ESTIMATING THE FAKE LEPTON BACKGROUNDS FOR THE DRELL-YAN DIFFERENTIAL CROSS SECTION MEASUREMENT USING 2016 CERN CMS DATA

Marijus Ambrozas, Andrius Juodagalvis

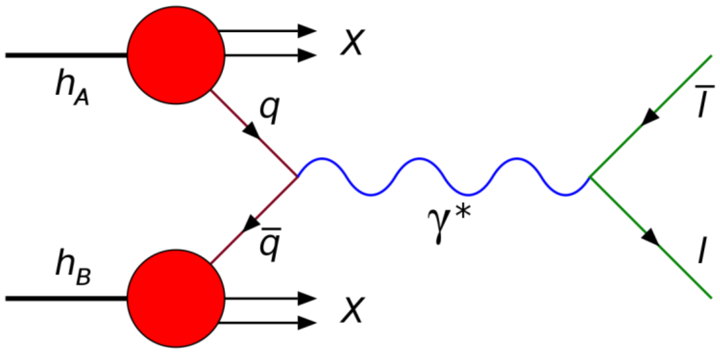
Institute of Theoretical Physics and Astronomy, Faculty of Physics, Vilnius University

marijus.ambrozas@gmail.com

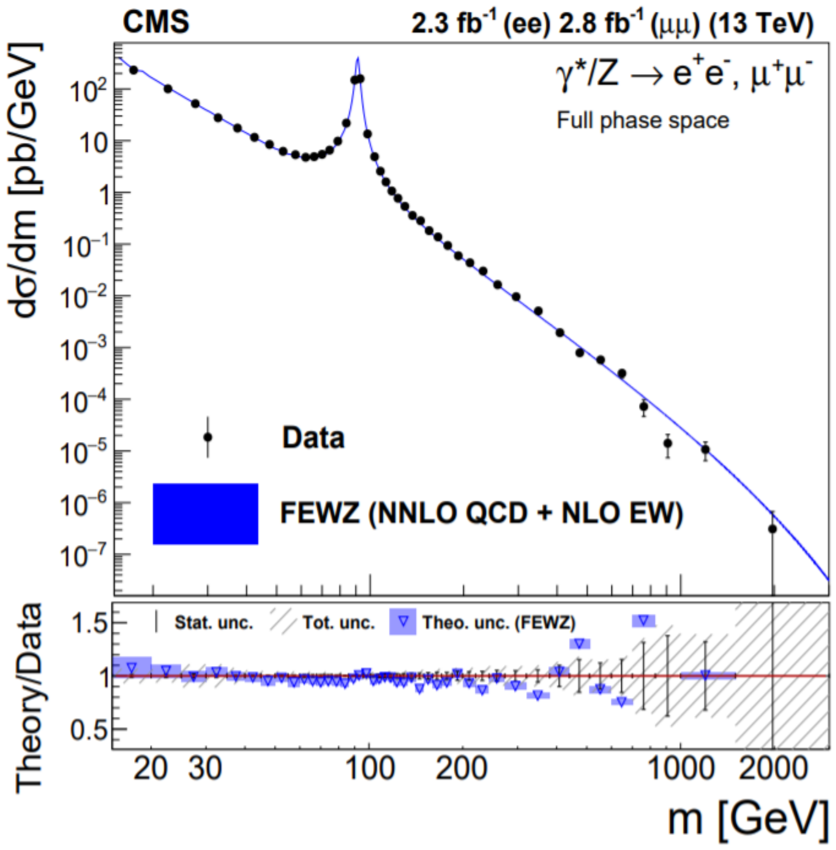
CERN BALTIC CONFERENCE 2021

Drell-Yan process (DY)

- A quark and an antiquark annihilate producing a lepton-antilepton pair
- Important for constraining parton distribution functions (PDF) and testing the precision of the standard model (SM)
- Dominant background in the analyses of other processes
- Final states under investigation:
  - Electron-positron pair (electron channel)
  - Muon-antimuon pair (muon channel) ← presented in this talk
- Main measurement variable – dilepton invariant mass m_{ll}

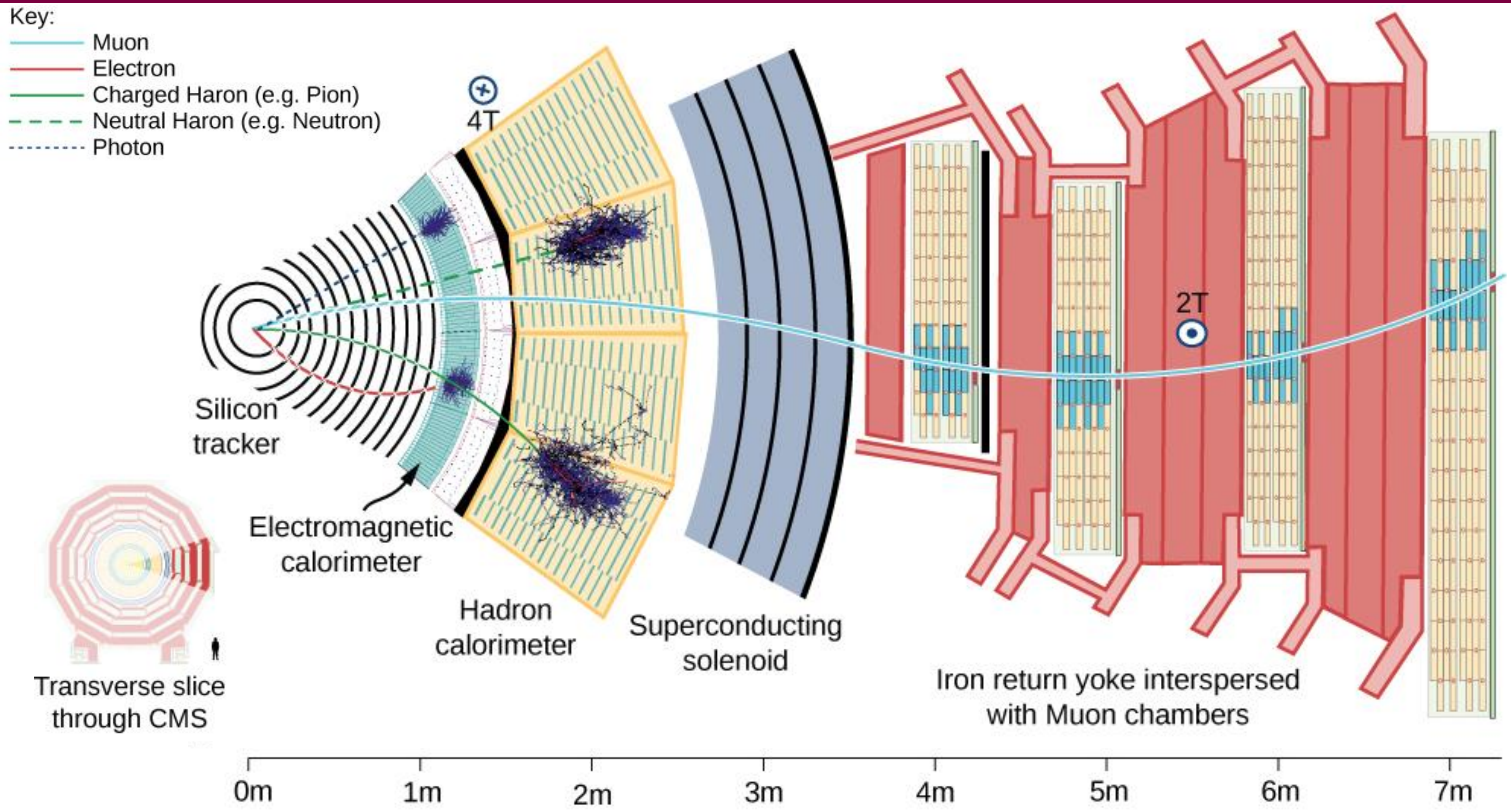


DY differential cross section $\mathrm{d}\sigma/\mathrm{d}m$



The CMS Collaboration. JHEP 12 059, 2019

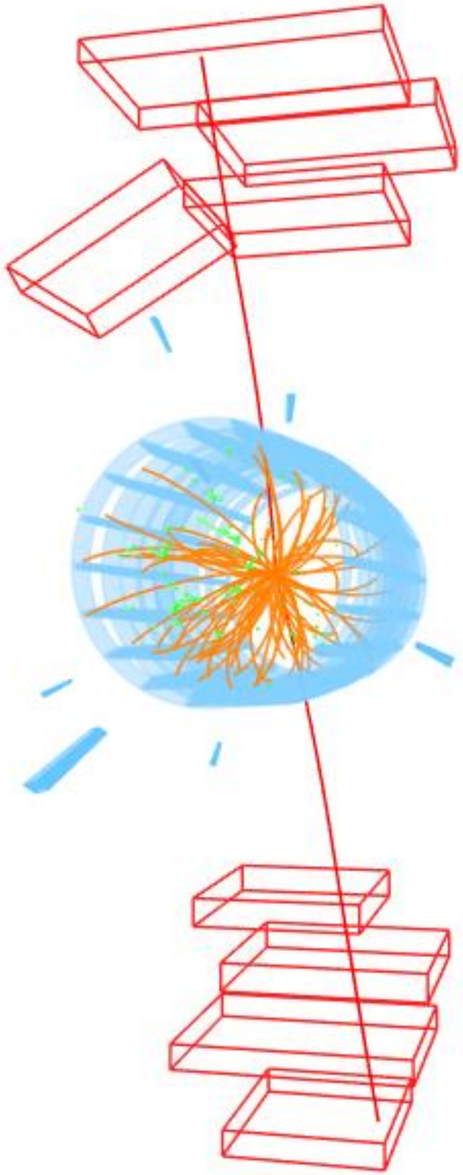
Compact Muon Solenoid (CMS)



