Study of the Vector Boson Fusion of the Z boson at CMS

21/09/2023

Giorgio Pizzati[a]

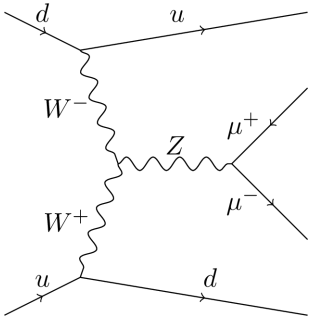
a: Università e INFN di Milano Bicocca

Goals of my Ph.D.

- Perform an analysis for VBF Z + jets for SM with Full Run II Data
  - Isolate as much as possible different types of backgrounds in Control Regions or Categories and perform a normalization fit
  - Isolate EW Zjj in a signal region with the use of kinematic variables cuts or with the use of a simple DNN

- Once the EW Zjj is under control:
  - Perform inclusive and differential cross-section measurements of the EW VBF-Z process
  - Perform EFT studies for dimension 6 operators in order to put constraints on Wilson coefficients with real data

Vector Boson Fusion



- VBF Topology can be identified by two jets with high invariant mass and high separation in pseudorapidity (opposite hemispheres)
- The two leptons originating from the Z are required to have an invariant mass under the Z mass peak
- Interested in Z decay into charged leptons → clean process (no MET)
- Inside the same sample (EW Zjj) one has VBF as other processes

Analysis strategy

- ★ At least 2 leptons of the same flavor opposite charge
- ★ $m_{jj} > 200$ GeV
- ★ $m_{ll} > 50$ GeV

- **0 b-jet**
- **$m_{ll} \in (M_Z-15, M_Z+15)$**

Preselections

0 b-jet

≥ 1 b-jet

VV

Top

$m_{ll} \notin (M_Z-15, M_Z+15)$

At least 1 jet with $p_T < 50$ GeV

DY PU enriched region

At least 2 jets with $p_T > 50$ GeV

DNN output ≥ 0.6

DNN output < 0.6

VBF Z inclusive

DY enriched region

With DNN output ≥ 0.6, in the VBF Z inclusive region, the signal is 10% of all the MCs

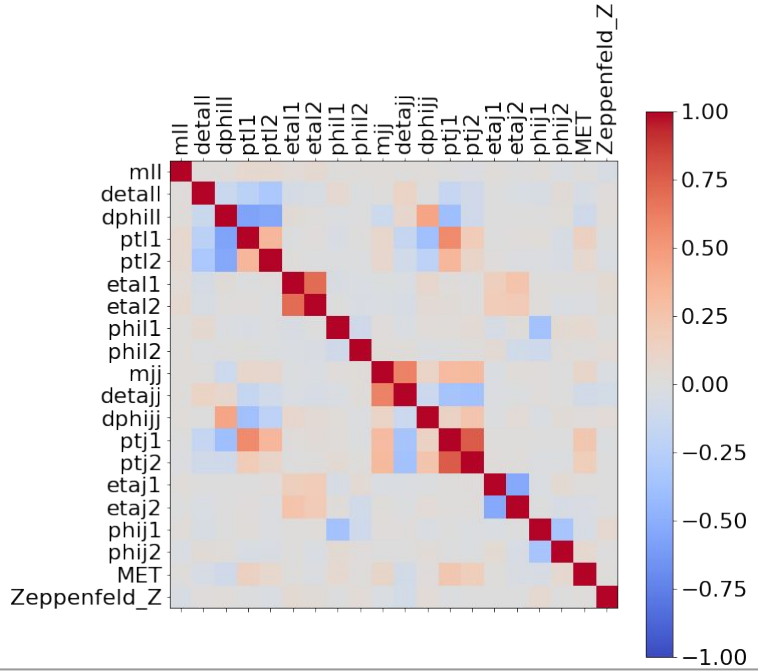
57% of signal efficiency
15% of background efficiency

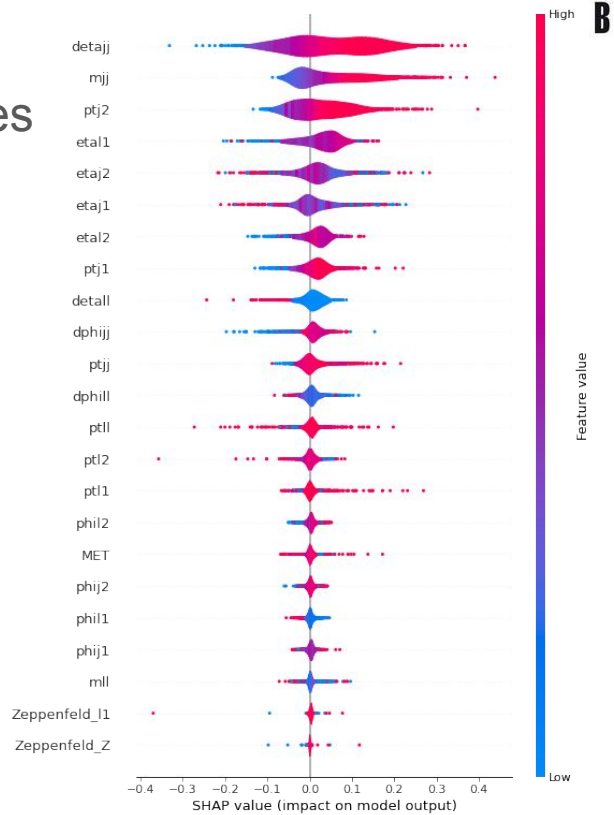
Deep Neural Network

- A Deep Neural Network has been trained to separate signal and background in a phase space with loose and generic cuts
- The chosen model was a DNN with 4 dense layers each with 128 neurons and the
- Training was performed with the Binomial Cross Entropy as loss function
- Since EW VBF-Z and DY appear to be the most difficult to separate, the model is trained only with these two samples
- The separation of the signal with the all the other backgrounds is reached even with the above training
- The DNN output, i.e. the score given by the DNN to each event during the evaluation step, is transformed with a simple function in order to have a flat signal and all the backgrounds peaking at 0
- The function chosen that satisfied this requirement is the cumulative of the DNN output evaluated on the signal

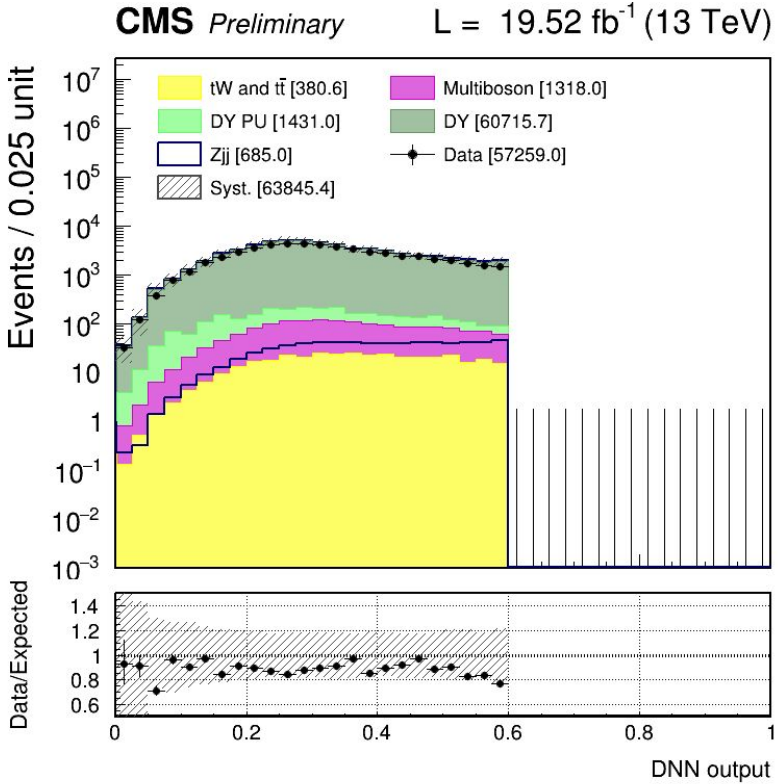
DNN input variables

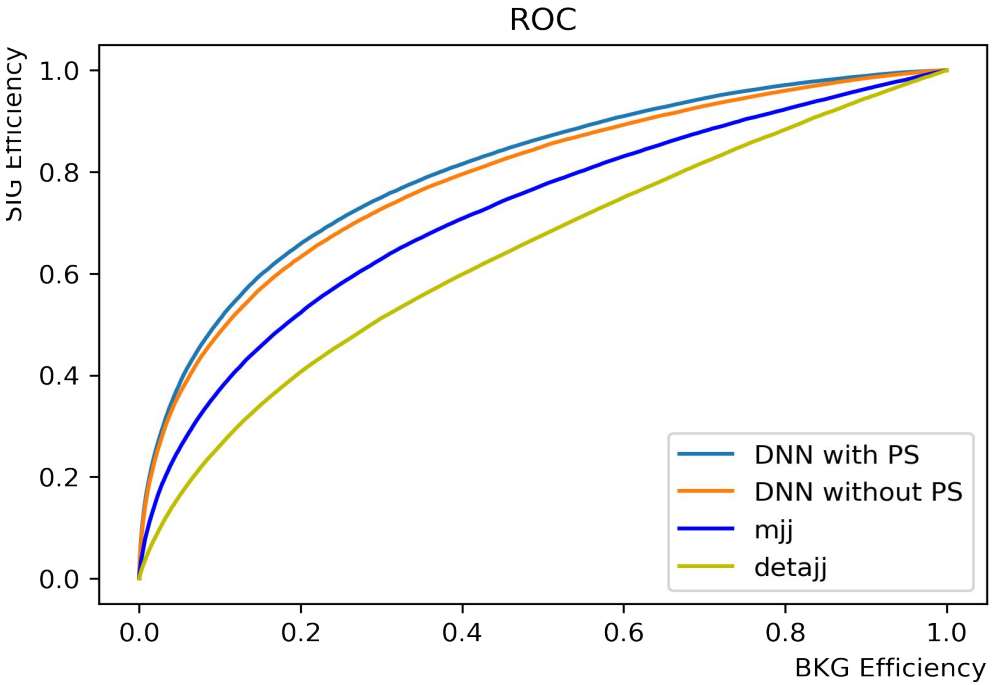
- Do not use Parton Shower sensible variables





