



# **Validation of 2011 CMS Open Data: measurement of associated W + charm production**

Oleksandr Kot<sup>a</sup>

*Supervised by:* Oleksandr Zenaiev<sup>b</sup>

<sup>a</sup> Taras Shevchenko National University of Kyiv, Ukraine

<sup>b</sup> Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany

September 8, 2016

## **Abstract**

The CMS Open Data 2011 were successfully validated by reproducing the measurement of the associated production of a W boson and a charmquark jet (W + c) in pp collisions at a center-of-mass energy of 7 TeV. The analysis is conducted with a data sample corresponding to a total integrated luminosity of  $2.5 \text{ fb}^{-1}$ , which equals one half of data collected by the CMS detector at the LHC in 2011. Measured cross sections of W + c associated production are consistent with cross sections obtained by the CMS Collaboration.

## Contents

1.Introduction.....	3
1.1 CMS Open Data.....	3
1.2 Theory .....	4
2.CMS detector.....	4
3.Measurement.....	5
3.1 Analysis strategy.....	5
3.2. Data samples and signal definition.....	6
3.3. Event selection.....	6
3.3.1. Lifetime tagging.....	7
3.3.2. Selection of exclusive $D^\pm$ decays.....	8
3.3.3. Selection of exclusive $D^{*\pm}$ decays.....	9
3.3.4. Selection of semileptonic charm decays.....	10
3.3.5. Characterization of W + c kinematics.....	11
3.4. Measurement of the W + c cross section.....	12
4. Conclusions.....	13
5. Acknowledgement.....	13
6.References.....	13

# 1. Introduction

## 1.1 CMS Open Data

This year the CMS collaboration has made 300 TB of high-quality data from the LHC available to the public through the CERN Open Data Portal[1]. These include 2.5 inverse femtobarns ( $\text{fb}^{-1}$ ) of data from proton collisions at 7 TeV making up half the data collected at the LHC by the CMS detector in 2011. This follows a previous release from November 2014, which made available around 27 TB of research data collected in 2010.

The collision data come in two types. The so-called “primary datasets” are in the same format used by the CMS Collaboration to perform research. The “derived datasets” on the other hand require a lot less computing power and can be readily analysed by university or even high-school students, and CMS has provided a limited number of datasets in this format.

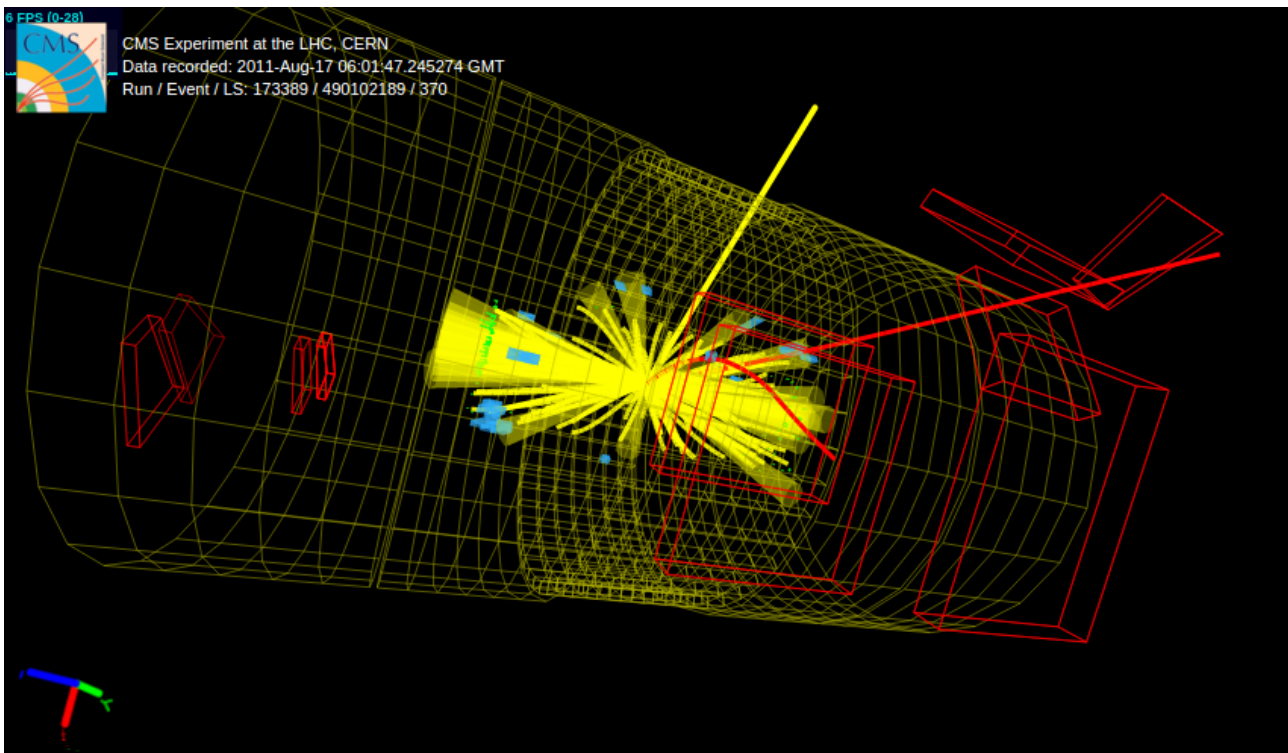


Figure 1 : Visualised event from Open Data Sample 2011 on the website <http://opendata.cern.ch/visualise/events/CMS>

Notably, CMS is also providing the simulated data generated with the same software version that should be used to analyse the primary datasets. Simulations play the crucial role in particle-physics research and CMS is also making available the protocols for generating the simulations that are provided. The data release is accompanied by analysis tools and code examples tailored to the datasets. These data are being made public in accordance with CMS’s commitment to a long-term data preservation and as part of the collaboration’s open-data policy.

A virtual-machine image based on CernVM, which comes preloaded with the software environment needed to analyse the CMS data, can also be downloaded from the portal[1,2]. Visualised event from CMS Open Data 2011 SingleMu[3] data sample is shown in Figure 1.

The goal of the current work is to validate the 2011 CMS Open Data by reproducing results for comparison with the published article of the CMS collaboration[4] using a subset of the same data.