David Barney. CMS slice.

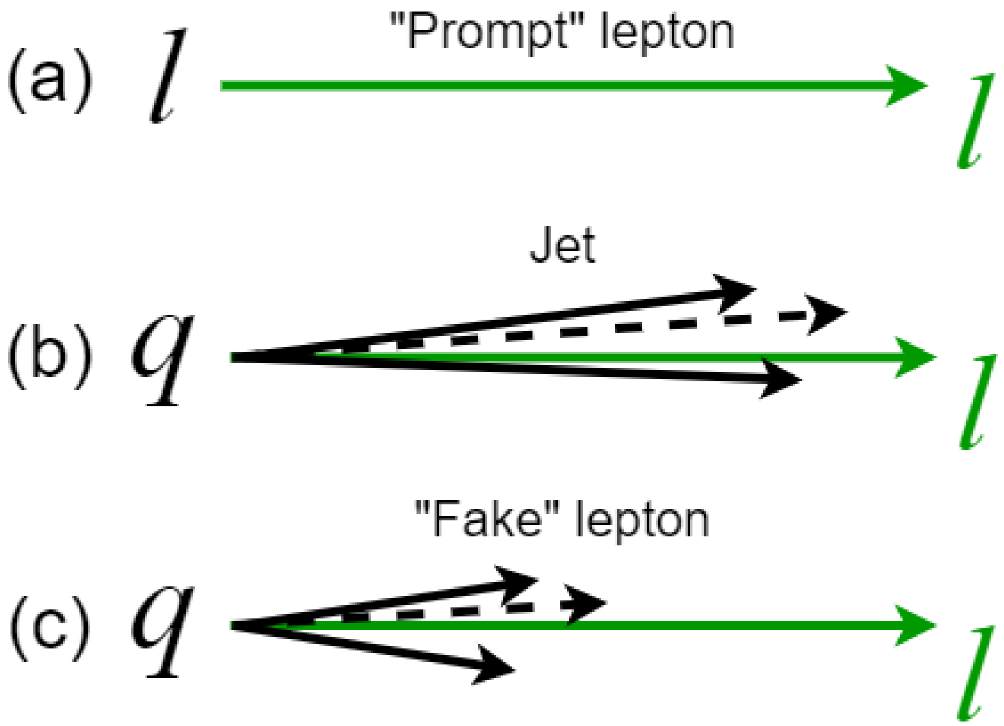
Drell-Yan backgrounds

- Only the **final state** can be observed for any high-energy physics process
- In case of the Drell-Yan process, we are searching for a **pair of isolated leptons**
- Impossible to tell whether the signal was produced by a pair of leptons originating from a prompt or delayed emission process
- Prompt leptons could originate from the Drell-Yan process
  - We call leptons from delayed emissions “fake” from now on
- Need to estimate the number of **background** (non-Drell-Yan) events in the selected event sample
- Most significant DY backgrounds: ZZ, $\bar{t}W$, tW, WZ, WW, $t\bar{t}$, $\mathrm{DY} \rightarrow \tau\tau$, $\boldsymbol{W}$**+Jets, QCD multijet**
- Background contribution can be estimated from simulation (MC), but **data-driven methods** are believed to be more accurate



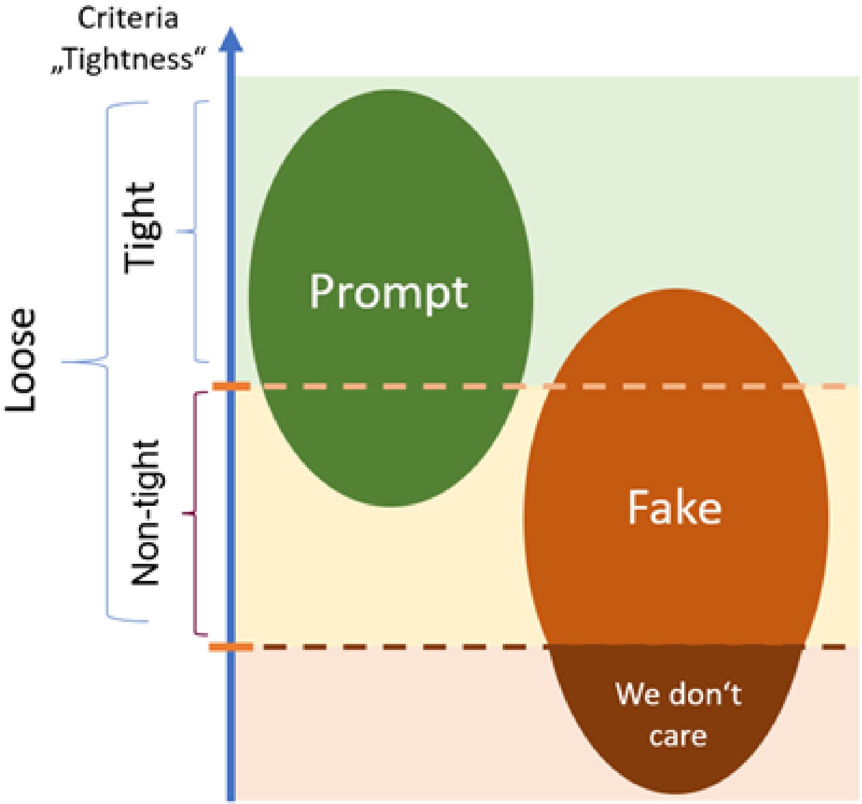
Jet events

- **Jet** is a cone-shaped particle stream produced by a final-state quark or gluon in a hadronization process
  - Leptons may be produced in delayed emission processes inside jets
- On rare occasions a lepton-dominated jet can be **misidentified** as an isolated lepton (we then call it a “**fake**” lepton)
- Dilepton final state sample can be contaminated with one (W+Jets) or two (QCD multijet) fake lepton events
- The “fake rate” and “matrix” methods are data-driven methods used to estimate misidentified particle background yields

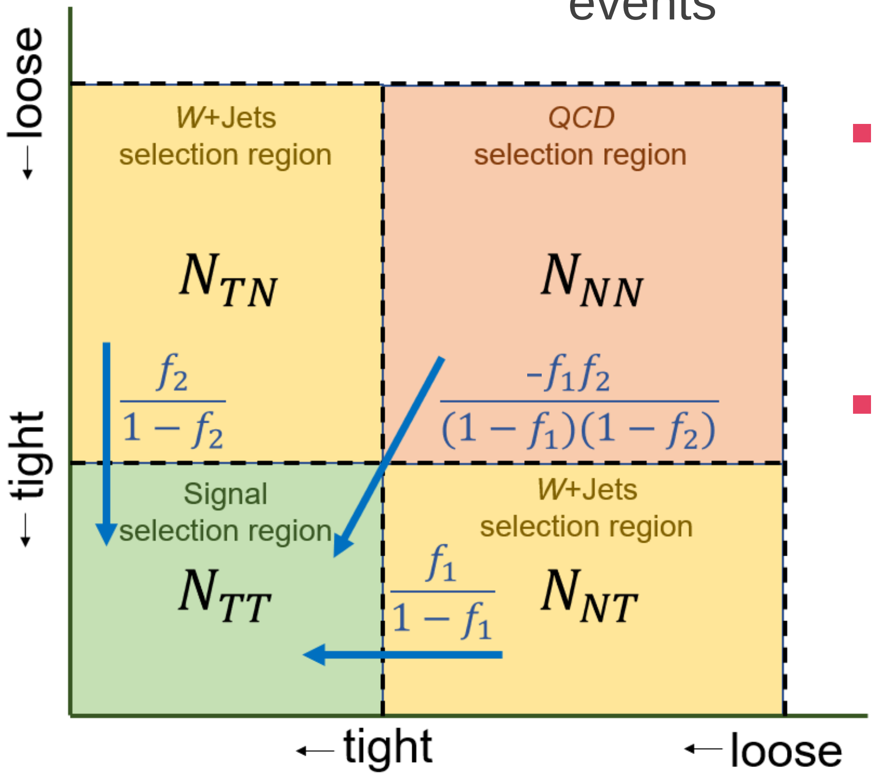


The "fake rate" method

- We calculate the probability that a **fake lepton** from the "loose" sample will pass the "tight" selection criteria (fake lepton selection efficiency)
  - "loose" and "tight" samples are defined by constraints on some analysis variables



