Fake Lepton Backgrounds

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

·1579·

Marijus Ambrozas marijus.ambrozas@cern.ch

What are "fake" leptons?



- In Drell-Yan analysis, we are interested in signal leptons coming from the primary vertex
 - These are often called "prompt" leptons
- In some rare cases, RECO leptons that do not originate from the primary vertex still manage to pass the lepton cuts
- There are a few types of such leptons:
 - 1. "Non-prompt leptons" (dominant contribution):
 - > Real leptons can be produced in heavy-flavor decay chains inside jets
 - When the lepton has dominant momentum inside the jet, it can pass the isolation cuts
 - Alternatively (but rarely), photon conversions can produce electrons and in-flight π/K decays can produce muons
 - 2. "Instrumental fakes" (nearly-negligible contribution):
 - Muon channel: jet hadronic punch-through may create signals in muon detectors and be reconstructed as a muon
 - \triangleright Electron channel: π⁰-rich jets may have high electromagnetic content and be misreconstructed as electrons
- When we say "fake lepton backgrounds", we usually mean all of the above
 - Though, "fake leptons" is a slang and should not be used in official documents
 - "Non-prompt" or "misidentified" are the most common terms in publications

(a)
$$l \xrightarrow{\text{Real "prompt" lepton}} l$$

(b) $q \xrightarrow{\text{Jet}} l$

"Fake" ("non-prompt") lepton

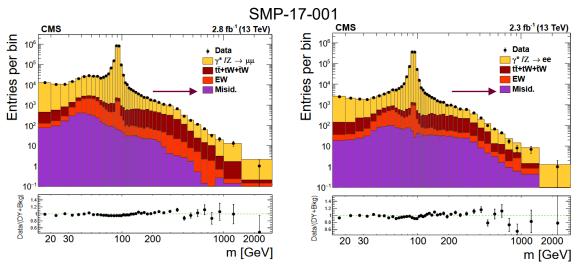
(c1)
$$q \longrightarrow 1$$

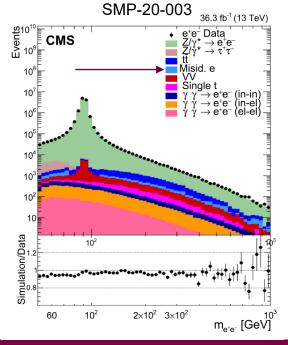
$$(c2) q \longrightarrow l$$

Fake lepton backgrounds



- Drell-Yan final state has 2 leptons, so the fake lepton background consists of 2 main types
 - 1. 1 prompt lepton, 1 fake lepton
 - 2. 2 fake leptons
- In principle, any process could contribute to the fake lepton background
 - > Even DY itself, if there are ISR jets
- However, there are two (three) most dominant ones:
 - 1. W+jets (1 fake lepton)
 - QCD multijet (2 fake leptons)
 - 3. γ+jets (2 fake leptons, only in the electron channel, almost negligible)
- Fake lepton background is the most prominent in the low mass and low lepton p_T region
- It is negligible in the Z peak region but should be estimated (at least) outside it to achieve high precision



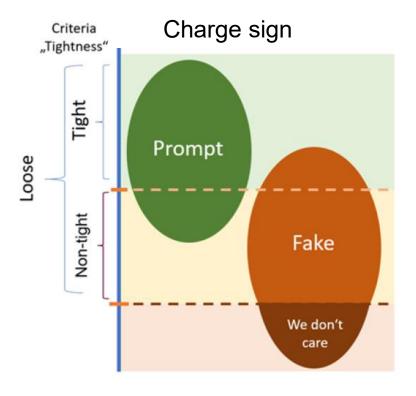




- Fake lepton background cannot be estimated from MC
 - Very large cross section and very low fake lepton production probability
 - Extremely large MC datasets should be produced to collect sufficient statistics after event selection
 - Also, the simulation accuracy is not great
- Therefore, we rely on data-driven methods
- There are many data-driven methods available:



- Fake lepton background cannot be estimated from MC
 - Very large cross section and very low fake lepton production probability
 - Extremely large MC datasets should be produced to collect sufficient statistics after event selection
 - Also, the simulation accuracy is not great
- Therefore, we rely on data-driven methods
- There are many data-driven methods available:
 - Matrix method
 - In theory, the most sophisticated, the least reliant on MC
 - In practice, the most difficult to properly implement (the matrix is ill-conditioned)

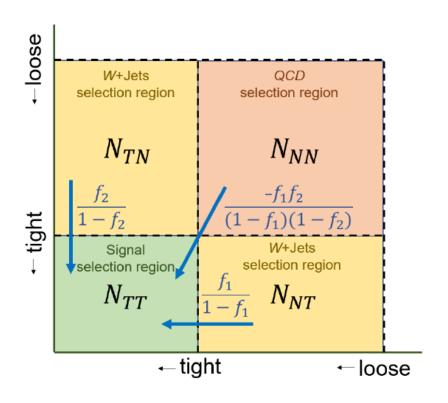


$$\begin{pmatrix} N_{PP} \\ N_{PF} \\ N_{FP} \\ N_{FF} \end{pmatrix} = \frac{1}{(f_1 - p_1)(f_2 - p_2)} \begin{pmatrix} \tilde{f}_1 \tilde{f}_2 & -\tilde{f}_1 f_2 & -f_1 \tilde{f}_2 & f_1 f_2 \\ -\tilde{f}_1 \tilde{p}_2 & \tilde{f}_1 p_2 & f_1 \tilde{p}_2 & -f_1 p_2 \\ -\tilde{p}_1 \tilde{f}_2 & \tilde{p}_1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



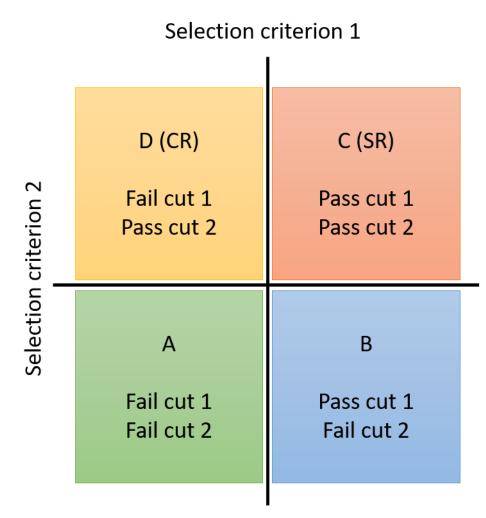
6

- Fake lepton background cannot be estimated from MC
 - Very large cross section and very low fake lepton production probability
 - Extremely large MC datasets should be produced to collect sufficient statistics after event selection
 - Also, the simulation accuracy is not great
- Therefore, we rely on data-driven methods
- There are many data-driven methods available:
 - Matrix method
 - In theory, the most sophisticated, the least reliant on MC
 - In practice, the most difficult to properly implement (the matrix is ill-conditioned)
 - Fake rate method
 - A simplification of the matrix method
 - Does not need the "prompt rate" but more reliant on MC
 - Still not easy to implement, needs different datasets and prescaled triggers



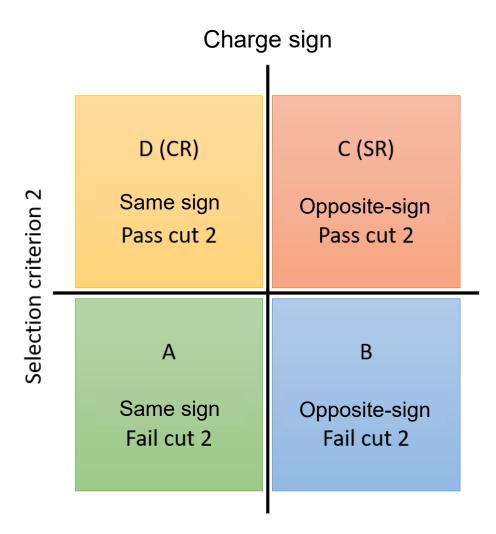


- Fake lepton background cannot be estimated from MC
 - Very large cross section and very low fake lepton production probability
 - Extremely large MC datasets should be produced to collect sufficient statistics after event selection
 - Also, the simulation accuracy is not great
- Therefore, we rely on data-driven methods
- There are many data-driven methods available:
 - Matrix method
 - In theory, the most sophisticated, the least reliant on MC
 - In practice, the most difficult to properly implement (the matrix is ill-conditioned)
 - 2. Fake rate method
 - A simplification of the matrix method
 - Does not need the "prompt rate" but more reliant on MC
 - > Still not easy to implement, needs different datasets and prescaled triggers
 - ABCD method
 - Relatively easy to implement
 - Performance depends on the chosen control regions



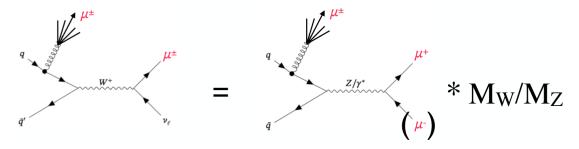


- Fake lepton background cannot be estimated from MC
 - Very large cross section and very low fake lepton production probability
 - Extremely large MC datasets should be produced to collect sufficient statistics after event selection
 - Also, the simulation accuracy is not great
- Therefore, we rely on data-driven methods
- There are many data-driven methods available:
 - Matrix method
 - In theory, the most sophisticated, the least reliant on MC
 - In practice, the most difficult to properly implement (the matrix is ill-conditioned)
 - 2. Fake rate method
 - A simplification of the matrix method
 - Does not need the "prompt rate" but more reliant on MC
 - > Still not easy to implement, needs different datasets and prescaled triggers
 - ABCD method
 - Relatively easy to implement
 - Performance depends on the chosen control regions
 - 4. Same-sign method
 - A subset of ABCD method





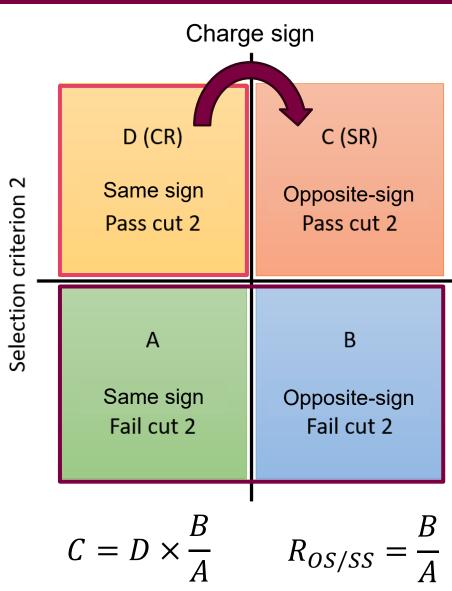
- Fake lepton background cannot be estimated from MC
 - Very large cross section and very low fake lepton production probability
 - Extremely large MC datasets should be produced to collect sufficient statistics after event selection
 - Also, the simulation accuracy is not great
- Therefore, we rely on data-driven methods
- There are many data-driven methods available:
 - Matrix method
 - In theory, the most sophisticated, the least reliant on MC
 - In practice, the most difficult to properly implement (the matrix is ill-conditioned)
 - 2. Fake rate method
 - A simplification of the matrix method
 - Does not need the "prompt rate" but more reliant on MC
 - Still not easy to implement, needs different datasets and prescaled triggers
 - ABCD method
 - Relatively easy to implement
 - Performance depends on the chosen control regions
 - 4. Same-sign method
 - A subset of ABCD method
 - Other methods
 - > Z→W?