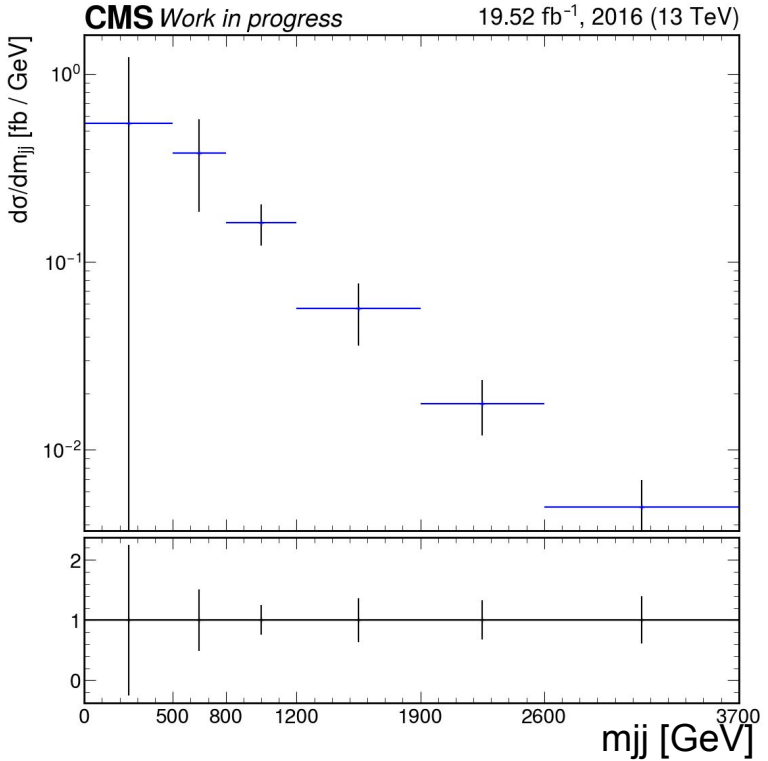
DNN evaluation and performances





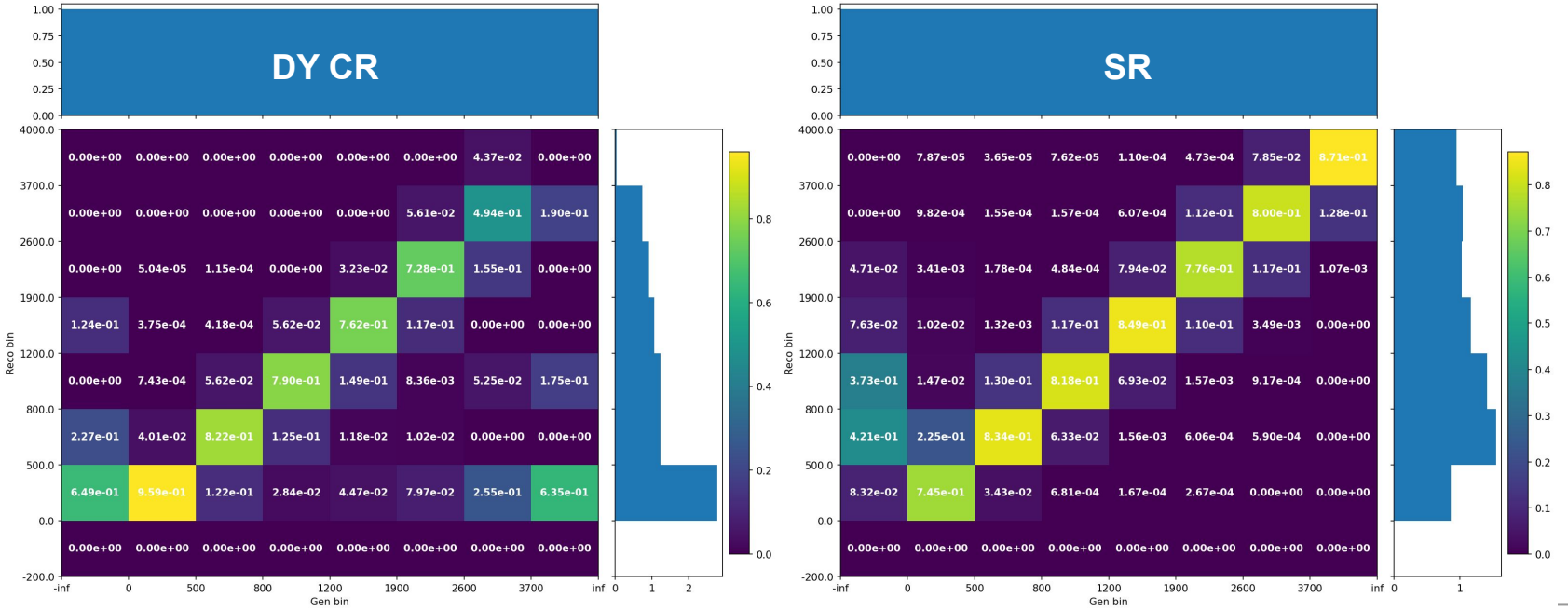
Differential cross-section measurements

- In order to perform differential xs measurements, the signal is split in different generator level bins (based on the variable to unfold) leading to many different signals
- Each signal is then let free to float in the final fit and should be mostly constrained based on the associated reco level bin
- An example of the unfolded result is reported on the right

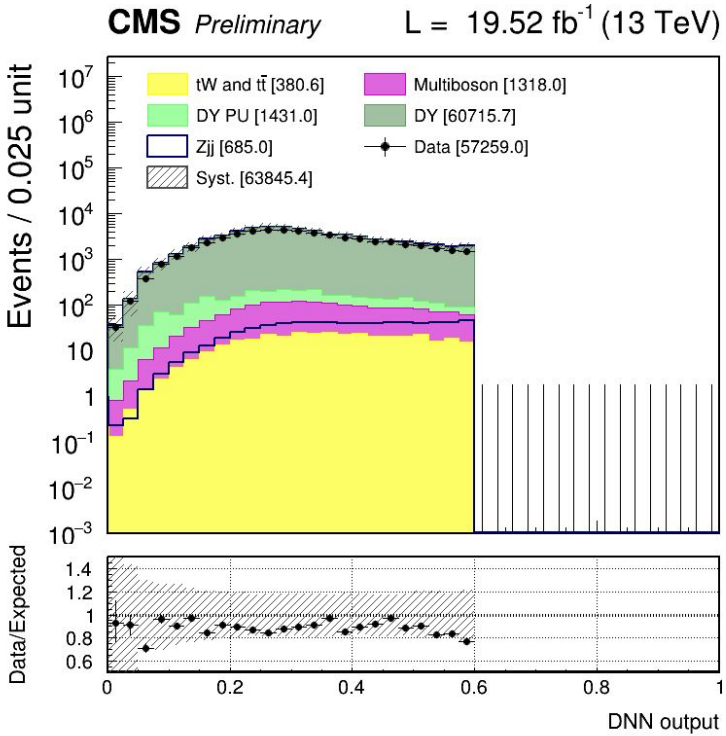


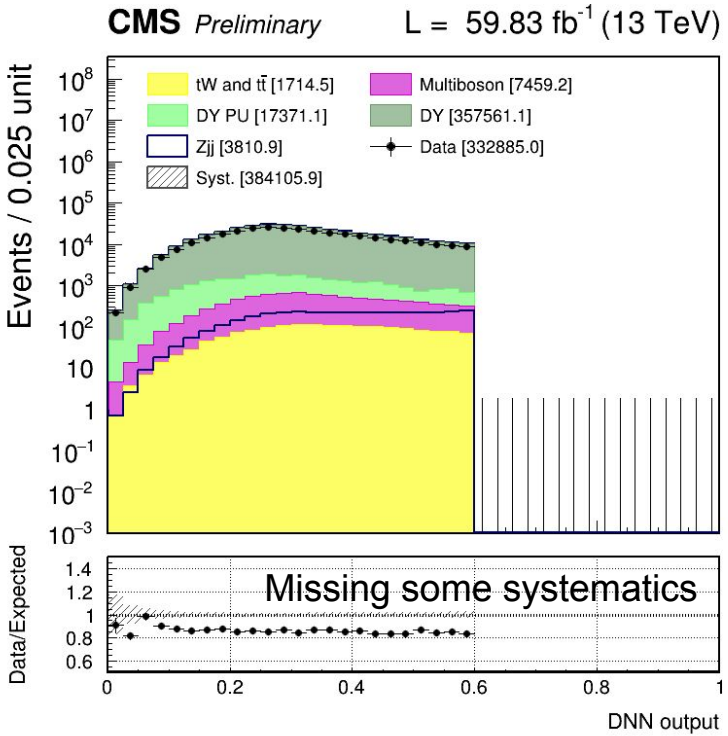
Comparison of response matrices: DY CR vs SR

- strong correlation between gen and reco mjj bins in both regions, larger for high score
- First gen bin (mjj < 0) are events where one gen jet is lost

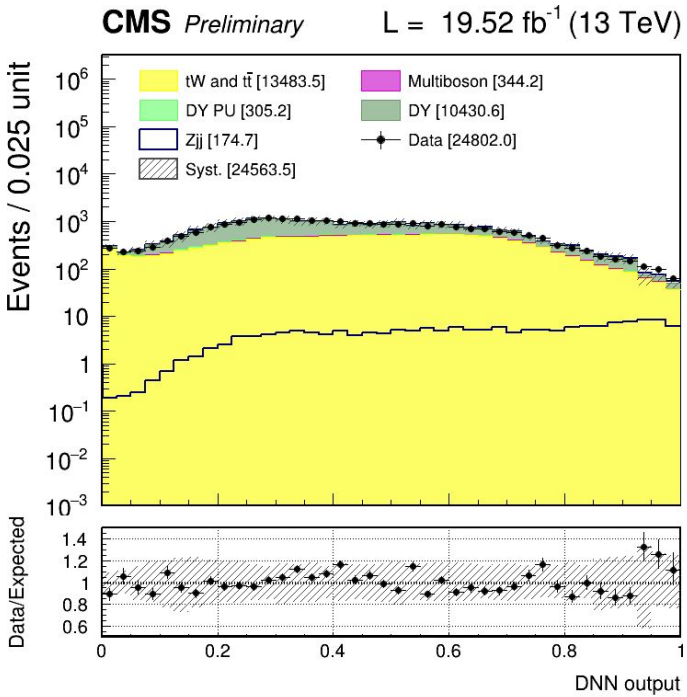


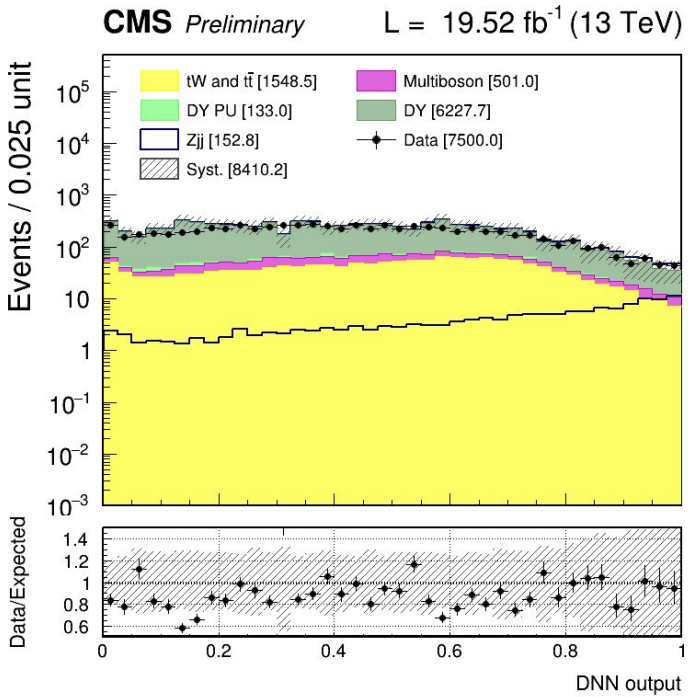
DNN output in DY cr: ratio is roughly flat





Top CR and VV CR are ok





Further studies for DY cr and SR

- Since the ratio plot is quite stable we don’t expect the DY rateParam to fluctuate going to higher values of DNN
- A **quantitative test is ongoing**: we plan to split furthermore the DY cr in to two regions, for test purposes only, (0.0 < DNN < 0.3, 0.3 < DNN < 0.6) and checking the rateParams in the two regions. We expect them to be stable getting closer to the signal region
- This is to check that a simple rateParameter is enough to constraint and fit the DY background in the two regions, despite the events falling in the two categories will have very different kinematics

Systematics uncertainties

Jet Calibration and Resolution

- Jets Energy Correction (JEC) is the result of the calibration of the Jet Energy Scale (JES)
- JEC are computed with data driven methods and detailed MC simulations: they account for offset energy from pileup, detector response to hadrons and differences between data and MC
- Jet Energy Resolution (JER) is calculated after JEC. Since the result is that JER is worse in data than in MC, a smearing method is applied to MC jets
- A statistical analysis should include systematic variations of JEC that affect the JES determination

Jet Energy Scales: Missing Transverse Energy impact

Type 1 MET correction: use Jets pt corrected with JEC instead of raw Jets pt

$$\vec{\not\mathbf{E}}_T^{\text{correct}} = -\sum_{i\in\text{jets}} \vec{\mathbf{p}}_{Ti}^{\text{JEC}} - \sum_{i\in\text{uncl}} \vec{\mathbf{p}}_{Ti}$$

