Opportunities and challenges of Standard Model production cross section measurements in proton-proton collisions at $\sqrt{s} = 8$ TeV using CMS Open Data

A. Apyan, W. Cuozzo, M. Klute, Y. Saito, M. Schott and B. Sintayehu

$^a$Fermilab, Batavia, IL, U.S.A.$^b$Massachusetts Institute of Technology, Cambridge, MA, U.S.A.$^c$Johannes Gutenberg-University, Mainz, Germany

E-mail: matthias.schott@cern.ch

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

KEYWORDS: Analysis and statistical methods; Simulation methods and programs
1 Introduction

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

The CMS Collaboration has published significant amounts of recorded and simulated proton-proton collisions at a center-of-mass energy of 8 TeV within the context of the Open Data initiative [6]. In principle, the availability of these data sets allows physicists who are not member of the LHC Collaborations to perform measurements. With this work, we systematically evaluate the physics potential of the available CMS Open Data for cross section measurements and broaden
the perspective of previous studies using CMS Open Data [7, 8]. Special focus is drawn on the limitations of such measurements and possible future improvements.

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

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

2 The CMS detector and CMS Open Data

2.1 The CMS detector and reconstructed objects

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

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

Muons are reconstructed from a global fit of hits in the MS and the ID, seeded by tracks in the muon system [14]. In this analysis, we typically require each muon to have a transverse momentum
of \( p_T > 25 \text{ GeV} \) within a pseudorapidity range of \( |\eta| < 2.1 \), corresponding to the single muon trigger coverage. In addition, following previous CMS measurements [2], standard quality requirements on the numbers of hits in the ID and in the MS, on the \( \chi^2 \) of the fit, and on the track impact parameters are applied. A relative isolation variable is computed as described for the electrons for a loose and tight isolation definition, respectively, however, computed with a cone-radius of \( \Delta R = 0.4 \). Given the small mass of muons and electrons compared to their energies relevant in the studied processes within this work, their transverse momentum and their transverse energy can be assumed to be equal, i.e. \( E_T = p_T \).

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

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

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

It is additionally required that selected reconstructed objects are not considered for further analysis if they are close to other reconstructed objects. Electron candidates are not further considered if a muon candidate with \( p_T > 20 \text{ GeV} \), passing standard quality criteria, has been reconstructed within \( \Delta R < 0.3 \) of the electron candidate.

### 2.2 Software and infrastructure

The CMS Open Data Software Framework (Release CMSSW_5_3_32), available in [6], builds the basis of this analysis. A dedicated open-source framework, \textsc{Bacon} [18], which was used for several published studies of the CMS Collaboration, e.g. [2], is used to read the Analysis Object Data (AOD) [19], extracting information on reconstructed objects as well as particle-level data.\(^1\) If available. The \textsc{Bacon} software framework is also used to apply a selection of events which have been recorded under stable detector conditions, known as \textit{GoodRun-List} [20], provided by the CMS Open Data project, leading to a reduced output-format based on the \textit{Root} software framework [21]. The typical size of one simulated top-quark pair event in the \textsc{Bacon}-output format amounts to 5

---

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

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

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

2.3 Selected Open Data

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

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

2.4 MC simulated samples

An overview of the various signal and background samples used in this analysis is given in table 2, indicating the underlying physics process, the dataset name, and the corresponding inclusive cross section at next-to-leading order (NLO) or next-to-next-to-leading order (NNLO). The Drell-Yan processes (\(W/Z\)) in the electron and muon decay channels were generated with the PowhegBox v.1.0 MC program [27, 28] interfaced to the Pythia v.6.4.26 parton shower model [29]. All other processes are modeled with the tree-level matrix element event generator MadGraph v5.1.3.30 [30] interfaced with Pythia 6.4.26. The CT10 parton distribution functions (PDFs) [31] and the \(Z2^*\) Pythia6 tune [32, 33] are used. The decays of tau-leptons are modeled using the Tauola program [34]. The Pythia v.6.4.26 is used for the modeling of photon radiation off final state particles. The strong coupling constant \(\alpha_s\) has been set to 0.130 at the Z boson mass scale for all matrix element calculations. The effect of pile-up has been simulated by overlaying MC-generated minimum bias events. The Geant4 program was used to simulate the passage of particles through the CMS detector [35].
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 $l$.