## 1.2 Theory

The study of associated production of a W boson and a charm (c) quark at hadron colliders (hereafter referred to as W + c production) provides direct access to the strange-quark content of the proton at an energy scale of the order of the W-boson mass ( $Q^2 \sim (100 \text{ GeV})^2$ ). As processes  $sg \rightarrow W^- + c$  and  $\bar{s}g \rightarrow W^+ + \bar{c}$  dominate at the hard-scattering level (see Figure 2), precise measurements of these processes at the Large Hadron Collider (LHC) may significantly reduce the uncertainties in the strange quark and antiquark parton distribution functions (PDFs) and help resolve existing ambiguities and limitations of low-energy neutrino deep-inelastic scattering (DIS) data. More precise knowledge of the PDFs is essential for many present and future precision analyses, such as the measurement of the W-boson mass. W + c production receives contributions at a few percent level from the processes  $d\bar{g} \rightarrow W^- + c$  and  $\bar{d}g \rightarrow W^+ + \bar{c}$ , which are Cabibbo suppressed. Overall, the  $W^- + c$  yield is expected to be slightly larger than the  $W^+ + \bar{c}$  yield at the LHC because of the participation of down valence quarks in the initial state. A key property of the  $qg \rightarrow W + c$  reaction is the presence of a charm quark and a W boson with opposite-sign charges [4].



Figure 2: Main diagrams at the hard-scattering level for associated W + c production at the LHC.

## 2. CMS detector

CMS is a multipurpose particle detector designed to see a wide range of particles and phenomena produced in high energy pp-collisions at the LHC. The central feature of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T surround the interaction point. Within the field volume are a silicon pixel and strip tracker, an electromagnetic calorimeter (ECAL), and a scintillator hadron calorimeter (HCAL). The superconducting magnet is surrounded by muon system, where muons are detected in gas-ionization detectors. Figure 3 shows slice of CMS.

The coordinate system adopted by CMS has the origin centered at the nominal collision point inside the experiment, the y-axis pointing vertically upward, and the x-axis pointing radially inward toward the center of the LHC. Thus, the z-axis points along the beam direction, the azimuthal angle  $\phi$  is measured from the x-axis in the x-y plane and the radial coordinate in this plane is denoted by r. The polar angle  $\theta$  is measured from the z-axis. Pseudorapidity is defined as  $\eta = -\ln(\tan(\theta/2))$ . Thus, the momentum and energy transverse to the beam direction, denoted by  $p_T$  and  $E_T$ , respectively, are computed from the x and y components. The imbalance of energy measured in the transverse plane is denoted by  $E_T^{miss}$ .

The tracker measures charged-particle trajectories in the pseudorapidity range  $|\eta| \leq 2.5$ . It consists of 1440 silicon pixel and 15 148 silicon strip detector modules. It provides an impact parameter resolution of 15  $\mu\text{m}$  and a  $p_T$  resolution of about 1% for charged particles with  $p_T$

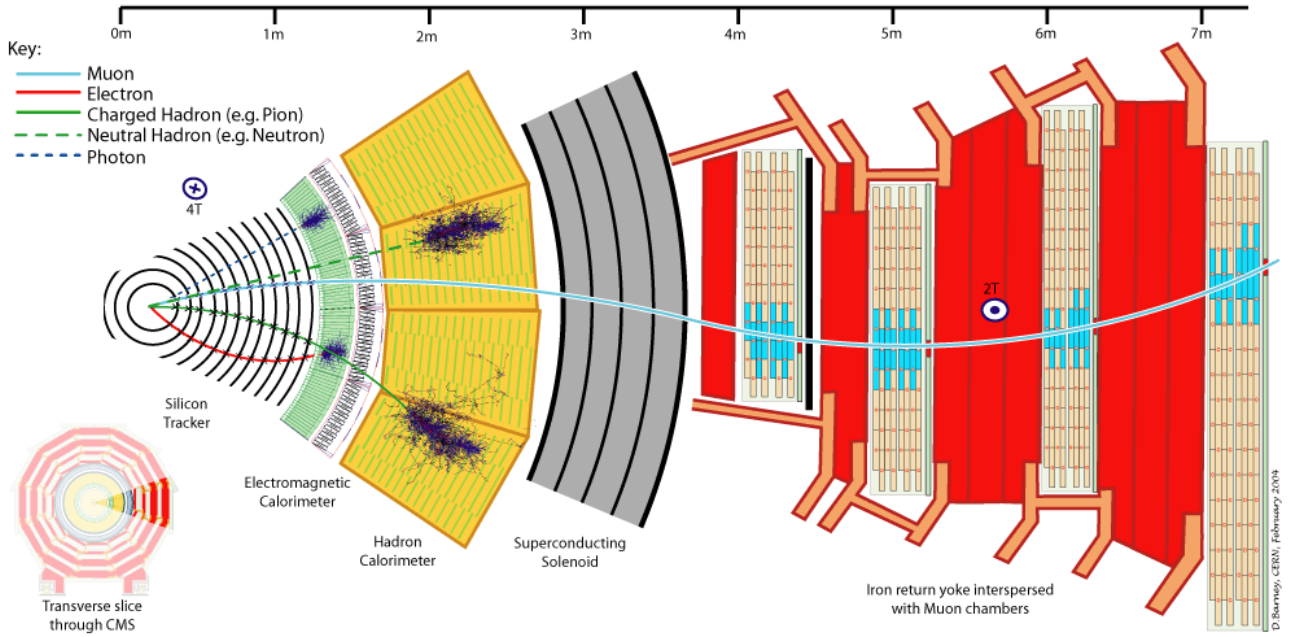


Figure 3: A transverse slice of the CMS detector

around 40 GeV. The ECAL consists of nearly 76 000 lead tungstate crystals, with coverage in pseudorapidity  $|\eta| \leq 1.479$  in a cylindrical barrel region and  $1.479 \leq |\eta| \leq 3.0$  in two endcap regions (EE). The ECAL is surrounded by the HCAL which is a sampling device with brass as passive material and scintillator as active material with coverage up to  $|\eta| < 3.0$ . Muons are detected in the pseudorapidity range  $|\eta| \leq 2.4$ , with detection planes based on three technologies: drift tubes, cathode strip chambers, and resistive-plate chambers [4,5].