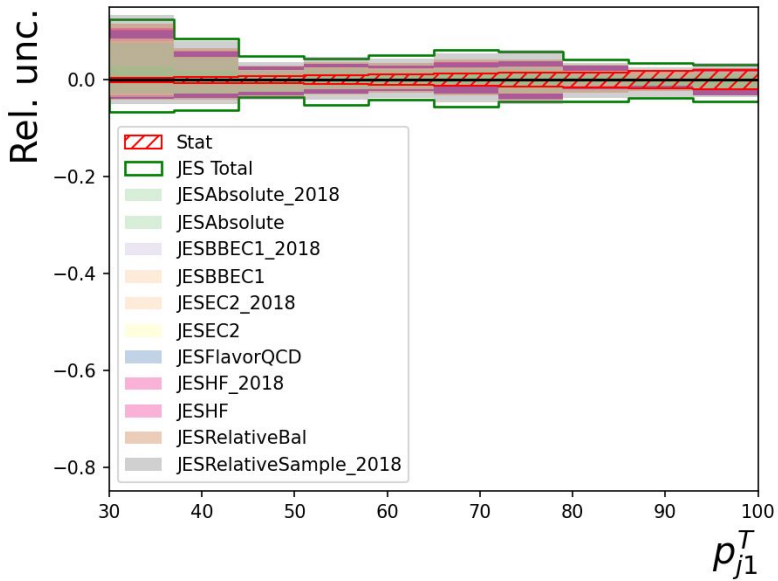
Now Jet pt should be varied for Jet Energy Scales → also MET is affected by JES

$$\vec{\not\mathbf{E}}_T^{\text{JES up}} = -\sum_{i\in\text{jets}} \vec{\mathbf{p}}_{Ti}^{\text{JES up}} - \sum_{i\in\text{uncl}} \vec{\mathbf{p}}_{Ti} \longrightarrow \vec{\not\mathbf{E}}_T^{\text{JES up}} = \vec{\not\mathbf{E}}_T^{\text{correct}} + \sum_{i\in\text{jets}} \vec{\mathbf{p}}_{Ti}^{\text{JEC}} - \vec{\mathbf{p}}_{Ti}^{\text{JES up}}$$

where $\vec{\mathbf{p}}_{Ti}^{\text{JES up}} = \vec{\mathbf{p}}_{Ti}^{\text{JEC}}\,(1+\Delta(p_{Ti}^{\text{JEC}}, \eta_i))$

JES Systematics

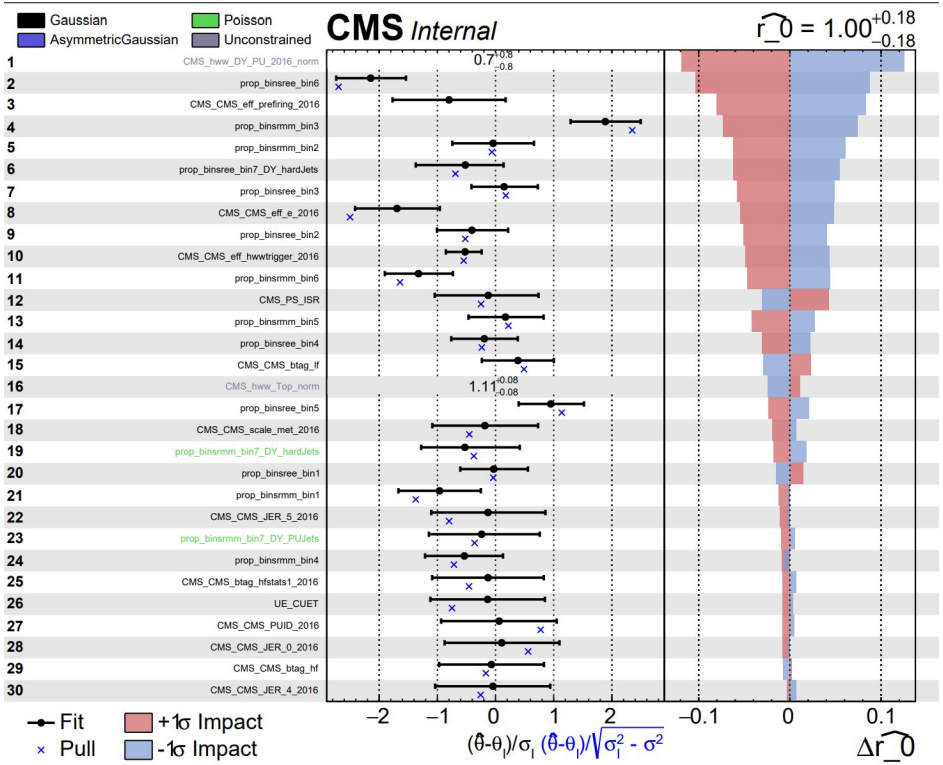
- During the last months I took care of computing the JES for the Latinos Working Group (Higgs Physics and Standard Model Physics)
- An example of the impact of the different JES sources is reported on the right for the Z+jets sample



Final fit (inclusive): 2016 HIPM (first era of 2016)

- Data-asimov fit
- We know that JER/JES together with QCD scales represents important contribution
- DY normalization factor for now is the leading systematic
- Counting the shifts of + sigma and - sigma of the systematics fit, the result is reasonable

Chi2: 19.57
Chi2/ndof: 0.4164



Effective Field Theory

Effective Field Theory

- We work with dimension 6 operators: Warsaw Basis with 59 operators
- Neglect odd dimension operators (they violated important accidental symmetries)
- Agnostic study of new physics with indirect searches

$$\mathcal{L}_{SMEFT} = \mathcal{L}_{SM} + \frac{1}{\Lambda^2}\sum_i c_i^{(6)}\mathcal{O}_i^{(6)}$$

Coefficients c_i called Wilson coefficients are one of the objects of this study (we actually use $\frac{c_i}{\Lambda^2}$ with $\Lambda = 1$ TeV)

$$\mathcal{L}_{SMEFT} = \mathcal{L}_{SM} + c_i^{(6)}\mathcal{O}_i^{(6)} \longrightarrow |A_{EFT}|^2 = |A_{SM}|^2 + c_i^2\,|A_{op}|^2 + 2\,c_i\ \mathrm{Re}(A_{SM}A_{op}^*)$$

The Lagrangian when only one operator is active

List of EFT operators the analysis is sensitive to

- Some EFT dimension 6 operators do not enter in the Feynman diagrams of the signal, therefore are excluded from the analysis

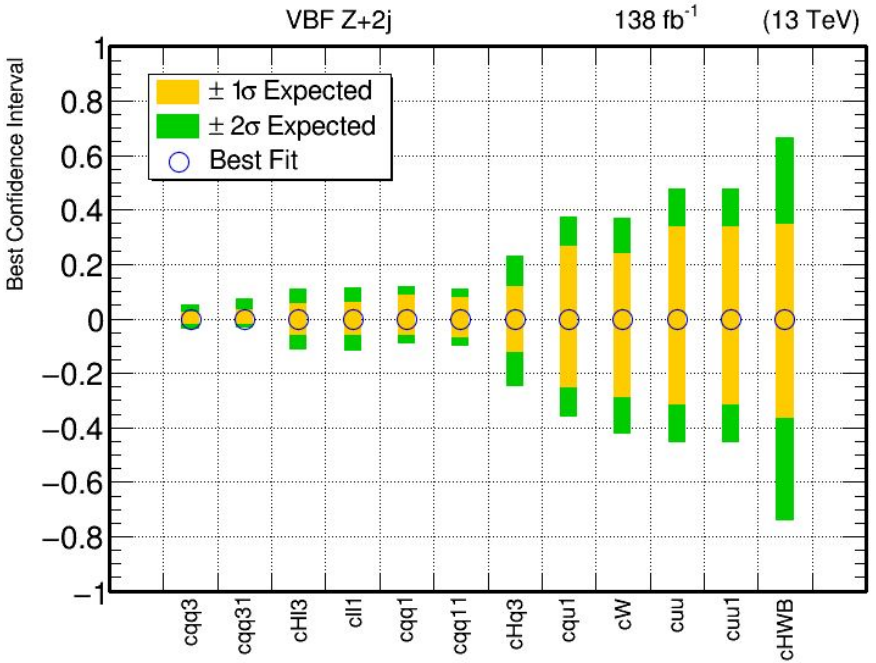
cG	No
cW	Yes
cH	No
cHbox	Yes
cHDD	Yes
cHG	Yes
cHW	Yes
cHB	Yes
cHWB	Yes
ceHRe	No
cuHRe	No
cdHRe	Yes
ceWRe	No
ceBRe	No
cuGRe	No
cuWRe	No
cuBRe	No
cdGRe	No
cdWRe	Yes
cdBRe	Yes

cHl1	Yes
cHl3	Yes
cHe	Yes
cHq1	Yes
cHq3	Yes
cHu	Yes
cHd	Yes
cHudRe	No
cll	No
cll1	Yes
cqq1	Yes
cqq11	Yes
cqq3	Yes
cqq31	Yes
clq1	Yes
clq3	Yes
cee	No
cuu	Yes
cuu1	Yes
cdd	Yes

cdd1	Yes
ceu	Yes
ced	Yes
cud1	Yes
cud8	Yes
cle	No
clu	Yes
cld	Yes
cqe	Yes
cqu1	Yes
cqu8	Yes
cqd1	Yes
cqd8	Yes
cledqRe	No
cquqd1Re	No
cquqd11Re	No
cquqd8Re	No
cquqd81Re	No
clequ1Re	No
clequ3Re	No

EFT Fits with Full Run 2

- A real EFT fit was already performed with one active operator at a time, using all the CRs and the SR but discarding systematics
- Once the SM cross-section measurements will be ready, we will move to a full EFT fit, considering many operators active (profiled fit) and considering all the systematics



Technical tasks

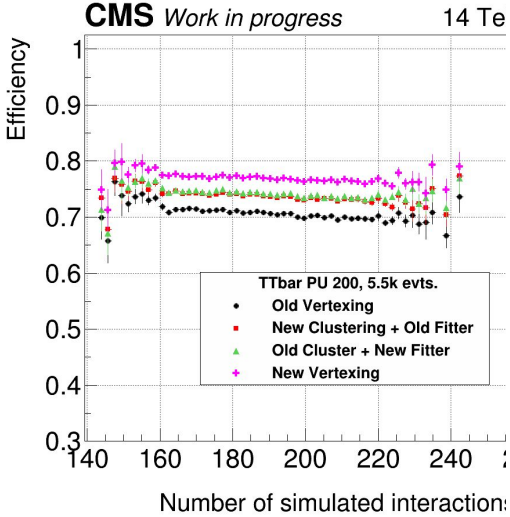
The New Framework for Latinos (NFL?)

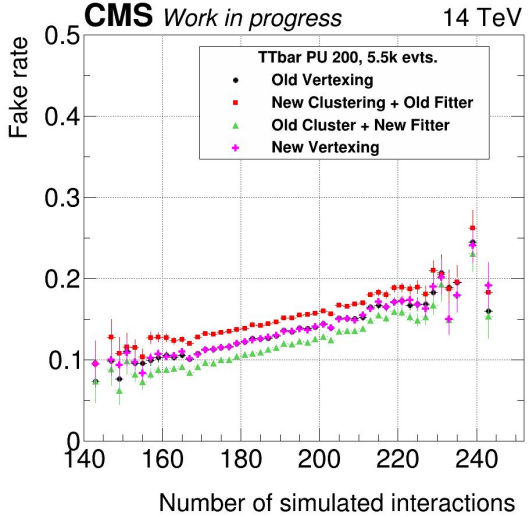
- The Latinos working group is the working group where historically part of the bicocca CMS group worked
- The previous framework was slow and not ready to analyze the new re-proccessed and corrected data and mc (called Ultra Legacy production)
- Based purely on python 2 made it take a few weeks for post processing of a single year of data taking and a few days for the analyst to produce histograms
- The new framework I've contributed to develop is entirely based on C++ bindings for python and on natively MultiThreaded RDataFrames making it blazing fast (a single day for post processing and a couple of hours to produce histograms)

Primary Vertex Reconstruction

- Project was to develop a faster and GPU compliant Primary Vertex Reconstruction

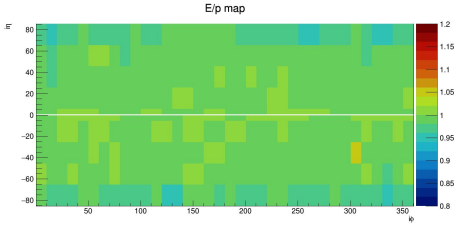
- I’ve developed a CPU version that runs **6x faster** than the official CMS one for Phase II (but no timing)
- Higher efficiency: 7%
- But also higher Fake Rate **0.5%**
- We moved also to Run 3 to cope with the higher Pile Up rate

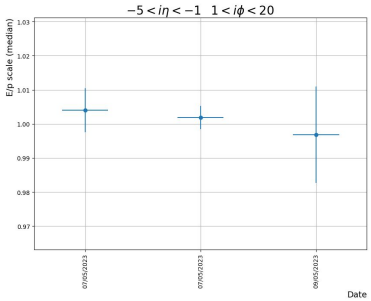




ECAL Calibration and Monitoring Automation: E/p

- ECAL calibration and monitoring is being moving to an automated chain, where for each fill/run jobs are automatically run in sequence and resubmitted by the automation tool
- E/p calibration was missing its automatic implementation and I took care of it together with Flavia
- Now the E/p jobs are run for each 2 /fb, we split the 324 harnesses (group of ECAL modules) into 12 jobs run in parallel