| Process                                      | Dataset Name                                                                 | Inclusive $\sigma$ [pb] | order in $\alpha_s$ |
|----------------------------------------------|-------------------------------------------------------------------------------|--------------------------|---------------------|
| $pp \rightarrow Z\gamma^* + X \rightarrow e^+e^- + X$ | DYToEE_M-20_C10_TuneZ2star_v2_8TeV \cite{37}                                | 1916                     | NNLO                |
| $pp \rightarrow Z\gamma^* + X \rightarrow \mu^+\mu^- + X$ | DYToMuMu_M-20_C10_TuneZ2star_v2_8TeV \cite{38}                              | 1916                     | NNLO                |
| $pp \rightarrow Z\gamma^* + X \rightarrow \ell^+\ell^- + X$ | DYJetsToLL_M-50_TuneZ2Star_v2_8TeV \cite{39}                                | 3533                     | NNLO                |
| $pp \rightarrow W^+ + X \rightarrow \mu^+\nu + X$          | WplusToMuNu_C10_8TeV \cite{40}                                               | 7322                     | NNLO                |
| $pp \rightarrow W^- + X \rightarrow \mu^-\bar{\nu} + X$    | WminusToMuNu_C10_8TeV \cite{41}                                              | 5181                     | NNLO                |
| $pp \rightarrow W^+ + X \rightarrow \tau^+\nu + X$         | WplusToTauNu_C10_8TeV \cite{42}                                              | 7322                     | NNLO                |
| $pp \rightarrow W^- + X \rightarrow \tau^-\bar{\nu} + X$   | WminusToTauNu_C10_8TeV \cite{43}                                             | 5181                     | NNLO                |
| $pp \rightarrow t\bar{t} + X \rightarrow 2\ell 2\nu 2b + X$ | TTJets_FullLeptMGDecays_TuneP11TeV_8TeV \cite{44}                            | 112.3                    | NLO                 |
| $pp \rightarrow t\bar{t} + X \rightarrow 1\ell 2q 2b + X$   | TTJets_SemiLeptMGDecays_8TeV \cite{45}                                       | 107.2                    | NLO                 |
| $pp \rightarrow t\bar{t} + X \rightarrow 4q 2b + X$         | TTJets_HadronicMGDecays_TuneP11mpiHi_8TeV \cite{46}                         | 25.8                     | NLO                 |
| $pp \rightarrow WW + X \rightarrow 2\ell 2\nu + X$         | WWJetsTo2L2Nu_TuneZ2star_8TeV \cite{47}                                       | 5.8                      | NLO                 |
| $pp \rightarrow WZ + X \rightarrow 3\ell 1\nu + X$          | WZJetsTo3Lnu_8TeV_TuneZ2Star \cite{48}                                        | 1.1                      | NNLO                |
| $pp \rightarrow ZZ + X \rightarrow 4\mu + X$               | ZZTo4mu_8TeV \cite{49}                                                       | 0.077                    | NLO                 |

Figure 1. Distribution of the $\rho$ 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.

The simulated event samples are reweighted to describe the distribution of the number of pile-up events in the data by reweighting the $\rho$ parameter distribution, where $\rho$ denotes the diffuse offset energy density \cite{36}. Moreover, a reweighting of the longitudinal position of the primary $pp$ collision vertex of the MC samples to data has been performed. The resulting $\rho$ distributions for data and simulated Z boson samples in the electron and muon decay channels are shown in figure 1. The differences in MC predictions with and without reweighting is taken as a systematic uncertainty.
3 Calibration

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

3.1 Muon performance

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

$$p_{T\text{Reco}} = p_{T\text{Truth}} + \beta \cdot (\alpha \cdot p_{T\text{Reco}} - p_{T\text{Truth}}),$$