Same-sign method – the basic idea



- In our DY analysis, we used the fake rate method initially but abandoned the idea
 - Some information important for the fake rate measurement was missing in nanoAOD
- Currently, we are using the same-sign method
 - Mainly because of its simplicity
- The method works as follows:
 - Perform the usual event selection but require the leptons to have the same charge
 - 2. Obtained sample should be (almost) signal-free and enriched with fake lepton backgrounds
 - 3. Estimate the fake lepton background by MC subtraction in the same-sign sample
 - 4. Extrapolate the result to the opposite-sign region by using the transfer ratio $R_{OS/SS}(\sim 1)$ taken from the sideband region
- Considering only QCD, we could expect the ratio to be ≈1 in the electron channel and ≈1.75 in the muon channel (CMS-AN-13-180)



Same-sign method – the nuances



- Implementing the same-sign method in a real (precision) analysis has a few more challenges:
- Same-sign method does not differentiate between singlefake and double-fake events
 - Estimates both QCD and W+Jets at the same time
 - If the R_{OS/SS} is different between QCD and W+Jets, one needs to evaluate the average R_{OS/SS} that takes into account relative QCD and W+Jets fractions
 - Solution: very carefully choose the sideband (A and B) regions to have the relative QCD/W+Jets fraction well represented
- 2. R_{OS/SS} might be a function of some variables
 - ➤ **Solution:** find the event variables R_{OS/SS} is the most sensitive to and measure R_{OS/SS} as a function of those variables
 - ➤ Then, apply the R_{OS/SS} event-by-event on the D region
 - > Validity of this method must be checked with a closure test

- Non-negligible electron charge misidentification probability (only in electron channel)
 - Same-sign electron sample "contaminated" by signal
 - Need to reweight the signal MC to account for charge misidentification differences between data and MC
 - Still the background estimation error around the Z peak is very high, making the result unreliable
 - Solution: discard the result around the Z peak and interpolate from outside it into the Z peak region instead
- We now proceed with the actual implementation of these methods

Same-sign sample: muon channel



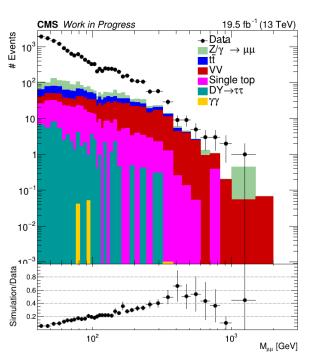
- The big difference between data and MC is attributed to the fake muon background
- We can clearly see that the fake muon background is the most dominant

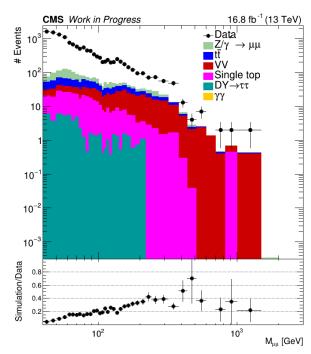
2016preAPV

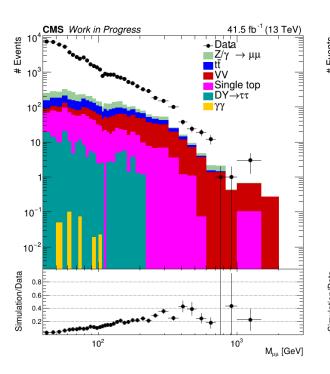
2016postAPV

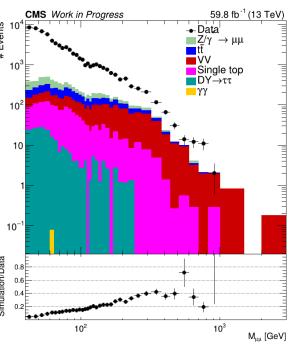
2017

2018









Same-sign sample: electron channel



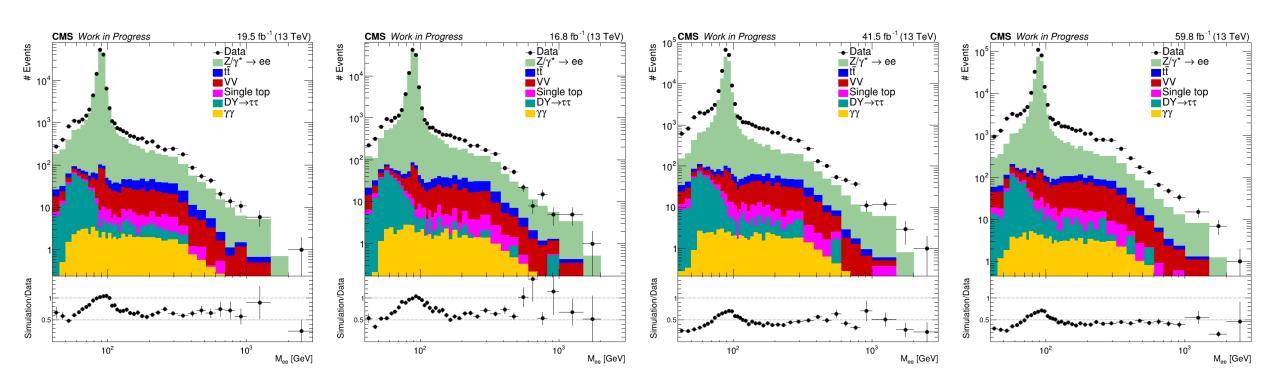
- The big difference between data and MC is attributed to the fake electron background
- We can clearly see that the largest contribution comes from DY because of the electron charge misidentification
- Charge misidentification is very different between data and MC in 2017 and 2018 → need to correct MC (next slide)

2016preAPV

2016postAPV

2017

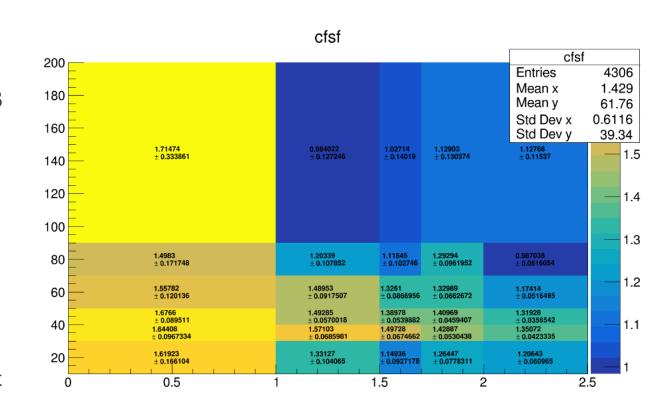
2018



Electron charge misidentification



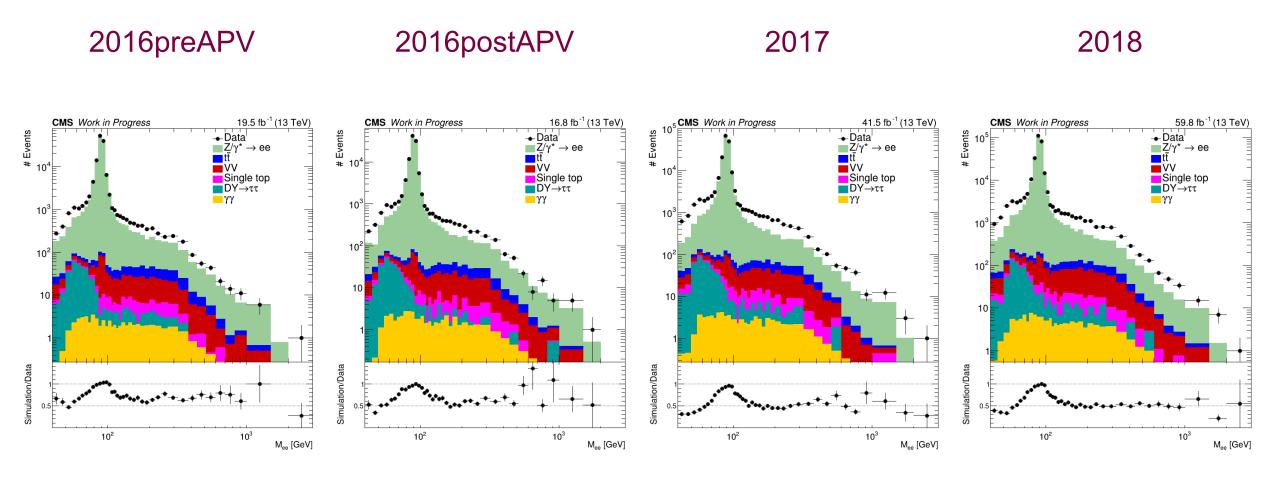
- Charge misidentification SFs were used to reweight MC so its charge misidentification probability better matches that of real experiment
- We are using charge misID SFs provided by DY AFB group
- The SFs are applied event-by-event as follows:
 - 1. Match the electrons with their GEN counterparts
 - Check which electrons have their RECO charge ≠ GEN charge
 - 3. For those electrons, pick the corresponding SF value, depending on its p_T and $|\eta|$
 - 4. Multiply the SF value with the total weight of the event
- Corrected result in next slide



Same-sign sample: electron channel (corrected)



 We can notice the MC/Data agreement at the Z peak becoming a lot better than before (especially for 2017 and 2018)



Same-sign sample: eµ channel



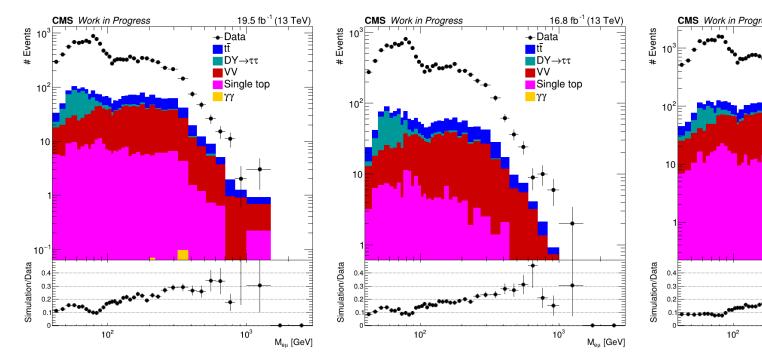
- Fake lepton backgrounds are also important for the eμ method (used to improve the top quark background estimation)
- Therefore, we perform the same procedure on eµ events
- The charge misidentification correction was also applied on the electron here