### 3. Measurement

#### 3.1 Analysis strategy

We are looking for  $W + c$  associated production in final state containing a  $W$ -boson which decays to a lepton ( $e$  or  $\mu$ ) and corresponding neutrino and a leading jet with charm-quark content. Jets originating from a  $c$  ( $\bar{c}$ ) parton are identified using one of the three following signatures: a displaced secondary vertex with three tracks and an invariant mass consistent with  $D^+ \rightarrow K^- \pi^+ \pi^+$  ( $D^- \rightarrow K^+ \pi^- \pi^-$ ) decay; a displaced secondary vertex with two tracks consistent with a  $D^0 \rightarrow K^- \pi^+$  ( $\bar{D}^0 \rightarrow K^+ \pi^-$ ) decay and associated with a previous  $D^{*+} \rightarrow D^0 \pi_s^+$  ( $D^{*-} \rightarrow \bar{D}^0 \pi_s^-$ ) decay at the primary vertex; or a semileptonic decay leading to a well-identified muon inside the jet. In total, since both electron and muon channels are considered in the  $W$ -boson decay, six different final states are explored.

The  $D^\pm$ ,  $D^{*\pm}$  (2010), and  $c \rightarrow lv + X$  decays provide a direct measurement of the charm-quark jet charge, which is a powerful tool to disentangle the  $W + c$  signal component from most of the background processes. We define two types of distributions: opposite-sign distributions, denoted by OS, are built on samples containing a  $W$  boson and a charm-quark jet with an opposite-charge sign; same-sign distributions, denoted by SS, are built from samples where the  $W$  boson and the charm-quark jet have the same charge sign. The final distributions used in the analysis are obtained by subtracting the SS distribution from the OS distribution (referred to as

OS – SS) for any given variable. This subtraction has no effect on the signal at leading order. In contrast,  $W+c\bar{c}$  and  $W+b\bar{b}$  events provide the same OS and SS contributions and are suppressed in OS – SS distributions. Moreover, any OS – SS asymmetry present in  $t\bar{t}$ , single-top-quark, or  $W$  + light-quark jet backgrounds is found to be negligible according to simulations[4]. As a consequence, OS – SS distributions are largely dominated by the  $W + c$  component, allowing for many detailed studies of the  $pp \rightarrow W+c+X$  process. Using displaced secondary vertices is a simple way to suppress backgrounds, such as Drell–Yan events,  $W$  + light-quark jet, and multijet final states with no heavy-flavour content. It also reduces backgrounds containing b-hadron decays, which often lead to secondary vertices with a higher track multiplicity than a typical D-meson decay [4].

### 3.2. Data samples and signal definition

This analysis was performed with a half of the data sample of proton-proton collisions at  $\sqrt{s}=7$  TeV collected with the CMS detector in 2011. A data set available for the analysis corresponds to an integrated luminosity  $L = 2.5 \text{ fb}^{-1}$ . As we are looking for muons and electrons originating from W-boson we use data samples triggered by muon or electrons triggers. Candidate events for the muon decay channel of the W boson are selected online by one of a single-muon trigger or single-electron triggers. These events are available in single-muon[3] and single-electron[6] data samples at CMS Open Data Portal.

At the hard-scattering level we identify  $W+c$  signal events as those containing an odd number of charm partons in the final state. This choice provides a simple operational definition of the process and ensures that pure QCD splittings of the  $g \rightarrow c\bar{c}$  type are associated with the background[4].

### 3.3. Event selection

The leptonic decay of a W boson into a muon or an electron, and a neutrino is characterized by the presence of a high-transverse momentum, isolated lepton. The neutrino escapes detection causing an apparent imbalance in the transverse energy of the event. Experimentally, the magnitude of the vector momentum imbalance in the plane perpendicular to the beam direction defines the missing transverse energy of an event,  $E_T^{\text{miss}}$ . In W-boson events, this variable is an estimator of the transverse energy of the undetected neutrino.

In order to have high quality lepton tracks we consider only those which are highly isolated, are in pseudorapidity range  $|\eta| < 2.1$  and have  $p_T^\mu > 25$  GeV for muons and  $p_T^e > 35$  GeV for electrons. The background arising from Drell–Yan processes is reduced by removing events containing additional muons (electrons) with  $p_T > 25$  (20) GeV in the pseudorapidity region  $|\eta_\mu| < 2.4$  ( $|\eta_e| < 2.5$ ). By requirement of high transverse mass, which is built from the transverse momentum of the isolated lepton, and the missing transverse energy in the event, we reduce background from events where lepton does not come from W-boson decay. Further details can be found in the paper [4].

A  $W + \text{jet}$  sample is selected by demanding the presence of at least one jet with  $p_T > 25$  GeV in the pseudorapidity range  $|\eta_{\text{jet}}| < 2.5$ , thus ensuring that the jet passes through the tracker volume, and hence achieving the best possible jet  $p_T$  resolution. A  $W + c$  candidate sample is further selected by searching for a distinct signature of a charmed particle decay among the constituents of the leading jet associated with the W boson [4]. The transverse mass was reconstructed from momentum of muon and missing energy is consistent with mass of W-boson

(see Figure 4).