(3.1)

where $p_{T\text{Reco}}$ is the reconstructed muon momentum, $p_{T\text{Truth}}$ is the truth muon momentum on particle level, $\alpha$ is a momentum scale parameter, and $\beta$ is a resolution parameter. The parameters $\alpha$ and $\beta$ are determined for three different regions in $\eta$, corresponding to the two endcap and one barrel regions, by a $\chi^2$ minimization procedure. The $\chi^2$ is calculated between the invariant mass spectrum of the di-lepton system in Z boson events in data and simulation for different choices of $\alpha$ and $\beta$. The average values of $\alpha$ and $\beta$ were found to be 0.998 and 1.13, respectively. Uncertainties in the momentum scale of 0.002 in the barrel region and 0.003 in the endcap regions are applied, which cover the observed discrepancies with data. The uncertainty in the resolution parameter is 0.05. The size of the uncertainties have been additionally tested by varying the invariant mass window requirement. The comparison between data and MC of the invariant mass distribution of di-muon pairs after the calibration procedure is shown in figure 2.

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

3.2 Electron performance

The energy scale and resolution corrections for electrons are derived in a similar way as for the muons, however, an off-set parameter $\kappa$ is also used in addition to a multiplicative scale factor $\alpha$.
Figure 2. Comparison of the di-muon mass spectrum (left) and the muon $\eta$ 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.

The addition of the $\kappa$ off-set parameter leads to a significantly better description of the data with the MC simulated samples. Using

$$E_{\text{T}}^{\text{Reco}} = \kappa + E_{\text{T}}^{\text{Truth}} + \beta \cdot (\alpha \cdot E_{\text{T}}^{\text{Reco}} - E_{\text{T}}^{\text{Truth}}),$$

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

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

3.3 Jet energy scale and resolution

The official CMS calibration and corrections for particle jets, in particular, the jet energy scale (JES) and the jet energy resolution (JER), have been applied within the BACON framework. These jet corrections and uncertainties were derived from the simulation, and are confirmed with in situ measurements using the energy balance of dijet and photon+jet events [36]. A reduced set of systematic variations is used to estimate JES and JER uncertainties on the final measurement. In 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 JER is varied by 20% for $30 < E_{\text{T}}^{\text{jet}} < 100$ GeV, by 10% for $100 < E_{\text{T}}^{\text{jet}} < 1000$ GeV and by 5% for $E_{\text{T}}^{\text{jet}} > 1000$ GeV. We apply a JER uncertainty of 20% for jets with an absolute rapidity larger
than 2.1. This simplified treatment of the jet calibration does not allow for a correct evaluation of correlations between different phase space regions. However, the resulting systematic uncertainties in the inclusive cross section measurements are expected to be valid.

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

3.4 Tagging of b-quarks

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

3.5 Missing transverse energy

We apply the official CMS calibration constants and correction factors to the reconstructed $E_T^{\text{miss}}$ observable on an event-by-event basis. The assigned uncertainties in $E_T^{\text{miss}}$ are based on [52], where the scale uncertainty is taken to be 10% for $E_T^{\text{miss}} < 20$ GeV, 5% for $20 < E_T^{\text{miss}} < 100$ GeV and 2% for $E_T^{\text{miss}} > 100$ GeV. The uncertainty in the $E_T^{\text{miss}}$ resolution is applied as a function of the scalar sum of all transverse energies of all reconstructed hadronic objects in the event, $\Sigma E_T$, and is
taken to be 20% for $\sum E_T < 100$ GeV and 10% for $\sum E_T < 100$ GeV. In addition, we propagate all the jet energy scale and resolution uncertainties to $E_T^{\text{miss}}$ by studying the impact on a recalculated $E_T^{\text{miss}}$ observable, which is based on all reconstructed objects in the event.