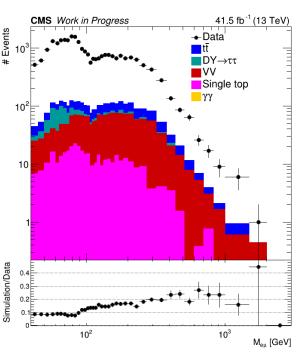
2016preAPV

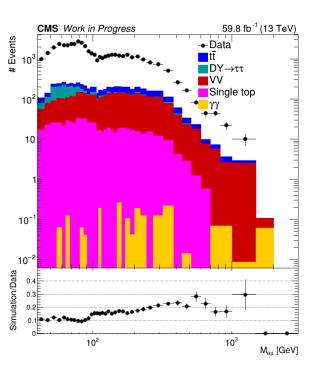
2016postAPV

2017

2018



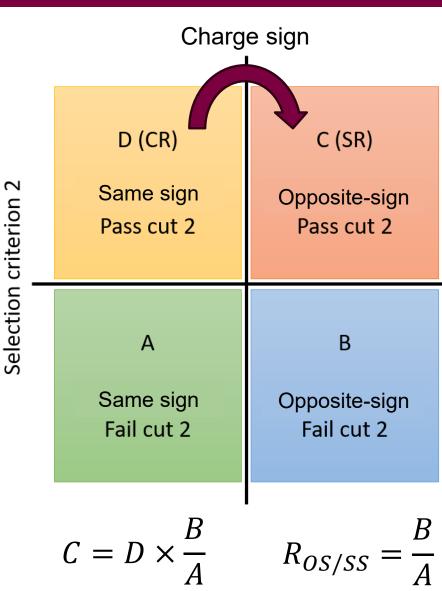




OS/SS ratio estimation



- We need to estimate the OS/SS ratio to extrapolate the fake lepton background from SS to OS region
- W+Jets and QCD have different R_{OS/SS} values (backup)
- Since we are estimating both W+Jets and QCD at the same time, we want to get the average R_{OS/SS} right
- We need a sideband that has the relative amount of W+Jets vs.
 QCD as close as possible to the signal region
- Two main sidebands were considered:
 - Regular event selection but **one** muon (electron) fails Tight PF ISO (Medium ID)
 - 2. Regular event selection but **both** muons (electrons) fail Tight PF ISO (Medium ID)
- For the eµ channel, we have two types of sideband 1 (one with failing muon, another with failing electron)

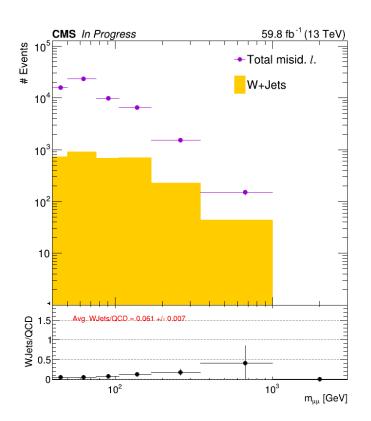


Choosing the best sideband: muon channel

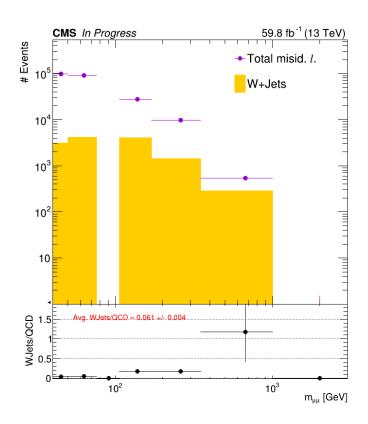


- Only 2018 shown, other eras similar
- Sideband 1 has very similar W+Jets/QCD ratio to the signal region
- Thus, we use sideband 1 to evaluate R_{os/ss} in the muon channel

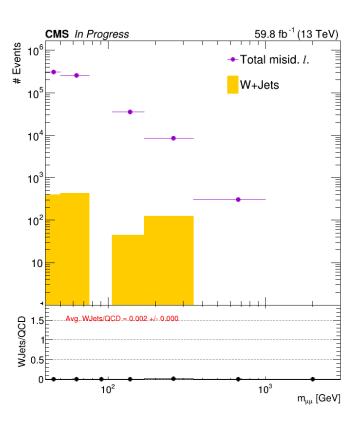
Signal region



Sideband 1



Sideband 2

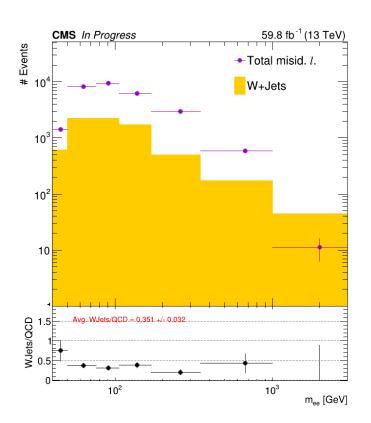


Choosing the best sideband: electron channel

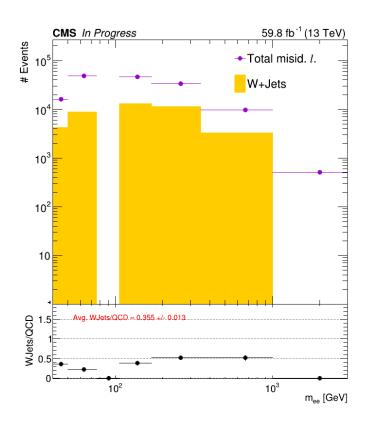


- Only 2018 shown, other eras similar
- Sideband 1 has very similar W+Jets/QCD ratio to the signal region
- Thus, we use sideband 1 to evaluate R_{OS/SS} in the electron channel

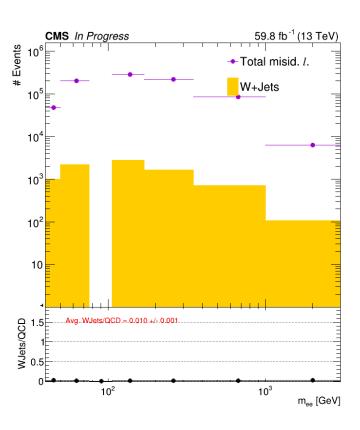
Signal region



Sideband 1



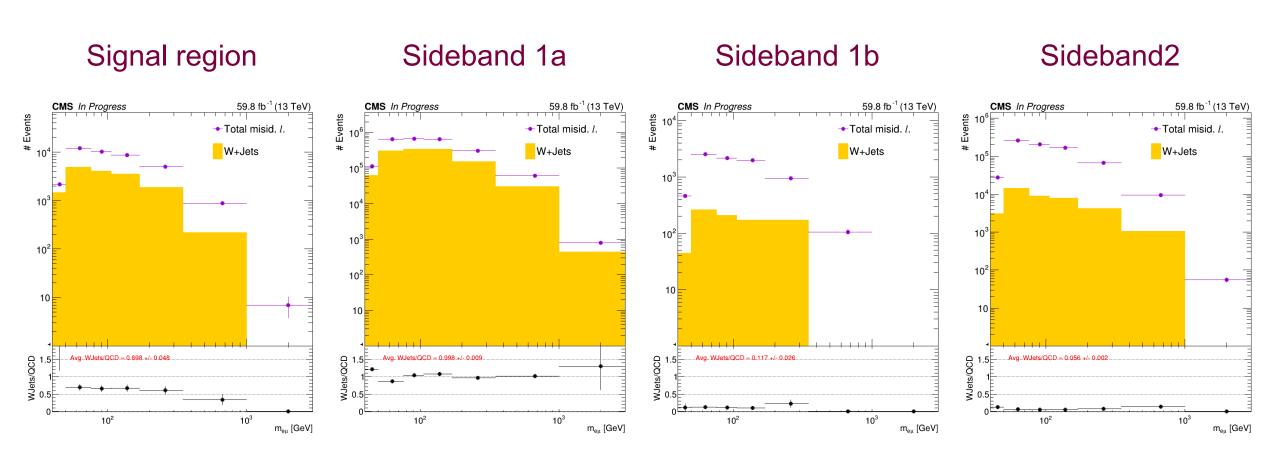
Sideband 2



Choosing the best sideband: eµ channel



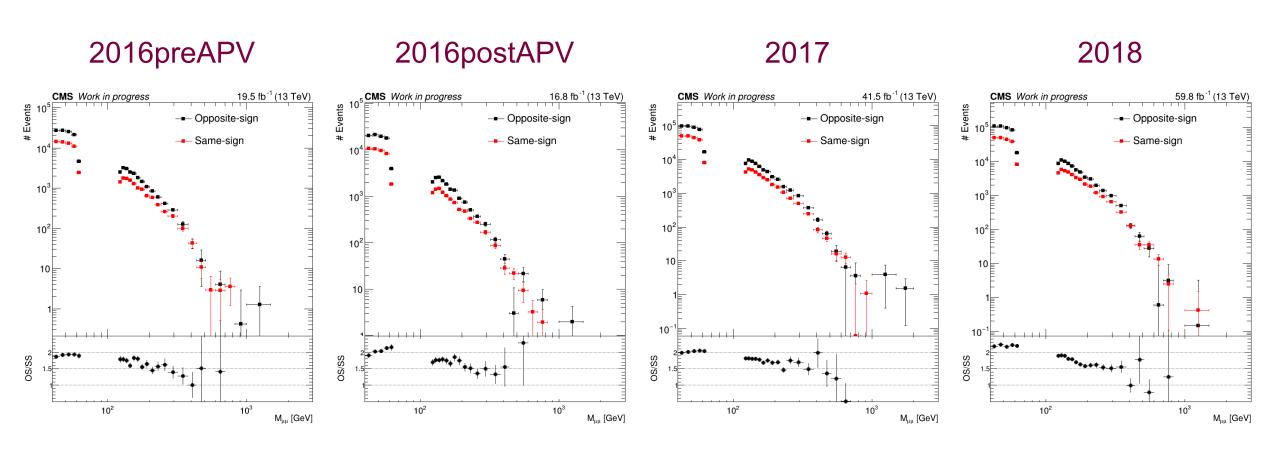
- Only 2018 shown, other eras similar
- Sideband 1a invert the electron cut, 1b invert the muon cut
- We choose sideband 1a to evaluate R_{os/ss} in the eμ channel as it most closely resembles the signal region