- $f_{Tight\,|\,Fake} = \dfrac{N^{QCD}_{Tight}}{N^{QCD}_{Tight}+N^{QCD}_{Non-tight}}$
- N_i^{QCD} is extracted from data (with some help from MC)
- After f is obtained, we can use it to estimate the number of background events



- However, W+Jets and QCD selection regions are contaminated with other processes as well
- Other contributions are subtracted using MC or fitting template distributions

T – tight, N – non-tight

The „matrix method“

- A more sophisticated method is the “matrix method”, which also uses prompt lepton efficiency
- Prompt efficiency should help estimating contamination with other processes in W+Jets and QCD selection regions (ideally no need to use MC)
  - T – tight; N – non-tight; P – prompt; F – fake
  - f – fake lepton selection efficiency; p – prompt lepton selection efficiency

Measured Yields →

Unknown (hidden) numbers

$$\begin{pmatrix} N_{TT} \\ N_{TN} \\ N_{NT} \\ N_{NN} \end{pmatrix} = \begin{pmatrix} p_1p_2 & p_1f_2 & f_1p_2 & f_1f_2 \\ p_1\tilde{p}_2 & p_1\tilde{f}_2 & f_1\tilde{p}_2 & f_1\tilde{f}_2 \\ \tilde{p}_1p_2 & \tilde{p}_1f_2 & \tilde{f}_1p_2 & \tilde{f}_1f_2 \\ \tilde{p}_1\tilde{p}_2 & \tilde{p}_1\tilde{f}_2 & \tilde{f}_1\tilde{p}_2 & \tilde{f}_1\tilde{f}_2 \end{pmatrix} \begin{pmatrix} N_{PP} \\ N_{PF} \\ N_{FP} \\ N_{FF} \end{pmatrix}$$

here $\tilde{x} = 1 - x$

- We can invert the matrix to find the hidden values like this:

$$\begin{pmatrix} N_{PP} \\ N_{PF} \\ N_{FP} \\ N_{FF} \end{pmatrix} = \frac{1}{(f_1-p_1)(f_2-p_2)} \begin{pmatrix} \tilde{f}_1\tilde{f}_2 & -\tilde{f}_1f_2 & -f_1\tilde{f}_2 & f_1f_2 \\ -\tilde{f}_1\tilde{p}_2 & \tilde{f}_1p_2 & f_1\tilde{p}_2 & -f_1p_2 \\ -\tilde{p}_1\tilde{f}_2 & \tilde{p}_1f_2 & p_1\tilde{f}_2 & -p_1f_2 \\ \tilde{p}_1\tilde{p}_2 & -\tilde{p}_1p_2 & -p_1\tilde{p}_2 & p_1p_2 \end{pmatrix} \begin{pmatrix} N_{TT} \\ N_{TN} \\ N_{NT} \\ N_{NN} \end{pmatrix}$$

N_{PP} – DY and prompt lepton bkg

$N_{PF} + N_{FP}$ – mostly Wjets (ideally)

N_{FF} – mostly QCD (ideally)

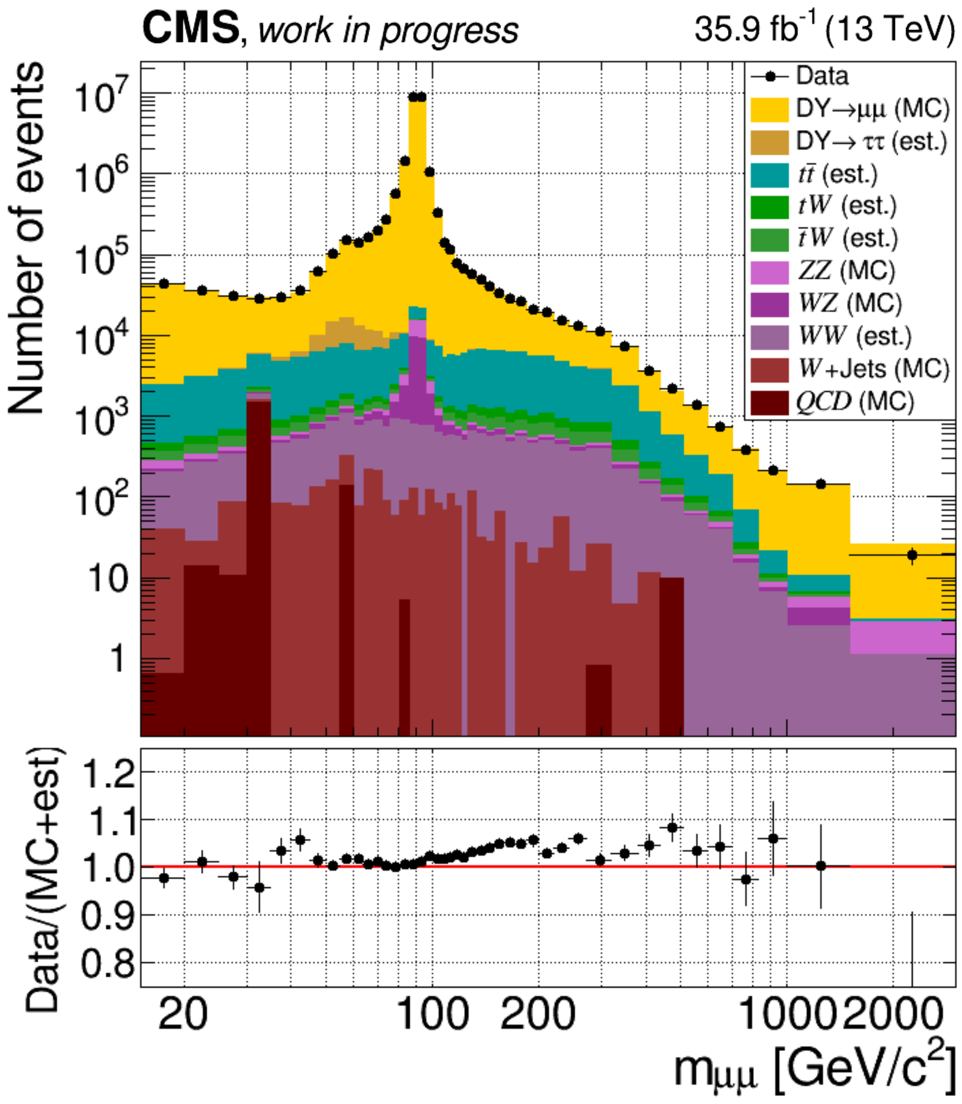
N_{TT} – events in “signal selection” region