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

4 Standard Model cross section measurements

4.1 Standard processes and signal selection

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

The inclusive production cross section for a given process can be experimentally determined
via
\[
\sigma_V^{\text{incl.}} = \frac{N_{\text{Signal}}}{\epsilon \cdot \mathit{BR} \cdot \int L dt} = \frac{N_{\text{Signal}}}{A \cdot C \cdot \mathit{BR} \cdot \int L dt}. \tag{4.1}
\]
The number of signal events is given by \( N_{\text{Signal}} = N_{\text{Data}} - N_{\text{Bkg}} \), where \( N_{\text{Data}} \) is the number of selected
events in data and \( N_{\text{Bkg}} \) is the number of background events surviving the signal selection. The factor \( \epsilon \) is the efficiency of the signal events passing the signal selection criteria. The efficiency correction
\( \epsilon \) can further be decomposed as the product of a fiducial acceptance, \( A \), and a detector-induced
correction factor, \( C \), i.e. \( \epsilon = A \cdot C \), defined below. To correct the cross section for the choice of a
specific decay channel, a branching ratio factor \( \mathit{BR} \) is applied, which is known to high precision [53]. Finally, the event yield is normalized by the integrated luminosity \( \int L dt \) of the data sample analyzed.

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

The fiducial acceptance, \( A \), and a detector-induced correction factor, \( C \), are defined as \( A = N_{\text{selected}}^{\text{particle}} / N_{\text{all}}^{\text{particle}} \) and \( C = N_{\text{selected}}^{\text{detector}} / N_{\text{selected}}^{\text{particle}} \), respectively. The fiducial acceptance
\( A \) is therefore the ratio of the number of events that pass the geometrical and kinematic cuts of
an analysis at particle level, \( N_{\text{selected}}^{\text{particle}} \) over the total number of generated events in a simulated
sample of signal process, \( N_{\text{all}}^{\text{particle}} \). These selection cuts on particle level require geometrical and
kinematic constraints close to the cuts applied on the reconstructed objects at detector level, e.g.
muons in the final state should fulfill \( p_T > 25 \text{ GeV} \) and \( |\eta| < 2.1 \). The dominant uncertainties on the
fiducial acceptance are renormalisation and factorisation scale uncertainties as well as uncertainties
due to the limited knowledge of the parton distribution function (PDF) of the proton. The detector
response correction factor \( C \) is the ratio of selected events on detector level, \( N_{\text{selected}}^{\text{detector}} \), over all
events that pass the kinematic selection on particle level, i.e. \( N_{\text{selected}}^{\text{particle}} \). The fiducial cross section
for a given process can therefore be measured by
\[ \sigma_{V}^{\text{fid}} = \frac{N_{\text{Signal}}}{C \cdot BR \cdot \int L dt}. \] (4.2)

Since this definition is independent of \( A \), modelling uncertainties are minimized. Once the fiducial cross section is known, one can extrapolate to the inclusive cross section using the acceptance correction factor, i.e.
\[ \sigma_{V}^{\text{incl}} = \frac{\sigma_{V}^{\text{fid}}}{A}. \]

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

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

The selection of positively and negatively charged \( W \) bosons is only applied in the muon decay channel, since no simulated \( W^\pm \rightarrow e\nu \) samples at \( \sqrt{s} = 8 \text{ TeV} \) were available on the CERN Open Data Portal. The \( W \) boson candidate events are selected at detector level by requiring exactly one reconstructed, tightly isolated muon with \( p_T > 25 \text{ GeV} \) and within \( |\eta| < 2.1 \). A minimum \( E_T^{\text{miss}} \) of 25 GeV is required and a minimum transverse mass requirement of \( m_T = \sqrt{2 \cdot p_T^l \cdot E_T^{\text{miss}} \cdot (1 - \cos(\Delta\phi))} > 40 \text{ GeV} \) is applied, where \( \Delta\phi \) is the opening angle of the reconstructed muon and the missing transverse momentum vector in the transverse plane. This selection differs from the chosen approach in [2] and is closer to [54] in order to reduce multijet background contributions. Similar kinematic constraints are applied at particle level on the charged decay lepton, the neutrino and the derived quantities (table 3). In total, 3,631,170 \( W^+ \) and 2,629,480 \( W^- \) candidate events have been selected, respectively.

