimeter (ECAL), while energies of hadrons
75 or hadronic particle jets are determined in a brass/scintillator hadron calorimeter (HCAL). Muons
76 are identified and measured in the muon system (MS), based on gaseous detectors, which surround
77 the hadronic calorimeter and are embedded in the steel flux-return yoke of the magnet system. CMS
78 uses a right-handed coordinate system. Its origin is defined at the interaction point of the proton
79 collisions, the x axis is pointing towards the center of the LHC, the y axis pointing upwards and the
80 z axis along the counterclockwise-beam direction. The polar angle θ is measured from the positive
81 z axis, however, mostly expressed in terms of the pseudorapidity η , defined by $\eta = -\ln(\tan \theta/2)$.
82 The azimuthal angle ϕ is measured in the $x - y$ plane. We refer to [9] for a detailed description of
83 the CMS experiment. CMS employs a particle-flow algorithm that provides a complete description
84 of the event and identifies electrons, muons, photons, charged hadrons, and neutral hadrons [10].

85 Electrons are identified as reconstructed energy clusters in the ECAL, which have been matched
86 to tracks measured in the ID [11–13]. In this analysis, we typically require the transverse energy
87 of electrons to be $E_T > 25 \text{ GeV}$ within $|\eta| < 1.44$ (barrel) or $1.57 < |\eta| < 2.5$ (endcap); the gap
88 between barrel and endcap is determined by the detector layout. In addition, standard electron
89 identification requirements, e.g. on the energy ratio measured in the ECAL and the HCAL or on the
90 track impact parameters, are applied, following previous CMS measurements [2]. An electron passes
91 a loose/tight isolation requirement if the vectorial sum of momenta of all reconstructed charged
92 particles, stemming from the primary vertex, within a cone-size of $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.3$
93 normalized by the E_T of the electron is smaller than 0.15 (0.10) in the barrel (endcap).

94 Muons are reconstructed from a global fit of hits in the MS and the ID, seeded by tracks in the
95 muon system [14]. In this analysis, we typically require each muon to have a transverse momentum

96 of $p_T > 25$ GeV within a pseudorapidity range of $|\eta| < 2.1$, corresponding to the single muon trigger
 97 coverage. In addition, following previous CMS measurements [2], standard quality requirements on
 98 the numbers of hits in the ID and in the MS, on the χ^2 of the fit, and on the track impact parameters
 99 are applied. A relative isolation variable is computed as described for the electrons for a loose and
 100 tight isolation definition, respectively, however, computed with a cone-radius of $\Delta R = 0.4$. Given
 101 the small mass of muons and electrons compared to their energies relevant in the studied processes
 102 within this work, their transverse momentum and their transverse energy can be assumed to be
 103 equal, i.e. $E_T = p_T$.

104 Hadronic jets are reconstructed using an anti- k_T algorithm with a radius parameter of 0.5 based
 105 on particle-flow objects [10, 15], where the clustering algorithm rejects objects that are coming
 106 from vertices of additional interactions per bunch crossing (pile-up). A jet area method [16] is used
 107 to correct for the remaining pile-up contributions.

108 In this analysis, we focus our study on jets with a transverse momentum of $p_T > 30$ GeV and
 109 a rapidity of $|y| < 2.4$, since this region allows for a good jet resolution and pile-up rejection.
 110 In addition, certain quality criteria on the reconstructed jet properties, such as energy fraction in
 111 the ECAL and HCAL or the number of particle-flow objects, are applied following standard CMS
 112 recommendations. Moreover, jets are required to have a distance of $\Delta R > 0.5$ to all reconstructed
 113 electron, muon, and photon candidates with $p_T > 20$ GeV. The three jets with the largest recon-
 114 structed p_T are denoted as j_1 , j_2 , and j_3 in order of decreasing p_T values. The origin of a jet from a
 115 bottom quark is identified via a combined secondary-vertex algorithm, which uses track impact pa-
 116 rameter and secondary-vertex information [17]. In this analysis, we use a 'medium' working-point
 117 for identified b-jets with an average efficiency of 85%.

118 Neutrinos leave the CMS detector undetected and hence cause an imbalance in the vectorial
 119 momentum sum of all final-state particles in the plane transverse to the beam axis. CMS defines the
 120 missing transverse momentum as the negative vector sum of all p_T of reconstructed particle flow
 121 objects, i.e. $\vec{p}_T^{\text{miss}} = -\sum_{\text{PF}} \vec{p}_T$. The magnitude of \vec{p}_T^{miss} is denoted as E_T^{miss} .

122 It is additionally required that selected reconstructed objects are not considered for further anal-
 123 ysis if they are close to other reconstructed objects. Electron candidates are not further considered
 124 if a muon candidate with $p_T > 20$ GeV, passing standard quality criteria, has been reconstructed
 125 within $\Delta R < 0.3$ of the electron candidate.

126 2.2 Software and Infrastructure

127 The CMS Open Data Software Framework (Release CMSSW_5_3_32), available in [6], builds
 128 the basis of this analysis. A dedicated open-source framework, BACON [18], which was used for
 129 several published studies of the CMS Collaboration, e.g. [2], is used to read the Analysis Object
 130 Data (AOD) [19], extracting information on reconstructed objects as well as particle-level data¹, if
 131 available. The BACON software framework is also used to apply a selection of events which have
 132 been recorded under stable detector conditions, known as *GoodRun-List* [20], provided by the CMS
 133 Open Data project, leading to a reduced output-format based on the ROOT software framework [21].
 134 The typical size of one simulated top-quark pair event in the BACON-output format amounts to 5

¹Particle-level information denotes the available information on final states after the MC event generation step, i.e. prior to the detector simulation

Data Stream	Trigger Name(s)	Dataset Name	$\int L dt$ [pb ⁻¹]
single muon trigger	HLT_IsoMu24,	/SingleMu/Run2012C-22Jan2013-v1 [22]	1,828
	HLT_IsoMu24_eta2p1	/SingleMu/Run2012B-22Jan2013-v1 [23]	
single electron trigger	HLT_Ele27_WP80	/SingleElectron/Run2012B-22Jan2013-v1 [24]	1,776
		/SingleElectron/Run2012C-22Jan2013-v1 [25]	

Table 1. Overview of data samples used in this analysis together with the corresponding integrated luminosity and the triggers, which have been used during the data taking.

135 kB. For this work, we developed an additional software package, which reduces the output files
136 of BACON further and transforms them into a plain ROOT-NTuple, denoted as ODNTUPLE in the
137 following with an average event size of 0.8 kB. Our analysis is based on these ODNTUPLE data.

138 2.3 Selected Open Data

139 The data acquisition system of CMS records only the event information of collisions with dedicated
140 signatures due to the high-collision rate and the limited bandwidth for data-processing. The data
141 used in this analysis has been collected when one of the triggers *HLT_IsoMu24*, *IsoMu24_eta2p1*
142 or *HLT_Ele27_WP80* has fired. These triggers are unprescaled for the full 2012 data-set and aim to
143 collect events with at least one isolated muon candidate within $|\eta| < 2.4$ and $p_T > 24$ GeV or with
144 at least one electron candidate within $|\eta| < 2.5$ and $E_T > 27$ GeV.

145 In total, muon-triggered (electron-triggered) events corresponding to 1.83 fb^{-1} (1.78 fb^{-1}) [22,
146 23] of integrated luminosity from CMS Open Data [24, 25] have been processed (Table 1). We
147 only studied roughly 10% of the full available dataset due to limitations on the available computing
148 resources during this project as well as the fact that our final results are already dominated by
149 systematic uncertainties. The integrated luminosity has been calculated using the publicly available
150 *GoodRun-List*. We assume an uncertainty of 2.5% in the integrated luminosity following the official
151 CMS recommendation ([26]).

152 2.4 MC Simulated Samples

153 An overview of the various signal and background samples used in this analysis is given in Table 2,
154 indicating the underlying physics process, the dataset name, and the corresponding inclusive cross
155 section at next-to-leading order (NLO) or next-to-next-to-leading order (NNLO). The Drell–Yan
156 processes (W/Z) in the electron and muon decay channels were generated with the POWHEGBOX
157 v.1.0 MC program [27, 28] interfaced to the PYTHIA v.6.4.26 parton shower model [29]. All other
158 processes are modeled with the tree-level matrix element event generator MADGRAPH v5.1.3.30
159 [30] interfaced with PYTHIA 6.4.26. The CT10 parton distribution functions (PDFs) [31] and the
160 Z2* PYTHIA6 tune [32, 33] are used. The decays of tau-leptons are modeled using the TAUOLA
161 program [34]. The PYTHIA v.6.4.26 is used for the modeling of photon radiation off final state
162 particles. The strong coupling constant α_s has been set to 0.130 at the Z boson mass scale for all
163 matrix element calculations. The effect of pile-up has been simulated by overlaying MC-generated
164 minimum bias events. The GEANT4 program was used to simulate the passage of particles through
165 the CMS detector [35].