Ph.D. Courses and Schools

- Interdisciplinary courses:
  - Basic principles of public relations and media relations for academics 1.5 CFU
  - The appropriate leadership. a sustainable approach to inclusive leadership in diverse contexts 1.5 CFU
- Physics courses:
  - EFT fitting in Standard Model measurements 2 CFU
  - Deep Learning for Physicists 2 CFU (next week)

- Physics Schools:
  - XIV INFN International School on Efficient Scientific Computing (ESC23) (beginning of october) -> 2 CFU

Backup

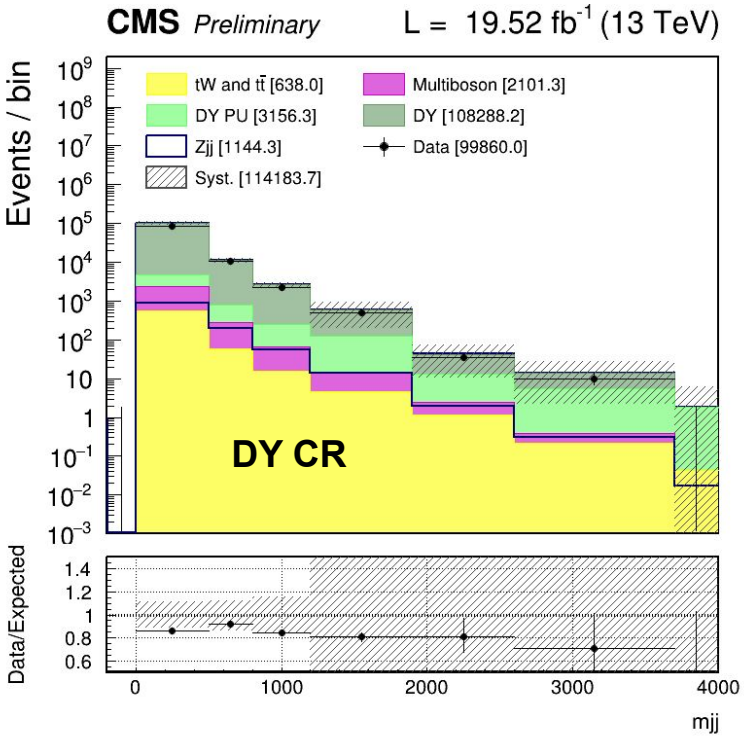
Studies for SR and DY CR

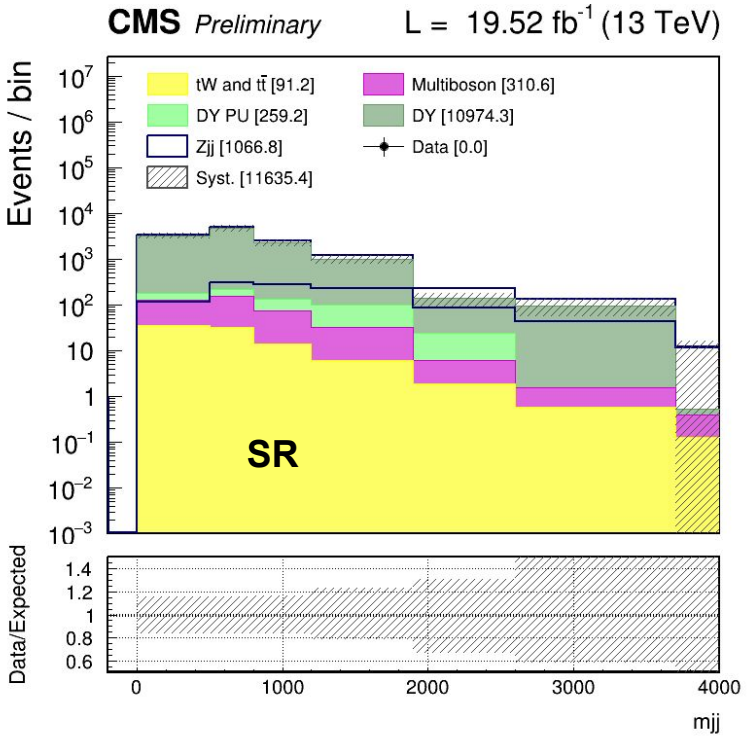
- We’re not losing any event: we use both regions to fit, all events at low DNN will contribute to estimate the DY rateParams, the high DNN events will be primarily used for the unfolding measurement (i.e. measuring the different signal strength modifiers) where the significance is much higher
- We don’t expect high sensitivity at low gen mjj values (i.e. first bin of gen mjj, < 500 GeV)
- It’s true that the DNN cut at 0.6, moves the majority of low mjj values to the DY CR, but that’s not really part of what we would like to measure

Difference of response matrix

- Gen bin with mjj < 500 GeV will be determined by the DY CR, while for all the others the SR is dominating
- Gen bin with mjj < 500 GeV measurements will be difficult due to the high contamination of DY background, but those signal events, hardly are Vector Boson Fusion like

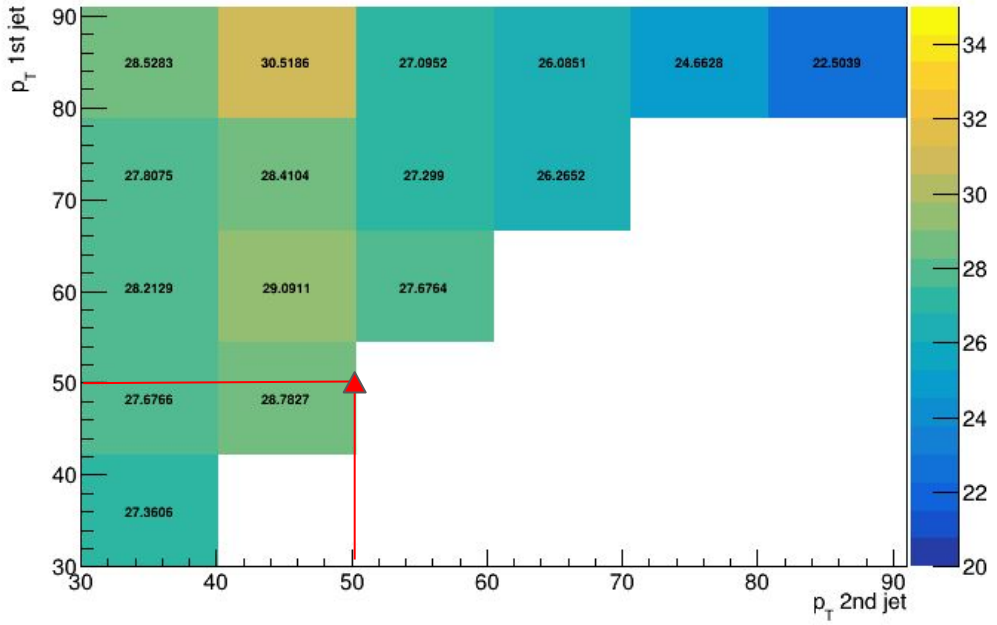
Variable to fit: mjj unfold



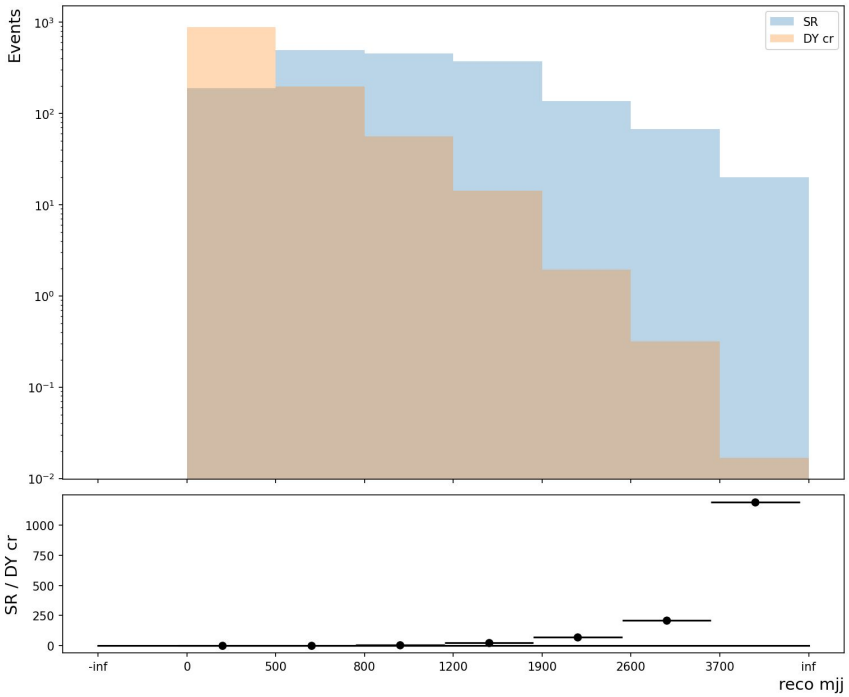


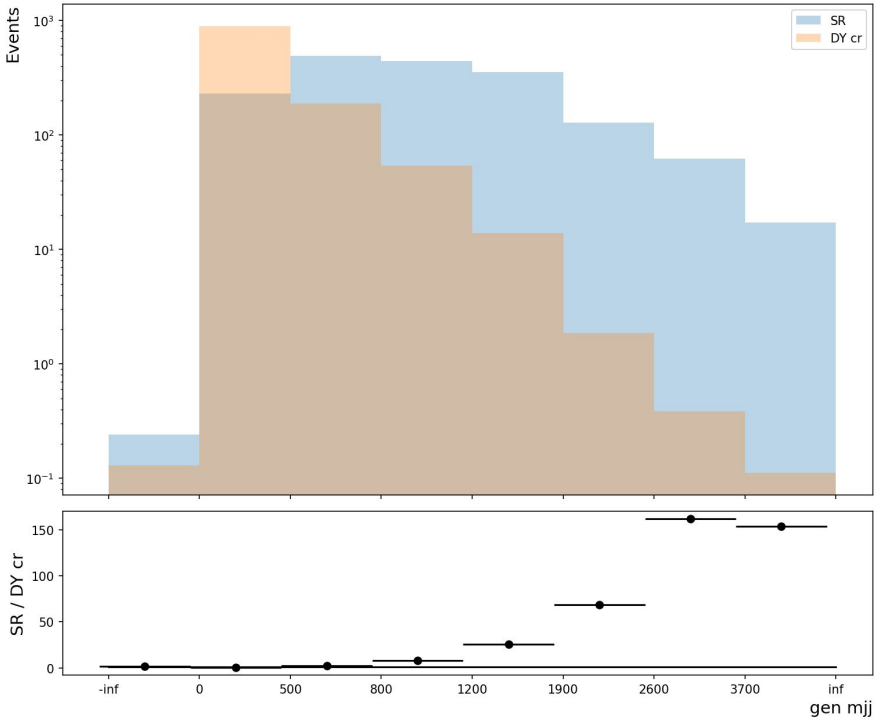
Optimization of the cuts on pt of leading jets

- Maximize proxy of sensitivity together with statistics
- S/sqrt(B) is actually the squared sum of S/sqrt(B) for different bins of mjj (from 1000 GeV and above)
- Statistical significance limit