$N_{TN} + N_{NT}$ – events in “W+Jets selection” region

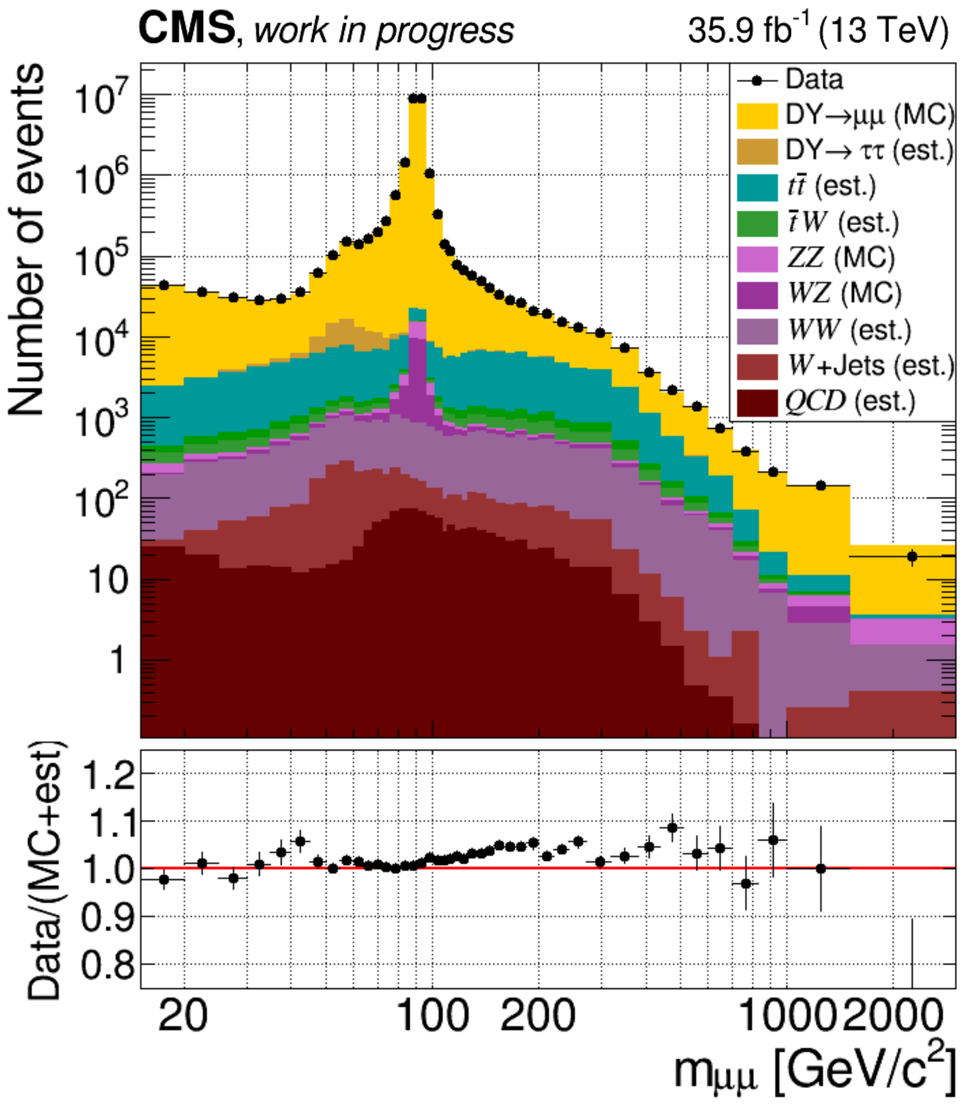
N_{NN} – events in “QCD selection” region

