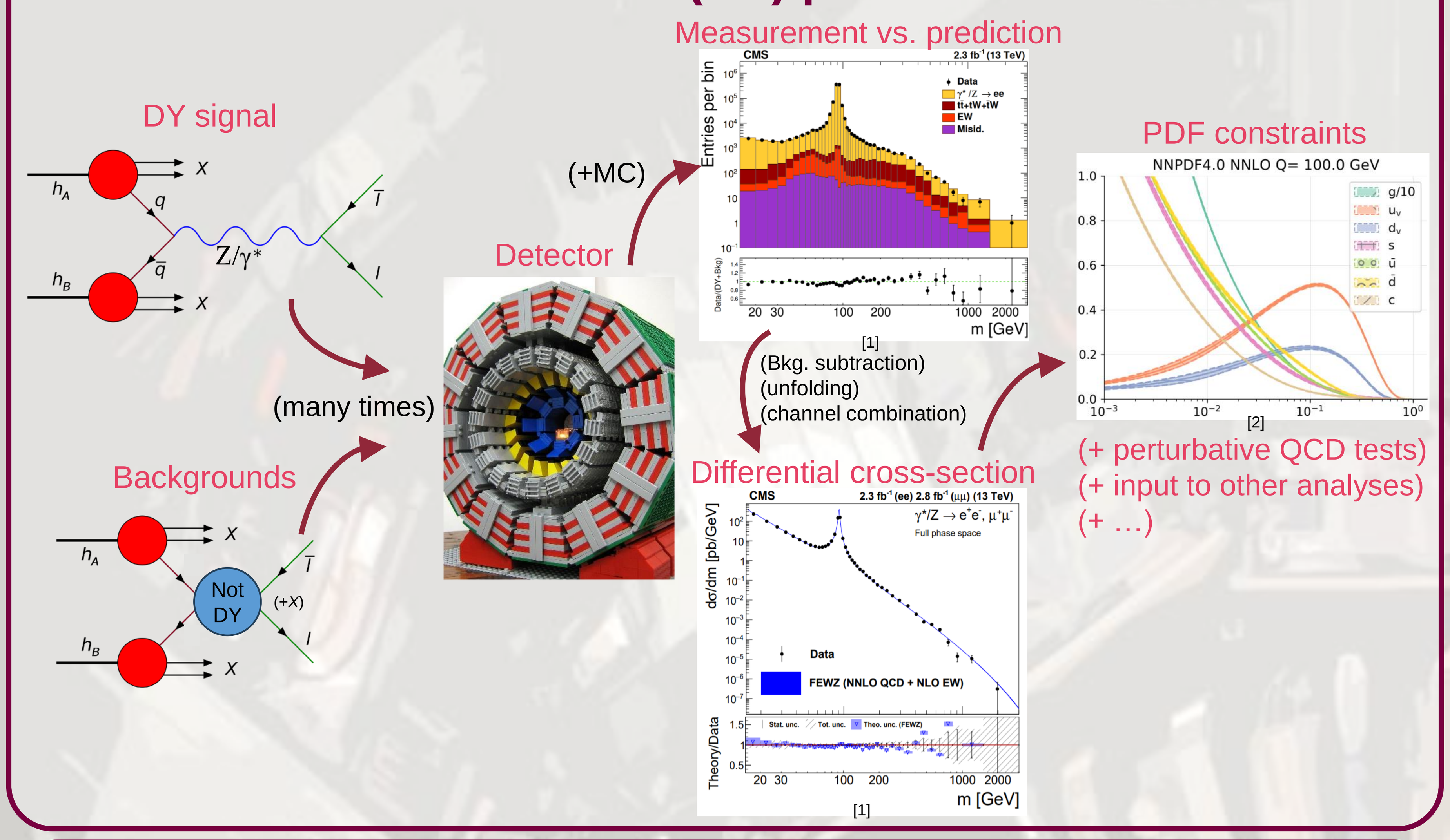
Data-driven dilepton background estimation improvements for the Drell-Yan analysis at CMS