The selection of top-quark pair events is performed only in the electron-muon final state, i.e. focusing on \( t\bar{t} \rightarrow W^+ b W^- \bar{b} \rightarrow (\mu^+ \nu) b (\nu^\prime \bar{\nu}^\prime) \bar{b} \) due to its small background contributions, using data that is triggered by the single muon trigger. Only events with exactly one loose isolated muon (within \( |\eta| < 2.1 \)) and exactly one oppositely charged loose isolated electron (within \( |\eta| < 1.44 \) or \( 1.57 < |\eta| < 2.5 \) are selected. The minimum transverse energy/momentum requirements for both

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

| 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 |
| $Z/\gamma^* \rightarrow \mu^+\mu^- + \geq 1$jet | 61,447 | $(1\mu^+1\mu^-), 70 < m_{\mu\mu} < 110$ GeV, $p_T^{\mu} > 20$ GeV, $|\eta^{\mu}| < 2.4$, $p_T^{\mu} > 30$ GeV, | $|y^{\mu}| < 2.4$, $\Delta R(j, l) > 0.5$ | 0.428 ± 0.029 |
| $W^+ \rightarrow \mu^+\nu$ | 3,631,170 | $(1\mu^+), p_T^\mu > 25$ GeV, $|\eta^\mu| < 2.4$, $p_T^\mu > 25$ GeV, $m_{\tau} > 40$ GeV | 0.593 ± 0.017 |
| $W^- \rightarrow \mu^-\nu$ | 2,629,480 | $(1\mu^-), p_T^\mu > 25$ GeV, $|\eta^\mu| < 2.4$, $p_T^\mu > 25$ GeV, $m_{\tau} > 40$ GeV | 0.611 ± 0.018 |
| $t\bar{t} \rightarrow \mu^+e^+\nu\bar{\nu}b\bar{b}$ | 1495 | $1\mu^+e^+\nu, p_T^{\mu} > 20$ GeV, $|\eta^{\mu}|, |\eta^{\nu}| < 2.4$, | $|y^{\mu}|, |y^{\nu}| < 2.4$, | 0.177 ± 0.012 |
| $W^+Z \rightarrow l^+\nu l^-\nu$ | 97 | $(e^+e^-\mu^+\mu^-),(\mu^+\mu^-e^+e^-),(\mu^+\mu^-\mu^+\mu^-), p_T^\mu > 25$ GeV, $|\eta^{\mu}| < 2.5, 80 < m_{ll} < 100$ GeV, $m_{\tau} > 40$ GeV | 0.363 ± 0.011 |

leptons is 25 GeV and the minimum $E_T^{miss}$ requirement is 40 GeV. Moreover, it is required that the candidate events contain at least two reconstructed jets with $p_T > 40$ GeV within $|y| < 2.4$. At least one of the reconstructed jets on detector level in the event has to be b-tagged. The number of candidate events passing this selection is 1,495. The requirements imposed at particle level are significantly loosened, e.g. no cuts on the neutrinos or jets are applied (see table 3).

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

4.2 Background estimations

The contribution of background processes other than multijet processes is estimated using fully simulated MC samples detailed in table 2. Each of the signal selections is applied to those samples and the corresponding yields are evaluated and weighted by the corresponding cross section of the

---

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.
processes and data luminosity. Following previous analyses, we assume a conservative uncertainty of 5% in the cross sections of all relevant background processes. The contributions of those background processes that are known to have only a small impact in the signal region, i.e. below the uncertainty of the largest background contribution, are neglected.

Multijet backgrounds, as well as backgrounds involving non-prompt leptons or jets that are wrongly identified as leptons, are estimated in data. A so called ABCD method is used where two orthogonal properties of events, separating signal from multijet background processes, are used to define four regions in phase space, of which one region (A) is the signal region. The events in regions B and C pass one signal selection criterion, but fail the second, while events in region D fail both signal selection requirements. Signal contributions, as well as contributions from background processes that have been determined via full MC simulations in the regions B, C 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 $E_T^{\text{miss}}$, 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. The difference of the resulting correction factor after a particular variation $i$, $C'_i$, to the nominal $C$ factor, is then taken as systematic $\Delta C = C'_i - C$. When applicable, these differences are symmetrized for up- and down-variations. The systematic uncertainties due to pile-up are estimated by comparing the selection with and without the $\rho$-parameter reweighting. All relevant sources of the uncertainties are treated as independent from each other and hence the total systematic uncertainty on $\Delta C_{\text{tot}}$ is given by the Gaussian sum of the individual uncertainties $\Delta C_i$. The systematic uncertainties in $C$
Table 4. Overview of the relative contribution of background processes to the signal region.