Dimuon mass distributions in the signal region

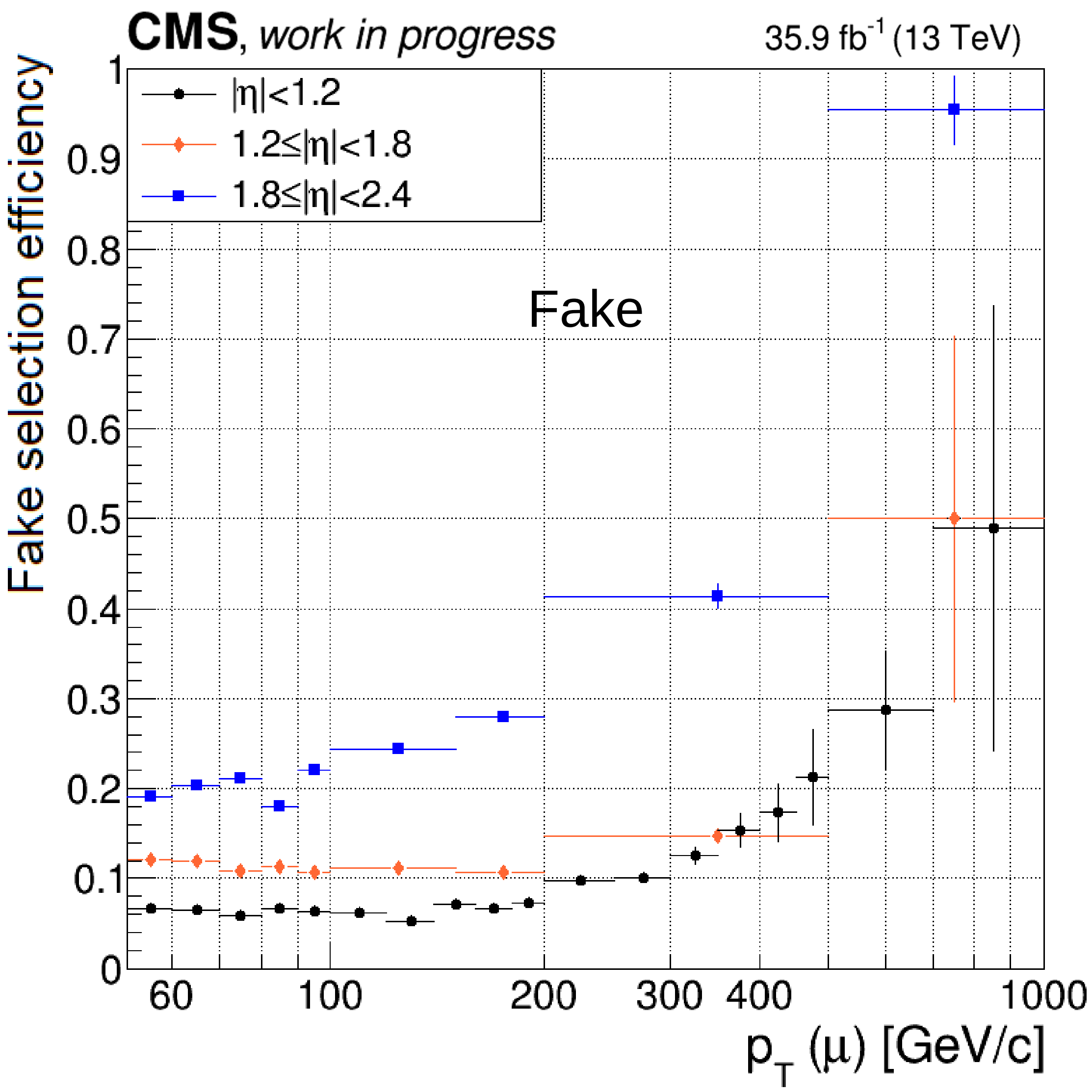
Using MC for fake muon backgrounds

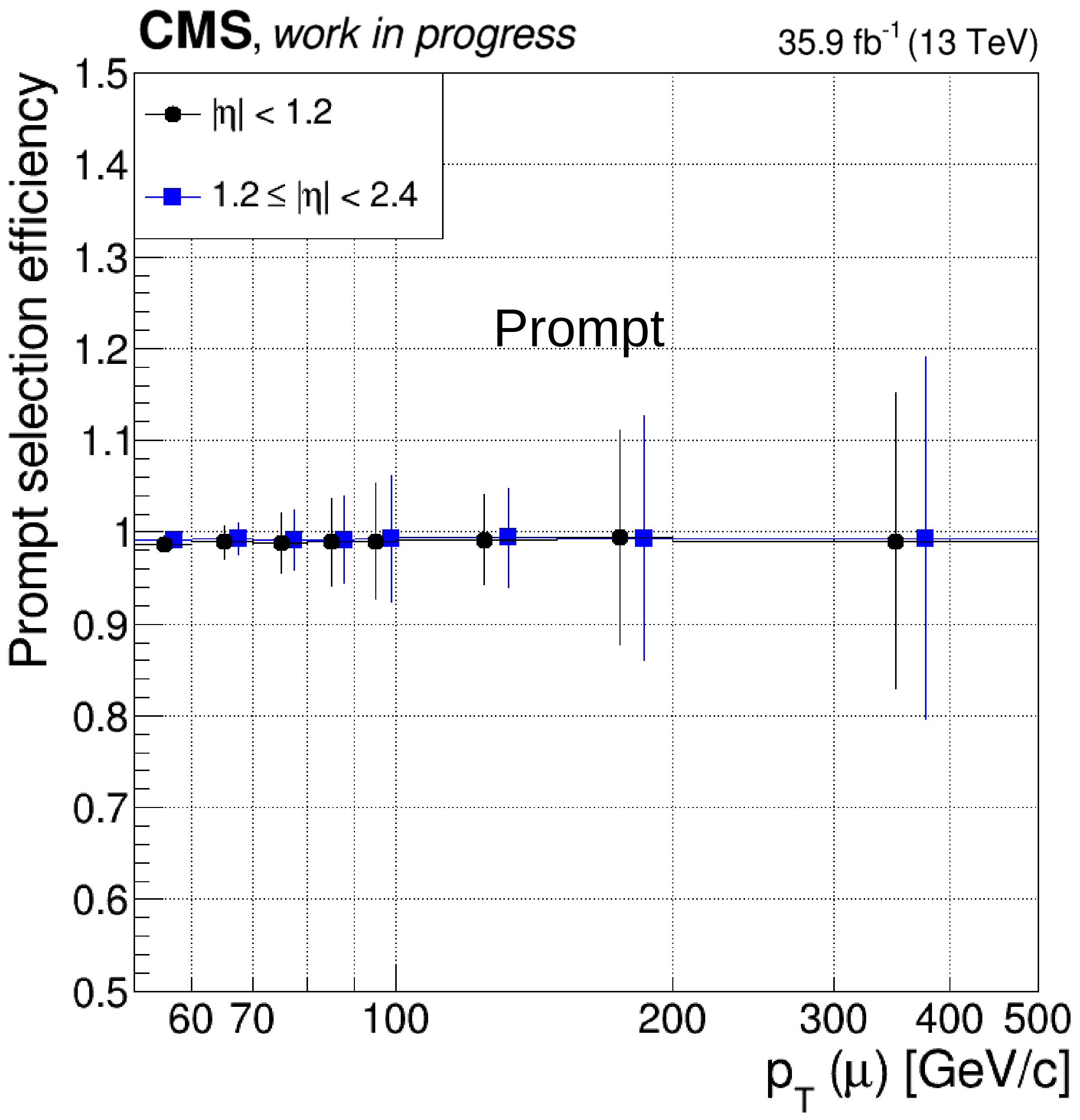


Expected result after applying a data-driven method



Estimated fake and prompt muon selection efficiencies





Blue markers are shifted to the right for clarity

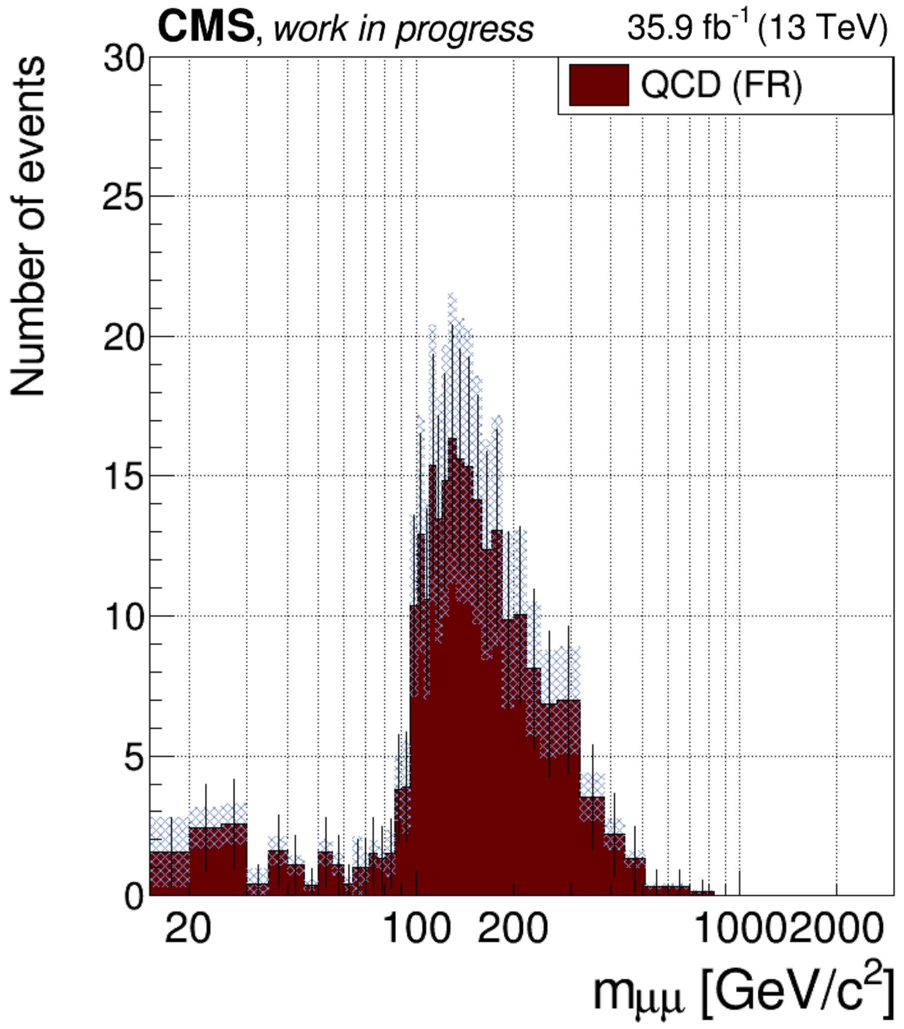
Matrix method validity and expectation regions

- MC tests have shown that each process has a slightly different prompt and fake selection efficiencies
- We cannot differentiate between different processes in real data
- Consequently, some processes get weighted incorrectly in the matrix method
- We need to subtract the incorrectly weighted distributions using MC in order to get a meaningful result
  - The argument that no MC is needed for matrix method is no longer valid
  - The matrix method procedure becomes very similar to the fake rate method
- In order to compare the performance of the matrix and fake rate methods, we apply all the same procedures on them and assign expectation bars that we estimate from:
  - Sensitivity to the statistical uncertainty of fake and prompt selection efficiencies
  - Sensitivity to the binning of fake and prompt selection efficiencies
  - Sensitivity to jet multiplicity in the event

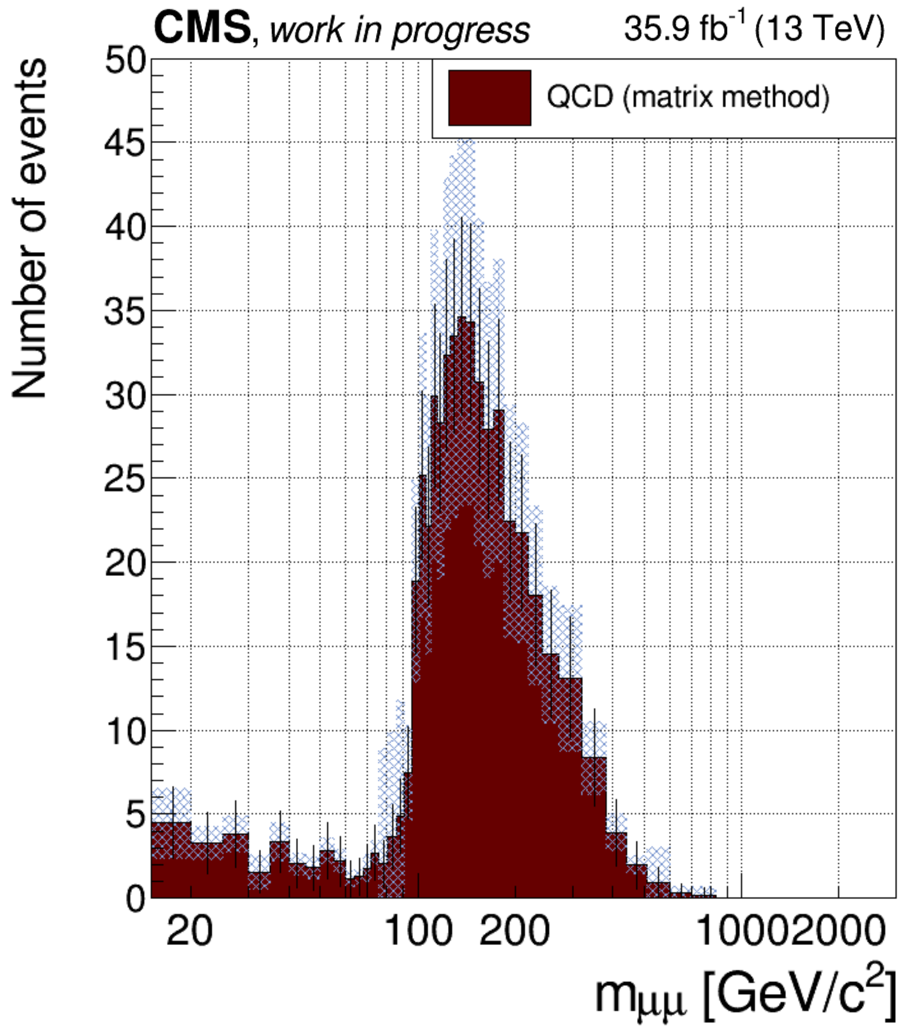
QCD estimation results

The expectation bars are relatively similar in width (~30%) but the matrix method suggests higher background yield. The difference could be assigned to systematic uncertainty.