Process	Dataset Name	Inclusive σ [pb]	order in α_s
$pp \rightarrow Z/\gamma^* + X \rightarrow e^+e^- + X$	DYToEE_M-20_CT10_TuneZ2star_v2_8TeV [37]	1916	NNLO
$pp \rightarrow Z/\gamma^* + X \rightarrow \mu^+\mu^- + X$	DYToMuMu_M-20_CT10_TuneZ2star_v2_8TeV [38]	1916	NNLO
$pp \rightarrow Z/\gamma^* + X \rightarrow l^+l^- + X$	DYJetsToLL_M-50_TuneZ2Star_8TeV [39]	3533	NNLO
$pp \rightarrow W^+ + X \rightarrow \mu^+\nu + X$	WplusToMuNu_CT10_8TeV [40]	7322	NNLO
$pp \rightarrow W^- + X \rightarrow \mu^-\nu + X$	WminusToMuNu_CT10_8TeV [41]	5181	NNLO
$pp \rightarrow W^+ + X \rightarrow \tau^+\nu + X$	WplusToTauNu_CT10_8TeV [42]	7322	NNLO
$pp \rightarrow W^- + X \rightarrow \tau^-\nu + X$	WminusToTauNu_CT10_8TeV [43]	5181	NNLO
$pp \rightarrow t\bar{t} + X \rightarrow 2l2\nu2b + X$	TTJets_FullLeptMGDecays_TuneP11TeV_8TeV [44]	112.3	NLO
$pp \rightarrow t\bar{t} + X \rightarrow 1l1\nu2q2b + X$	TTJets_SemiLeptMGDecays_8TeV [45]	107.2	NLO
$pp \rightarrow t\bar{t} + X \rightarrow 4q2b + X$	TTJets_HadronicMGDecays_TuneP11mpiHi_8TeV [46]	25.8	NLO
$pp \rightarrow WW + X \rightarrow 2l2\nu + X$	WWJetsTo2L2Nu_TuneZ2star_8TeV [47]	5.8	NLO
$pp \rightarrow WZ + X \rightarrow 3l1\nu + X$	WZJetsTo3LNu_8TeV_TuneZ2Star [48]	1.1	NNLO
$pp \rightarrow ZZ + X \rightarrow 4\mu + X$	ZZTo4mu_8TeV [49]	0.077	NLO

Table 2. Overview of simulated event samples used in this analysis together with the corresponding inclusive cross sections. Inclusive charged leptons (e, μ, τ) are denoted with l .

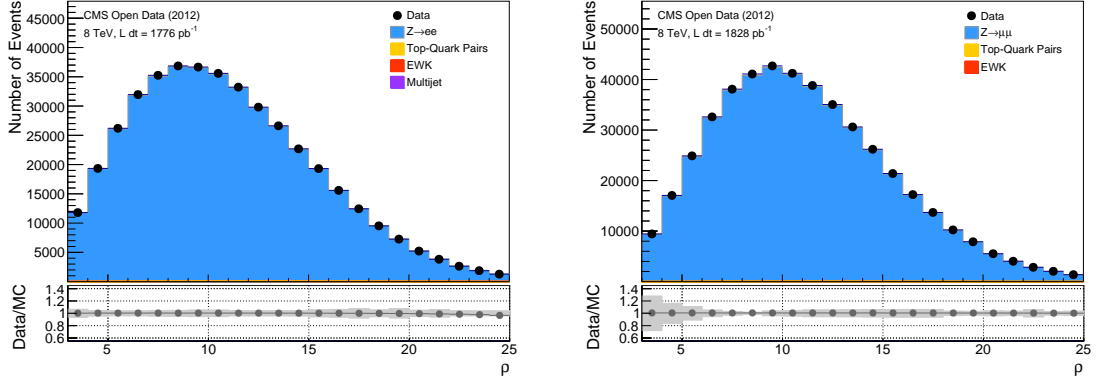


Figure 1. Distribution of the ρ parameter, sensitive to the pile-up activity, per event for electron (left) and muon (right) events as well as reweighted simulated Drell–Yan events in electron and muon decay channels, respectively. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

166 The simulated event samples are reweighted to describe the distribution of the number of
167 pile-up events in the data by reweighting the ρ parameter distribution, where ρ denotes the diffuse
168 offset energy density [36]. Moreover, a reweighting of the longitudinal position of the primary
169 pp collision vertex of the MC samples to data has been performed. The resulting ρ distributions
170 for data and simulated Z boson samples in the electron and muon decay channels are shown in
171 Figure 1. The differences in MC predictions with and without reweighting is taken as a systematic
172 uncertainty.

173 3 Calibration

174 Although the detector simulation of CMS experiments provides a very good description of the
 175 expected event signatures, some remaining differences in reconstruction, trigger, and isolation
 176 efficiencies, as well as in the momentum and energy scales and resolutions are present between
 177 MC and data. Dedicated corrections are applied to minimize these differences and are discussed
 178 in the following. The Z boson candidate events in the electron and muon decay channels in data
 179 and simulated samples are used to validate our corrections. The $Z \rightarrow \mu^+\mu^-$ candidate events are
 180 selected by requiring events with exactly two oppositely charged, isolated muons with $p_T > 25$ GeV
 181 and $|\eta| < 2.1$. The $Z \rightarrow e^+e^-$ candidate events are selected by requiring events with exactly two
 182 oppositely charged, isolated electrons with $E_T > \text{GeV}$ and $|\eta| < 1.44$ or $1.57 < |\eta| < 2.5$. The
 183 invariant mass of these two lepton candidates has to be between 60 and 120 GeV. This selection
 184 ensures a nearly background-free sample of Z boson candidates.

185 3.1 Muon Performance

186 The momentum scale and resolution corrections for muons are derived by comparing the recon-
 187 structed invariant mass spectrum of Z boson candidates between data and simulation. The transverse
 188 momenta of the reconstructed muons can be modified via

$$p_T^{\text{Reco}} = p_T^{\text{Truth}} + \beta \cdot (\alpha \cdot p_T^{\text{Reco}} - p_T^{\text{Truth}}), \quad (3.1)$$

189 where p_T^{Reco} is the reconstructed muon momentum, p_T^{Truth} is the truth muon momentum on particle
 190 level, α is a momentum scale parameter, and β is a resolution parameter. The parameters α and
 191 β are determined for three different regions in η , corresponding to the two endcap and one barrel
 192 regions, by a χ^2 minimization procedure. The χ^2 is calculated between the invariant mass spectrum
 193 of the di-lepton system in Z boson events in data and simulation for different choices of α and β .
 194 The average values of α and β were found to be 0.998 and 1.13, respectively. Uncertainties in the
 195 momentum scale of 0.002 in the barrel region and 0.003 in the endcap regions are applied, which
 196 cover the observed discrepancies with data. The uncertainty in the resolution parameter is 0.05.
 197 The size of the uncertainties have been additionally tested by varying the invariant mass window
 198 requirement. The comparison between data and MC of the invariant mass distribution of di-muon
 199 pairs after the calibration procedure is shown in Figure 2.

200 The corrections for reconstruction and trigger efficiencies for single muons as well as their
 201 uncertainties were taken from official CMS publications [2, 50]. The average correction weights of
 202 the muon reconstruction and trigger efficiencies are found to be 0.985 ± 0.006 and 0.950 ± 0.008 ,
 203 respectively. The muon isolation is well described by the simulation, i.e. the correction weight is
 204 set to 1.000 and an uncertainty of 0.002 is applied. A comparison of the η distribution of muons
 205 from Z boson candidates between data and MC is shown in Figure 2, where all the corrections
 206 above have been applied. The remaining differences are covered by the systematic uncertainties.

207 3.2 Electron Performance

208 The energy scale and resolution corrections for electrons are derived in a similar way as for the
 209 muons, however, an off-set parameter κ is also used in addition to a multiplicative scale factor α .

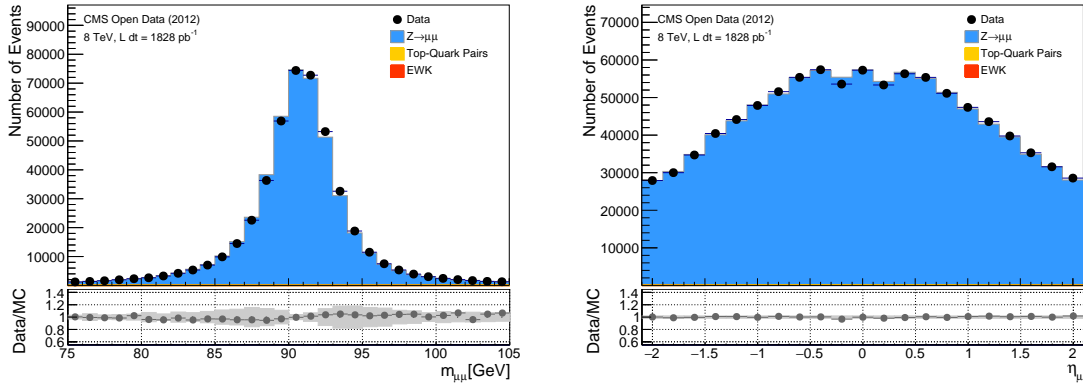


Figure 2. Comparison of the di-muon mass spectrum (left) and the muon η distribution (right) for data and MC for Z boson candidate events, after all corrections have been applied. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

210 The addition of the κ off-set parameter leads to a significantly better description of the data with
 211 the MC simulated samples. Using

$$E_T^{\text{Reco}} = \kappa + E_T^{\text{Truth}} + \beta \cdot (\alpha \cdot E_T^{\text{Reco}} - E_T^{\text{Truth}}), \quad (3.2)$$

212 we find an average value of the energy offset parameter of $\kappa = -0.5 \pm 0.1$ GeV and values of α and β
 213 consistent with 1. The systematic uncertainty in the scale parameter α is 0.003, while the resolution
 214 uncertainties on β range from 0.017 (barrel) to 0.045 (endcap) for electrons with $E_T < 80$ GeV, and
 215 are 0.005 for $E_T > 80$ GeV [13].