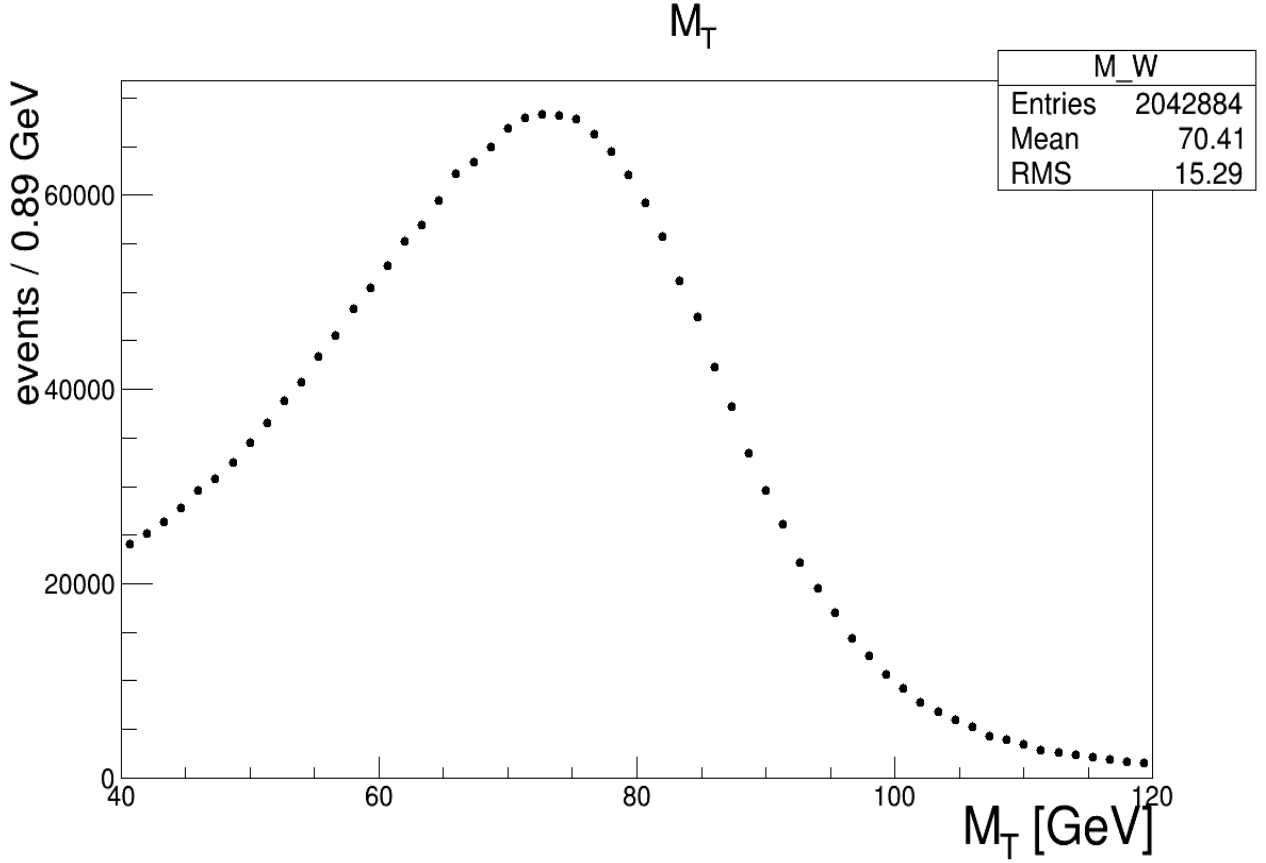


Figure 4: Transverse mass of  $\mu + E_T^{miss}$

### 3.3.1. Lifetime tagging

To employ the long lifetime of  $D^\pm$  and  $D^0$  mesons we reconstruct a secondary vertex, which is point where the D-meson decays. The location of the secondary vertex allows us to calculate the decay length ( $\vec{dl}$ ) considering the beamspot (the beamspot is the luminous region produced by the collisions of proton beams) as the primary vertex. In practice it is more expedient to use projected decay length, which equal transversal component of scalar product of decay length on particle momentum divided into transversal particle momentum (see Figure 5, Expression 1). This quantity indicates whether the vectors of reconstructed particle momentum and decay length are codirectional or not. The opposite directionality of momentum and decay length means that reconstructed candidate is likely wrong and is rejected. The same directionality of momentum and  $\vec{dl}$  vectors indicates that reconstructed candidate is likely the particle we are looking for. The second important parameter is decay length significance. It is defined by relation of the projected decay length and its uncertainty. In analysis we require positive projected decay length less than 2 cm and decay length significant more than 3 for both  $D^\pm$  and  $D^0$  candidates.

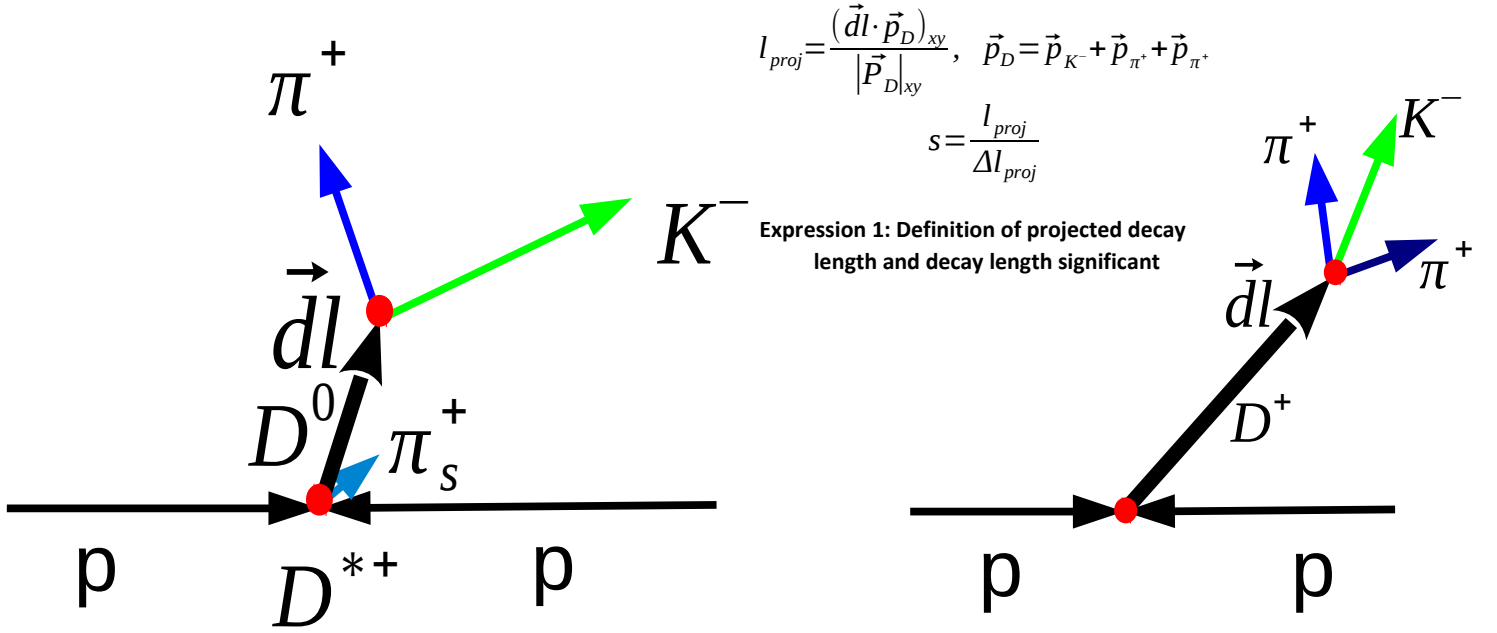


Figure 5: To explanation of decay length and decay length significant.