Fake rate method



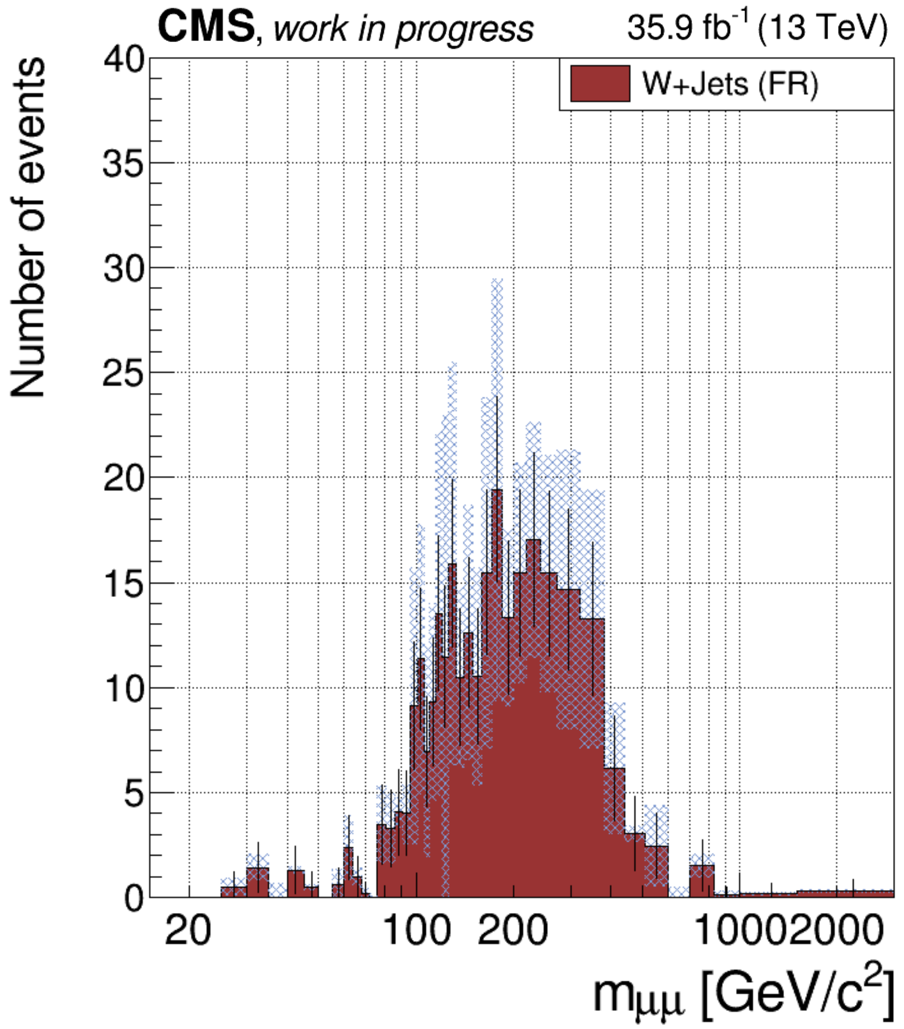
Matrix method



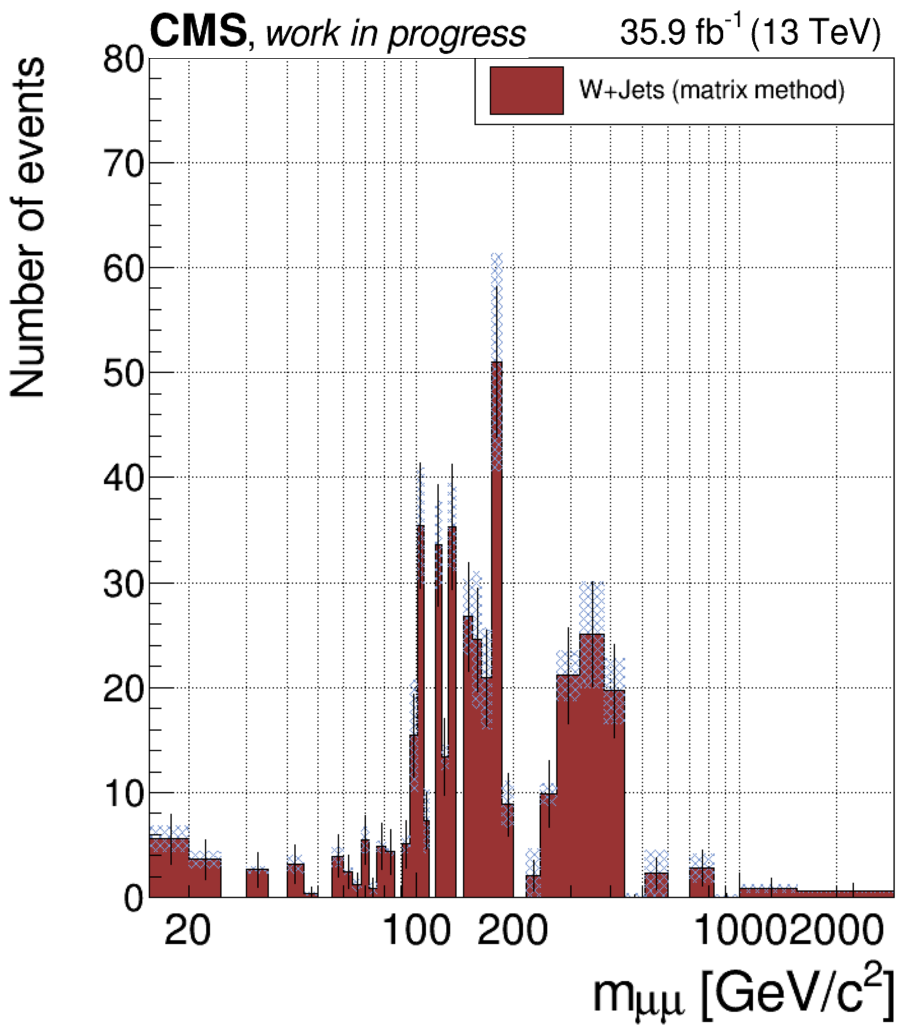
W+Jets estimation results

The expectation bars are significantly wider for the fake rate method but the matrix method provides a very discontinuous spectrum, making the fake rate result look more reliable in the case of DY $d\sigma/dm_{ll}$ measurement.

Fake rate method



Matrix method



Summary and conclusions

- The “fake rate” method and “matrix method” are equivalent methods for fake lepton background estimation but have different procedures:
  - “Fake rate” method relies on fake lepton selection efficiency and MC subtraction of prompt dilepton events to estimate the backgrounds from fake-enriched selection regions
  - “Matrix method” relies on both fake and prompt lepton selection efficiencies and makes use of fake-enriched selection region together with signal selection region to be less reliant on MC
- Both “fake rate” method and “matrix method” have their own advantages and shortcomings:
  - “Fake rate” method heavily relies on MC accuracy but is less complicated
  - “Matrix method” is less reliant on MC but needs very precise tuning of fake and prompt selection efficiencies, making it hard to achieve reliable results for $d\sigma/dm_{ll}$ measurement where DY signal is very dominant
  - In the case of mixed states (e.g., electron+muon which were not discussed in this presentation), one cannot split the uncertainties per fake process (muon or electron) when using the “matrix method”
- Having both measurements allow us to pick one as a central value and use the difference between the two methods as an estimate of systematic uncertainty
- One possible improvement for the “matrix method” could be using maximum likelihood fit to obtain the most probable N_{PP}, N_{PF}, N_{FP}, N_{FF} values for the given data distributions and measured prompt and fake efficiencies
  - E.g., as described in JHEP 11 (2014) 031