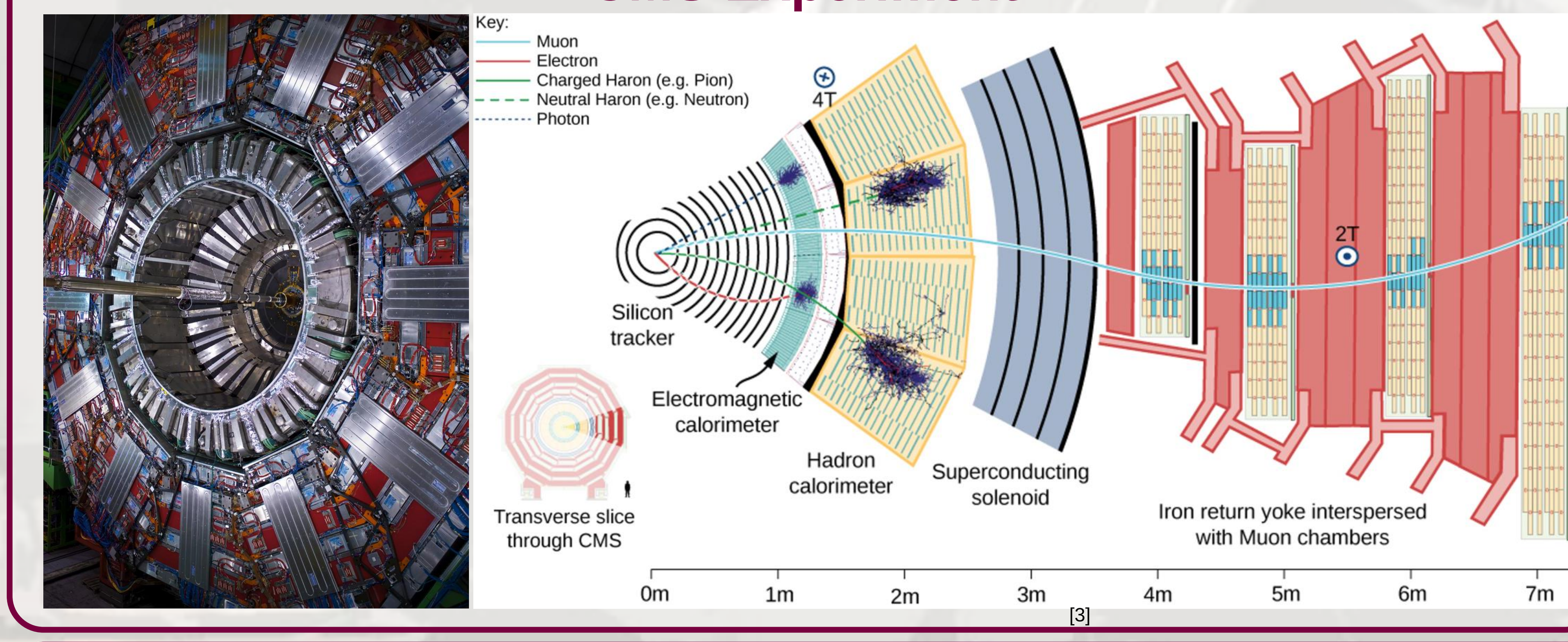
Marijus Ambrozas, on behalf of CMS Collaboration
Faculty of Physics, Vilnius University, Lithuania
marijus.ambrozas@cern.ch

ICHEP2024: 42nd International Conference on High Energy Physics. Prague, July 2024

Drell-Yan (DY) process



CMS Experiment



Drell-Yan backgrounds

Prompt lepton backgrounds: $t\bar{t}$, Single top, WW, WZ, ZZ, DY→ττ, γγ

Non-prompt (or misidentified) lepton backgrounds: W+Jets, γ+Jets, QCD

Estimation methods: MC; eμ method; same-sign method; Misidentification rate (pass-fail, loose-to-tight) method; Matrix method