| Process | $Z \rightarrow \tau \tau$ | $Z \rightarrow \mu \mu$ | $t \bar{t} \rightarrow 2V2b$ | $t \bar{t} \rightarrow 1V2b2q$ | 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% |
| $Z/\gamma^* \rightarrow \mu^+\mu^- + \geq 1$ 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% |
| $t \bar{t} \rightarrow \mu^\pm e^\mp \nu\bar{\nu}$ | 0.8% | — | 4.0% | <0.1% | <0.1% |
| $W^\pm Z \rightarrow l^\pm v l^\mp (l = e, \mu)$ | — | — | — | 8.1% | 1% |

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.

| 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% |
| $Z/\gamma^* \rightarrow \mu^+\mu^- + \geq 1$ 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% |
| $t \bar{t} \rightarrow \mu^+ e^- v \bar{b}b$ | 1.5% | 0.2% | 1.0% | 0.2% | 5.2% | 1.4% | 3% | 1.9% | 6.7% |
| $W^\pm Z \rightarrow l^\pm v l^\mp (l = e, \mu)$ | 1.3% | 0.2% | 1.5% | 0.3% | — | 1.3% | — | 1.8% | 3.0% |

range between 3% and 10% and dominate over the statistical uncertainties due to the size of the MC samples. An overview of the uncertainty breakdown on the $C$ factors for all validation samples is given in table 5.

Numerous control distributions between data and MC for all seven processes have been validated. Good agreement between data and MC has been observed in all of them. The normalized invariant mass and lepton rapidity distributions for the Drell-Yan processes have already been discussed in section 3. Two selected jet distributions of the $Z$+jets study are shown as an example in figure 6, where good agreement of the MC prediction and the data can be observed. The measurement of the $W$ boson production cross section is able to test the description of $E_{T}^{miss}$, hence figure 7 displays the comparison of data and MC for $E_{T}^{miss}$ and $m_T$. Similarly, figure 8 shows the comparison of the leading jet $p_T$ as well as $E_{T}^{miss}$ for the $t \bar{t}$ selection, with a similar conclusion. The study of $WZ$ diboson production is statistically limited. Control distributions of the invariant and transverse mass observables are shown in figure 9, where a good agreement within the limited statistical precision is observed.
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.

Figure 7. Normalized $m_T$ (left) and $E_T^{\text{miss}}$ 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.

4.4 Results and comparisons

The fiducial production cross sections for the seven validation processes are determined via eq. 4.2, using the detector correction factors (table 3) as well as the estimated background contributions
Figure 8. Normalized leading jet $p_T$ (left) and $E_T^{\text{miss}}$ 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.

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.

The resulting cross sections are summarized in table 6 and figure 10, together with their statistical and systematic uncertainties, where the latter are separated into detector-related and luminosity uncertainties. The results are also compared to previously published results from the CMS and ATLAS Collaborations, depending on which fiducial phase-space regions are closer to our own choices [2, 51, 54–56]. The $t\bar{t}$ and $WZ$ cross sections are compared to the published inclusive production cross sections. Apart from the diboson $WZ$ process, all cross section measurements are dominated by systematic uncertainties due to the detector effects. The uncertainties due the background processes are small for all channels, except the $W^\pm$ boson production, where the uncertainty in the multijet background is of a similar size as the uncertainties due to the missing transverse energy requirements. It should be noted that the uncertainties for the $Z/\gamma^* \rightarrow l^+l^-$ processes in our analysis appear to be smaller than the official measurements by CMS [2], however,

---

*No model uncertainties were considered when extrapolating from our fiducial cross section to the inclusive cross section.*
Table 6. Overview of measured cross sections of seven validation processes as well as previously published results together with theory predictions.