Comparison of Signal mjj distribution in SR and DY CR





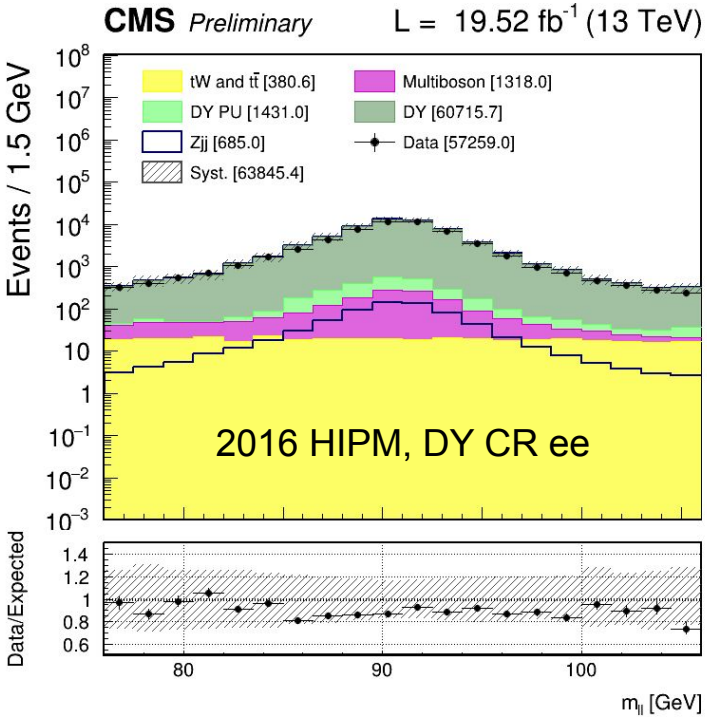
Moved to UL for Full Run 2

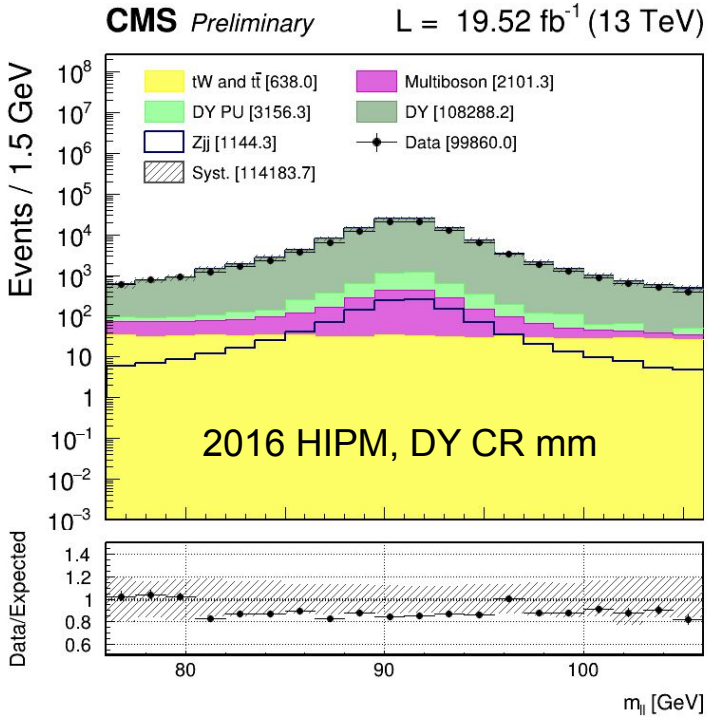
- We have the three years set up with the same configuration
- The set of nuisance and their correlation is under development
- Still missing JES for UL but will be ready by the end of July
- The key point is that we fully switched to UL and in a few weeks will have the full set of nuisances ready for the Full Run 2 fit

Our previous presentations at V+Jets:

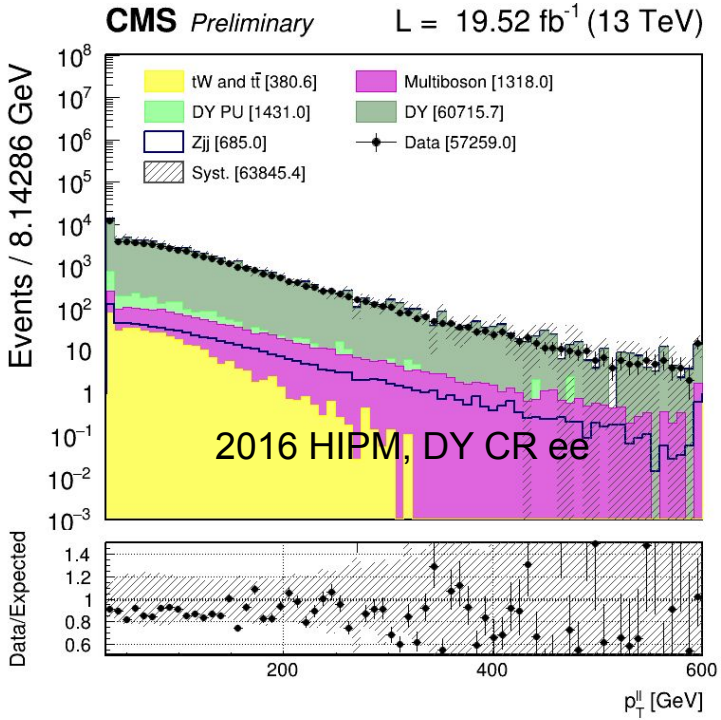
- https://indico.cern.ch/event/1252689/#3-vbf-z-jets-and-eft-analysis
- https://indico.cern.ch/event/1223034/#4-status-of-the-vbf-z-analysis
- https://indico.cern.ch/event/1186410/#8-updates-on-vbf-z-eft-dim-6

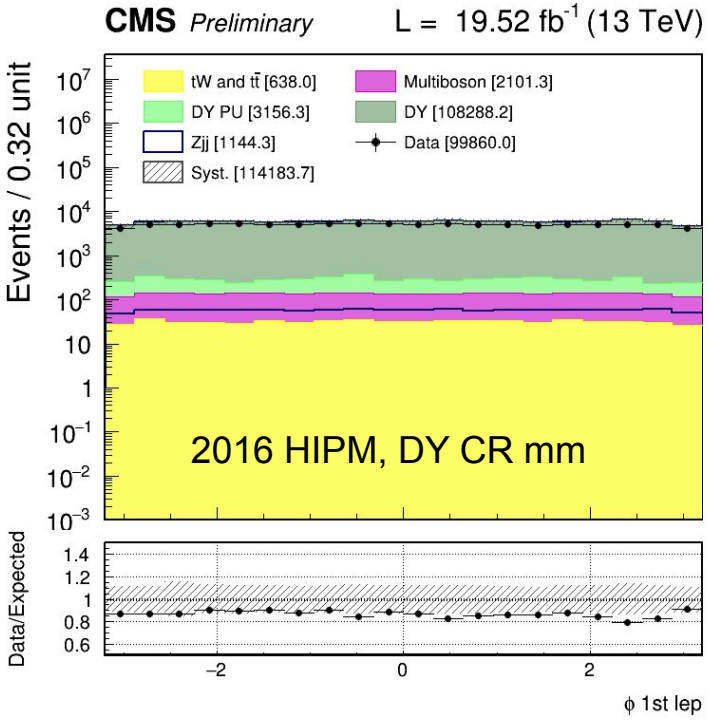
Data / MC Agreement





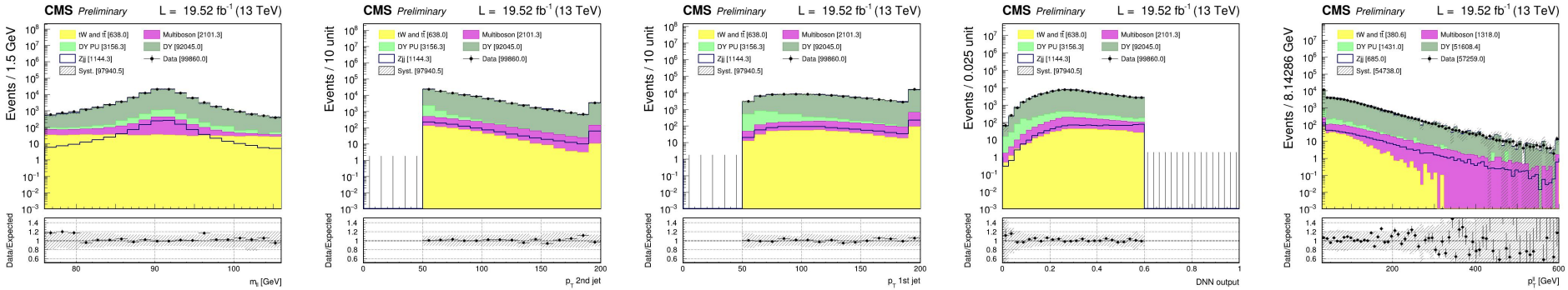
Data / MC Agreement





Data / MC Agreement

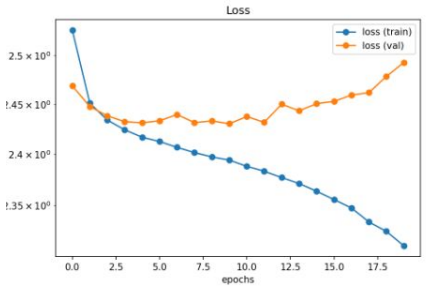
- Across all years the discrepancies between Data and MC are within uncertainties (main contribution from DY normalization uncertainty and JER, JES will also give important contribution)
- We have tested that a normalization fit for the main backgrounds is enough and discrepancies do not show shape deviations

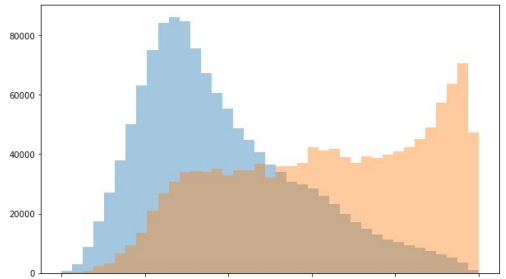


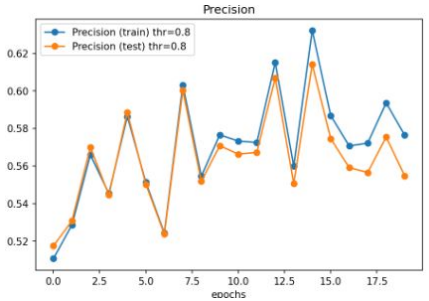
DNN training

DNN training

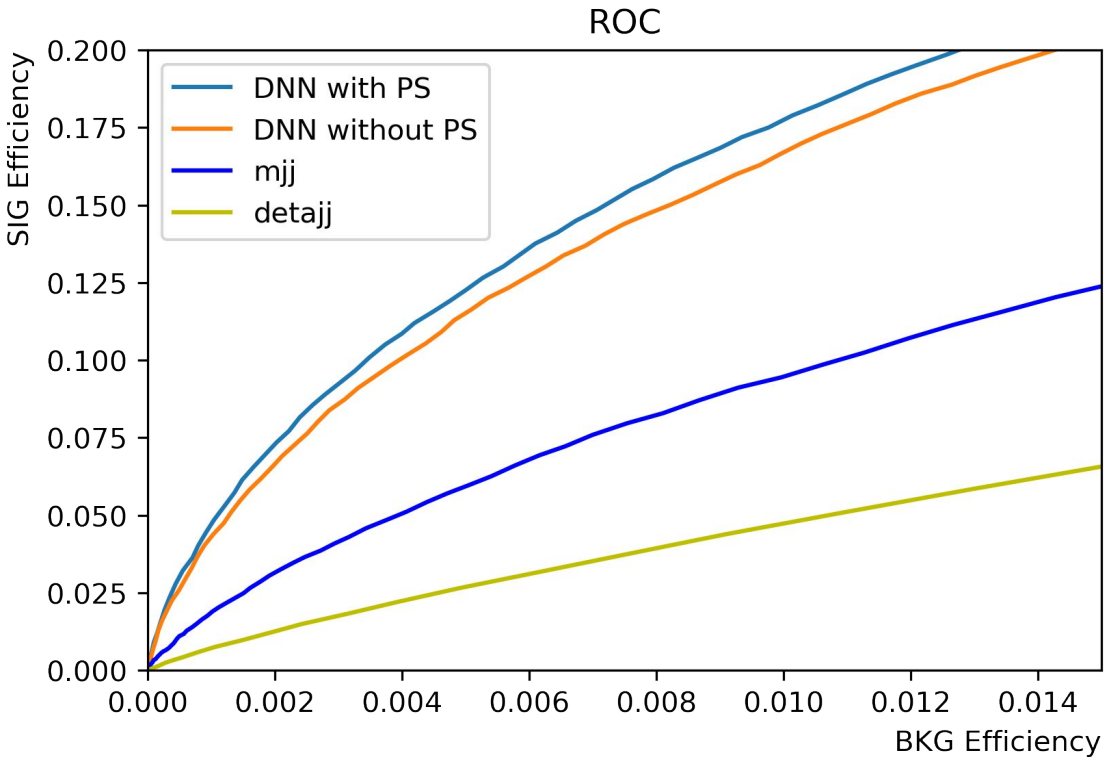
- A new DNN was trained (6 layers, higher density in the most hidden layers)
- Variables used: MET, Zeppenfeld_Z, Zeppenfeld_l1, detajj, detall, dphijj, dphill, etaj1, etaj2, etal1, etal2, mjj, mll, phij1, phij2, phil1, phil2, ptj1, ptj2, ptjj, ptl1, ptl2, ptll
- Observables sensitive to Parton shower: HT, QGL_1, QGL_2, Njet, dphi_js_met_min
- Two different training with and without PS variables
- Checked the impact of PS Weight on the two ROCs
- Trained the DNN to separate the signal from DY (both hard and PU)