"Same-sign" method

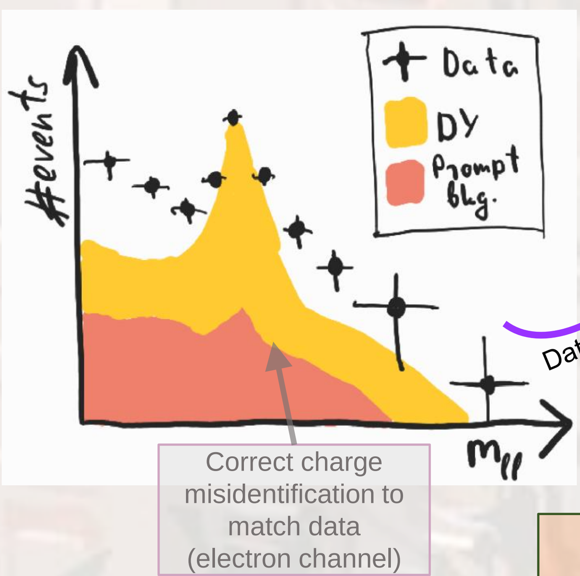
Basic idea

- $l^\pm l^\pm$ control region
  → enriched with non-prompt backgrounds
- Estimate the non-prompt yield by MC subtraction
- Reweight the result by OS/SS ratio $R_{OS/SS}$ to get a result for $l^\pm l^\mp$
- Typically, $1 < R_{OS/SS} < 2$ is chosen
- Most often used in the muon channel only due to non-zero electron charge misidentification probability

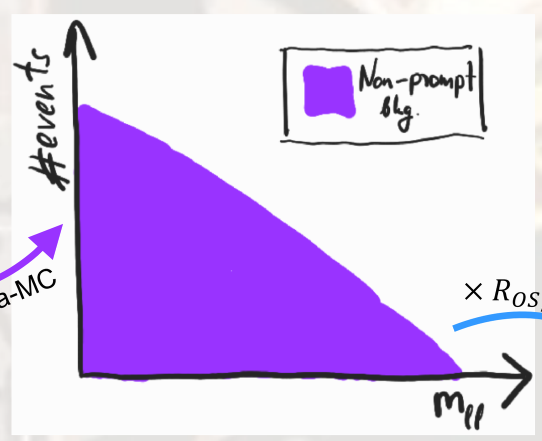
Improvements

- Correcting the electron charge misidentification probability for MC is possible using Z peak events
  → method becomes robust even in the electron channel (outside Z peak)
- Precise $R_{OS/SS}$ values can be obtained by using data control regions
  → Even variable $R_{OS/SS}$ is possible to reduce systematic uncertainty