216 Corrections to the electron identification and the single-electron trigger efficiencies are taken
 217 from [2, 13] and are close to 1 in most regions, with larger corrections for electrons with $E_T < 30$
 218 GeV in parts of the endcap region. The applied systematic uncertainties are 1.4% and 2.4% for
 219 electron identification and trigger efficiencies, respectively. The isolation for electron is also well
 220 described by the MC simulations, so no reweighting is necessary for the isolation requirement
 221 efficiency. Given the more complicated nature of electron signatures compared to muons in the ID
 222 and ECAL, an uncertainty of 0.004 in the isolation requirement efficiency is applied. A comparison
 223 between data and MC of the invariant mass distribution and of the η distribution of electrons from
 224 Z boson candidates is shown in Figure 3, where all corrections have been applied. The remaining
 225 differences are covered by the systematic uncertainties.

226 3.3 Jet Energy Scale and Resolution

227 The official CMS calibration and corrections for particle jets, in particular, the jet energy scale
 228 (JES) and the jet energy resolution (JER), have been applied within the BACON framework. These
 229 jet corrections and uncertainties were derived from the simulation, and are confirmed with in situ
 230 measurements using the energy balance of dijet and photon+jet events [36]. A reduced set of
 231 systematic variations is used to estimate JES and JER uncertainties on the final measurement. In
 232 particular, the JES is varied by 2% for $|y^{\text{jet}}| < 1.3$ and by 3% for $|y^{\text{jet}}| > 1.3$, following [36]. The
 233 JER is varied by 20% for $30 < E_T^{\text{jet}} < 100$ GeV, by 10% for $100 < E_T^{\text{jet}} < 1000$ GeV and by 5%

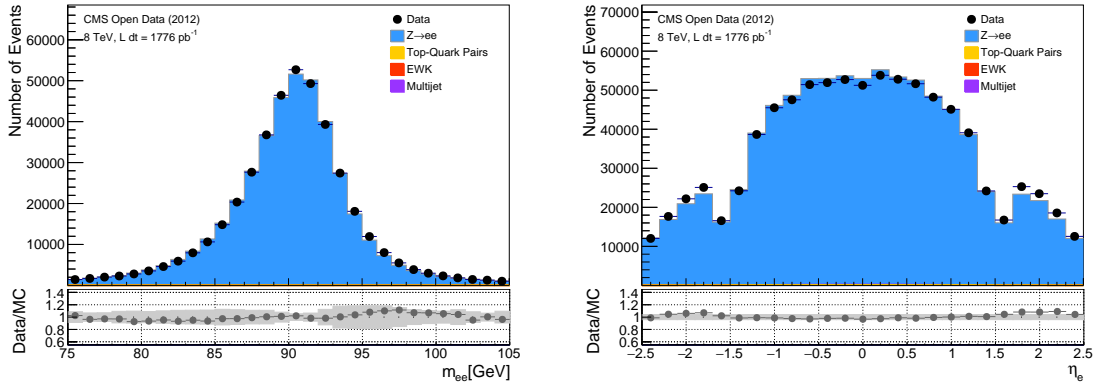


Figure 3. Comparison of the di-electron mass spectrum (left) and the electron η distribution (right) for data and MC for Z boson candidate events, after all corrections have been applied. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

234 for $E_T^{jet} > 1000$ GeV. We apply a JER uncertainty of 20% for jets with an absolute rapidity larger
 235 than 2.1. This simplified treatment of the jet calibration does not allow for a correct evaluation of
 236 correlations between different phase space regions. However, the resulting systematic uncertainties
 237 in the inclusive cross section measurements are expected to be valid.

238 The calibration of jets, as well as the assigned systematic uncertainties, is tested again using
 239 Z boson candidate events in the muon decay channel. For this, Z boson events with a transverse
 240 momentum, $p_T(Z)$, between 50 and 100 GeV with exactly one reconstructed jet with $E_T^{jet} > 30$
 241 GeV and $|y^{jet}| < 2.1$ are selected. The transverse momentum of the Z boson, precisely measured
 242 by its decay leptons, should be balanced in a first approximation by the transverse energy of this jet,
 243 hence the ratio of $p_T(Z)/E_T^{jet}$ should peak around 1. The comparison between data and MC of this
 244 ratio is shown in Fig. 4, where a good agreement within the assigned systematic uncertainties can
 245 be seen. This study has been repeated for higher values of $p_T(Z)$ and more jets in the final state, all
 246 indicating a good closure.

247 3.4 Tagging of b-Quarks

248 The b-tagging efficiency for the working point of the Combined Secondary Vertex algorithm used in
 249 this analysis is 70% for a light-quark misidentification probability of 1.5%. The efficiency has been
 250 measured in data and compared with the MC predictions [17, 51]. In general, a very good agreement
 251 has been found for jet energies between 30 and 500 GeV, where a systematic uncertainty of the order
 252 of 3% was assigned on the efficiency estimate in data. Hence, we do not apply additional b-tagging
 253 efficiency corrections. However, we assign an uncertainty of 5% in the b-tagging efficiency as we
 254 do not apply any kinematic dependent efficiency corrections.

255 3.5 Missing Transverse Energy

256 We apply the official CMS calibration constants and correction factors to the reconstructed E_T^{miss} ob-
 257 servable on an event-by-event basis. The assigned uncertainties in E_T^{miss} are based on [52], where
 258 the scale uncertainty is taken to be 10% for $E_T^{miss} < 20$ GeV, 5% for $20 < E_T^{miss} < 100$ GeV and

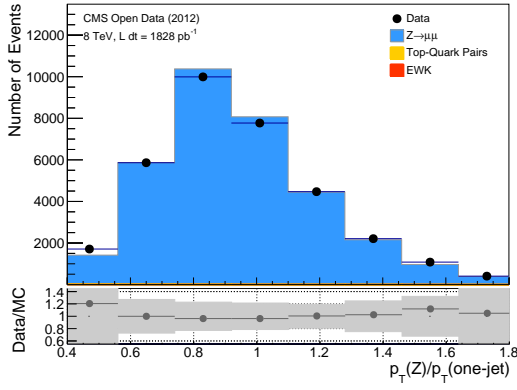


Figure 4. Comparison of the ratio of measured $p_T(Z)$ and the measured jet energy E_T for Z boson events in the muon decay channel with exactly one jet with $E_T > 30$ GeV and $50 < p_T(Z) < 100$ GeV for data and MC, after all corrections have been applied. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

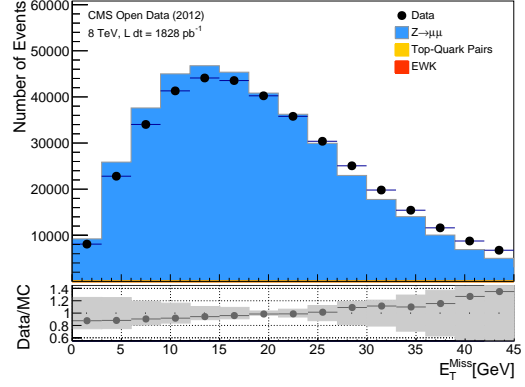


Figure 5. Comparison of the E_T^{miss} distribution for Z boson events in the muon decay channel with $p_T(Z) < 30$ GeV for data and MC, after all corrections have been applied. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

259 2% for $E_T^{\text{miss}} > 100$ GeV. The uncertainty in the E_T^{miss} resolution is applied as a function of the
 260 scalar sum of all transverse energies of all reconstructed hadronic objects in the event, $\sum E_T$, and is
 261 taken to be 20% for $\sum E_T < 100$ GeV and 10% for $\sum E_T > 100$ GeV. In addition, we propagate all
 262 the jet energy scale and resolution uncertainties to E_T^{miss} by studying the impact on a recalculated
 263 E_T^{miss} observable, which is based on all reconstructed objects in the event.

264 The E_T^{miss} observable is validated by studying the observed hadronic recoil, the vector sum of
 265 all hadronic energies, in events containing $Z \rightarrow \mu\mu$ candidates. By construction, the transverse
 266 momentum of the Z boson must be balanced by the hadronic recoil, $\vec{H}R$, i.e. $\vec{p}_T(Z) + \vec{H}R = 0$.
 267 Since the momenta of the decay muons of the Z boson can be measured with high precision,
 268 providing a high precision measurement of $p_T(Z)$, one can effectively probe the simulation of the
 269 detector response on the hadronic recoil. It should be noted that the description of the hadronic
 270 recoil is effectively the same as the description of E_T^{miss} . This can be seen most easily in W boson
 271 events in their leptonic decay channel. Momentum conservation in the transverse plane implies
 272 $\vec{p}_T(W) + \vec{H}R = \vec{p}_T^l + \vec{p}_T^\nu + \vec{H}R = 0$, hence $\vec{p}_T^\nu = -\vec{p}_T^l - \vec{H}R$. The E_T^{miss} description is therefore
 273 equivalent to the description of $|\vec{H}R|$ as the decay lepton can be measured with high precision.
 274 A comparison of the E_T^{miss} distribution in Z boson events, defined as $|\vec{p}_T(Z) + \vec{H}R|$, in the muon
 275 decay channel with $p_T(Z) < 30$ for data and MC is shown in Figure 5, together with the systematic
 276 uncertainties. The observed differences are covered by the systematic uncertainties.