### 3.3.2. Selection of exclusive $D^\pm$ decays

We identify  $D^\pm \rightarrow K^\mp \pi^\pm \pi^\pm$  decays in the selected W + jets sample using secondary vertices with three tracks and a reconstructed invariant mass within 50 MeV of the  $D^\pm$  mass,  $1869.5 \pm 0.4$  MeV. The kaon mass is assigned to the track that has opposite sign to the total charge of the three-prong vertex and the remaining tracks are assumed to have the mass of a charged pion.

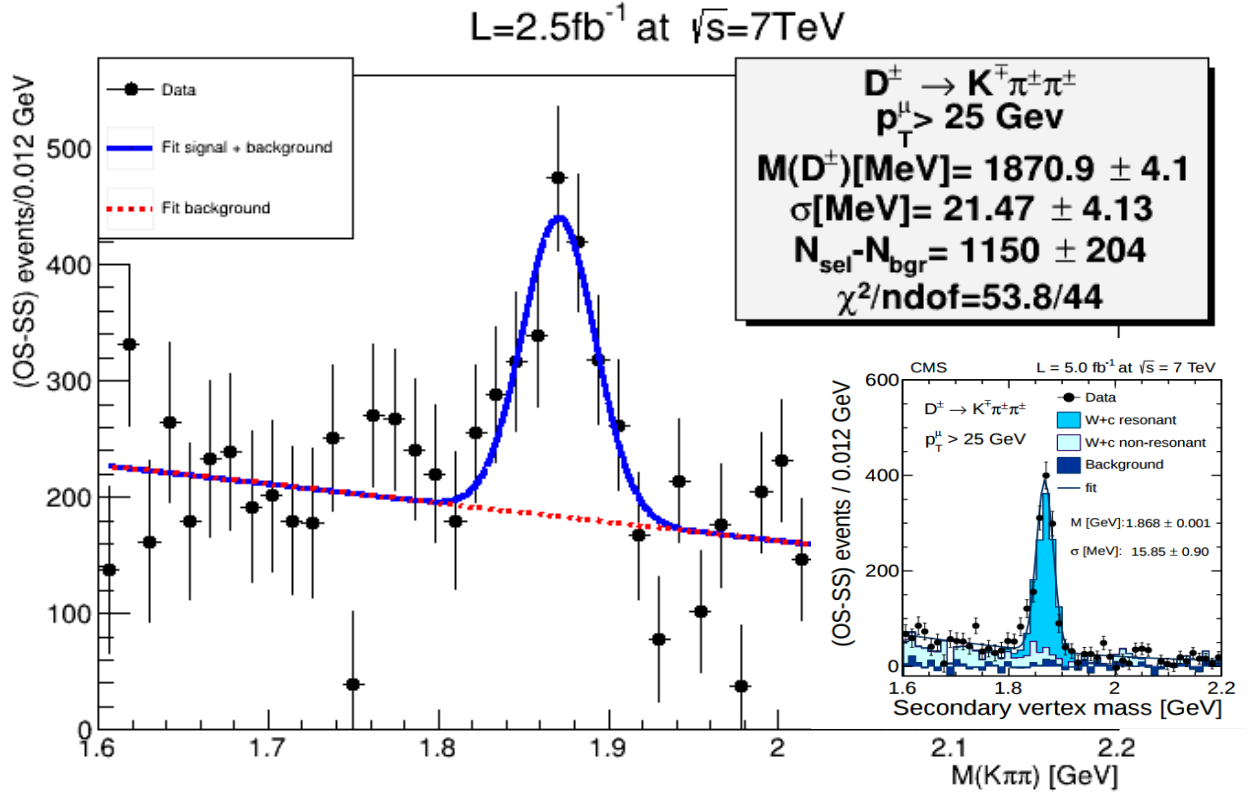


Figure 6: The invariant mass of three-prong secondary vertices in data after substruction of the SS component (big in the center) and the corresponding distribution from the paper [4] (small in right bottom corner)



Here and next paragraphs we'll consider events when W-boson decays into muon channel only. Figure 6 shows the OS – SS distributions of the reconstructed invariant mass for  $D^\pm$  candidates associated with  $W \rightarrow \mu\nu$  decay . The distribution is fitted by the sum of a Gaussian function for signal and a second degree polynomial for background. The obtained mass of  $D^+$  meson is in a good agreement with the PDG[7].

### 3.3.3. Selection of exclusive $D^{*\pm}$ decays

To identify  $D^{*\pm}$  firstly it is necessary to find opposite-charged kaon and pion originating and reconstruct the secondary vertex consistent with the mass of  $D^0$ . This two-track system is combined with a primary track having  $p_T > 0.3$  GeV found in a cone of  $\Delta R = 0.1$  around the direction of the  $D^0$  candidate momentum,  $\Delta R = \sqrt{(\eta_{D^0} - \eta_{\pi_s})^2 + (\phi_{D^0} - \phi_{\pi_s})^2}$ . The secondary track with charge opposite to the charge of the primary track is assumed to be the kaon in the  $D^0$  decay. Only combinations with a reconstructed mass differing from the  $D^0$  mass ( $1864.86 \pm 0.13$  MeV) by less than 70 MeV are kept. The  $D^{*\pm}$  signal is identified as a peak in the distribution of the difference between the reconstructed  $D^{*\pm}$  and  $D^0$  masses near the expected value,  $m_{\text{rec}}(D^{*\pm}) - m_{\text{rec}}(D^0) = 145.421 \pm 0.010$  MeV. The OS – SS distribution of the reconstructed mass difference  $m_{\text{rec}}(D^{*\pm}) - m_{\text{rec}}(D^0)$  is shown in Figure 7.