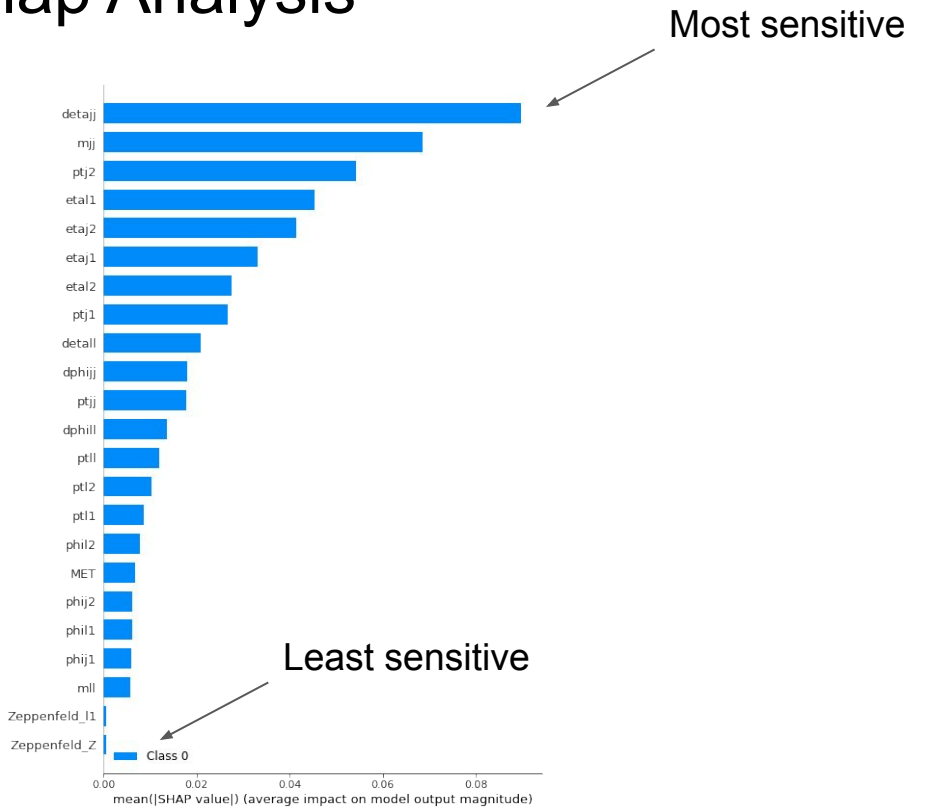


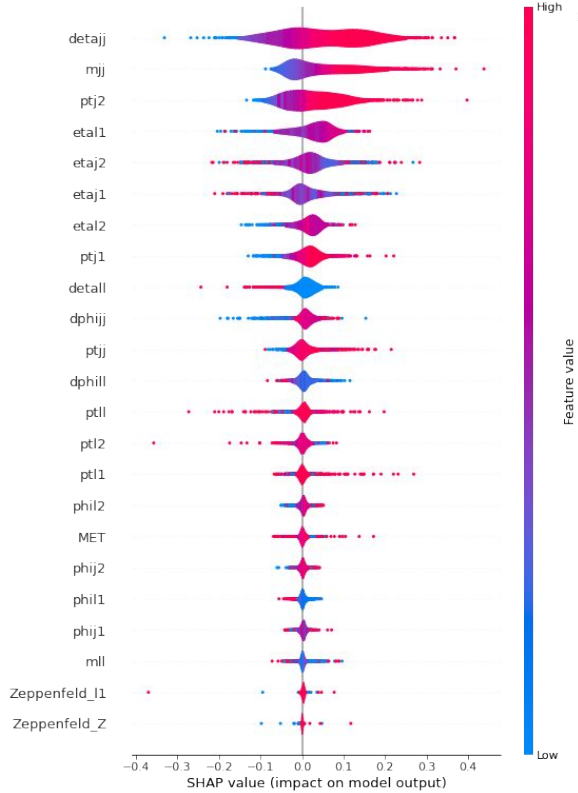


ROC performance comparison zoom



Shap Analysis

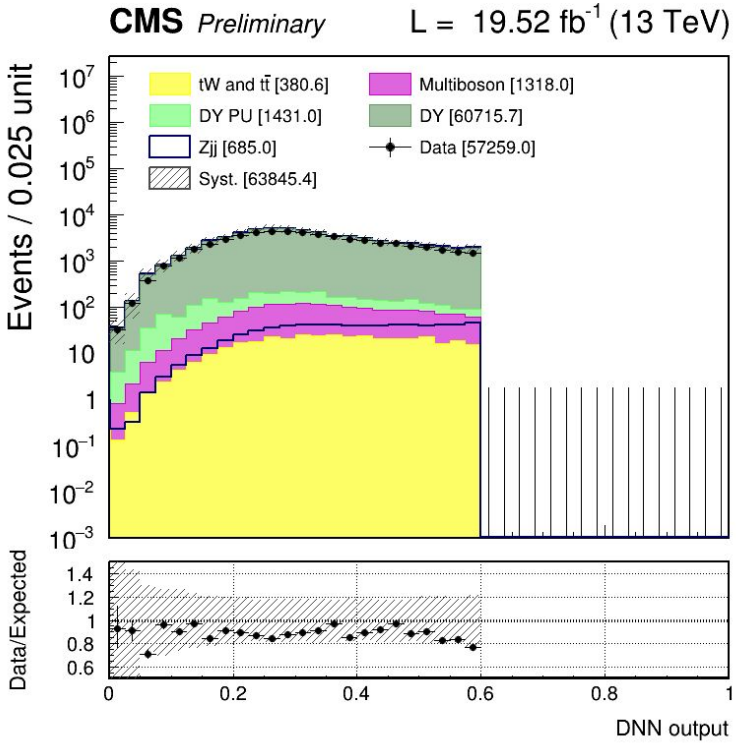


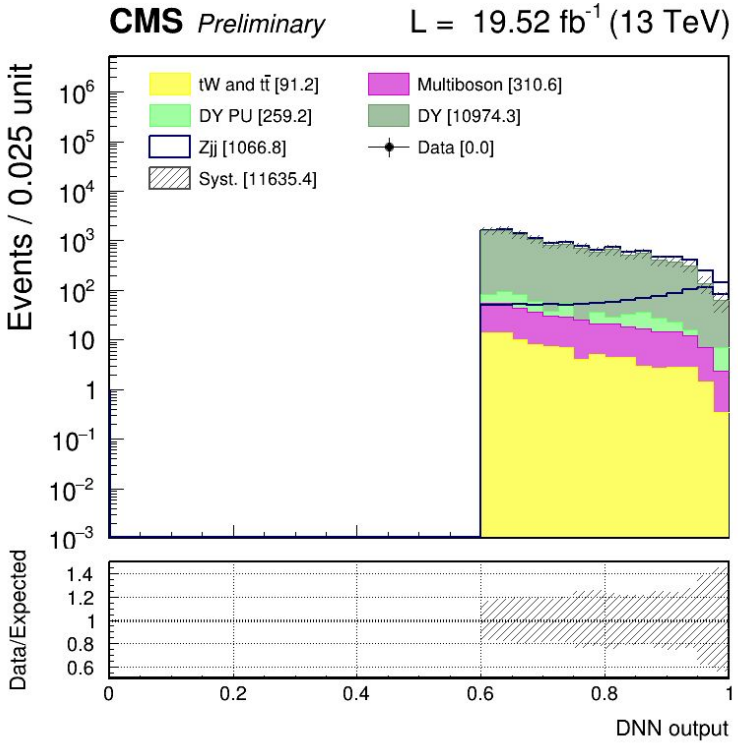


Performing Fit

- Fit performed using the 4 regions: mjj for the SR and Events for CRs
- EE and MM are separated and combined across all regions
- Fit performed with MC asimov (-t -1) and Data asimov (-t -1 --toysFreq)
- A Full Run 2 with a restrictive set of nuisances is performed together with a fit of 2016 HIPM with all nuisance (except for JES)
- Results as impacts in next slides