277 4 Standard Model Cross Section Measurements

278 4.1 Standard Processes and Signal Selection

279 To validate all aspects of our analysis framework, starting from the correct interpretation of the
 280 reconstructed objects to the estimation of systematic uncertainties, several inclusive cross section
 281 measurements of SM processes have been performed and compared to high-precision measurements
 282 of the LHC Collaborations as well as to the theoretical predictions. In particular, we have measured
 283 the fiducial cross section of the Drell–Yan process in the electron and muon decay channel, $pp \rightarrow$
 284 $Z/\gamma^* \rightarrow l^+l^-$ ($l = e, \mu$), the fiducial cross section of W^\pm boson production in the muon decay
 285 channels, $pp \rightarrow W^\pm \rightarrow \mu^\pm \nu$, the production cross section of the Z boson in association with
 286 exactly one energetic jet, the production of top-quark pairs in the electron-muon final state, and the
 287 WZ diboson production cross section in the fully leptonic decay channel. These processes probe
 288 different aspects of the analysis infrastructure. The Z boson production cross section mainly probes
 289 lepton identification and reconstruction efficiencies, as well as the jet calibration when requiring
 290 an additional energetic jet in the final state. The study of W bosons also probes the single lepton
 291 trigger performance. The top-quark pair production enables tests of the jet performance and the
 292 b -tagging performance. The study of the WZ diboson production is mainly sensitive to the lepton
 293 reconstruction performance.

294 The inclusive production cross section for a given process can be experimentally determined
 295 via

$$\sigma_V^{\text{incl.}} = \frac{N_{\text{Signal}}}{\epsilon \cdot BR \cdot \int L dt} = \frac{N_{\text{Signal}}}{A \cdot C \cdot BR \cdot \int L dt}. \quad (4.1)$$

296 The number of signal events is given by $N_{\text{Signal}} = N_{\text{Data}} - N_{\text{Bkg}}$, where N_{Data} is the number of
 297 selected events in data and N_{Bkg} is the number of background events surviving the signal selection.
 298 The factor ϵ is the efficiency of the signal events passing the signal selection criteria. The efficiency
 299 correction ϵ can further be decomposed as the product of a fiducial acceptance, A , and a detector-
 300 induced correction factor, C , i.e. $\epsilon = A \cdot C$, defined below. To correct the cross section for the
 301 choice of a specific decay channel, a branching ratio factor BR is applied, which is known to high
 302 precision [53]. Finally, the event yield is normalized by the integrated luminosity $\int L dt$ of the data
 303 sample analyzed.

304 The efficiency correction factor ϵ can be estimated with simulations of the signal process.
 305 These simulations include both a detailed description of the object reconstruction in the detector,
 306 called the *detector level*, and the final-state particle information of the generator calculations, called
 307 the *particle level*. The same signal selection cuts that are applied on data are applied to the simulated
 308 events at detector level. In addition, basic signal selection cuts, such as a minimum p_T cut, can
 309 also be applied to the final-state particles at the particle level. Following these definitions, ϵ can be
 310 defined as the ratio of all events which pass the signal selection on detector level $N_{\text{detector}}^{\text{selected}}$ over the
 311 number of all generated events $N_{\text{particle}}^{\text{all}}$.

312 The fiducial acceptance, A , and a detector-induced correction factor, C , are defined as $A =$
 313 $N_{\text{particle level}}^{\text{selected}}/N_{\text{particle level}}^{\text{all}}$ and $C = N_{\text{detector level}}^{\text{selected}}/N_{\text{particle level}}^{\text{selected}}$, respectively. The fiducial acceptance
 314 A is therefore the ratio of the number of events that pass the geometrical and kinematic cuts of
 315 an analysis at particle level, $N_{\text{particle level}}^{\text{selected}}$, over the total number of generated events in a simulated
 316 sample of signal process, $N_{\text{particle level}}^{\text{all}}$. These selection cuts on particle level require geometrical and

317 kinematic constraints close to the cuts applied on the reconstructed objects at detector level, e.g.
 318 muons in the final state should fulfill $p_T > 25$ GeV and $|\eta| < 2.1$. The dominant uncertainties on the
 319 fiducial acceptance are renormalisation and factorisation scale uncertainties as well as uncertainties
 320 due to the limited knowledge of the parton distribution function (PDF) of the proton. The detector
 321 response correction factor C is the ratio of selected events on detector level, $N_{\text{detector level}}^{\text{selected}}$, over all
 322 events that pass the kinematic selection on particle level, i.e. $N_{\text{particle level}}^{\text{selected}}$. The fiducial cross section
 323 for a given process can therefore be measured by

$$\sigma_V^{\text{fid.}} = \frac{N_{\text{Signal}}}{C \cdot BR \cdot \int L dt}. \quad (4.2)$$

324 Since this definition is independent of A , modelling uncertainties are minimized. Once the fiducial
 325 cross section is known, one can extrapolate to the inclusive cross section using the acceptance
 326 correction factor, i.e. $\sigma_V^{\text{incl.}} = \sigma_V^{\text{fid.}}/A$.

327 All recorded events that are kept for the analysis are required to fulfill the *GoodRun-List*
 328 requirements, contain at least one good primary vertex, and to be either triggered by a single
 329 electron or a single muon trigger (see Section 2.3).

330 The selection of the Z boson candidates has already been introduced in Section 3. The fiducial
 331 volume for the Drell-Yan process is defined at so called Born level (before final state photon radiation)
 332 by requiring the two decay leptons with a transverse momentum of $p_T > 25$ GeV within $|\eta| < 2.1$
 333 and $|\eta| < 2.4$ for the muon and electron decay channels, respectively, following the corresponding
 334 CMS publication [2]. When studying Z boson production in association with jets in the muon decay
 335 channel, the minimum lepton p_T requirement is lowered to 20 GeV and the pseudo rapidity range
 336 increased to $|\eta| < 2.4$ on particle level². Moreover, at least one reconstructed jet with a transverse
 337 energy of at least 30 GeV within $|y^{\text{jet}}| < 2.4$ at particle and detector level is required. Jets are
 338 also reconstructed using generator particles, by clustering final-state particles with decay length $c\tau$
 339 > 10 mm, using the anti- k_T algorithm with radius parameter $R = 0.5$. In total, 434,179 candidate
 340 events in the electron decay channel and 473,626 candidate events in the muon decay channel were
 341 selected, while 61,447 events with more than one reconstructed jet were found. The number of
 342 selected events as well as the fiducial cross section definitions for all Drell-Yan measurements are
 343 summarized in Table 3.

344 The selection of positively and negatively charged W bosons is only applied in the muon
 345 decay channel, since no simulated $W^\pm \rightarrow e\nu$ samples at $\sqrt{s} = 8$ TeV were available on the
 346 CERN Open Data Portal. The W boson candidate events are selected at detector level by requir-
 347 ing exactly one reconstructed, tightly isolated muon with $p_T > 25$ GeV and within $|\eta| < 2.1$.
 348 A minimum E_T^{miss} of 25 GeV is required and a minimum transverse mass requirement of $m_T =$
 349 $\sqrt{2 \cdot p_T^l \cdot E_T^{\text{miss}} \cdot (1 - \cos(\Delta\phi))} > 40$ GeV is applied, where $\Delta\phi$ is the opening angle of the recon-
 350 structed muon and the missing transverse momentum vector in the transverse plane. This selection
 351 differs from the chosen approach in [2] and is closer to [54] in order to reduce multijet background
 352 contributions. Similar kinematic constraints are applied at particle level on the charged decay

²We did not alter the requirements on muons at the detector level, as our calibration has only been validated for muons with $p_T > 25$ GeV within $|\eta| < 2.1$. The modelling uncertainties due to the extrapolation are assumed to be small.

353 lepton, the neutrino and the derived quantities (Table 3). In total, 3,631,170 W^+ and 2,629,480 W^-
 354 candidate events have been selected, respectively.

355 The selection of top-quark pair events is performed only in the electron-muon final state, i.e.
 356 focusing on $t\bar{t} \rightarrow W^\pm b W^\mp \bar{b} \rightarrow (\mu^\pm \nu) b (e^\mp \nu) \bar{b}$ due to its small background contributions, using data
 357 that is triggered by the single muon trigger. Only events with exactly one loose isolated muon
 358 (within $|\eta| < 2.1$) and exactly one oppositely charged loose isolated electron (within $|\eta| < 1.44$ or
 359 $1.57 < |\eta| < 2.5$) are selected. The minimum transverse energy/momentum requirements for both
 360 leptons is 25 GeV and the minimum E_T^{miss} requirement is 40 GeV. Moreover, it is required that the
 361 candidate events contain at least two reconstructed jets with $p_T > 40$ GeV within $|y| < 2.4$. At
 362 least one of the reconstructed jets on detector level in the event has to be b-tagged. The number
 363 of candidate events passing this selection is 1,495. The requirements imposed at particle level are
 364 significantly loosened, e.g. no cuts on the neutrinos or jets are applied (see Table 3).

365 The WZ diboson production cross section is studied in the fully leptonic final state, i.e.
 366 requiring at least three charged, loosely isolated electrons or muons with $p_T > 25$ GeV within
 367 $|\eta| < 2.1$ at detector level, but within $|\eta| < 2.5$ at particle level. We only use data that is triggered
 368 by the single-muon trigger as its performance could be cross-checked in the $W \rightarrow \mu\nu$ analysis.
 369 Hence, the $ee\nu\nu$ final state is not considered further ³. The missing transverse energy at detector
 370 level is required to be larger than 20 GeV. The oppositely charged leptons of the same flavor whose
 371 invariant mass, m_{ll} , is closest to the Z boson candidate mass are required to fulfill $66 < m_{ll} < 116$
 372 GeV. The third lepton is identified as the W boson decay lepton, and the resulting transverse mass
 373 is required to be above 40 GeV. Similar, but not exactly identical, requirements are applied at the
 374 particle level objects and summarized in Table 3 together with the number of selected WZ candidate
 375 events in data. The efficiency correction factors C for the seven different processes considered are
 376 also summarized in Table 3.