OS/SS ratio: muon channel



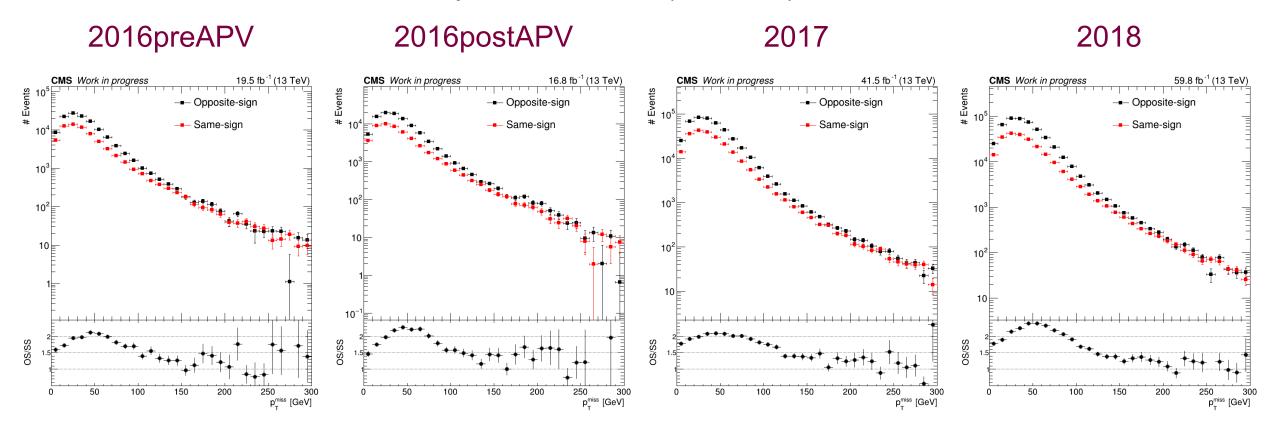
- Comparing OS and SS distributions in the muon channel sideband 1
- R_{OS/SS} changes as a function of dilepton mass
- Check if there are strong dependencies on other variables (next slide)



OS/SS ratio: muon channel (II)



- R_{OS/SS} also has strong dependence on missing p_T (possibly due to separation between W+Jets and QCD w.r.t. this variable)
- Therefore, we measure R_{OS/SS} in the muon channel as a 2D function of mass and missing p_T
- The 2D ratio is fitted with an analytical 2D function (next slide)



2D OS/SS ratio: muon channel



- Different colors show different dilepton mass values
 - But the function is not binned in mass, it is a smooth distribution

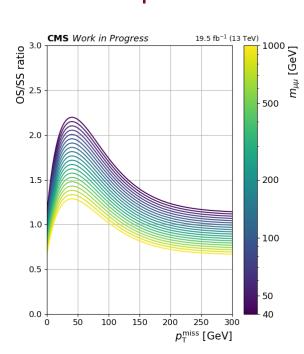
$$R_{\text{OS/SS}}(m_{\mu\mu}, E_{\text{T}}^{\text{miss}}) = \left[a_0 + a_1 \log_{10}(m_{\mu\mu})\right] \left[a_2 - a_3 E_{\text{T}}^{\text{miss}} \exp\left(-a_4 E_{\text{T}}^{\text{miss}}\right)\right]$$

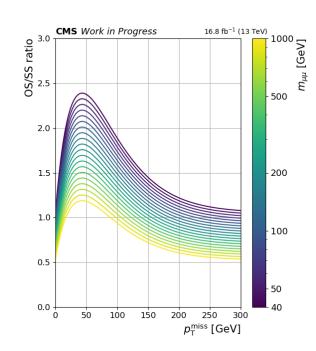
2016preAPV

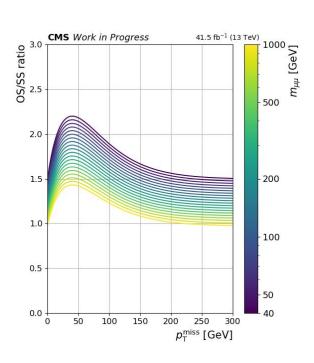
2016postAPV

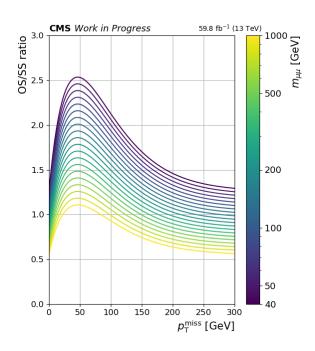
2017

2018





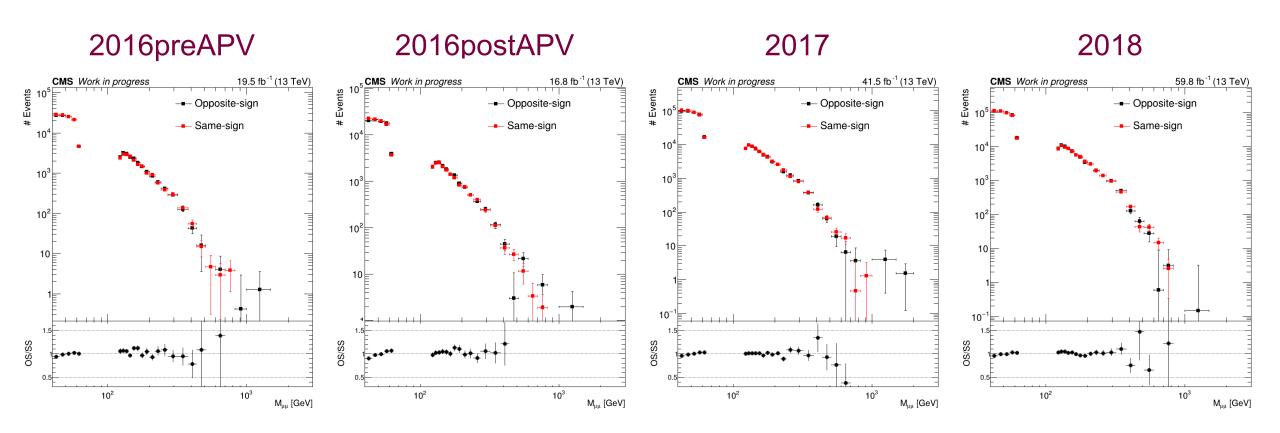




OS/SS ratio validation: muon channel



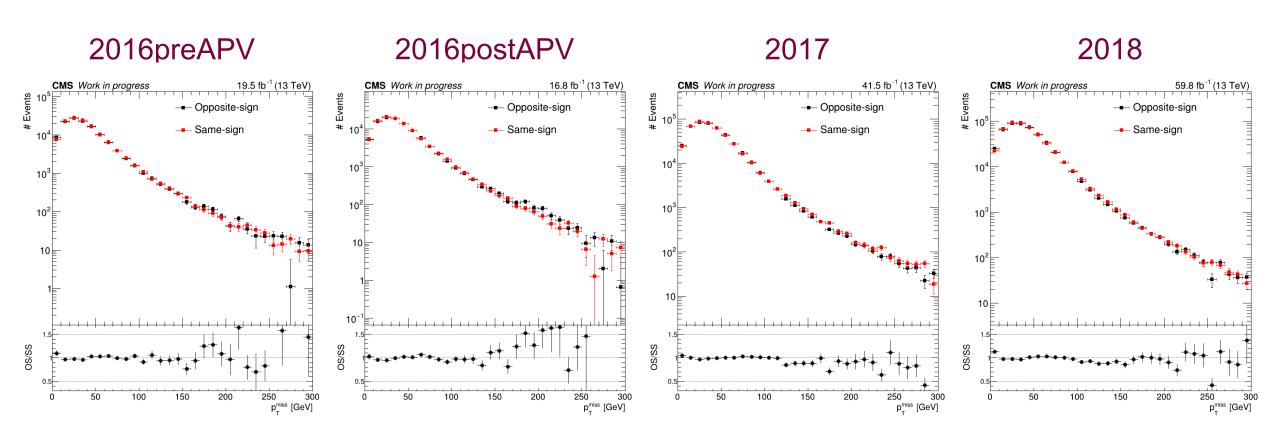
- We validate the obtained ratios by performing a closure test
 - Reweigh the SS events in the sideband itself and see if it matches the OS distribution there
- We can see that the closure is nearly perfect



OS/SS ratio validation: muon channel (II)



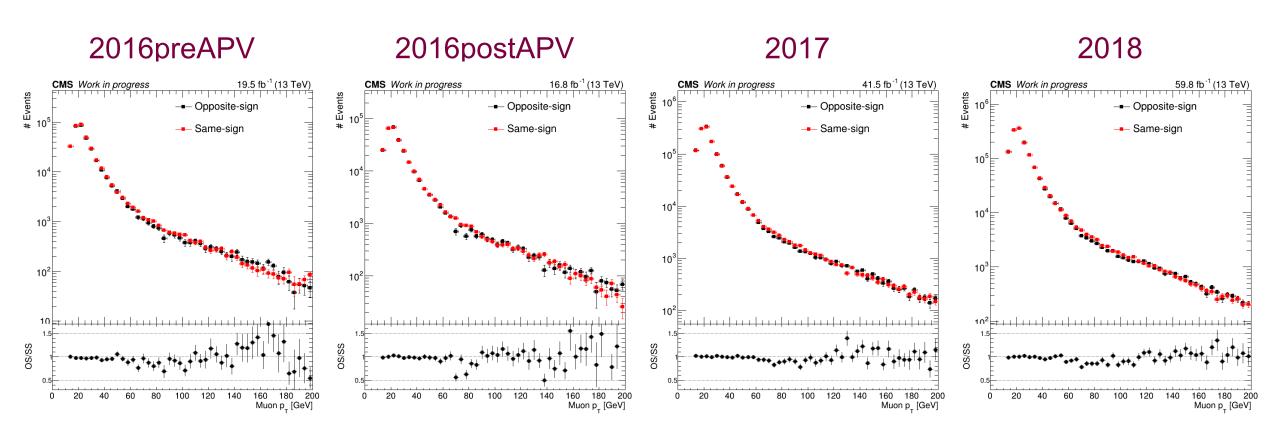
- We validate the obtained ratios by performing a closure test
 - Reweigh the SS events in the sideband itself and see if it matches the OS distribution there
- We can see that the closure is nearly perfect



OS/SS ratio validation: muon channel (III)



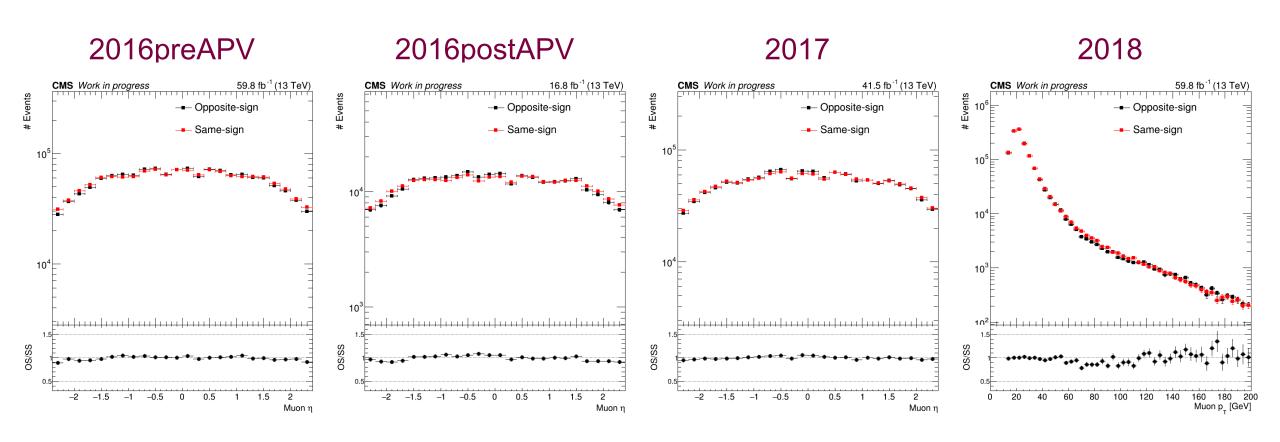
- We validate the obtained ratios by performing a closure test
 - Reweigh the SS events in the sideband itself and see if it matches the OS distribution there
- We can see that the closure is nearly perfect



