Modeling dilepton background using Boosted Decision Trees

Pierre Chatagnon

14th of March 2024

Outline

I Motivations

II Event mixing and reweighting

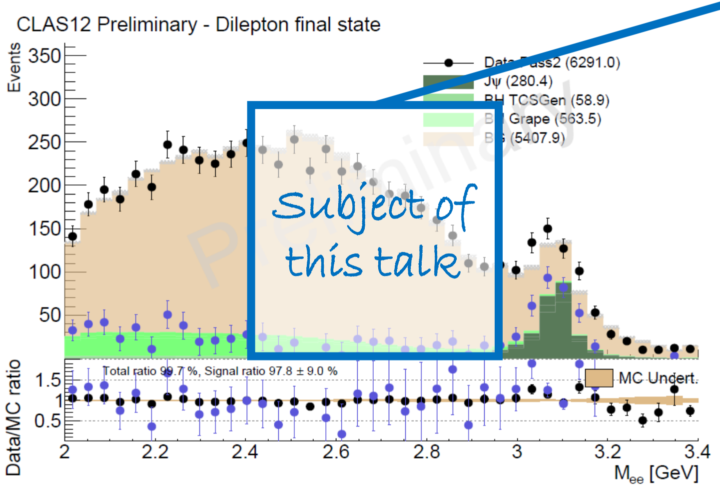
III Application: normalization factor for the J/ψ analysis

Motivation: Cross-section computation for the J/ψ photoproduction

$$ep \rightarrow (e')\gamma p \rightarrow (e')J/\psi\ p' \rightarrow (X)e^+e^-p'$$

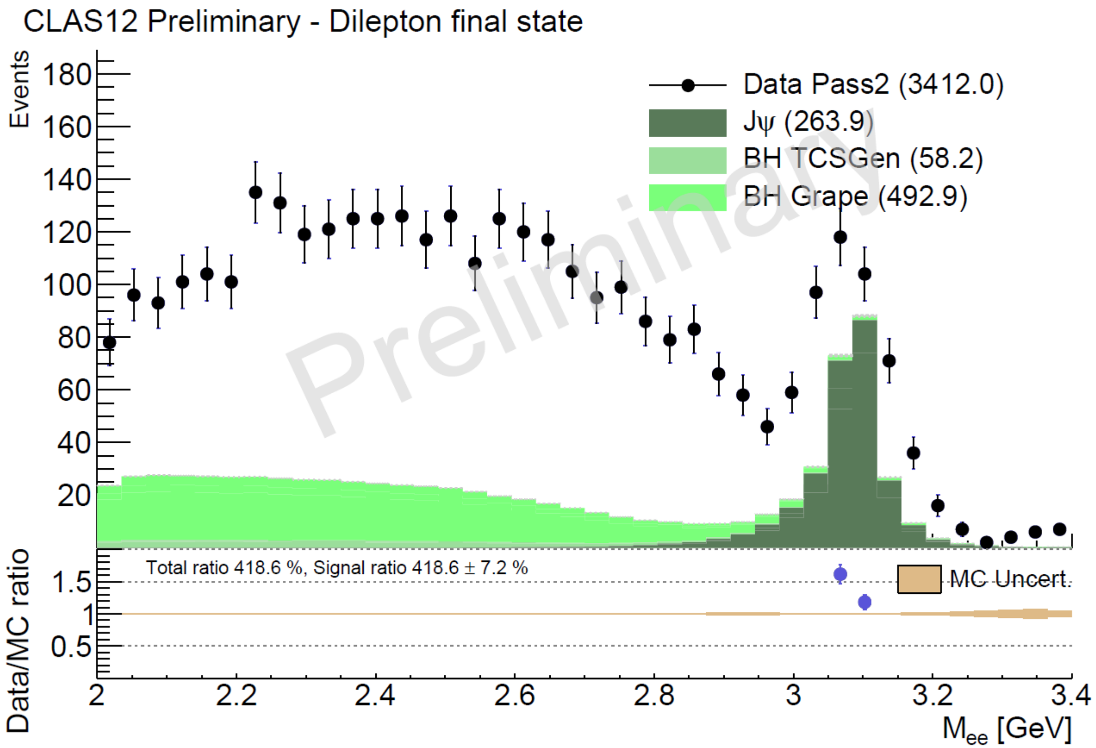
$$\sigma_j = \frac{N_{J/\psi j}}{\mathcal{F}_j \cdot \mathcal{L} \cdot \omega_{cj} \cdot B_r \cdot \epsilon_j \cdot \epsilon_{Rad/j}}$$