377 4.2 Background Estimations

378 The contribution of background processes other than multijet processes is estimated using fully
 379 simulated MC samples detailed in Table 2. Each of the signal selections is applied to those samples
 380 and the corresponding yields are evaluated and weighted by the corresponding cross section of the
 381 processes and data luminosity. Following previous analyses, we assume a conservative uncertainty
 382 of 5% in the cross sections of all relevant background processes. The contributions of those
 383 background processes that are known to have only a small impact in the signal region, i.e. below
 384 the uncertainty of the largest background contribution, are neglected.

385 Multijet backgrounds, as well as backgrounds involving non-prompt leptons or jets that are
 386 wrongly identified as leptons, are estimated in data. A so called $ABCD$ method is used where
 387 two orthogonal properties of events, separating signal from multijet background processes, are
 388 used to define four regions in phase space, of which one region (A) is the signal region. The
 389 events in regions B and C pass one signal selection criterion, but fail the second, while events in
 390 region D fail both signal selection requirements. Signal contributions, as well as contributions
 391 from background processes that have been determined via full MC simulations in the regions B, C

³No MC sample of the $W \rightarrow e\nu$ processes is available within the CMS Open Data initiative, which would allow the validation of the corresponding single electron trigger efficiency.

Process	# selected events	Definition of fiducial phase-space	C factor
$Z/\gamma^* \rightarrow e^+e^-$	434,179	$(1e^+1e^-)$, $60 < m_{ee} < 120$ GeV, $p_T^e > 25$ GeV, $ \eta^e < 2.1$	0.525 ± 0.015
$Z/\gamma^* \rightarrow \mu^+\mu^-$	473,626	$(1\mu^+1\mu^-)$, $60 < m_{\mu\mu} < 120$ GeV, $p_T^\mu > 25$ GeV, $ \eta^\mu < 2.1$	0.637 ± 0.010
$Z/\gamma^* \rightarrow \mu^+\mu^-$ $+ \geq 1jet$	61,447	$(1\mu^+1\mu^-)$, $70 < m_{\mu\mu} < 110$ GeV, $p_T^\mu > 20$ GeV, $ \eta^\mu < 2.4$, $p_T^{jet} > 30$ GeV, $ y^{jet} < 2.4$ $\Delta R(j, l) > 0.5$	0.428 ± 0.029
$W^+ \rightarrow \mu^+\nu$	3,631,170	$(1\mu^+)$, $p_T^\mu > 25$ GeV, $ \eta^\mu < 2.4$, $p_T^\nu > 25$ GeV, $m_T > 40$ GeV	0.593 ± 0.017
$W^- \rightarrow \mu^-\nu$	2,629,480	$(1\mu^-)$, $p_T^\mu > 25$ GeV, $ \eta^\mu < 2.4$, $p_T^\nu > 25$ GeV, $m_T > 40$ GeV	0.611 ± 0.018
$t\bar{t} \rightarrow \mu^\mp e^\pm \nu \bar{b}b$	1495	$1\mu^\pm, 1e^\mp$, $p_T^e, p_T^\mu > 20$ GeV, $ \eta^e , \eta^\mu < 2.4$,	0.177 ± 0.012
$W^\pm Z \rightarrow l^\pm \nu l^+ l^-$ ($l = e, \mu$)	79	$(e^\pm e^\mp \mu^\pm), (\mu^\pm \mu^\mp e^\pm), (\mu^\pm \mu^\mp \mu^\pm)$, $p_T^l > 25$ GeV, $ \eta^l < 2.5$, $80 < m_{ll} < 100$ GeV, $m_T > 40$ GeV	0.363 ± 0.011

Table 3. Overview of selected candidate events, the definition of the corresponding fiducial phase-space regions as well as detector correction (C) factors for seven chosen validation processes.

and D , are subtracted. Assuming no correlation between the two selection properties, the multi-jet background yield in region A can then be estimated by $N_A = N_B \cdot N_C / N_D$.

For Drell-Yan processes, the events are categorized as oppositely-charged and same-charged lepton pair events as well as in fully-isolated and semi- or non-isolated lepton pairs. When applying the $ABCD$ method described above, a multijet background contribution of 0.2% is found. A systematic 50% systematic uncertainty is applied to this contribution. This is validated by varying the degree of non-isolation of lepton pairs and repeating the multijet background estimation.

The definition of the $ABCD$ regions in W^\pm boson processes are also isolated and non-isolated leptons, as well as events with ($E_T < 25$ GeV, $m_T < 40$ GeV) and ($E_T > 25$ GeV, $m_T > 40$ GeV), where a muon trigger without an isolation requirement has been used. This choice leads to a multijet background estimate of 200000 events. The systematic uncertainty in this value is estimated again by varying the degree of the lepton non-isolation as well as the requirements on the E_T and m_T . An uncertainty of 30% covers all observed variations in the background yield. The same regions are used for the WZ diboson signal selection to estimate the multijet background as well as the Z +jets background contribution, where one jet is miss-identified as a lepton, yielding a background contribution of $< 1\%$.

The multijet contribution in the $t\bar{t}$ study is estimated by studying events where both leptons fail the isolation requirement and/or fail the requirement on E_T^{miss} , yielding to a relative contribution of below 1%. A careful analysis of the multijet background can certainly reduce the corresponding systematic uncertainties for all estimations, however, the approach we chose is fully justified in the context of this study with its limited precision focus.

An overview of the expected background contributions is given in Table 4.

4.3 Systematic Uncertainties

The systematic uncertainties in the detector correction factors C (see Section 3) have been evaluated within our analysis framework by varying each correction independently within its uncertainties.

Process	$Z \rightarrow \tau\tau$	$Z \rightarrow \mu\mu$	$t\bar{t} \rightarrow 2l2\nu2b$ $t\bar{t} \rightarrow 1l1\nu2b2q$	DiBoson	Multijet
$Z/\gamma^* \rightarrow e^+e^-$	<0.2%	-	0.2%	<0.1%	0.2%
$Z/\gamma^* \rightarrow \mu^+\mu^-$	0.1%	-	0.1%	<0.1%	<0.2%
$Z/\gamma^* \rightarrow \mu^+\mu^- + \geq 1 \text{ jet}$	0.1%	-	0.6%	0.7%	<0.2%
$W^+ \rightarrow \mu^+\nu$	-	5.3%	0.3%	<0.1%	2.8%
$W^- \rightarrow \mu^-\nu$	-	5.8%	0.4%	<0.1%	3.8%
$t\bar{t} \rightarrow \mu^\mp e^\pm \nu \bar{\nu} b\bar{b}$	0.8%	-	4.0%	<0.1%	<0.1%
$W^\pm Z \rightarrow l^\pm \nu l^+ l^- (l = e, \mu)$	-	-	-	8.1%	1%

Table 4. Overview of the relative contribution of background processes to the signal region

Process	Elec. Eff.	Elec. Scale/Res.	Muon Eff.	Muon Scale/Res.	JES/JER	E_T^{miss}	b-tagging	pile-up	Total
$Z/\gamma^* \rightarrow e^+e^-$	2.9%	0.2%	-	-	-	-	-	0.1%	2.9%
$Z/\gamma^* \rightarrow \mu^+\mu^-$	-	-	1.5%	0.3%	-	-	-	0.1%	1.6%
$Z/\gamma^* \rightarrow \mu^+\mu^- + \geq 1 \text{ jet}$	-	-	1.5%	0.4%	6.5%	-	-	0.3%	6.7%
$W^+ \rightarrow \mu^+\nu$	-	-	0.8%	0.2%	-	2.2%	-	1.8%	2.9%
$W^- \rightarrow \mu^-\nu$	-	-	0.8%	0.2%	-	2.2%	-	1.8%	2.9%
$t\bar{t} \rightarrow \mu^\mp e^\pm \nu \bar{\nu} b\bar{b}$	1.5%	0.2%	1.0%	0.2%	5.2%	1.4%	3%	1.9%	6.7%
$W^\pm Z \rightarrow l^\pm \nu l^+ l^- (l = e, \mu)$	1.3%	0.2%	1.5%	0.3%	-	1.3%	-	1.8%	3.0%

Table 5. Relative uncertainties on the detector correction factor C for all studied validation processes due to different systematic uncertainties of detector effects. The uncertainties on the electron efficiencies (Elec. Eff) as well as on the muon efficiencies (Muon. Eff.) summarize reconstruction, identification, isolation and trigger efficiencies. Scale and resolution effects (Scale/Res.) for electrons and muons, as well as jet energy scale and resolution uncertainties (JES/JER) are separated.

417 The difference of the resulting correction factor after a particular variation i , C'_i , to the nominal C
418 factor, is then taken as systematic $\Delta C = C'_i - C$. When applicable, these differences are symmetrized
419 for up- and down-variations. The systematic uncertainties due to pile-up are estimated by comparing
420 the selection with and without the ρ -parameter reweighting. All relevant sources of the uncertainties
421 are treated as independent from each other and hence the total systematic uncertainty on ΔC_{tot}
422 is given by the Gaussian sum of the individual uncertainties ΔC_i . The systematic uncertainties in C
423 range between 3% and 10% and dominate over the statistical uncertainties due to the size of the MC
424 samples. An overview of the uncertainty breakdown on the C factors for all validation samples is
425 given in Table 5.

426 Numerous control distributions between data and MC for all seven processes have been vali-
427 dated. Good agreement between data and MC has been observed in all of them. The normalized
428 invariant mass and lepton rapidity distributions for the Drell–Yan processes have been already
429 discussed in Section 3. Two selected jet distributions of the Z +jets study are shown as an example
430 in Figure 6, where good agreement of the MC prediction and the data can be observed. The
431 measurement of the W boson production cross section is able to test the description of E_T^{miss} , hence
432 Figure 7 displays the comparison of data and MC for E_T^{miss} and m_T . Similarly, Figure 8 shows the
433 comparison of the leading jet p_T as well as E_T^{miss} for the $t\bar{t}$ selection, with a similar conclusion.