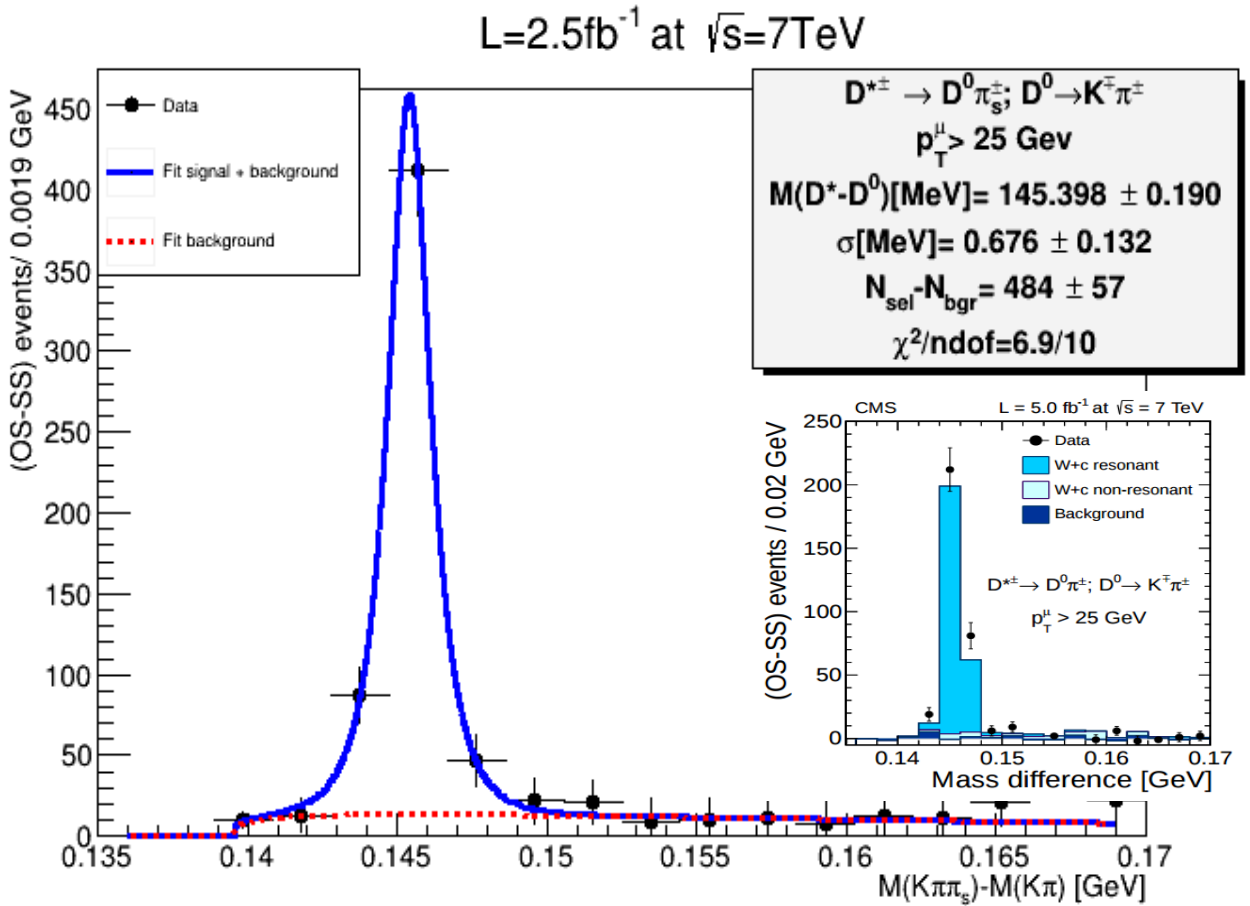


Figure 7: Distribution of the reconstructed mass difference between  $D^*$  and  $D^0$  candidates in the selected W + c sample, after subtraction of SS component (big in the center) and the corresponding distribution from the paper [4] (small in right bottom corner)

Distribution fitted by a sum of modified Gaussian function (Expression 2) for signal and step-like function for background (Expression 3). Obtained mass difference between  $D^{*\pm}$  and  $D^0$  mesons is consistent with PDG [7] values.

$$\frac{A}{\sqrt{(2\pi)}\sigma} \exp\left(-\frac{1}{2}\left(\frac{|x-\mu|}{\sigma}\right)^{1+\frac{1}{1+\frac{|x-\mu|}{2\sigma}}}\right)$$

Expression 2: Modified Gaussian function

$$B(x-m_\pi)^\alpha e^{-\beta x}$$

Expression 3: Step-like background function

### 3.3.4. Selection of semileptonic charm decays

In addition to the previous exclusive channels, we consider the identification of charm-quark jets via semileptonic decays of the  $c$  quark. Only jets containing semileptonic decays into muons are considered. Muons in jets are identified with the same criteria used for muon identification in  $W$ -boson decays, with the exception that the isolation requirements are not applied. Since the OS – SS strategy effectively suppresses all backgrounds except Drell–Yan processes, additional requirements are applied in order to reduce the Drell–Yan contamination to manageable levels without affecting the signal in an appreciable way. We require  $p_T^\mu < 25$  GeV,  $p_T^\mu / p_T^{\text{jet}} < 0.6$ , and  $p_T^{\text{rel}} < 2.5$  GeV, where  $p_T^\mu$  denotes here the transverse momentum of the muon identified require the invariant mass of the dilepton system to be above 12 GeV, in order to avoid the region of low-mass resonances. Finally, dimuon events with an invariant mass above 85 GeV are