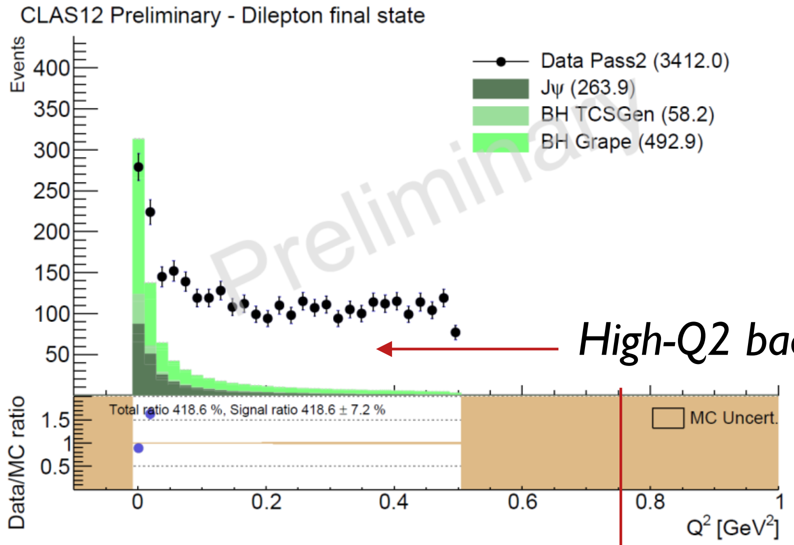
- $\mathcal{F}_j \cdot \mathcal{L}$: Number of photons and Number of targets
- $N_{J/\psi j}$: Number of J/Psi from data
- $\epsilon_{Rad/j}$: Radiative corrections from MC
- ω_{cj}: Subject of this talk
- B_r: Branching ratio: 6%
- ϵ_j: Reconstruction efficiency from MC



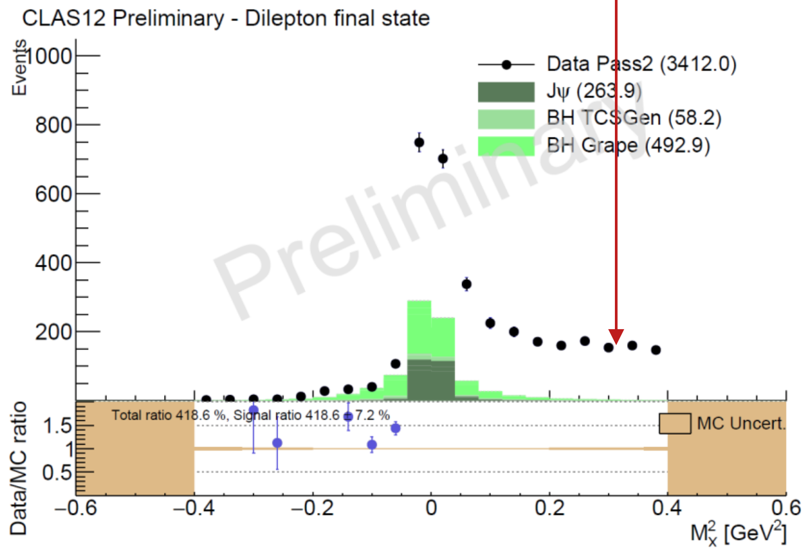
Comparison data/MC – Fall 2018 inbending

- Plotting conventions
  - Color-filled histograms are *stacked*, ie they show the total number of events with contributions for different channels "on top of each other"
  - Marker histograms are *not stacked* and simply superimposed