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

This is due to the smaller data-set used by CMS as well as the limited number of significant digits in the published result. The smaller systematic uncertainties in the diboson $WZ$ cross section measurement is due to the smaller lepton reconstruction uncertainties assumed in our analysis and motivated in section 3. We also compare the inclusive cross section to the available theoretical predictions, which have been previously published. Figure 11 shows the ratio of the theory predictions to our measured fiducial cross sections and to the previously published results by either ATLAS or CMS. All our measurements are in good agreement with the previously published results as well as the SM predictions. The measurement systematic uncertainties are between 1.6 and 6.7%.

5 Opportunities and challenges of the CMS Open Data initiative

The CMS Open Data initiative offers a unique opportunity to study and measure properties of the SM as long as a limited precision is sufficient. Measurements with higher precision currently seem not achievable, given the limited available information on the detector calibration as well as the systematic uncertainties of relevant observables. Clearly, these calibration efforts are one of the main areas of research within the Collaborations and the publications of the corresponding information in an easily accessible and understandable format for external physicists is highly challenging. One example is that experimental uncertainties in the energy scale of particle jets involve dozens of calibration parameters which have to be applied correctly. Another example are uncertainties in the lepton identification efficiencies, which are correlated in a complex manner, where the correlations matter for high precision measurements. Hence, precision measurements of (differential) cross sections as well as cross section ratios should, and can only, be performed by the LHC Collaborations.
As a first possible improvement of the CMS Open Data initiative, we suggest that some simplified baseline calibrations, as well as uncertainties, should become available. This is highly desirable to make legacy data analyses possible to confront future theories with data. As a second possible improvement of the CMS Open Data initiative, we suggest that the CMS Collaboration publishes dedicated baseline analyses, such as presented in section 4. This would allow external physicists to understand the technical details, e.g. how to apply the detector calibration or simply how to calculate the integrated luminosity for a given data sample.

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

**Acknowledgments**

We would like to thank the CMS Collaboration for providing the full 2012 data set as well as for the documentation on the CMS detector performance. This work would have not been possible without the excellent performance of the LHC as well as the existing computing infrastructure and the support from CERN. We would also like to thank Guillelmo Gomez-Ceballos and Frank Fiedler for the helpful comments during the revision of this paper. M.S. would like to thank in addition the Fulbright commission as well as the Volkswagen Foundation for the support of this work. Moreover, he would like to thank his colleagues at MIT, Philip Harris in particular, for answering all questions regarding the treatment of the CMS Open Data for this project as well as the pleasant environment during the Fulbright research scholarship.
References

[1] ATLAS collaboration, Precision measurement and interpretation of inclusive $W^+$, $W^-$ and Z/$\gamma^*$ production cross sections with the ATLAS detector, Eur. Phys. J. C 77 (2017) 367 [arXiv:1612.03016].

[2] CMS collaboration, Measurement of inclusive $W$ and $Z$ boson production cross sections in pp collisions at $\sqrt{s} = 8$ TeV, Phys. Rev. Lett. 112 (2014) 191802 [arXiv:1402.0923].

[3] CMS collaboration, Measurement of the $t\bar{t}$ production cross section in the dilepton channel in pp collisions at $\sqrt{s} = 8$ TeV, JHEP 02 (2014) 024 [Erratum ibid. 02 (2014) 102] [arXiv:1312.7582].

[4] ATLAS collaboration, Measurement of the angular coefficients in $Z$-boson events using electron and muon pairs from data taken at $\sqrt{s} = 8$ TeV with the ATLAS detector, JHEP 08 (2016) 159 [arXiv:1606.00689].

[5] ATLAS collaboration, Measurement of the transverse momentum and $\phi_{\eta}$ distributions of Drell-Yan lepton pairs in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector, Eur. Phys. J. C 76 (2016) 291 [arXiv:1512.02192].

[6] CMS collaboration, Software framework for CMS Open Data analysis, http://opendata.cern.ch/docs/about-cms.