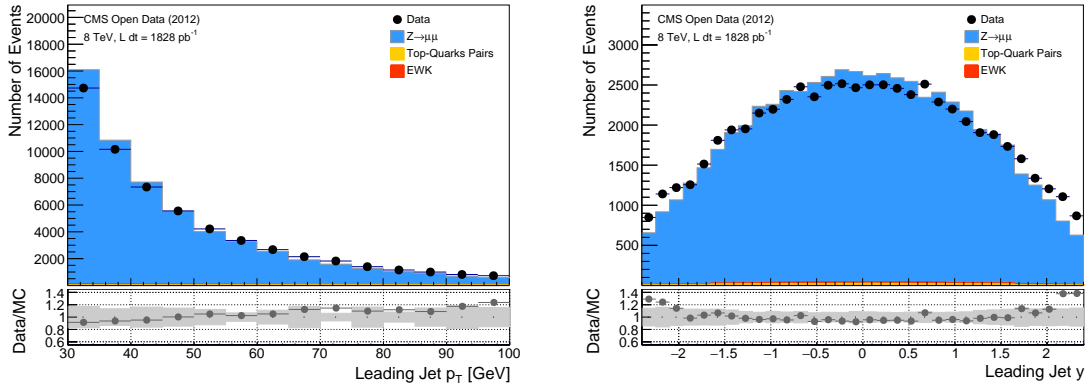


Figure 6. Normalized jet p_T (left) and jet-rapidity distributions (right) for data and MC in the Z+jets study, after all corrections have been applied. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

434 The study of WZ diboson production is statistically limited. Control distributions of the invariant
 435 and transverse mass observables are shown in Figure 9, where a good agreement within the limited
 436 statistical precision is observed.

437 4.4 Results and Comparisons

438 The fiducial production cross sections for the seven validation processes are determined via Eq. 4.2,
 439 using the detector correction factors (Table 3) as well as the estimated background contributions
 440 (Table 4). The resulting cross sections are summarized in Table 6 and Figure 10, together with
 441 their statistical and systematic uncertainties, where the latter are separated into detector-related and
 442 luminosity uncertainties. The results are also compared to previously published results from the
 443 CMS and ATLAS Collaborations, depending on which fiducial phase-space regions are closer to our
 444 own choices [2, 51, 54–56]. The $t\bar{t}$ and WZ cross sections are compared to the published inclusive
 445 production cross sections⁴. Apart from the diboson WZ process, all cross section measurements
 446 are dominated by systematic uncertainties due to the detector effects. The uncertainties due the
 447 background processes are small for all channels, except the W^\pm boson production, where the
 448 uncertainty in the multijet background is of a similar size as the uncertainties due to the missing
 449 transverse energy requirements. It should be noted that the uncertainties for the $Z/\gamma^* \rightarrow l^+l^-$
 450 processes in our analysis appear to be smaller than the official measurements by CMS [2], however,
 451 this is due to the smaller data-set used by CMS as well as the limited number of significant digits in the
 452 published result. The smaller systematic uncertainties in the diboson WZ cross section measurement
 453 is due to the smaller lepton reconstruction uncertainties assumed in our analysis and motivated in
 454 Section 3. We also compare the inclusive cross section to the available theoretical predictions,
 455 which have been previously published. Figure 11 shows the ratio of the theory predictions to our
 456 measured fiducial cross sections and to the previously published results by either ATLAS or CMS.

⁴No model uncertainties were considered when extrapolating from our fiducial cross section to the inclusive cross section.

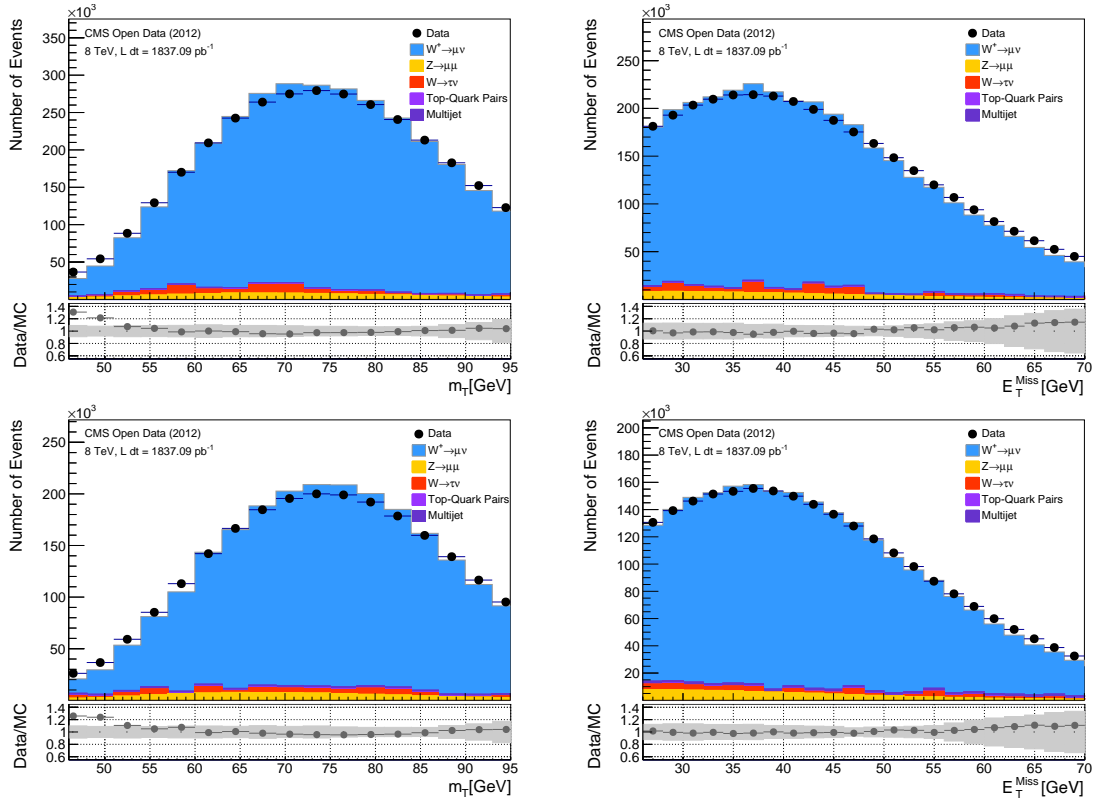


Figure 7. Normalized m_T (left) and E_T^{miss} distributions (right) for data and MC in W^+ events (upper row) and W^- events (lower row), after all corrections have been applied. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

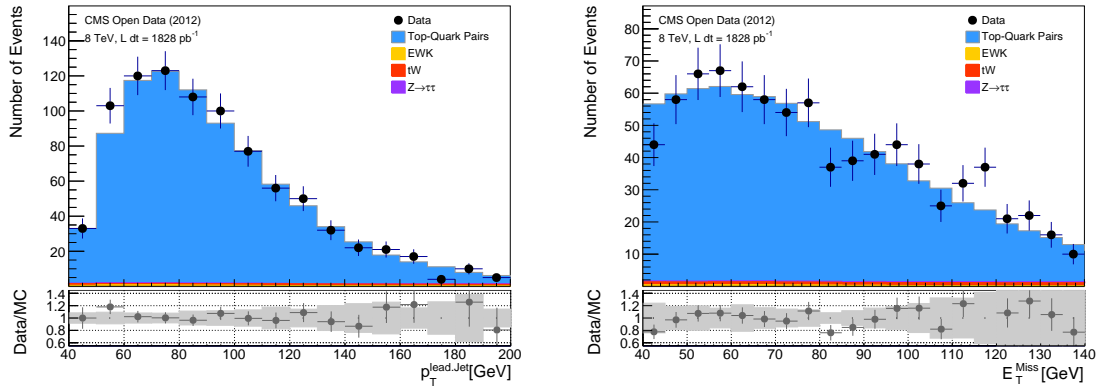


Figure 8. Normalized leading jet p_T (left) and E_T^{miss} distributions (right) for data and MC in the $t\bar{t}$ study, after all corrections have been applied. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

457 All our measurements are in good agreement with the previously published results as well as the
 458 SM predictions. The measurement systematic uncertainties are between 1.6 and 6.7%.

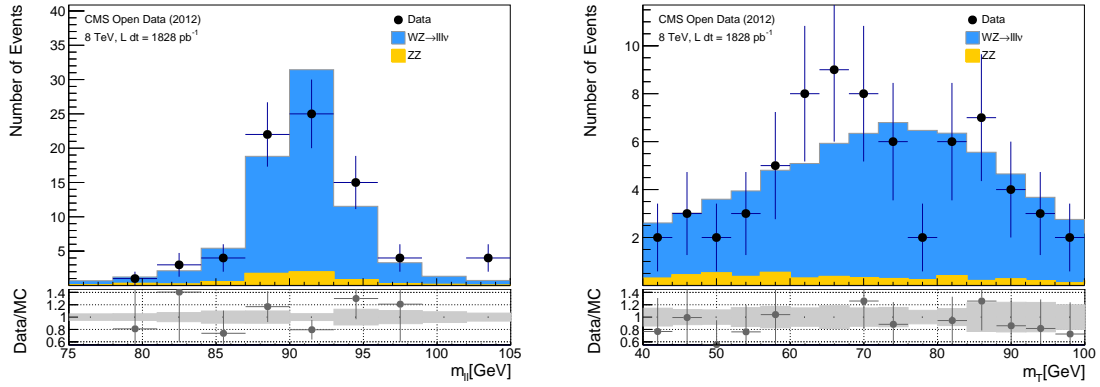


Figure 9. Normalized invariant mass m_{ll} (left) and transverse mass m_T distributions (right) for data and MC in the WZ study, after all corrections have been applied. The gray band in the ratio indicates the systematic uncertainties on the predictions without luminosity uncertainties.