OS/SS ratio validation: muon channel (IV)



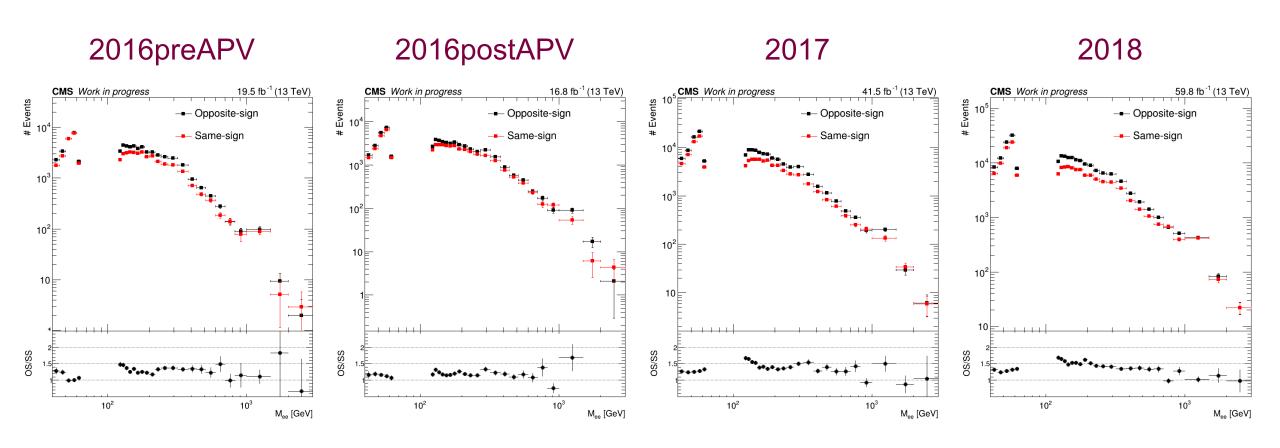
- We validate the obtained ratios by performing a closure test
 - Reweigh the SS events in the sideband itself and see if it matches the OS distribution there
- We can see that the closure is nearly perfect



OS/SS ratio: electron channel



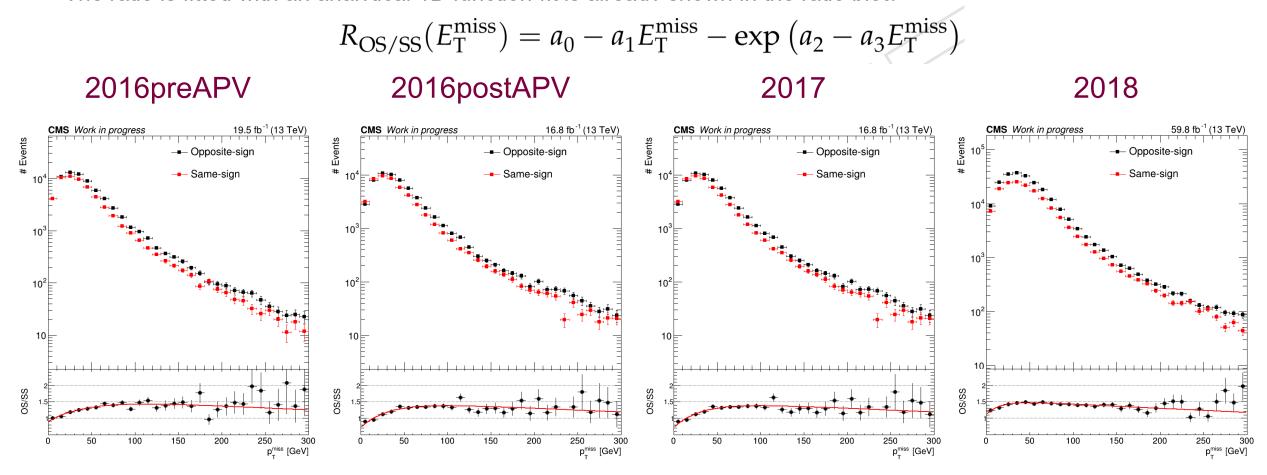
- Comparing OS and SS distributions in the electron channel sideband 1
- R_{OS/SS} almost flat w.r.t. dilepton mass
- Check if there are strong dependencies on other variables (next slide)



OS/SS ratio: electron channel (II)



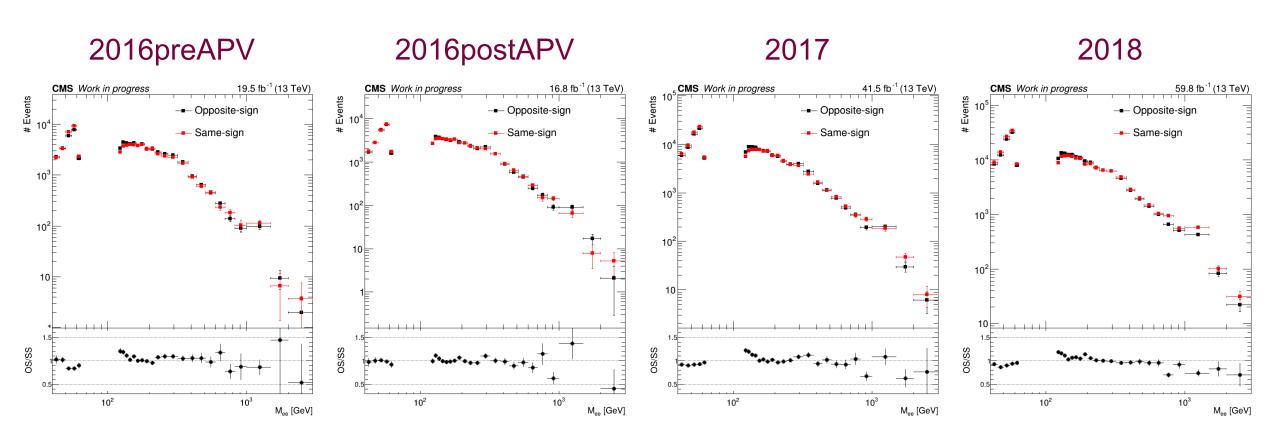
- $R_{OS/SS}$ also has strong dependence on missing p_T (possibly due to separation between W+Jets and QCD w.r.t. this variable)
- Therefore, we measure R_{OS/SS} in the electron channel as a 1D function of missing p_T
- The ratio is fitted with an analytical 1D function (it is already shown in the ratio plot)



OS/SS ratio validation: electron channel



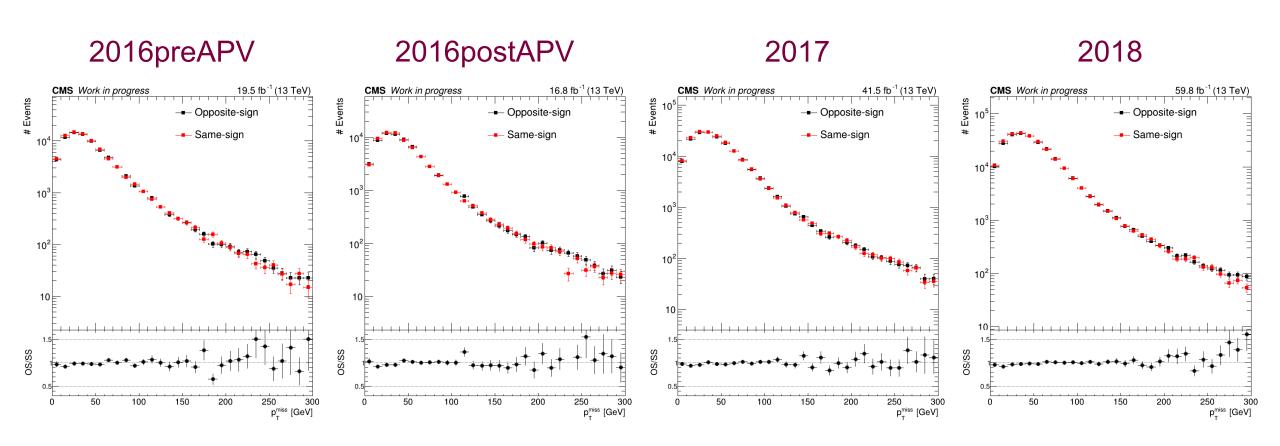
- We validate the obtained ratios by performing a closure test
 - Reweigh the SS events in the sideband itself and see if it matches the OS distribution there
- We can see that the closure is nearly perfect



OS/SS ratio validation: electron channel (II)



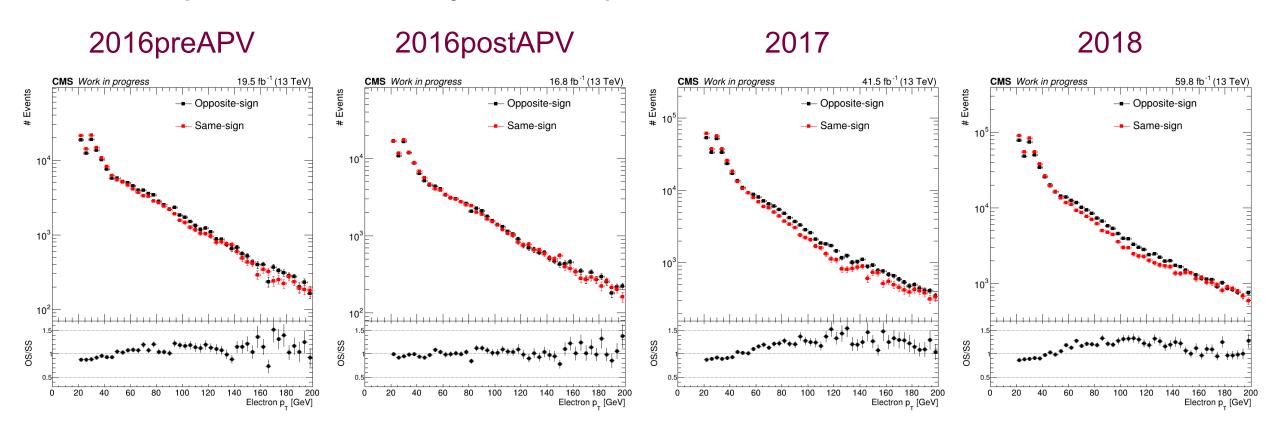
- We validate the obtained ratios by performing a closure test
 - Reweigh the SS events in the sideband itself and see if it matches the OS distribution there
- We can see that the closure is nearly perfect