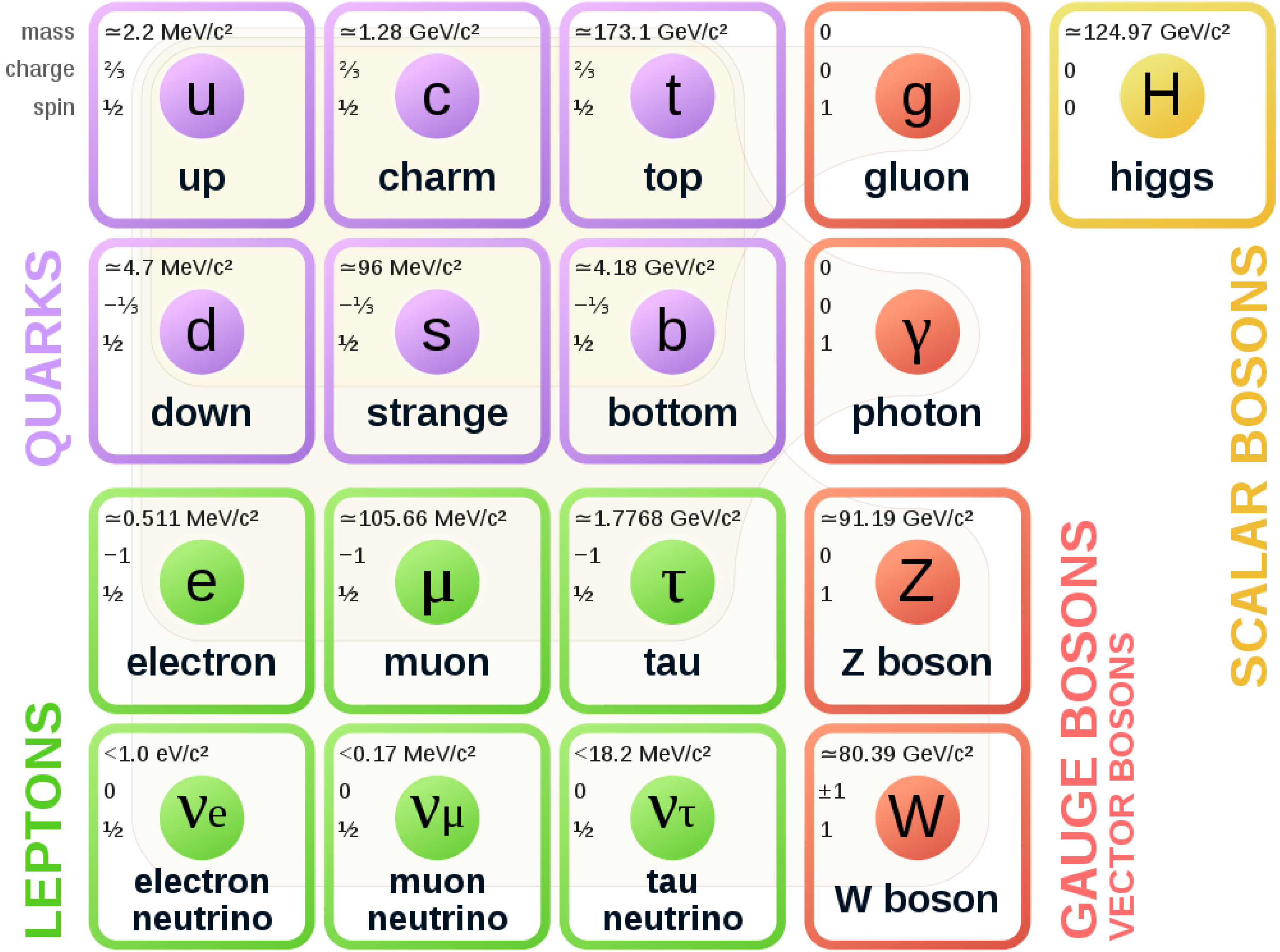
Thank you for your attention!

Marijus Ambrozas, Andrius Juodagalvis

marijus.ambrozas@gmail.com

Fizikos fakultetas

Standard model

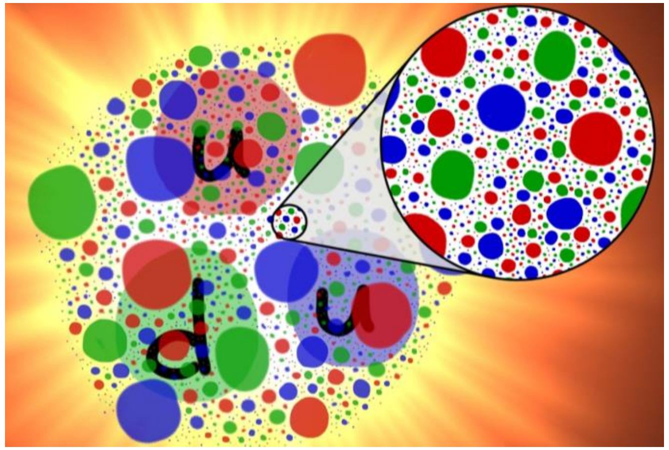


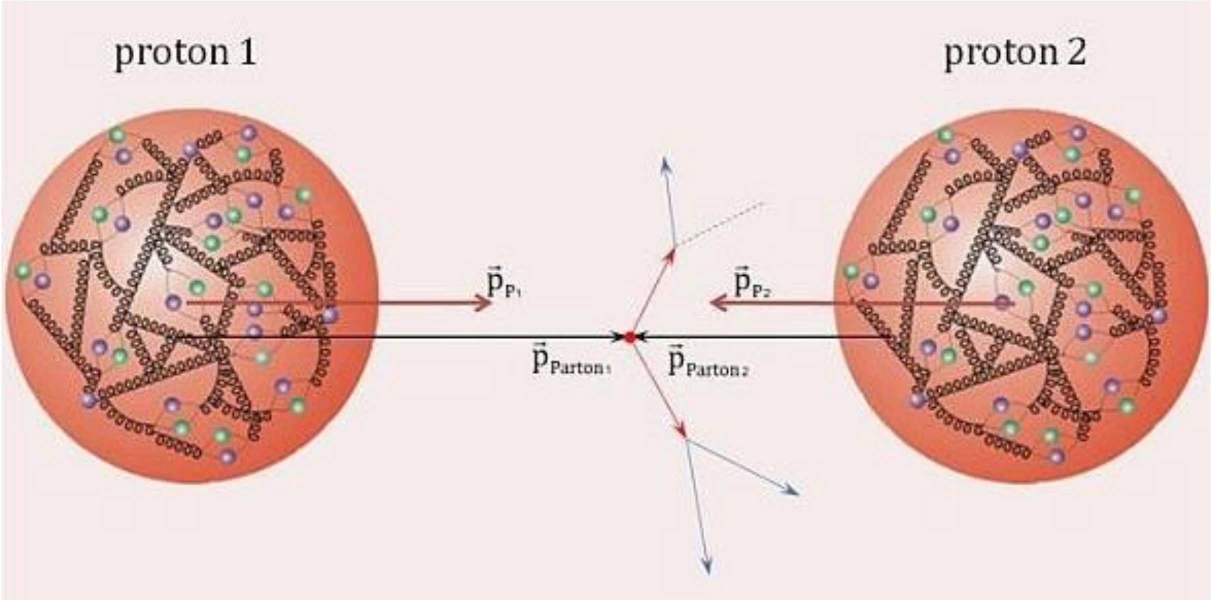
Wikimedia Commons: MissMJ

Proton structure and pp collisions

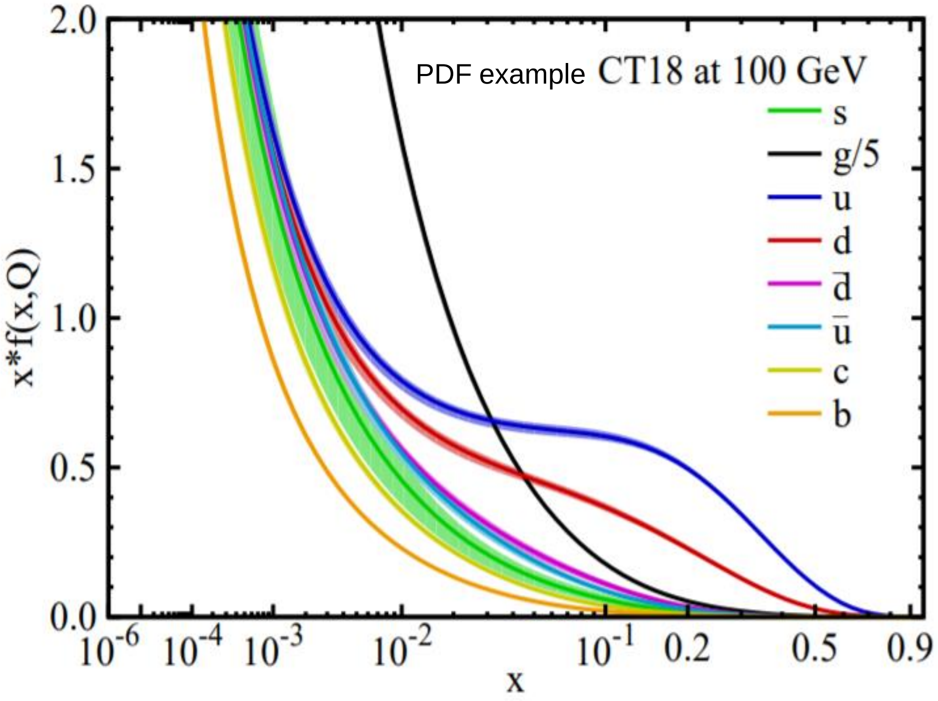
- Proton consists of three valence quarks and quark-gluon "sea"
- The structure of the proton is described using parton distribution functions (PDFs)
- Interactions between non-valence quarks are possible in proton-proton collisions