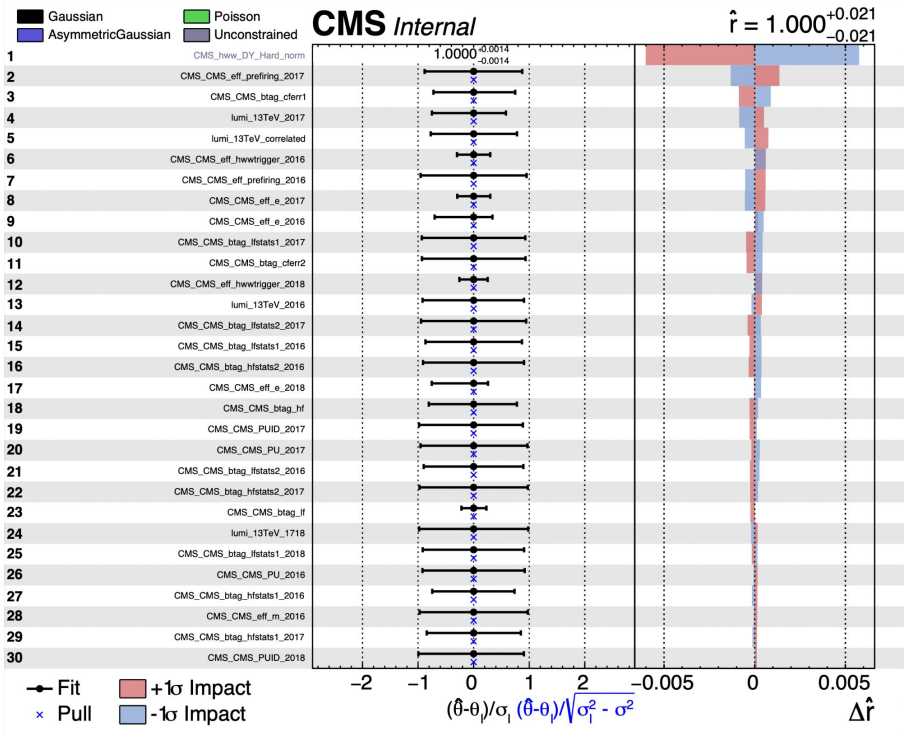
DNN output for the ee channel



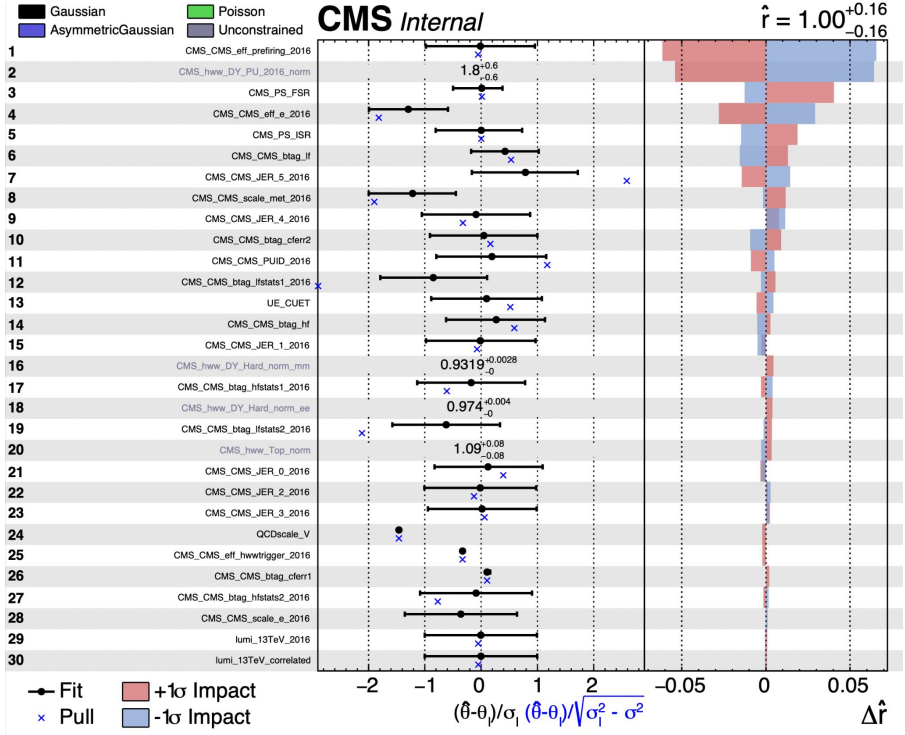


Impacts

Full Run 2 with few nuisances



2016 HIPM with most nuisances



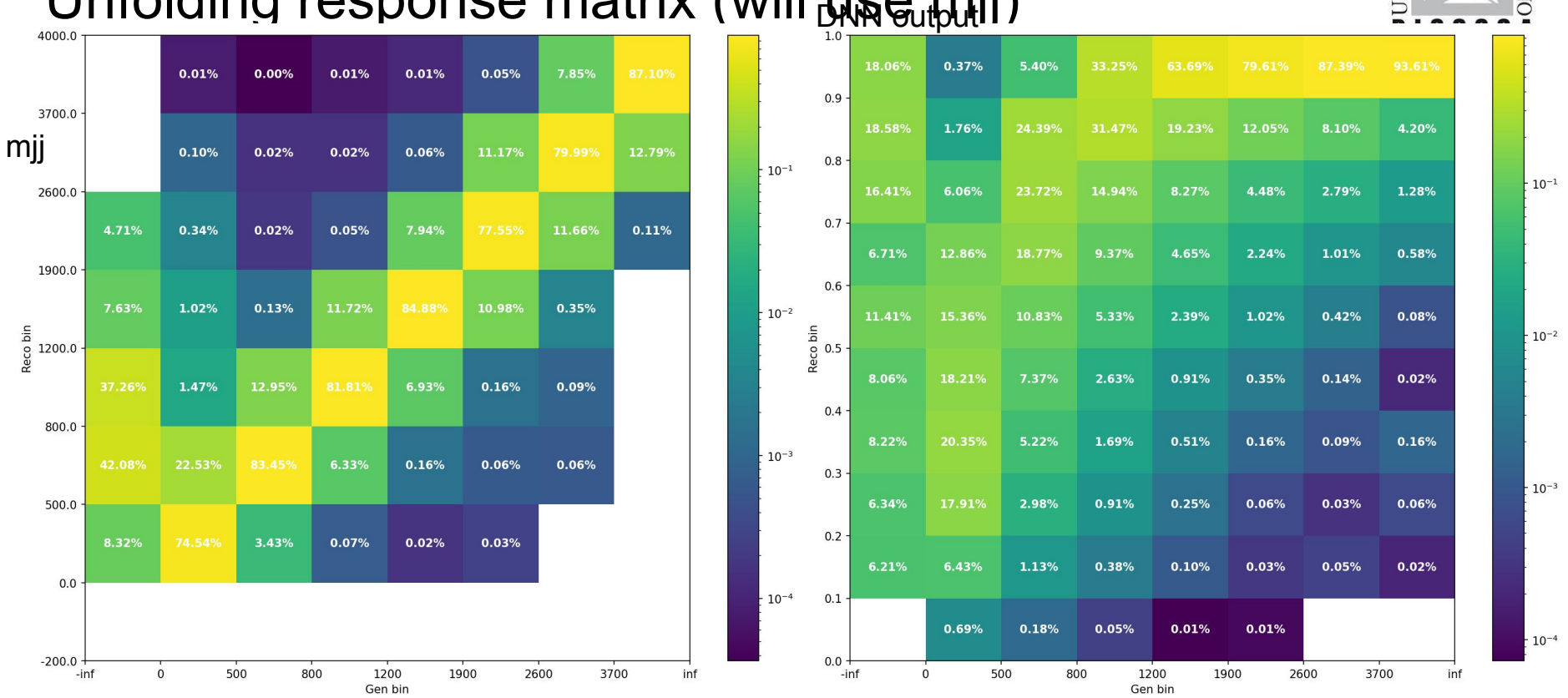
Unfolding

- Our goal is to perform differential cross section measurements
- Combine is used to fit the different signal strength modifiers for each Generator level bin
- The fiducial region is defined with the same preselections of the analysis, no bVeto / bTag and no DNN cut
- The definition of mjj at gen level takes the two leading Jets with at least 15 GeV and removes those that have dressed leptons in a cone of 0.3 (as it's done at reco level)

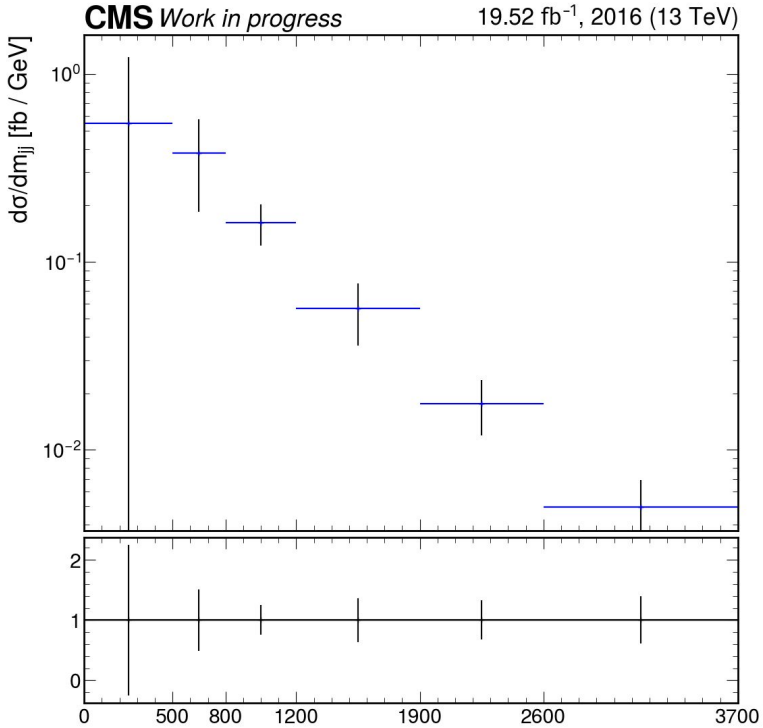
Fiducial Region

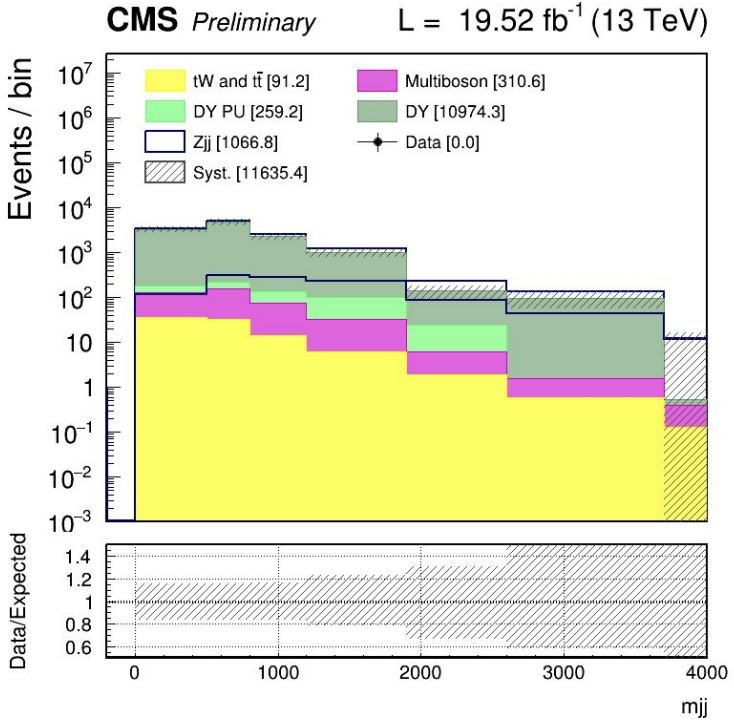
- ★ At least 2 leptons of the same flavor
- ★ At least 2 jets with $p_T > 50$ GeV
- ★ $m_{jj} > 200$ GeV
- ★ $m_{ll} > 50$ GeV
- ★ $m_{ll} \in (M_Z\text{-}15, M_Z\text{+}15)$

Unfolding response matrix (will use mjj)



Expected measurements





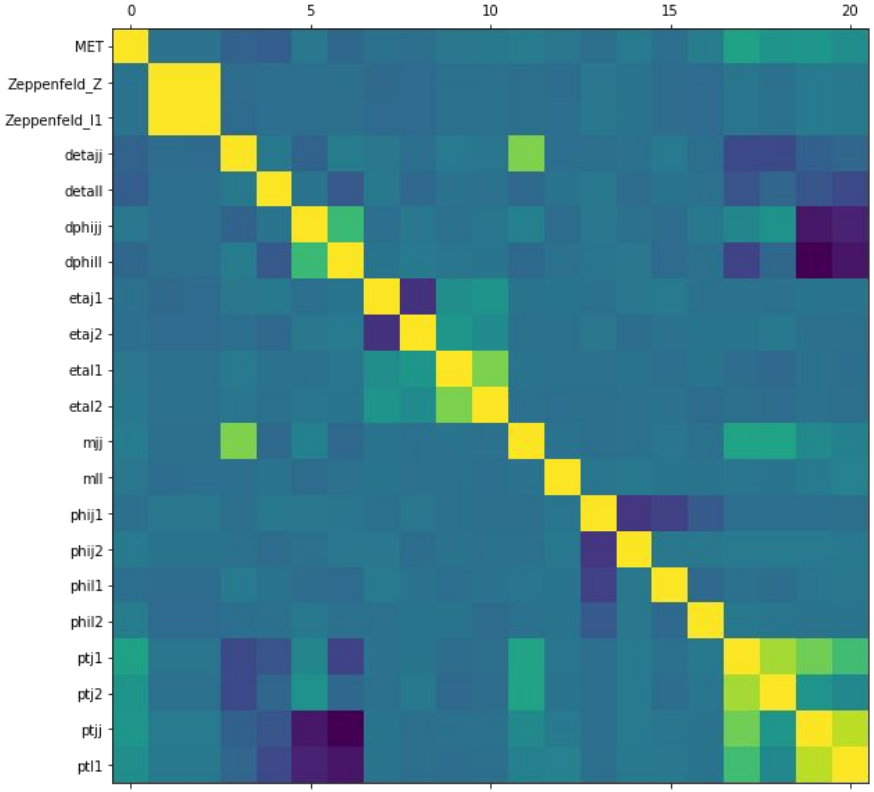
Summary and next steps

- We finally moved to UL for the three years
- We've defined an unfolding procedure for a single variable: mjj

Next steps:

- First and most important next step is to include the full set of nuisances on all the three years
- We're currently performing different tests for the fit procedure
- We will add the unfolding of other variables (ptll, dphijj and pseudo rapidity separation of the two jets)
- Once these first points will be under control we will test different parton shower (already have a sample with Herwig)
- We will also include EFT reweights in the analysis

Correlation matrix in DNN



Cut on ptjs

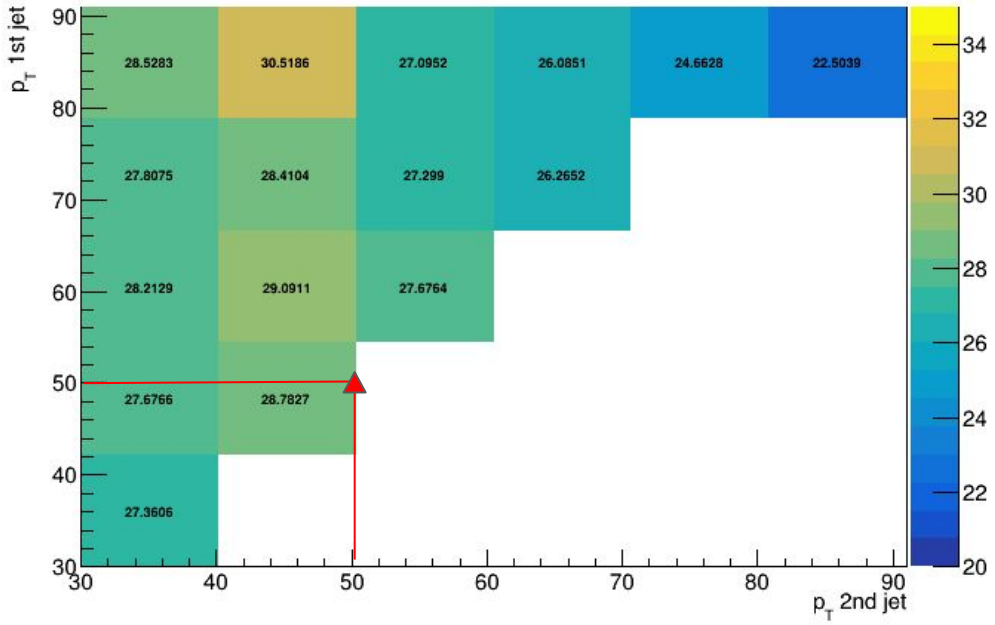
Are the high threshold on ptjs needed? (Over 50 GeV)