Process	Measurement cross section [pb] (<i>stat.</i> \pm <i>sys.</i> \pm <i>lumi.</i>)	Prediction cross section [pb]	Previous Result cross section [pb] (<i>stat.</i> \pm <i>sys.</i> \pm <i>lumi.</i>)
$Z/\gamma^* \rightarrow e^+e^-$	$\sigma^{fid} = 461 \pm 17$ ($1 \pm 13 \pm 11$)	$\sigma^{fid.} = 450 \pm 20$ [2]	$\sigma^{fid} = 450 \pm 20$ [2] ($10 \pm 10 \pm 10$)
$Z/\gamma^* \rightarrow \mu^+\mu^-$	$\sigma^{fid} = 406 \pm 12$ ($1 \pm 6 \pm 10$)	$\sigma^{fid.} = 400 \pm 10$ [2]	$\sigma^{fid} = 410 \pm 20$ [2] ($10 \pm 10 \pm 10$)
$Z/\gamma^* \rightarrow \mu^+\mu^- + \geq 1 \text{ jet}$	$\sigma^{fid} = 77.1 \pm 5.5$ ($0.4 \pm 5.1 \pm 1.9$)	$\sigma^{fid.} = 76.3 \pm 5.0$ [55]	$\sigma^{fid} = 75.5 \pm 4.0$ [55] ($0.1 \pm 3.7 \pm 1.4$)
$W^+ \rightarrow \mu^+\nu$	$\sigma^{fid} = 3052 \pm 124$ ($1 \pm 98 \pm 76$)	$\sigma^{fid.} = 3015 \pm 100$ [54]	$\sigma^{fid} = 3110 \pm 66$ [54] ($0.5 \pm 29 \pm 59$)
$W^- \rightarrow \mu^-\nu$	$\sigma^{fid} = 2103 \pm 86$ ($1 \pm 69 \pm 52$)	$\sigma^{fid.} = 2105 \pm 60$ [54]	$\sigma^{fid} = 2137 \pm 47$ [54] ($0.4 \pm 22 \pm 41$)
$t\bar{t} \rightarrow \mu^\mp e^\pm \nu \bar{b}b\bar{b}$	$\sigma^{incl.} = 4.54 \pm 0.35$ ($0.14 \pm 0.30 \pm 0.11$)	$\sigma^{incl.} = 4.37 \pm 0.35$ [51]	$\sigma^{incl.} = 4.23 \pm 0.14$ [51] ($0.02 \pm 0.10 \pm 0.10$)
$W^\pm Z \rightarrow l^\pm \nu l^+ l^-$	$\sigma^{incl.} = 28.1 \pm 3.3$ ($3.1 \pm 0.9 \pm 0.7$)	$\sigma^{incl.} = 23.7 \pm 0.4$ [56]	$\sigma^{incl.} = 24.09 \pm 1.8$ [56] ($0.87 \pm 1.6 \pm 0.6$)

Table 6. Overview of measured cross sections of seven validation processes as well as previously published results together with theory predictions.

459 5 Opportunities and Challenges of the CMS Open Data Initiative

460 The CMS Open Data initiative offers a unique opportunity to study and measure properties of
461 the SM as long as a limited precision is sufficient. Measurements with higher precision currently
462 seem not achievable, given the limited available information on the detector calibration as well
463 as the systematic uncertainties of relevant observables. Clearly, these calibration efforts are one
464 of the main areas of research within the Collaborations and the publications of the corresponding
465 information in an easily accessible and understandable format for external physicists is highly
466 challenging. One example is that experimental uncertainties in the energy scale of particle jets

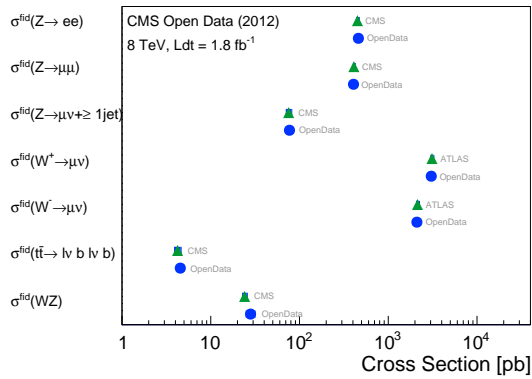


Figure 10. Overview of the measured SM production cross sections of seven validation processes.

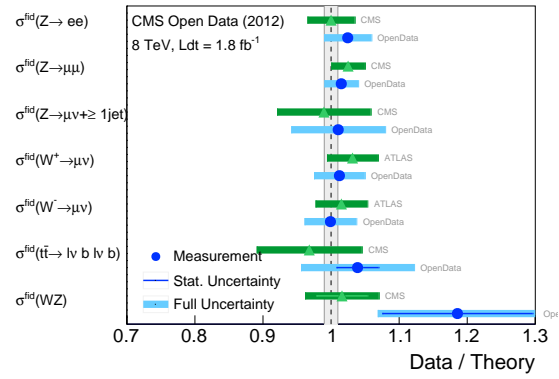


Figure 11. Ratio of measured cross sections, derived in this work as well as by the LHC Collaborations, to their corresponding theoretical predictions.

467 involve dozens of calibration parameters which have to be applied correctly. Another example are
 468 uncertainties in the lepton identification efficiencies, which are correlated in a complex manner,
 469 where the correlations matter for high precision measurements. Hence, precision measurements of
 470 (differential) cross sections as well as cross section ratios should, and can only, be performed by the
 471 LHC Collaborations.

472 As a first possible improvement of the CMS Open Data initiative, we suggest that some
 473 simplified baseline calibrations, as well as uncertainties, should become available. This is highly
 474 desirable to make legacy data analyses possible to confront future theories with data. As a second
 475 possible improvement of the CMS Open Data initiative, we suggest that the CMS Collaboration
 476 publishes dedicated baseline analyses, such as presented in Section 4. This would allow external
 477 physicists to understand the technical details, e.g. how to apply the detector calibration or simply
 478 how to calculate the integrated luminosity for a given data sample.

479 However, even given the mentioned shortcomings, we see a significant physics potential in
 480 the CMS Open Data initiative as illustrated in this work. We have demonstrated that it is possible
 481 to repeat measurements of the Z boson production cross section in the electron and muon decay
 482 channels, also in association with one jet in the final state, of the W^\pm boson production cross section
 483 in the muon decay channel, of the top-quark pair production cross section in the fully leptonic decay
 484 mode, and of the WZ diboson production cross section. Our validation measurements agree within
 485 less than 3% to the official measurements by the CMS and ATLAS Collaborations. The differences
 486 are within the statistical and systematic uncertainties of the measurements. This lays the foundation
 487 to extend cross section measurements to extreme phase-space regions, which have not been probed
 488 so far.

489 Acknowledgments

490 We would like to thank the CMS Collaboration for providing the full 2012 data set as well as for
 491 the documentation on the CMS detector performance. This work would have not been possible
 492 without the excellent performance of the LHC as well as the existing computing infrastructure and

493 the support from CERN. We would also like to thank Guillermo Gomez-Ceballos and Frank Fiedler
494 for the helpful comments during the revision of this paper. M.S. would like to thank in addition the
495 Fulbright commission as well as the Volkswagen Foundation for the support of this work. Moreover,
496 he would like to thank his colleagues at MIT, Philip Harris in particular, for answering all questions
497 regarding the treatment of the CMS Open Data for this project as well as the pleasant environment
498 during the Fulbright research scholarship.

499 References

- 500 [1] ATLAS collaboration, *Precision measurement and interpretation of inclusive W^+ , W^- and Z/γ^**
501 *production cross sections with the ATLAS detector*, *Eur. Phys. J. C* **77** (2017) 367, [1612.03016].
- 502 [2] CMS collaboration, *Measurement of inclusive W and Z boson production cross sections in pp*
503 *collisions at $\sqrt{s} = 8$ TeV*, *Phys. Rev. Lett.* **112** (2014) 191802, [1402.0923].
- 504 [3] CMS collaboration, *Measurement of the $t\bar{t}$ production cross section in the dilepton channel in pp*
505 *collisions at $\sqrt{s} = 8$ TeV*, *JHEP* **02** (2014) 024, [1312.7582].
- 506 [4] ATLAS collaboration, *Measurement of the angular coefficients in Z -boson events using electron and*
507 *muon pairs from data taken at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *JHEP* **08** (2016) 159,
508 [1606.00689].
- 509 [5] ATLAS collaboration, *Measurement of the transverse momentum and ϕ_η^* distributions of Drell-Yan*
510 *lepton pairs in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **76**
511 (2016) 291, [1512.02192].
- 512 [6] CMS, “Software Framework for CMS Open Data Analysis.”
513 <http://opendata.cern.ch/docs/about-cms>.
- 514 [7] C. Cesarotti, Y. Soreq, M. J. Strassler, J. Thaler and W. Xue, *Searching in CMS Open Data for Dimuon*
515 *Resonances with Substantial Transverse Momentum*, *Phys. Rev. D* **100** (2019) 015021, [1902.04222].
- 516 [8] A. Tripathy, W. Xue, A. Larkoski, S. Marzani and J. Thaler, *Jet Substructure Studies with CMS Open*
517 *Data*, *Phys. Rev. D* **96** (2017) 074003, [1704.05842].
- 518 [9] CMS collaboration, *The CMS Experiment at the CERN LHC*, *JINST* **3** (2008) S08004.
- 519 [10] CMS collaboration, *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and*
520 *MET*, Tech. Rep. CMS-PAS-PFT-09-001, CERN, Geneva, Apr, 2009.
- 521 [11] W. Adam, R. Fruehwirth, A. Strandlie and T. Todor, *Reconstruction of Electrons with the*
522 *Gaussian-Sum Filter in the CMS Tracker at the LHC*, Tech. Rep. CMS-NOTE-2005-001, CERN,
523 Geneva, Jan, 2005.
- 524 [12] CMS collaboration, *Energy Calibration and Resolution of the CMS Electromagnetic Calorimeter in*
525 *pp Collisions at $\sqrt{s} = 7$ TeV*, *JINST* **8** (2013) P09009, [1306.2016].
- 526 [13] CMS collaboration, *Performance of Electron Reconstruction and Selection with the CMS Detector in*
527 *Proton-Proton Collisions at 8 TeV*, *JINST* **10** (2015) P06005, [1502.02701].
- 528 [14] CMS collaboration, *Performance of muon identification in pp collisions at 7 TeV*, Tech. Rep.
529 CMS-PAS-MUO-10-002, CERN, Geneva, 2010.
- 530 [15] CMS collaboration, *Commissioning of the Particle-flow Event Reconstruction with the first LHC*
531 *collisions recorded in the CMS detector*, Tech. Rep. CMS-PAS-PFT-10-001, 2010.