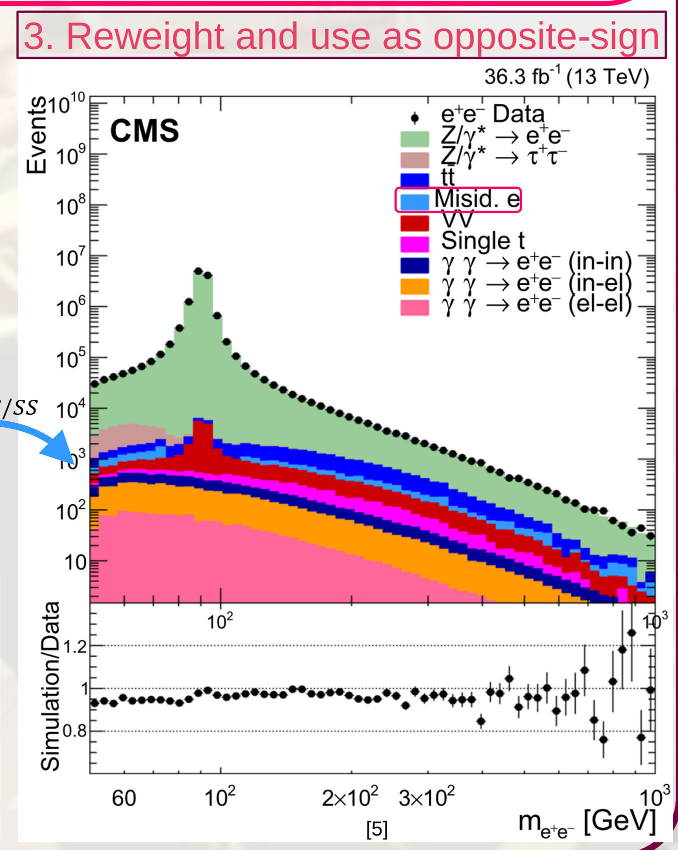
1. Select same-sign events



2. Subtract prompt lepton MC



3. Reweight and use as opposite-sign



✔
- Very simple
- Fairly robust with good reweighting

✘
- Needs careful treatment for electron charge misidentification
- Non-constant reweighting factors can make it more complicated

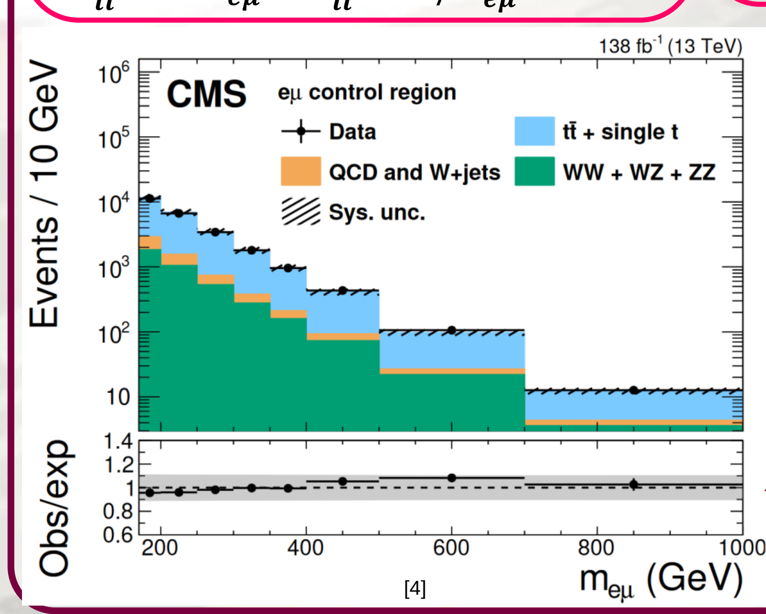
"eμ" method

Basic idea

- $e^\pm\mu^\mp$ sample → only prompt lepton backgrounds
- $N^{bkg}_{e^\pm\mu^\mp}/N^{bkg}_{l^\pm l^\mp} = 2$ (ideally)
- In reality, it is not exact
  → extract from MC
- $N^{bkg}_{ll} = N^{obs}_{e\mu} \cdot N^{bkg\ MC}_{ll}/N^{bkg\ MC}_{e\mu}$