- Atlas collaboration used a value of 85 GeV as a threshold for the leading Jet pt
- By splitting the samples in bins of mjj (see [(a)] for values) we wanted to check how the sensitivity scales
- Mjj splitting is at gen level -> the two leading jets were required to be gen matched (no PU jets are present)
- For the complete set of plots go to Zjj link and DY link

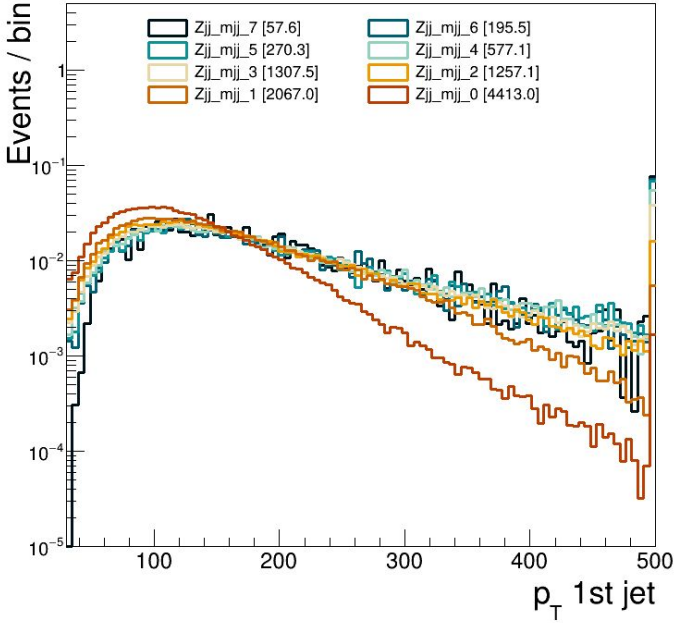
[(a)] [500,750,1000,1500,2000,2500,3500]

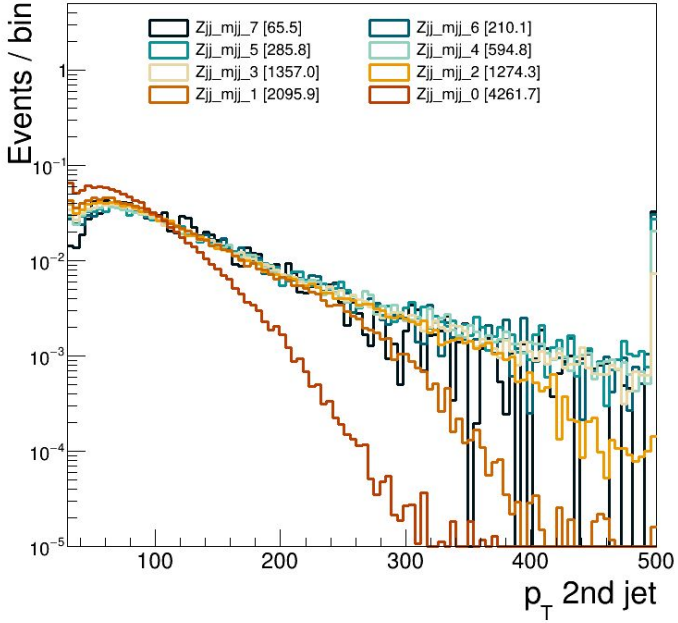
S/sqrt(B) with different thresholds of ptjs

- S/sqrt(B) is actually the squared sum of S/sqrt(B) for different bins of mjj (from 1000 GeV and above)
- Statistical significance limit



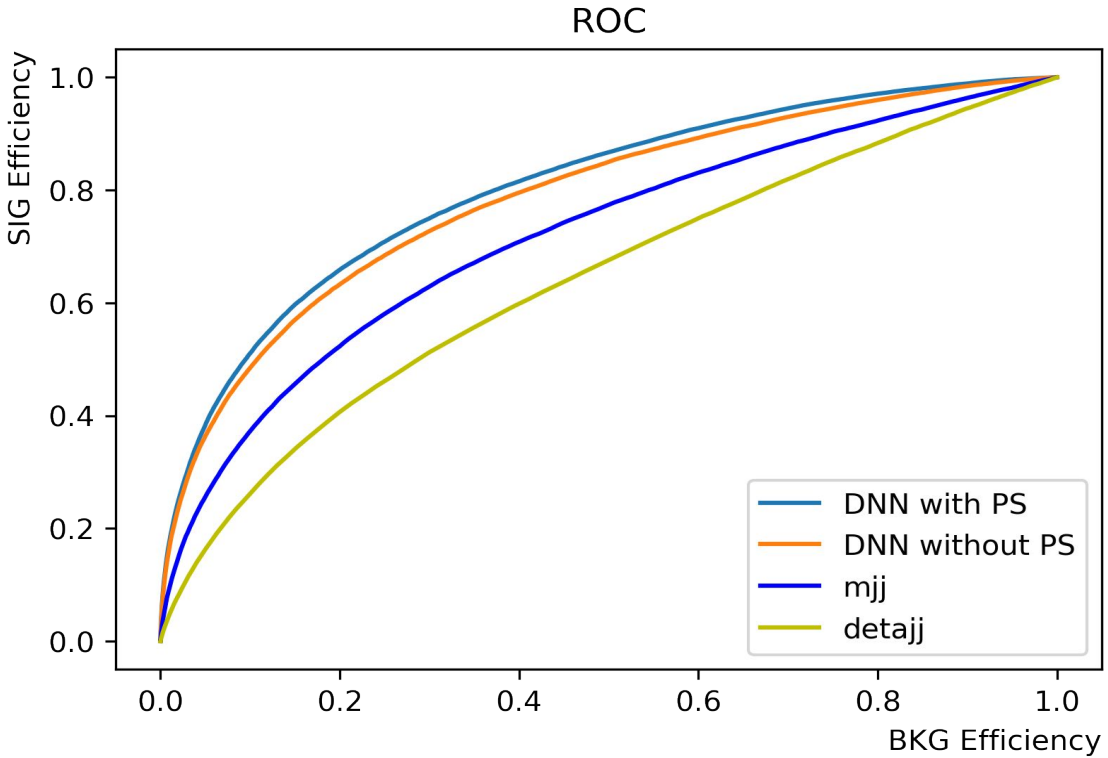
Ptjs for Zjj in bins of mjj (all at Gen)



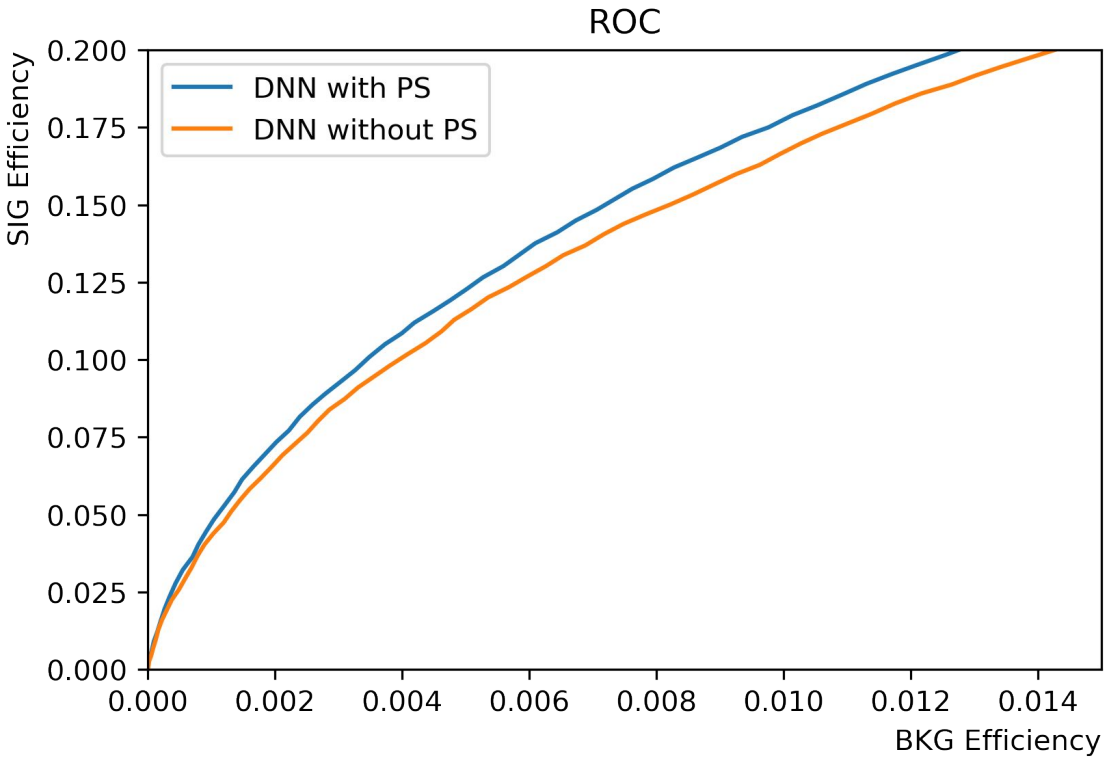


Plots are normalized, for the complete set of plots (normalized and stacked) refer to link for test3

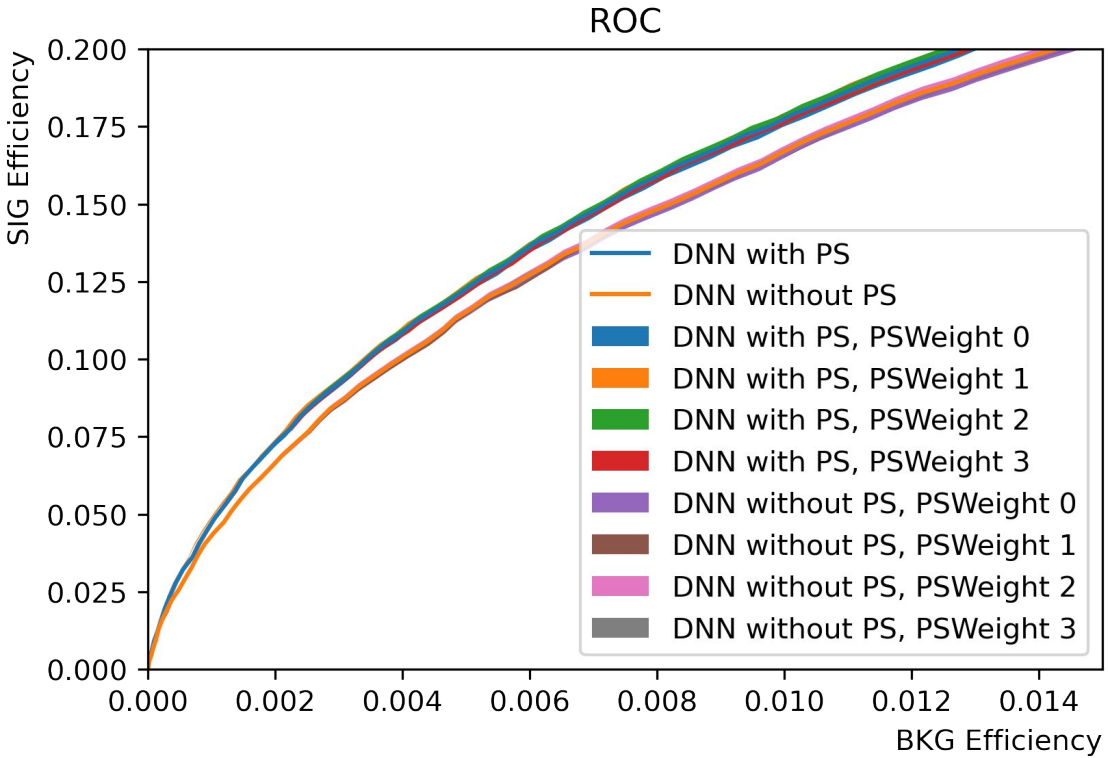
ROC performance comparison



ROC performance comparison



ROC performance comparison with PSweights variations



ROC performance comparison with PSweights variations separated