- 532 [16] CMS collaboration, *Determination of Jet Energy Calibration and Transverse Momentum Resolution*
533 *in CMS*, *JINST* **6** (2011) P11002, [[1107.4277](#)].
- 534 [17] CMS collaboration, *Identification of b-quark jets with the CMS experiment*, *JINST* **8** (2013) P04013,
535 [[1211.4462](#)].
- 536 [18] C. V. Philip Harris, Cristina Ana Mantilla Suarez et al., “Bacon analysis framework.”
537 <https://github.com/ksung25/BaconProd/tree/Run1>.
- 538 [19] A. Hinzmann, *Tools for physics analysis in CMS*, *J. Phys. Conf. Ser.* **331** (2011) 032042.
- 539 [20] CMS collaboration, *CMS list of validated runs for primary datasets of 2012 data taking*, *CERN Open*
540 *Data Portal*, tech. rep. 10.7483/OPENDATA.CMS.C00V.SE32.
- 541 [21] R. Brun and F. Rademakers, *ROOT: An object oriented data analysis framework*, *Nucl. Instrum. Meth.*
542 **A389** (1997) 81–86.
- 543 [22] CMS collaboration, “Single mu primary dataset in aod format from run of 2012
544 (/singlemu/run2012c-22jan2013-v1/aod).” CERN Open Data Portal: <http://opendata.cern.ch>.
- 545 [23] CMS collaboration, “Single mu primary dataset in aod format from run of 2012
546 (/singlemu/run2012b-22jan2013-v1/aod).” CERN Open Data Portal: <http://opendata.cern.ch>.
- 547 [24] CMS collaboration, “Single electron primary dataset in aod format from run of 2012
548 (/singleelectron/run2012c-22jan2013-v1/aod).” CERN Open Data Portal:
549 <http://opendata.cern.ch>.
- 550 [25] CMS collaboration, “Single electron primary dataset in aod format from run of 2012
551 (/singleelectron/run2012b-22jan2013-v1/aod).” CERN Open Data Portal:
552 <http://opendata.cern.ch>.
- 553 [26] CMS collaboration, *CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update*,
554 **CMS-PAS-LUM-13-001**.
- 555 [27] S. Alioli, P. Nason, C. Oleari and E. Re, *NLO vector-boson production matched with shower in*
556 *POWHEG*, *JHEP* **07** (2008) 060, [[0805.4802](#)].
- 557 [28] S. Alioli, P. Nason, C. Oleari and E. Re, *A general framework for implementing NLO calculations in*
558 *shower Monte Carlo programs: the POWHEG BOX*, *JHEP* **06** (2010) 043, [[1002.2581](#)].
- 559 [29] T. Sjostrand, S. Mrenna and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, *JHEP* **05** (2006) 026,
560 [[hep-ph/0603175](#)].
- 561 [30] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, *MadGraph 5 : Going Beyond*, *JHEP*
562 **06** (2011) 128, [[1106.0522](#)].
- 563 [31] J. Gao, M. Guzzi, J. Huston, H.-L. Lai, Z. Li, P. Nadolsky et al., *CT10 next-to-next-to-leading order*
564 *global analysis of QCD*, *Phys. Rev.* **D89** (2014) 033009, [[1302.6246](#)].
- 565 [32] CMS collaboration, *Study of the Underlying Event at Forward Rapidity in pp Collisions at $\sqrt{s} = 0.9$,*
566 *2.76, and 7 TeV*, *JHEP* **04** (2013) 072, [[1302.2394](#)].
- 567 [33] CMS collaboration, *Event generator tunes obtained from underlying event and multiparton scattering*
568 *measurements*, *Eur. Phys. J.* **C76** (2016) 155, [[1512.00815](#)].
- 569 [34] N. Davidson, G. Nanava, T. Przedzinski, E. Richter-Was and Z. Was, *Universal Interface of TAUOLA*
570 *Technical and Physics Documentation*, *Comput. Phys. Commun.* **183** (2012) 821–843, [[1002.0543](#)].
- 571 [35] GEANT4 collaboration, S. Agostinelli et al., *GEANT4: A Simulation toolkit*, *Nucl. Instrum. Meth.*
572 **A506** (2003) 250–303.

- 573 [36] CMS collaboration, *Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV*,
574 [JINST 12 \(2017\) P02014](#), [1607.03663].
- 575 [37] CMS collaboration, “Dytoee_m-20_ct10_tunez2star_v2_8tev-powheg-pythia6 in aodsim format for
576 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 577 [38] CMS collaboration, “Dytomumu_m-20_ct10_tunez2star_v2_8tev-powheg-pythia6 in aodsim format
578 for 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 579 [39] CMS collaboration, “Simulated dataset
580 dyjetstoll_m-50_tunez2star_8tev-madgraph-tarball-tauola-taupolaroff in aodsim format for 2012
581 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 582 [40] CMS collaboration, “Simulated dataset wplustomunu_ct10_8tev-powheg-pythia6 in aodsim format
583 for 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 584 [41] CMS collaboration, “Simulated dataset wminustomunu_ct10_8tev-powheg-pythia6 in aodsim format
585 for 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 586 [42] CMS collaboration, “Simulated dataset wplustotaunu_ct10_8tev-powheg-pythia6-tauola in aodsim
587 format for 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 588 [43] CMS collaboration, “Simulated dataset wminustotaunu_ct10_8tev-powheg-pythia6-tauola in aodsim
589 format for 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 590 [44] CMS collaboration, “Simulated dataset ttjets_fulleptmgdecays_tunep11tev_8tev-madgraph-tauola in
591 aodsim format for 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 592 [45] CMS collaboration, “Simulated dataset ttjets_semleptmgdecays_8tev-madgraph in aodsim format for
593 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 594 [46] CMS collaboration, “Simulated dataset
595 ttjets_hadronicmgdecays_tunep11mpihi_8tev-madgraph-tauola in aodsim format for 2012 collision
596 data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 597 [47] CMS collaboration, “Simulated dataset wwjetsto2l2nu_tunez2star_8tev-madgraph-tauola in aodsim
598 format for 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 599 [48] CMS collaboration, “Simulated dataset wzjetsto3lnu_8tev_tunez2star_madgraph_tauola in aodsim
600 format for 2012 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 601 [49] CMS collaboration, “Simulated dataset zzto4mu_8tev-powheg-pythia6 in aodsim format for 2012
602 collision data..” CERN Open Data Portal: <http://opendata.cern.ch>.
- 603 [50] CMS collaboration, *The CMS trigger system*, [JINST 12 \(2017\) P01020](#), [1609.02366].
- 604 [51] CMS collaboration, *Measurement of the t-tbar production cross section in the e-mu channel in
605 proton-proton collisions at sqrt(s) = 7 and 8 TeV*, [JHEP 08 \(2016\) 029](#), [1603.02303].
- 606 [52] CMS collaboration, *Performance of the CMS missing transverse momentum reconstruction in pp data
607 at sqrt(s) = 8 TeV*, [JINST 10 \(2015\) P02006](#), [1411.0511].
- 608 [53] PARTICLE DATA GROUP collaboration, M. Tanabashi et al., *Review of Particle Physics*, [Phys. Rev. D98
609 \(2018\) 030001](#).
- 610 [54] ATLAS collaboration, *Measurement of the cross-section and charge asymmetry of W bosons
611 produced in proton-proton collisions at sqrt(s) = 8 TeV with the ATLAS detector*, [1904.05631](#).
- 612 [55] CMS collaboration, *Measurements of differential production cross sections for a Z boson in
613 association with jets in pp collisions at sqrt(s) = 8 TeV*, [JHEP 04 \(2017\) 022](#), [1611.03844].

614 [56] CMS collaboration, *Measurement of the WZ production cross section in pp collisions at $\sqrt{s} = 7$ and 8*
615 *TeV and search for anomalous triple gauge couplings at $\sqrt{s} = 8$ TeV*, *Eur. Phys. J. C* **77** (2017) 236,
616 [[1609.05721](#)].