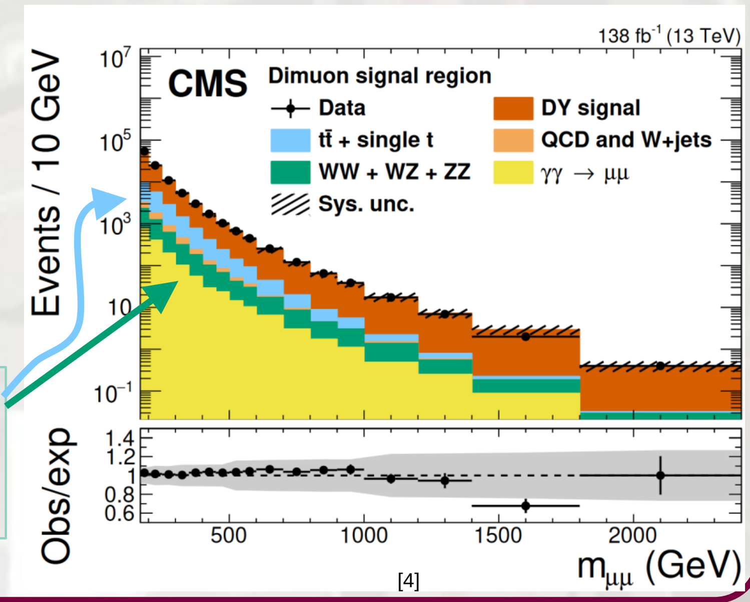
Improvements

- $N^{obs}_{e\mu}$ contains non-prompt contributions
  → need to be subtracted
  → non-prompt estimation getting improved (see other methods)
- MC improvements allow using eμ sample only for MC validation
  → if $N^{obs}_{e\mu}/N^{bkg\ MC}_{e\mu} \approx 1$, then $N^{bkg}_{ll} \approx N^{bkg\ MC}_{ll}$
  → The difference from 1 can be used as systematic uncertainty



Use this ratio (after non-prompt subtraction)

To reweight (or validate) these backgrounds



"Misidentification rate" method

Basic idea

- Control regions with inverted lepton track "quality" cuts (ID working point, isolation, etc.)
  → enriched with non-prompt backgrounds
- Estimate non-prompt by MC subtraction in regions with inverted cuts for one lepton (N_{PF}) and both leptons (N_{FF})
- Reweight the events using the "misidentification rate" f to obtain the result in the signal region:

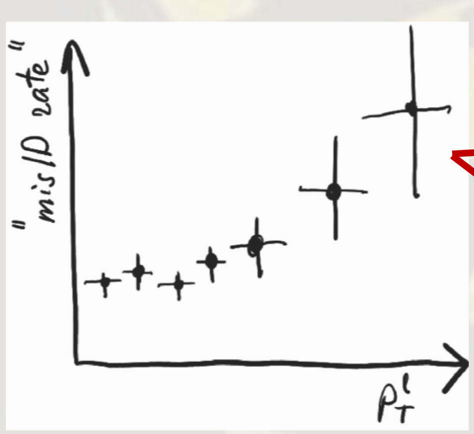
$$N_{PP} = \frac{f}{1-f}N_{PF} - \frac{f_1 f_2}{(1-f_1)(1-f_2)}N_{FF}$$

- The "misID rate" is measured in a separate control region as a function of lepton (p_T, η)
  → Usually, single lepton events or DY Z peak events with a third additional lepton are used