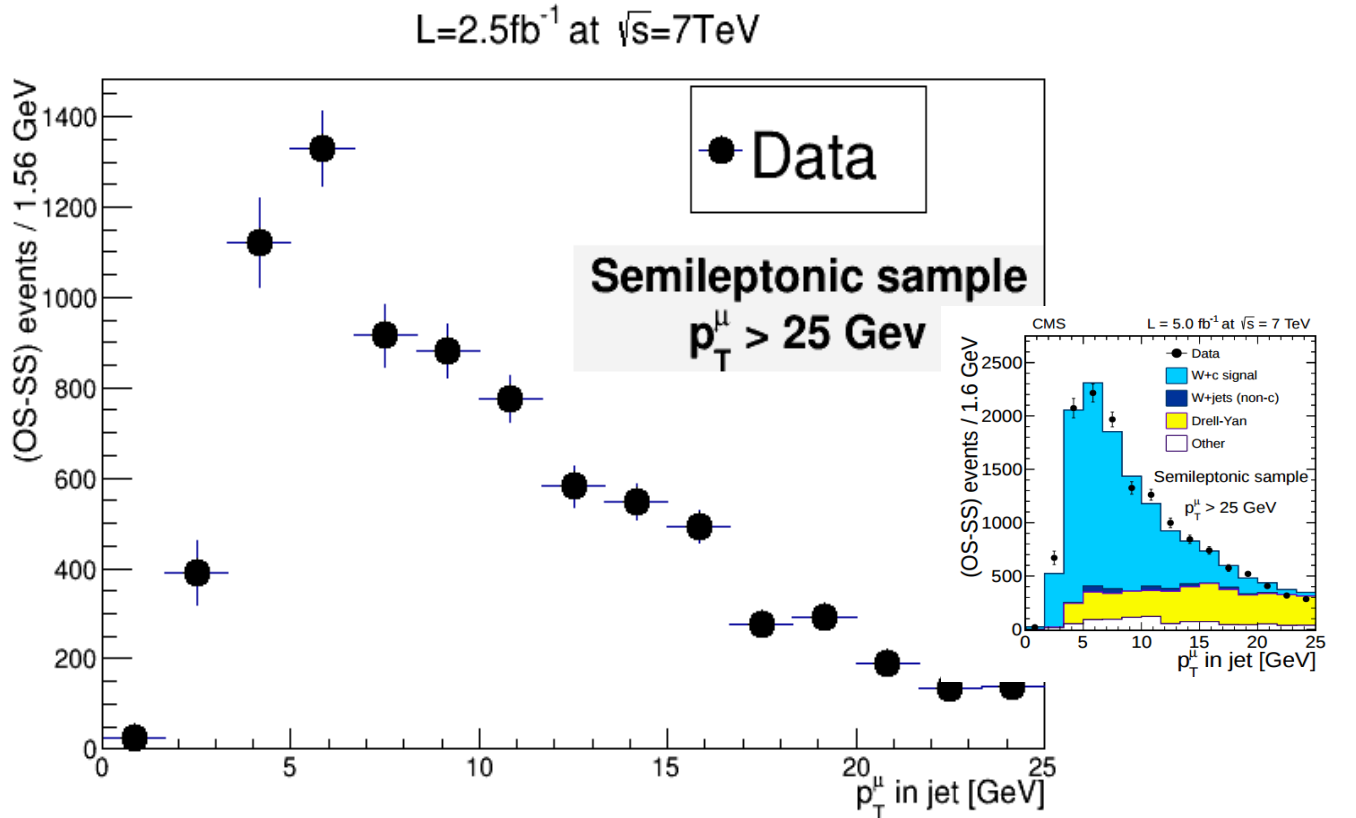


Figure 8: Distributions of the transverse momentum of the muon inside the leading jet of the event, after subtraction of the SS component (big in the center) and the corresponding distribution from the paper [4] (small in right corner)

rejected. The latter requirement is not applied to the sample with W-boson decays into electrons, which is minimally affected by high-mass dilepton contamination.

Figure 8 shows the resulting transverse momentum distribution of the selected muons inside the leading jet after the OS – SS subtraction procedure.

### 3.3.5. Characterization of W + c kinematics

Figure 9 shows the distributions of the jet pseudorapidity and the jet momentum fraction carried by the  $D^\pm$  candidates (top row of plots) and the  $D^{\pm*}$  candidates (middle row of plots), while the jet pseudorapidity and the jet momentum fraction carried by the muon is shown for the semileptonic candidates (bottom row of plots).

The obtained distributions for semileptonic decay of charm have similar shape and comparable statistics with distributions from paper[4], what testify similar event selections.

The observed difference in the distributions for  $D^\pm$  and  $D^{\pm*}$ -mesons indicates different secondary vertex reconstruction and selection of D-mesons.

In order to reduce background in final state with  $K^\mp \pi^\pm \pi^\pm (D^\pm)$  it must be worthwhile to do secondary vertex fitting without track weighting.