OS/SS ratio validation: electron channel (III)



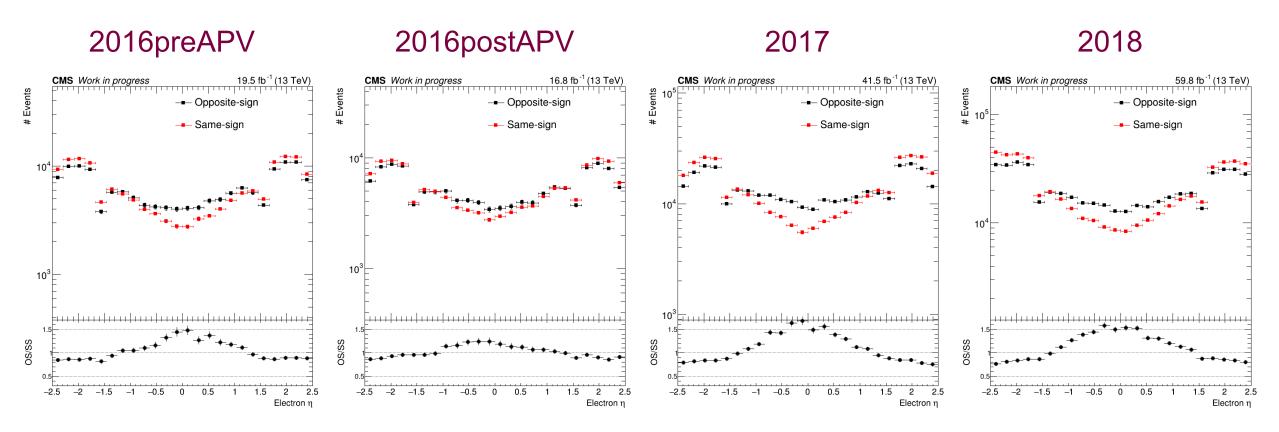
- We validate the obtained ratios by performing a closure test
 - Reweigh the SS events in the sideband itself and see if it matches the OS distribution there
- There is some bias w.r.t electron p_T
 - Acceptable for the current study, should be improved for the 2D measurement



OS/SS ratio validation: electron channel (IV)



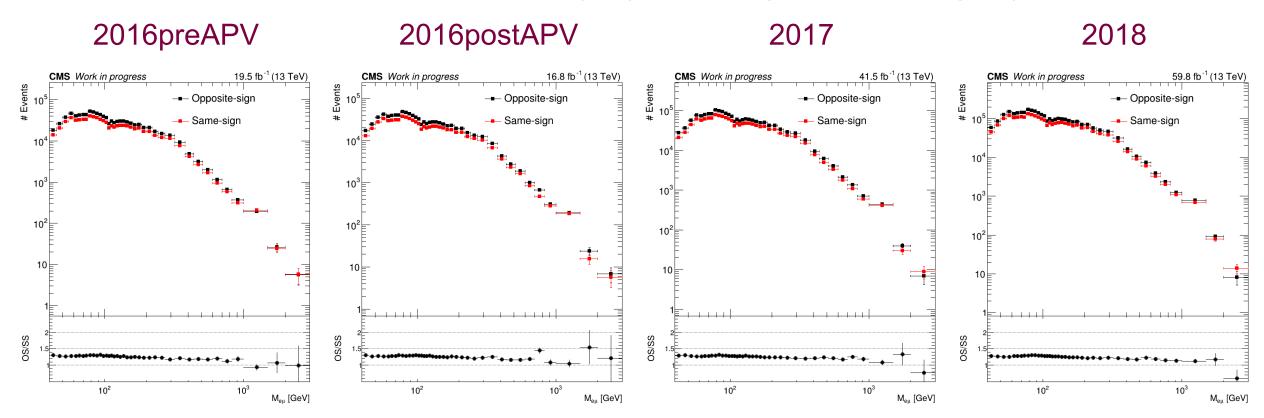
- We validate the obtained ratios by performing a closure test
 - Reweigh the SS events in the sideband itself and see if it matches the OS distribution there
- There is some relatively strong bias w.r.t electron η
 - Acceptable for the current study, must be improved for the 2D measurement



OS/SS ratio: eµ channel

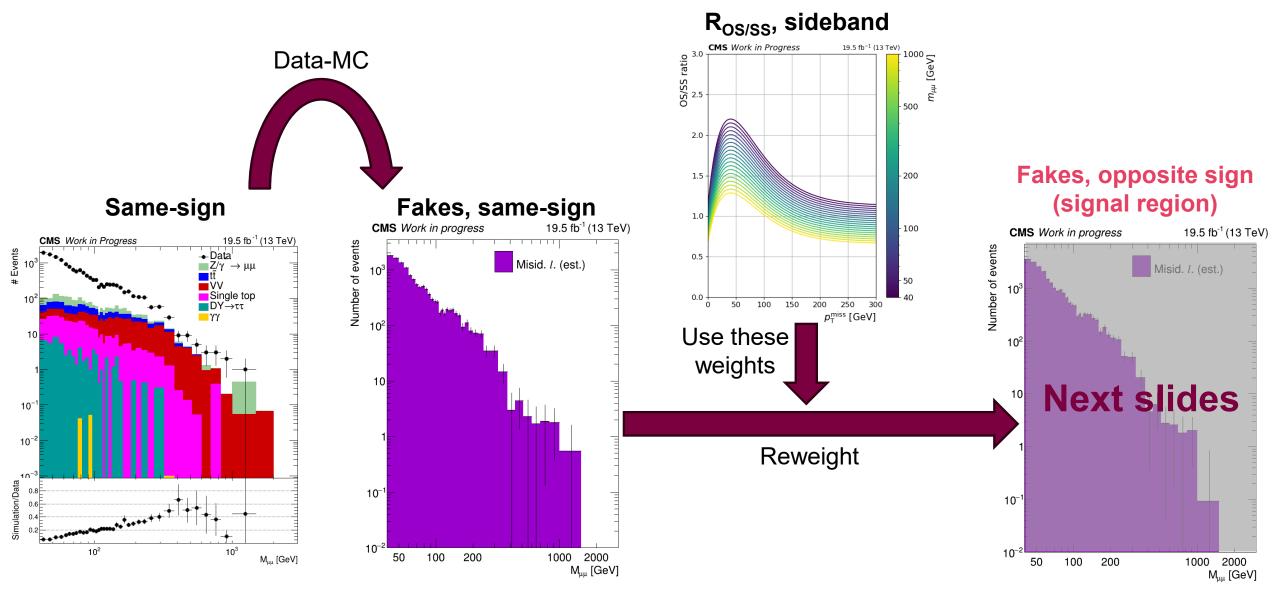


- The ratio is almost perfectly flat w.r.t dilepton mass
 - Similar with other distributions
- Therefore, we use a constant value R_{OS/SS}=1.27 for all eras
- No closure test is needed for a constant ratio (it is just rescaling the whole histogram)



Fake lepton background estimation

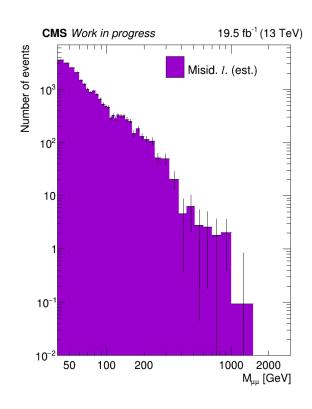




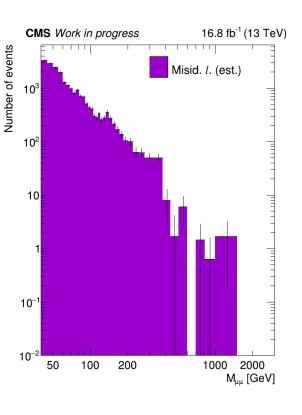
Final fake lepton background (opposite-sign): muon channel



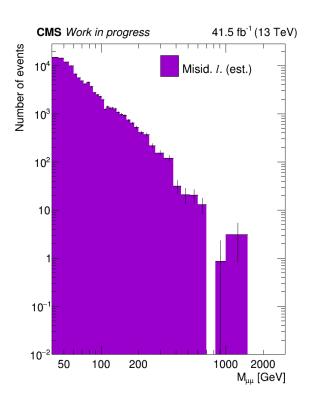
2016preAPV



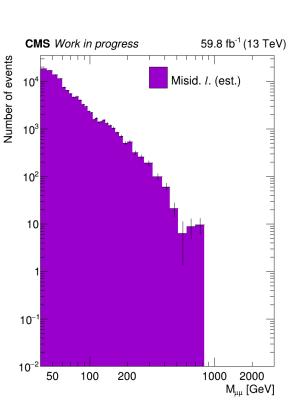
2016postAPV



2017



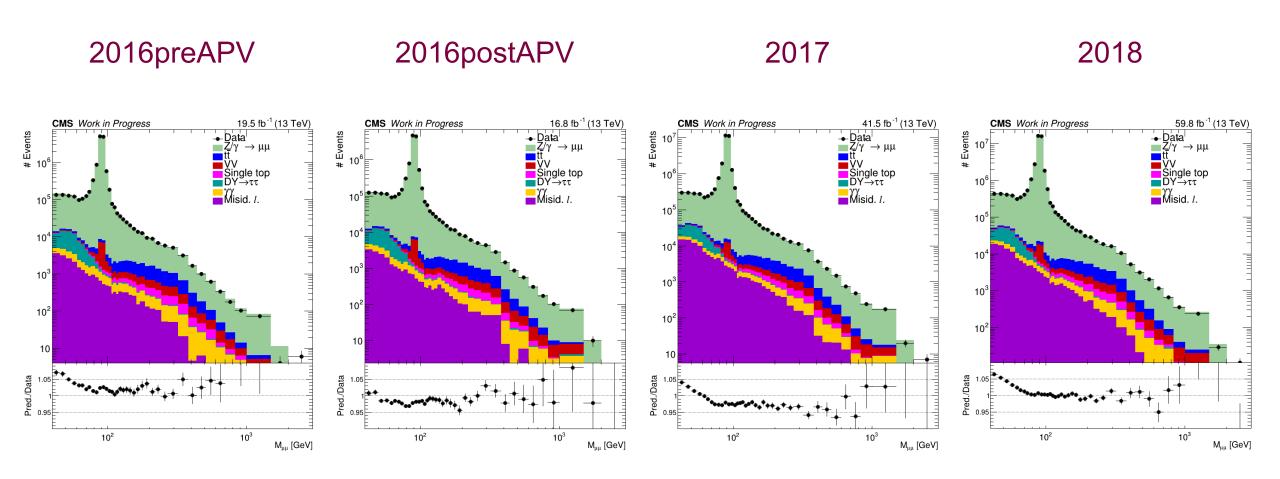
2018



Final fake lepton background (opposite-sign): muon channel (II)