https://news.fnal.gov/2012/05/quarks-and-gluons-and-partons-oh-my/





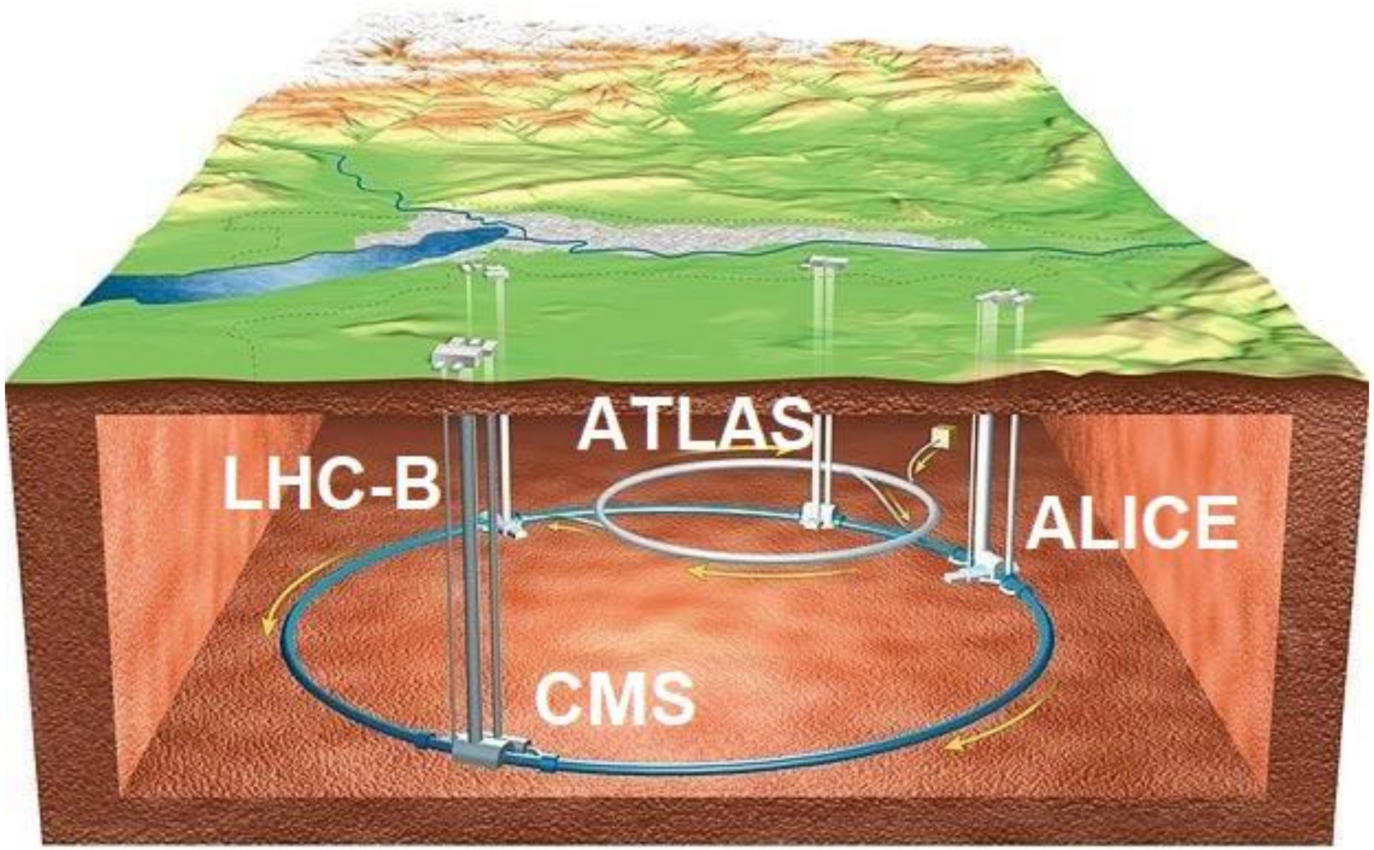
https://atlas.physicsmasterclasses.org/en/zpath_protoncollisions.htm



T.J. Hou et al. MSUHEP-19-025, 2019.

Large Hadron Collider

- Large Hadron Collider (LHC) is the largest particle collider ever made
- It produces 13 TeV proton collisions; it is the highest collision energy achieved by humanity so far
- Compact Muon Solenoid (CMS) is one of the four largest experiments at LHC



swissinfo.ch

Invariant mass

- For a particle, invariant mass is its rest mass, calculated as $m = \sqrt{E^2 - |\vec{p}|^2}$
- Invariant mass for particle systems is calculated as follows (example for 2 particles)

$$m_{12}^2 = (E_1 + E_2)^2 - |\vec{p}_1 + \vec{p}_2|^2$$

- If two particles are the only decay products of a single mother particle, their invariant mass m_{12} will be equal to the mass of the mother particle

MC normalization

- Simulated event distributions from different processes must be normalized to observed integrated luminosity $\mathcal{L}_{int}$ (proportional to the produced number of collisions)
  - Otherwise you cannot compare data and simulation quantitively
- Simulated events can have different individual weights ω_i^{GEN} assigned to them by MC event generator
- For a process with a cross section σ at integrated luminosity $\mathcal{L}_{int}$ we expect $N_{exp} = \sigma\mathcal{L}_{int}$ events
- If we want to make the effective number of events in the simulated dataset equal to the expected number of events, we must assign a specific weight ω_i to each event:

$$\omega_i = \omega_i^{GEN} \frac{\sigma\mathcal{L}_{int}}{\sum_{j=1}^{N} \omega_j^{GEN}}$$

Event selection criteria (main analysis region)

Electron channel	Muon channel
Trigger: HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	**Trigger:** HLT_IsoMu24 OR HLT_IsoTkMu24 (used trigger matching)
$p_T^{Lead} > 28\ GeV$, $p_T^{Sublead} > 17\ GeV$, $\lvert\eta_{SC}\rvert < 2.4$, **Excluding** $1.4442 < \lvert\eta_{SC}\rvert < 1.566$	$p_T^{Lead} > 28\ GeV$, $p_T^{Sublead} > 17\ GeV$, $\lvert\eta\rvert < 2.4$
Electron MediumID	Muon TightID, $I_{PF}^{rel} < 0.15$
Exactly two electrons passing the event selection	Two muons with smallest vertex fit $\chi^2 < 20$, Opposite-sign, 3D angle $< \pi - 0.005$ rad

Event selection criteria (fake efficiency estimation)

Electron channel

Tight	Loose
Triggers: HLT_PhotonX_v, where X is an OR of 22, 30, 36, 50, 75, 90, 120, 175	
$p_T > 25$ GeV, $\lvert\eta_{SC}\rvert < 2.4$, excluding $\lvert\eta_{SC}\rvert \in (1.4442, 1.566)$	
Barrel: $\sigma_{i\eta i\eta} < 0.013$, $H/E < 0.13$, $\lvert\Delta\eta_{in}^{seed}\rvert < 0.01$, $\lvert\Delta\phi_{in}\rvert < 0.07$	
Endcap: $\sigma_{i\eta i\eta} < 0.035$, $H/E < 0.13$	
Number of missing hits <= 1	
Electron MediumID	–
Veto events with more than 1 electron passing MediumID	

Muon channel

Tight	Loose
Trigger: HLT_Mu50	
$p_T > 52$ GeV, $\lvert\eta\rvert < 2.4$	
Muon TightID	
$I_{PF}^{rel} < 0.15$	I_{PF}^{rel} = any

Event selection criteria (prompt efficiency estimation)

The working principle is to select the muon emerging from W decay in W+Jets event.

Muon channel
HLT_Mu50
Single tight muon in the event
$p_T > 52\ GeV$, $\lvert\eta\rvert < 2.4$
$MET > 20\ GeV$, $M_T > 60\ GeV$
One jet with $p_T > 40\ GeV$
Veto all additional jets with $p_T > 17\ GeV$