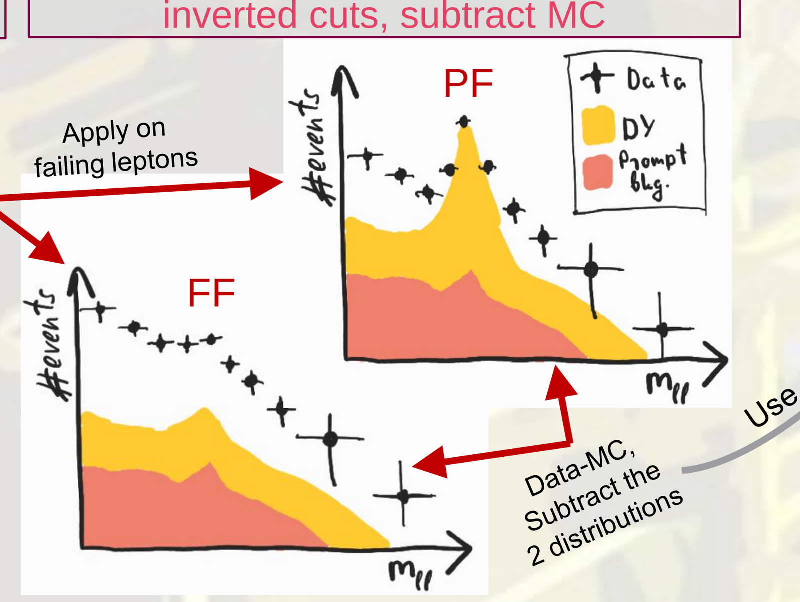
Improvements

- Template fits can help achieving better estimates of the "misID rate" and/or the background distribution
- Function fits are often used to obtain smooth functional forms for the "misID rate"
  → symbolic regression tools (e.g., pySR) can be useful
- More statistics allow for more sophisticated "misID rate" binning and categorization
- More control regions used to evaluate, apply and validate the "misID rate" in certain analyses
  → E.g., [6] splits the electron phase space into 3 parts and makes use of eμ event samples to get a total of 71 control regions (+8 signal regions)!

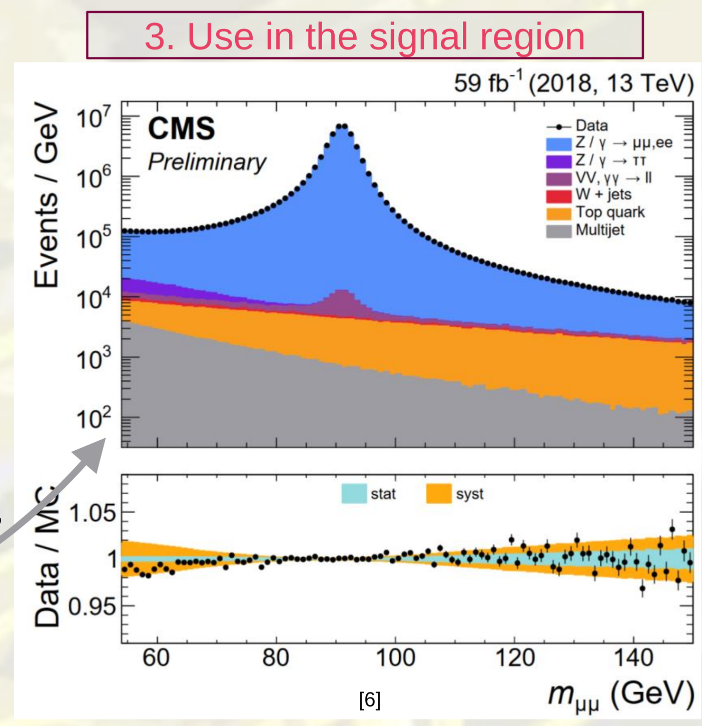
1. Measure the "misID rate"



2. Apply the "misID rate" on events with inverted cuts, subtract MC



3. Use in the signal region



✔
- Works well in both ee and μμ channels
- Simple validation (same-sign)

✘
- Somewhat complicated procedure
- May be tricky to define the phase space for "misidentification rate" definition
- Often relies on prescaled triggers

"Matrix" method

Basic idea

- Very rarely used in DY analyses, but technically possible
- Extension of the "misidentification rate" method
  → additionally uses the prompt lepton efficiency ϵ to be less reliant on MC
- Events with prompt (R) and non-prompt (N) leptons are related to events with reconstructed leptons that pass (P) or fail (F) the quality cuts as follows:

$$\begin{pmatrix} N_{PP} \\ N_{PF} \\ N_{FP} \\ N_{FF} \end{pmatrix} = \begin{pmatrix} \epsilon_1\epsilon_2 & \epsilon_1 f_2 & f_1\epsilon_2 & f_1 f_2 \\ \epsilon_1\tilde{\epsilon}_2 & \epsilon_1\tilde{f}_2 & f_1\tilde{\epsilon}_2 & f_1\tilde{f}_2 \\ \tilde{\epsilon}_1\epsilon_2 & \tilde{\epsilon}_1 f_2 & \tilde{f}_1\epsilon_2 & \tilde{f}_1 f_2 \\ \tilde{\epsilon}_1\tilde{p}_2 & \tilde{\epsilon}_1\tilde{f}_2 & \tilde{f}_1\tilde{\epsilon}_2 & \tilde{f}_1\tilde{f}_2 \end{pmatrix} \begin{pmatrix} N_{RR} \\ N_{RN} \\ N_{NR} \\ N_{NN} \end{pmatrix}$$

here, $\tilde{x} \equiv 1 - x$

- Inverting this matrix allows us to obtain the prompt and non-prompt contributions from signal + control region distributions:

$$\begin{pmatrix} N_{RR} \\ N_{RN} \\ N_{NR} \\ N_{NN} \end{pmatrix} = \frac{1}{(f_1-\epsilon_1)(f_2-\epsilon_2)} \begin{pmatrix} \tilde{f}_1\tilde{f}_2 & -\tilde{f}_1 f_2 & -f_1\tilde{f}_2 & f_1 f_2 \\ -\tilde{f}_1\tilde{\epsilon}_2 & \tilde{f}_1\epsilon_2 & f_1\tilde{\epsilon}_2 & -f_1\epsilon_2 \\ -\tilde{\epsilon}_1\tilde{f}_2 & \tilde{\epsilon}_1 f_2 & \epsilon_1\tilde{f}_2 & -\epsilon_1 f_2 \\ \tilde{\epsilon}_1\tilde{\epsilon}_2 & -\tilde{\epsilon}_1\epsilon_2 & -\epsilon_1\tilde{\epsilon}_2 & \epsilon_1\epsilon_2 \end{pmatrix} \begin{pmatrix} N_{PP} \\ N_{PF} \\ N_{FP} \\ N_{FF} \end{pmatrix}$$

- The prompt lepton efficiency ϵ can be taken from MC or measured using the regular tag-and-probe technique with appropriate baseline selection

Improvements

- All possible improvements to the "misidentification rate" also can apply to the "matrix" method
- Additionally, using an extended likelihood method, including all data used to make the measurement, can be employed to improve the matrix method result [7]

✔
- Ideally, the most accurate method (true with very large statistics and good efficiency definitions)
- Can be applied well outside the scope of non-prompt lepton background estimation

✘
- Very complicated
- The matrix is often ill-conditioned
- Requires very accurate and precise tuning of "misidentification rate" and prompt lepton efficiency
  → otherwise, may lead to very large or negative results

References

[1] CMS Collaboration. Measurement of the differential Drell-Yan cross section in proton-proton collisions at $\sqrt{s} =$ 13 TeV. JHEP 12 059, 2019.
[2] NNPDF Collaboration. The path to proton structure at 1% accuracy. EPJC 82 428, 2022.
[3] David Barney. CMS Slice. CMS-OUTREACH-2018-017.
[4] CMS Collaboration. Measurement of the Drell-Yan forward-backward asymmetry at high dilepton masses in proton-proton collisions at $\sqrt{s} =$ 13 TeV. JHEP 08 063, 2022.
[5] CMS Collaboration. Measurement of the mass dependence of the transverse momentum of lepton pairs in Drell–Yan production in proton–proton collisions at $\sqrt{s} =$ 13 TeV. EPJC 83 628, 2023.
[6] CMS Collaboration. Measurement of the Drell-Yan forward-backward asymmetry and of the effective leptonic weak mixing angle using proton-proton collisions at 13 TeV. CMS-PAS-SMP-22-010.
[7] T. Gillam, Ch. Lester, Improving estimates of the number of 'fake' leptons and other mis-reconstructed objects in hadron collider events: BoB's your UNCLE. JHEP 11 031, 2014.

[1] [2] [3] [4] [5] [6] [7]