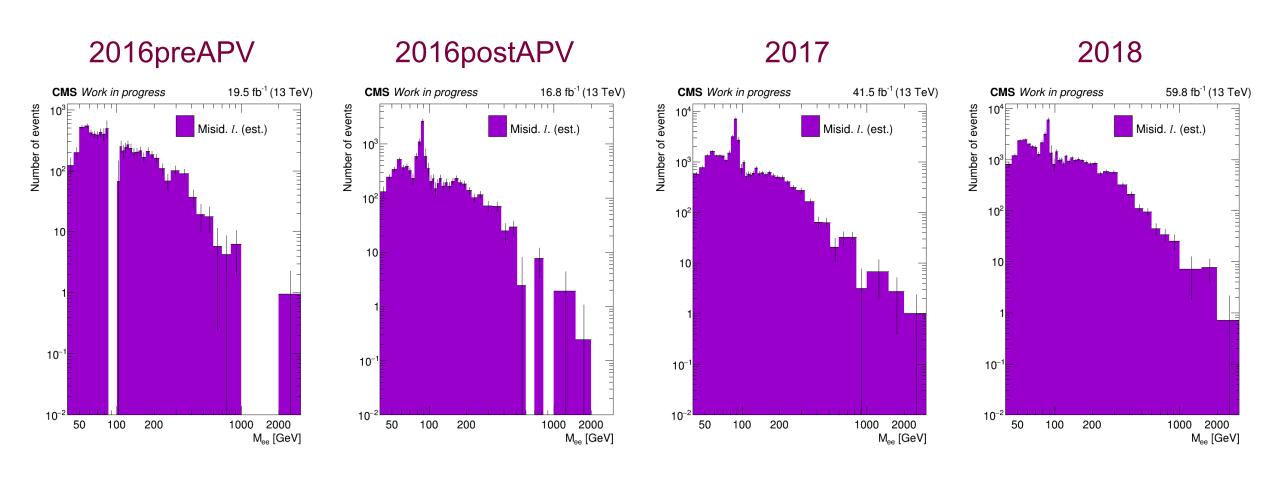
- View of the whole signal region
- Discrepancy at low mass was already apparent before adding fake lepton backgrounds



Fake lepton background (opposite-sign): electron channel



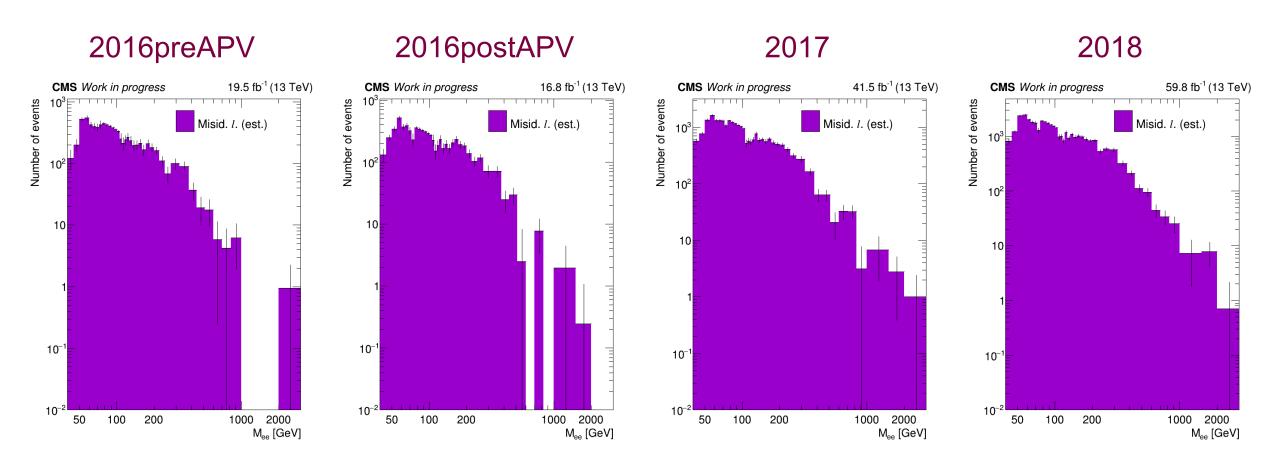
- Result unreliable in the Z peak region
- Fakes are negligible compared to DY in that region
- Nevertheless, we improve the result by linearly interpolating from outside into the Z peak region



Final fake lepton background (opposite-sign): electron channel



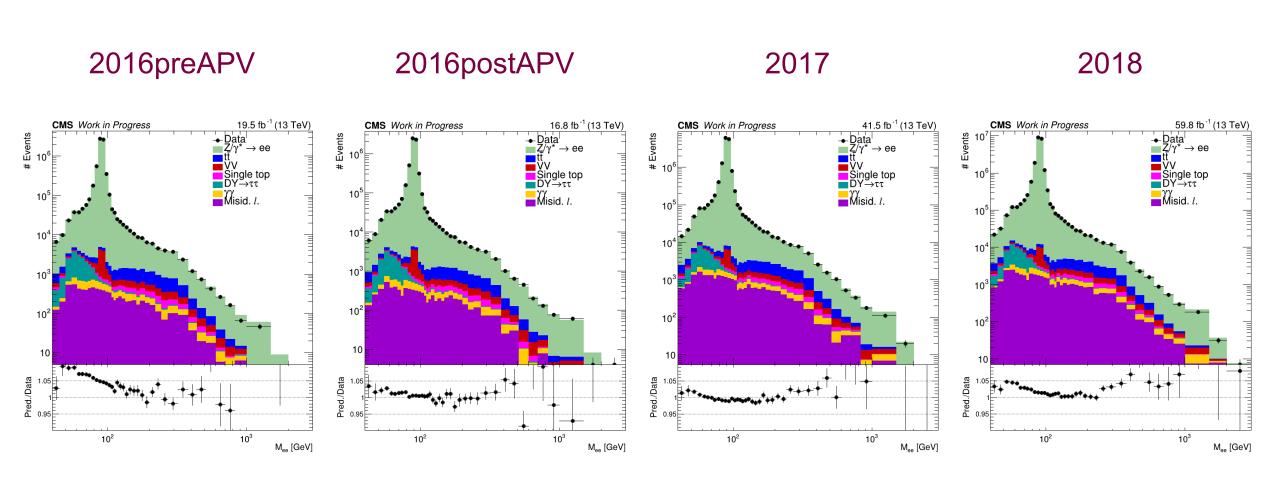
- The apparent "jump" where the interpolated result starts is mainly due to binning
 - I am interpolating the density plot and then reverting it back into a simple histogram.



Final fake lepton background (opposite-sign): electron channel (II)



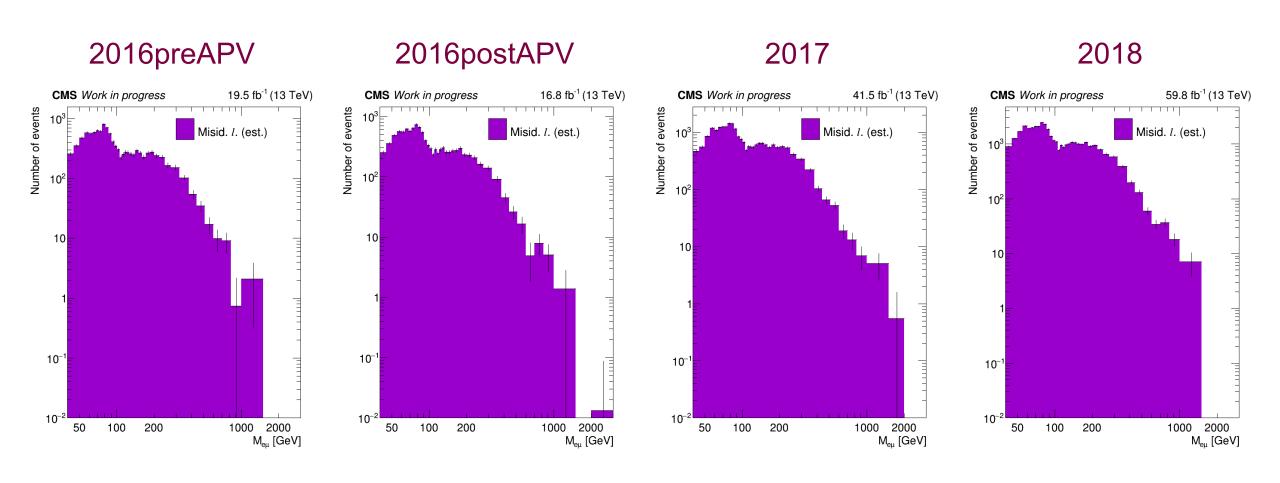
View of the whole signal region



Final fake lepton background (opposite-sign): eµ channel



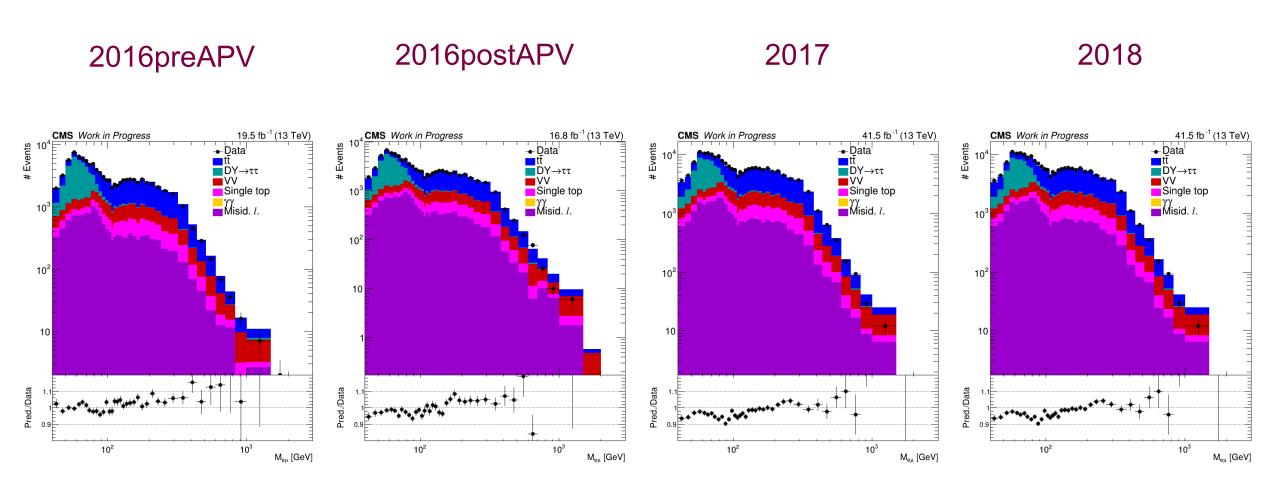
- The shape is reminiscent of the electron channel fakes
- Maybe could be used as a template in electron channel?