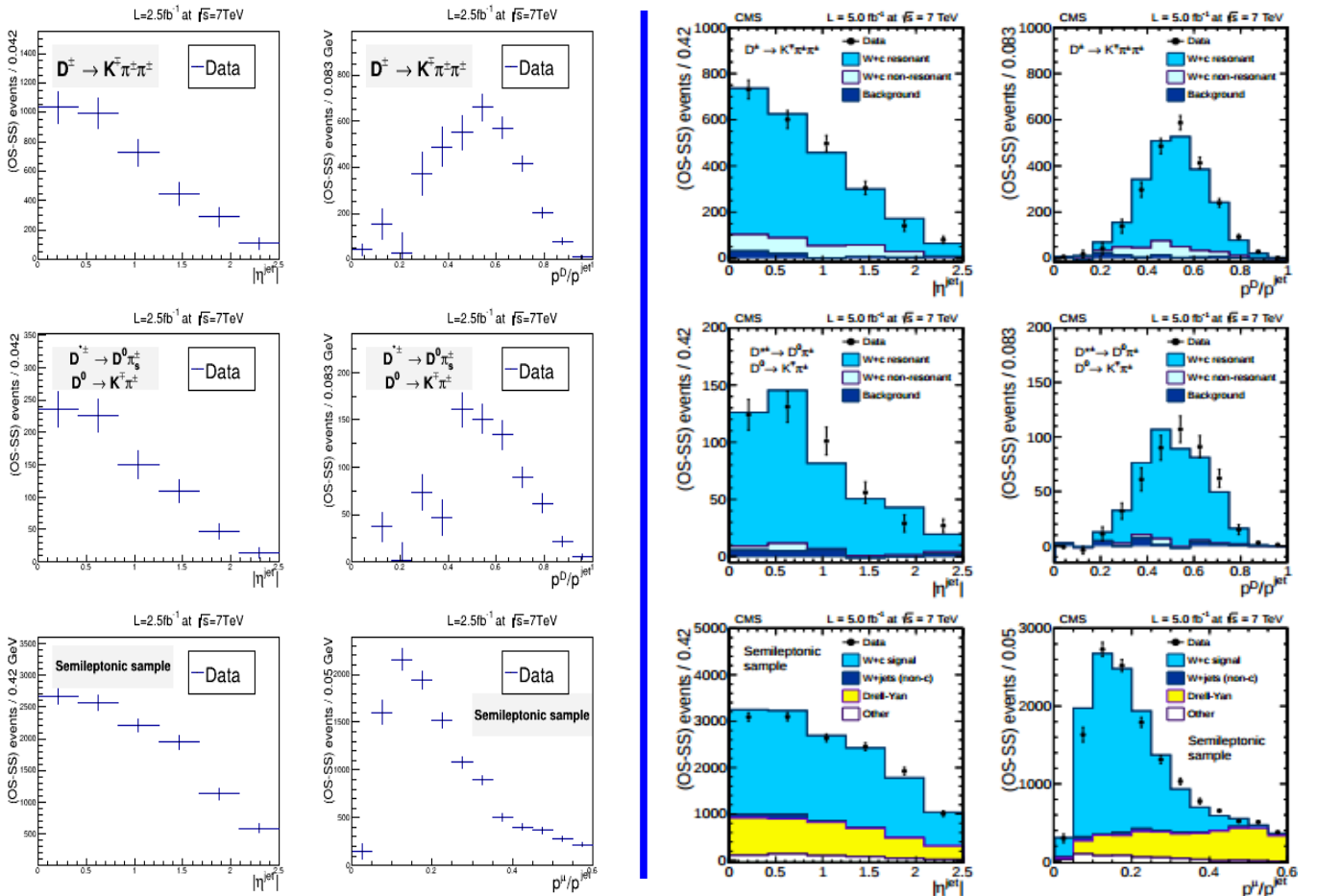


Figure 9: Distributions of W + c selected events in the different charm decay channels as a function of the jet pseudorapidity (left) and the jet momentum fraction (right) carried by the D meson or by the muon inside the jet. The top row corresponds to the  $D^\pm$  decay channel, the middle row corresponds to the  $D^{\pm*}$  decay channel, and the bottom row corresponds to semileptonic charm decays into muons. Histograms which are on the left side from blue bars belong to the present analysis, from right side to the paper [4].

### 3.4. Measurement of the W + c cross section

The measurement of the W + c cross section is finalised with two final states containing a well-identified  $W \rightarrow \mu\nu$  decay plus a leading jet with charm content. We use the exclusive  $D^\pm$  and  $D^{*\pm}$ (2010) samples, described before. The W + c cross section is determined in the fiducial region

$p_T^\mu > 25 \text{ GeV}$ ,  $|\eta^\mu| < 2.1$ ,  $p_T^{\text{jet}} > 25 \text{ GeV}$ ,  $|\eta^{\text{jet}}| < 2.5$  using the following expression:

$$\sigma(W+c) = \frac{N_{\text{peak}}}{L_{\text{int}} \cdot B(c \rightarrow \text{FinalState}) \cdot A} \quad \text{Expression 4: Cross section definition}$$

where  $N_{\text{peak}}$  is the number of OS – SS events in corresponding peak,  $L_{\text{int}}$  is the integrated luminosity,  $B$  is the relevant charm branching fraction and  $A$  is acceptance. It's determined using Monte Carlo (MC) simulations, generated by MadGraph + Pythia and equals number of reconstructed MC events (in peak) over number of generated MC events in the corresponding channel. For reconstructing of MC events we apply the same procedures as for data. The number of events in the peak is determined by the parameter  $A$  of Gaussian function (for modified Gaussian additionally multiplied by factor 1.2177) divided into bin width.

Using Expression 4 we calculate cross sections. Calculated values are given in Table 1. In Table 2 the results from the paper are given.

Calculation of the average cross section and its uncertainty in Table 1 was done with Expressions 5 [8].

$$\bar{x} = \frac{\sum_j \frac{x_j}{(\Delta x_j)^2}}{\sum_k \frac{1}{(\Delta x_k)^2}} \quad \Delta \bar{x} = \frac{1}{\sqrt{\sum_k \frac{1}{(\Delta x_k)^2}}} \quad \text{Expression 5: Average waighted mean and its uncertainty}$$

Table 1: Cross sections for two final states and average cross section

W → μν, pt(μ)>25 GeV		OPENDATA
Final state	C [%]	σ(W+c) [pb]
D <sup>±</sup>	22.6 ± 2.2	97.7 ± 19.8 (stat)
D <sup>*±</sup>	26.8 ± 1.1	116.2 ± 14.5 (stat)
<b>Average</b>		<b>109.7 ± 11.7 (stat)</b>

Table 2: Cross sections for three final states and average cross section from paper [4]

W → μν, pt(μ)>25 GeV		CMS
Final state	C [%]	σ(W+c) [pb]
D <sup>±</sup>	11.4 ± 0.3	103.6 ± 7.8 (stat)
D <sup>*±</sup>	8.5 ± 0.4	116.9 ± 8.7 (stat)
μ	20.4 ± 0.2	106.5 ± 2.6 (stat.)
<b>Average</b>		<b>107.7 ± 3.3 (stat)</b>

## 4. Conclusions

The CMS Open Data 2011 have been successfully validated by reproducing the measurement of associated  $W + c$ . Obtained result is consistent with world averages and produced by CMS collaboration before.

The signature of  $W$ -boson production together with a charm-quark jet is observed by identifying the leptonic decay of the  $W$  boson into a muon and a neutrino and the reconstruction of exclusive and inclusive final states from the decay of charm hadrons. In total, distinct  $W + c$  signals are observed independently in six final states. Reconstructed  $D^\pm$ ,  $D^{*\pm}$  mesons associated with high transverse momentum jets and  $W$ -bosons decaying to muons are used to calculate the cross section. The determined cross section is consistent with the result published by the CMS collaboration using the full 2011 data set.

## 5. Acknowledgement

I have spent my time very usefully working with CMS data and I wish to thank DESY summer school organisers who gave this opportunity to me. I am very grateful to my supervisor Oleksandr Zenaiev for the great efforts he has invested, for sufficient patience and helpfulness, explanation of numerous complication points, corrections of my code and a lot of time invested. Nazar Stefaniuk who helped me a lot to start with my project by sharing his experience with working on CMS Open Data. Achim Geiser for discussions and comments. Olaf Bahnke for helping of leisure organising. Lecturers for interesting lectures. Finally I express my gratitude to everyone working on the DESY for their friendliness and helpfulness.

## 6. References

- [1] <http://opendata.cern.ch/>
- [2] <http://cms.web.cern.ch/news/cms-releases-new-batch-research-data-lhc>
- [3] <http://opendata.cern.ch/record/32>
- [4] CMS Collaboration, "Measurement of associated  $W + c$  production in  $pp$  collisions  $\sqrt{s}=7$  TeV", JHEP 02 (2014) 013, doi:10.1009/JHEP02(2014)013.
- [5] CMS Collaboration, "The CMS experiment at the CERN LHC", JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.
- [6] <http://opendata.cern.ch/record/31>
- [7] K.A. Olive et al. (Particle Data Group), Chin. Phys. C38, 090001 (2014)
- [8] "Charm Production and QCD Analysis at HERA and LHC", Oleksandr Zenaiev, DESY-THESIS-2015-012