[7] C. Cesarotti, Y. Soreq, M.J. Strassler, J. Thaler and W. Xue, Searching in CMS Open Data for dimuon resonances with substantial transverse momentum, Phys. Rev. D 100 (2019) 015021 [arXiv:1902.04222].

[8] A. Tripathee, W. Xue, A. Larkoski, S. Marzani and J. Thaler, Jet substructure studies with CMS Open Data, Phys. Rev. D 96 (2017) 074003 [arXiv:1704.05842].

[9] CMS collaboration, The CMS experiment at the CERN LHC, 2008 JINST 3 S08004.

[10] CMS collaboration, Particle-flow event reconstruction in CMS and performance for jets, taus and MET, CMS-PAS-PFT-09-001, CERN, Geneva, Switzerland (2009).

[11] W. Adam, R. Fruehwirth, A. Strandlie and T. Todor, Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC, CMS-NOTE-2005-001, CERN, Geneva, Switzerland (2005).

[12] CMS collaboration, Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV, 2013 JINST 8 P09009 [arXiv:1306.2016].

[13] CMS collaboration, Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV, 2015 JINST 10 P06005 [arXiv:1502.02701].

[14] CMS collaboration, Performance of muon identification in pp collisions at $\sqrt{s} = 7$ TeV, CMS-PAS-MUO-10-002, CERN, Geneva, Switzerland (2010).

[15] CMS collaboration, Commissioning of the particle-flow event reconstruction with the first LHC collisions recorded in the CMS detector, CMS-PAS-PFT-10-001, CERN, Geneva, Switzerland (2010).

[16] CMS collaboration, Determination of jet energy calibration and transverse momentum resolution in CMS, 2011 JINST 6 P11002 [arXiv:1107.4277].

[17] CMS collaboration, Identification of b-quark jets with the CMS experiment, 2013 JINST 8 P04013 [arXiv:1211.4462].

[18] C.V. Philip Harris et al., Bacon analysis framework, https://github.com/ksung25/BaconProd/tree/Run1.
[19] A. Hinzmann, Tools for physics analysis in CMS, J. Phys. Conf. Ser. 331 (2011) 032042.
[20] CMS collaboration, CMS list of validated runs for primary datasets of 2012 data taking, CERN Open Data portal, (2017).
[21] R. Brun and F. Rademakers, ROOT: an object oriented data analysis framework, Nucl. Instrum. Meth. A 389 (1997) 81.
[22] CMS collaboration, Single mu primary dataset in aod format from run of 2012 (/singlemu/run2012c-22jan2013-v1/aod), CERN Open Data portal, http://opendata.cern.ch.
[23] CMS collaboration, Single electron primary dataset in aod format from run of 2012 (/singleelelectron/run2012c-22jan2013-v1/aod), CERN Open Data portal, http://opendata.cern.ch.
[24] CMS collaboration, CMS luminosity based on pixel cluster counting — Summer 2013 update, CMS-PAS-LUM-13-001, CERN, Geneva, Switzerland (2013).
[25] 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 [arXiv:1002.2581].
[26] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, MadGraph 5: going beyond, JHEP 06 (2011) 128 [arXiv:1106.0522].
[27] J. Gao et al., CT10 next-to-next-to-leading order global analysis of QCD, Phys. Rev. D 89 (2014) 033009 [arXiv:1302.6246].
[28] CMS collaboration, Study of the underlying event at forward rapidity in pp collisions at √s = 2.76 and 7 TeV, JHEP 04 (2013) 072 [arXiv:1302.2394].
[29] N. Davidson, G. Nanava, T. Przedzinski, E. Richter-Was and Z. Was, Universal interface of TAUOLA technical and physics documentation, Comput. Phys. Commun. 183 (2012) 821 [arXiv:1002.0543].
[30] GEANT4 collaboration, GEANT4: a simulation toolkit, Nucl. Instrum. Meth. A 506 (2003) 250.
[31] CMS collaboration, Event generator tunes obtained from underlying event and multiparton scattering measurements, Eur. Phys. J. C 76 (2016) 155 [arXiv:1512.00815].
...