Final fake lepton background (opposite-sign): eµ channel (II)



View of the whole "signal" region



Systematic uncertainty

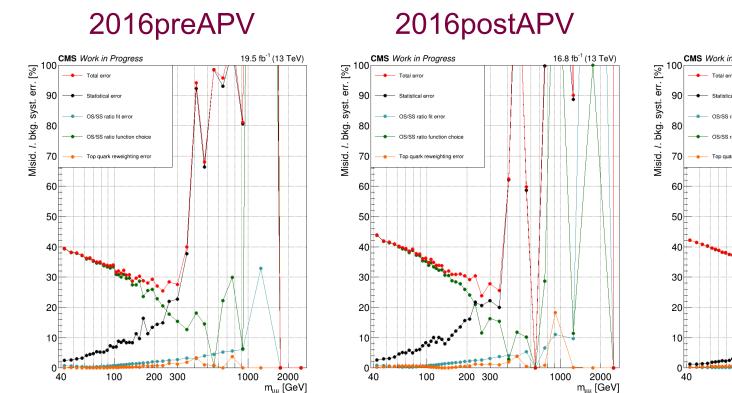


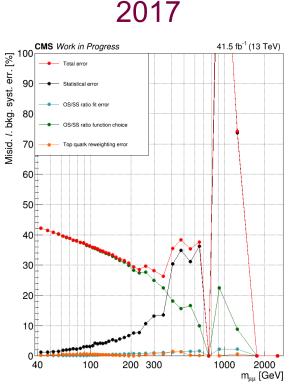
- The following systematic uncertainty sources were considered for the fake background estimation (grey – only for electron channel):
 - 1. R_{OS/SS} fit parameter uncertainties
 - Estimated using fit parameter variations by the error values given from the fit
 - 2. R_{OS/SS} fit function choice
 - Estimated using an alternative R_{OS/SS} value which is just a constant average R_{OS/SS}
 - 3. Same-sign top quark background reweighting using the eµ method
 - > Estimated as the difference between the nominal result and the one where MC top background is not reweighted
 - 4. Electron charge misidentification scale factor uncertainty
 - > Estimated from up/down variation of charge misID SFs within their provided errors
 - 5. Z peak interpolation choice
 - Estimated using an alternative cubic interpolation
- For the eµ channel, considering that no fitting, interpolation, or top quark background reweighting was done to evaluate the fake lepton backgrounds, a constant systematic uncertainty of 10% was assigned

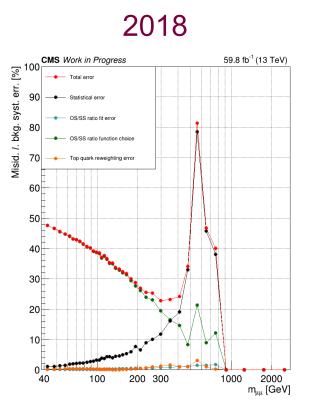
Systematic uncertainty: muon channel



- The total error does not exceed 50% for masses below 400 GeV
- R_{OS/SS} function choice error dominant at low mass
- Statistical error dominant at high mass







Systematic uncertainty: electron channel



- The total error does not exceed 40% for masses below 400 GeV
- R_{OS/SS} function choice error dominant at low mass
- Interpolation error dominant in the Z peak region

2000

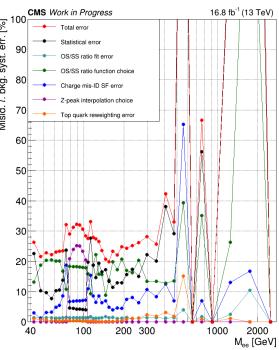
1000

Statistical error dominant at high mass

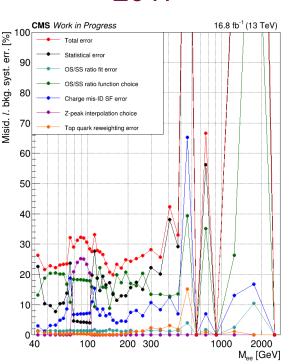
2016preAPV CMS Work in Progress 19.5 fb⁻¹ (13 TeV Total error Statistical error OS/SS ratio function choice Charge mis-ID SF error Z-peak interpolation choice Top quark reweighting error 40 30 20 10

200 300

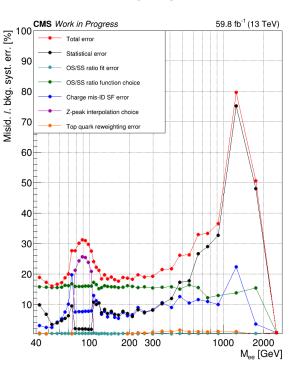




2017



2018



Wrap-up and discussion



- "Fake" lepton backgrounds mainly from non-prompt decays inside jets
- Dominant fake lepton backgrounds in DY analysis: W+Jets, QCD multijet, (γ+jets almost negligible)
- Data-driven methods needed, many different are available
- Same-sign method is the choice for our analysis, mainly due to simplicity
- Same-sign method: estimate the background in the same-sign region, reweight using the transfer ratio estimated from the sideband
- The best sideband where one muon (electron) fails the ISO (Medium ID) cut
- OS/SS transfer ratio estimated as a function of event variables
- Closure test shows that the method is robust but should be improved in the electron channel for the 2D analysis
- Electron channel uses electron charge misidentification correction and histogram interpolation to the Z peak region to obtain a more reliable result
- The final result has errors <50% in the low mass region (where the background is the most relevant)

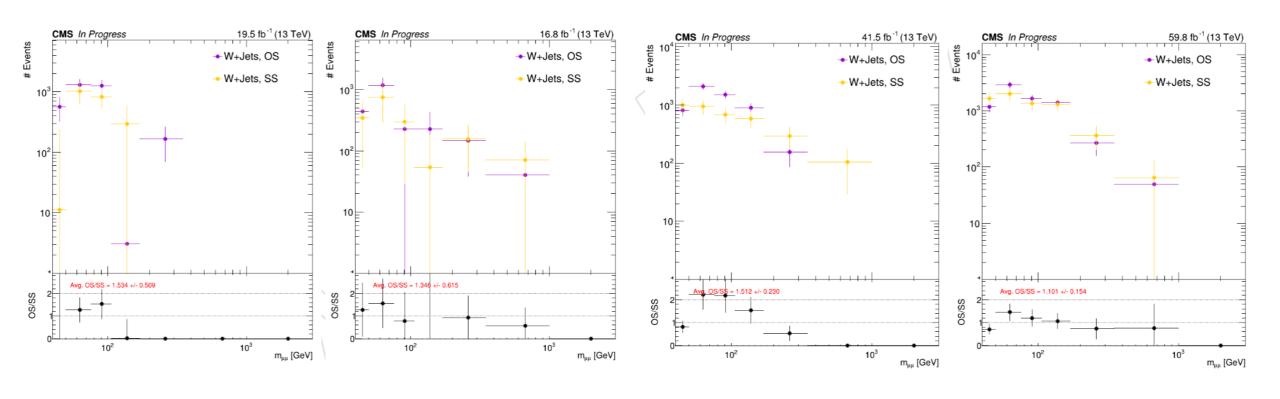


Backup

M. Ambrozas (VU) Fake Lepton Backgrounds 2025-11-05 4

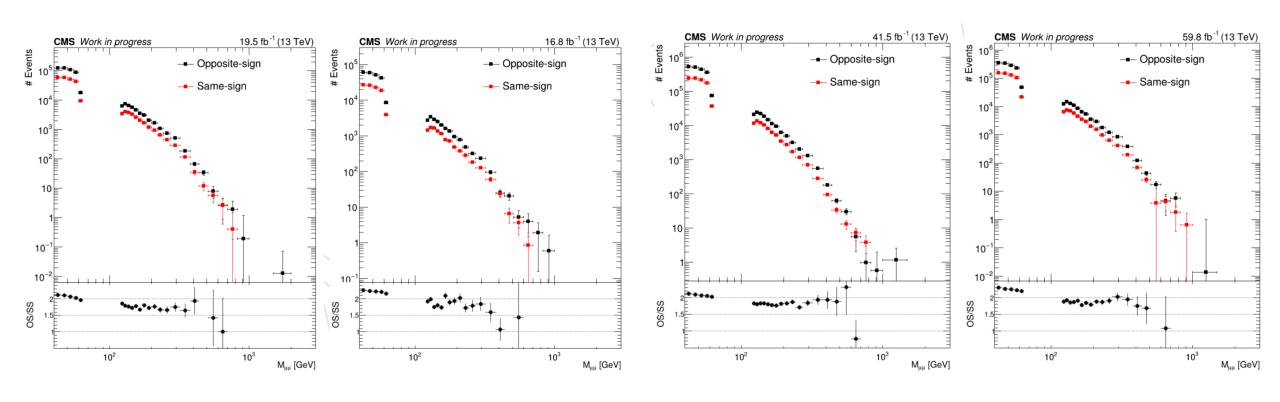
OS/SS ratio for W+Jets → μμ





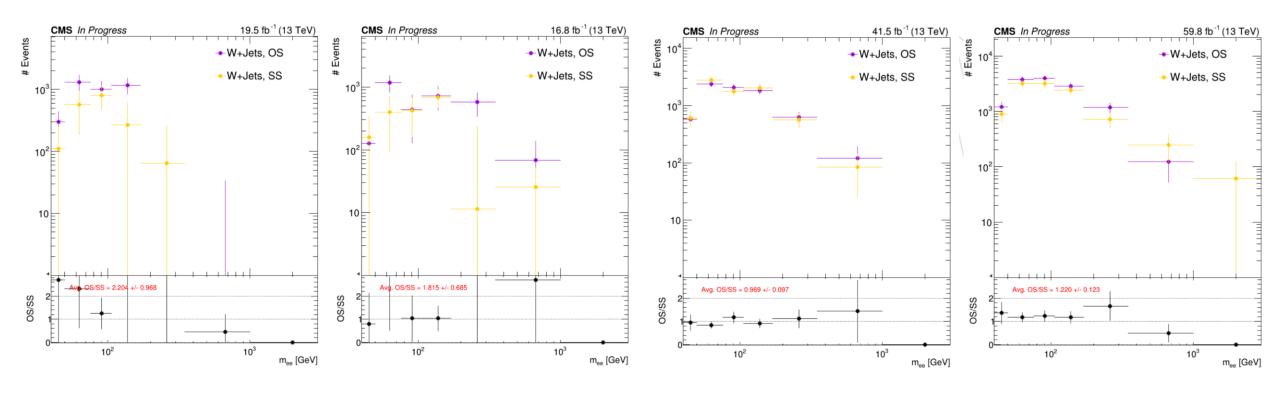
OS/SS ratio for QCD → μμ





OS/SS ratio for W+Jets → ee





OS/SS ratio for QCD → ee