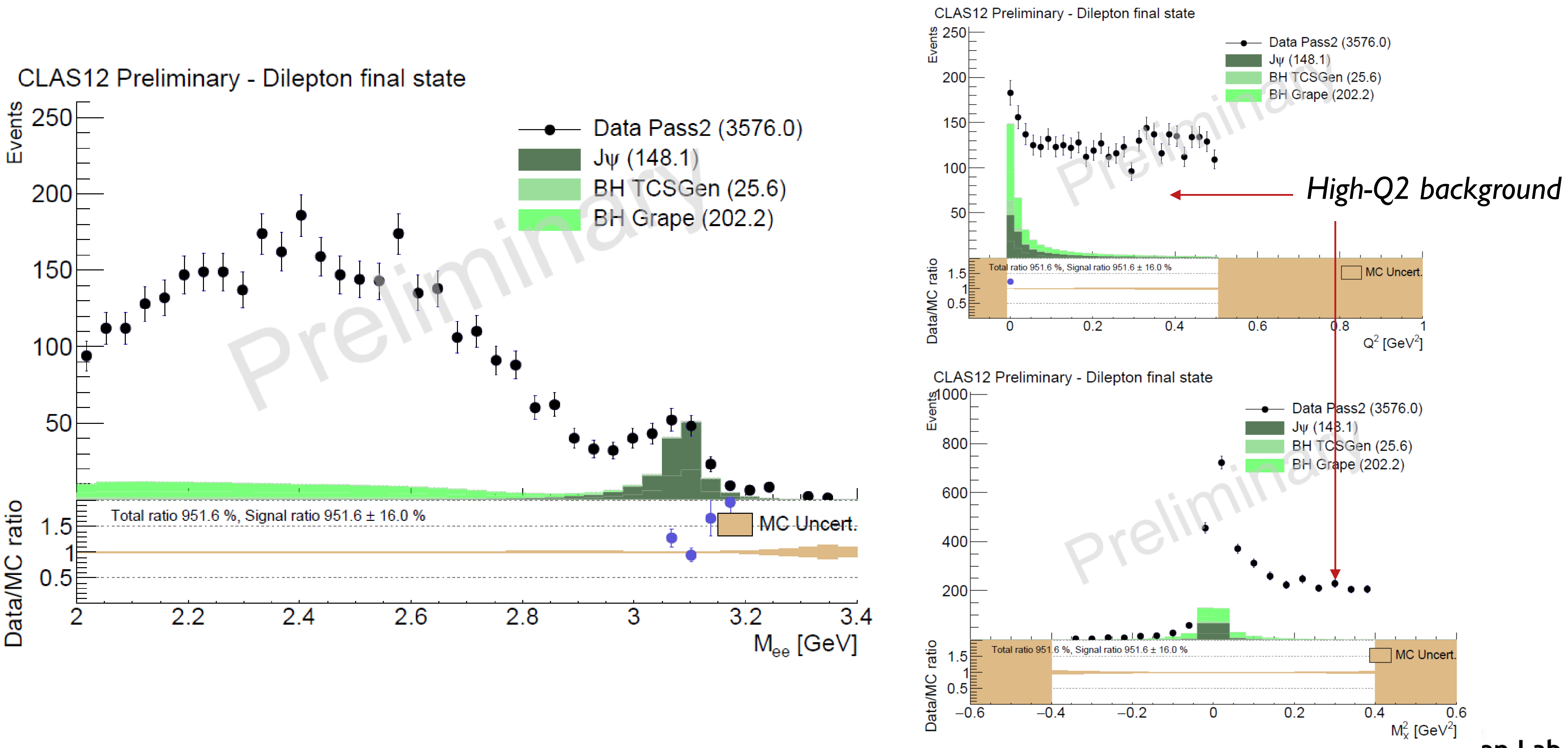




High-Q2 background



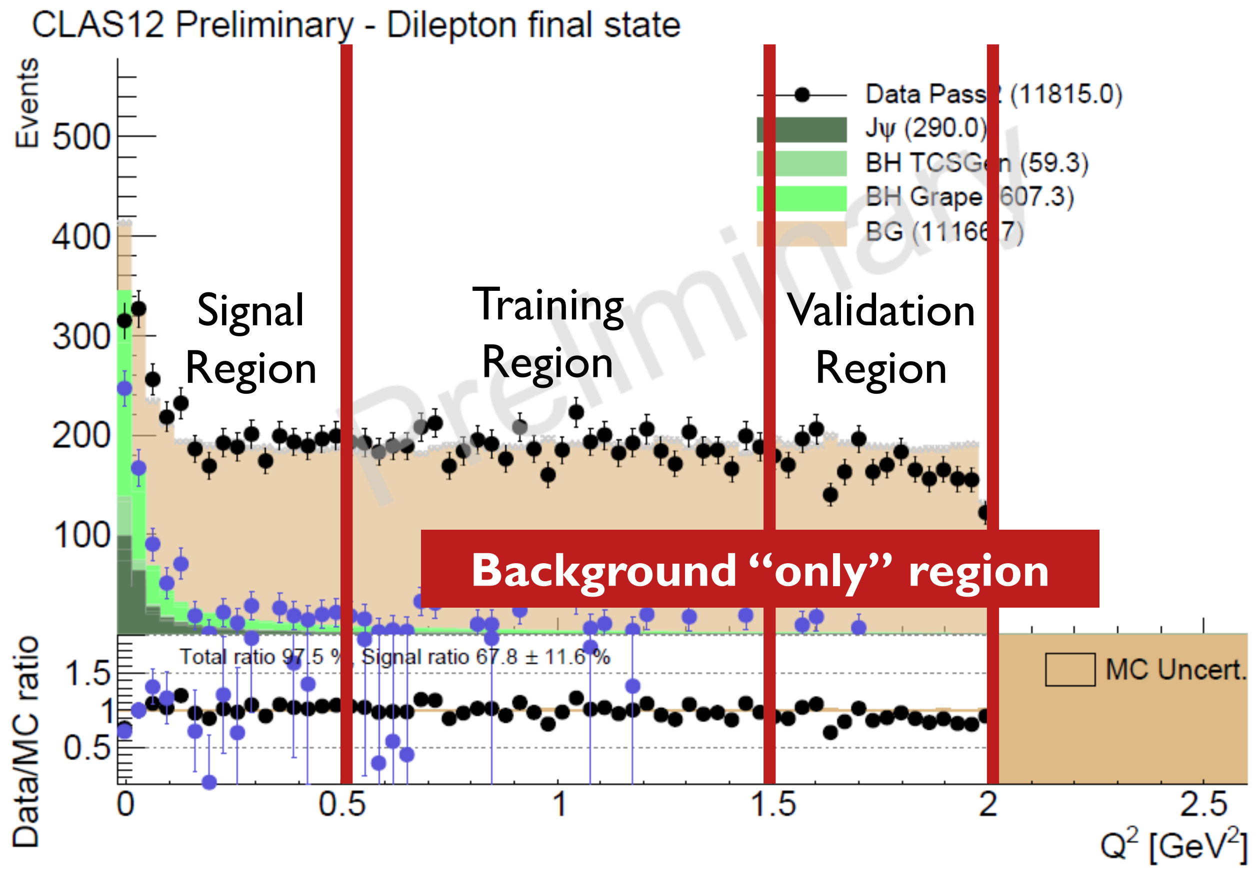
Comparison data/MC – Fall 2018 outbending



Overall strategy for background modelization

1) Event mixing
   - From data randomly select electron, positron, proton (from different events)
   - Construct kinematics and make sure they are within the region of interest (M_{ee}>2 GeV, $|MM|^2$<0.4 GeV^2, Q^2<2 GeV^2)

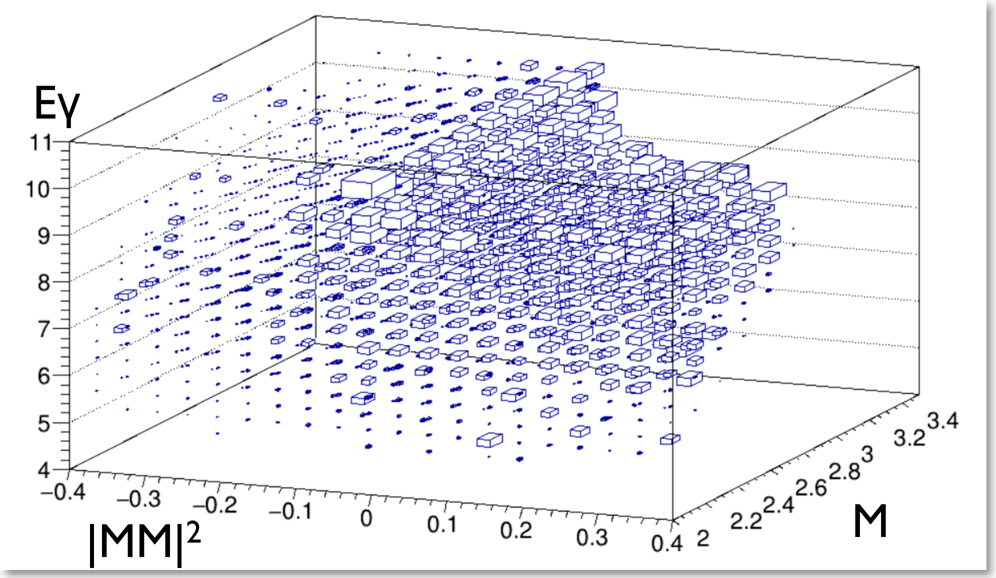
2) Reweight events to match data in the training region
3) Validate the weights on the validation region.
4) Apply weights on the signal region and obtained BG-subtracted yields



Reweighting methods

Binned weights

- Compute ratio $\omega = \frac{N_{target}|_{bin}}{N_{source}|_{bin}}$ and apply to event from the mixed BG sample.
- Inconvenient method
  1) Need to track bin indices
  2) Which variable to use ?
  3) Curse of dimensionality: the more variable, the less events per bins



Boosted decision trees

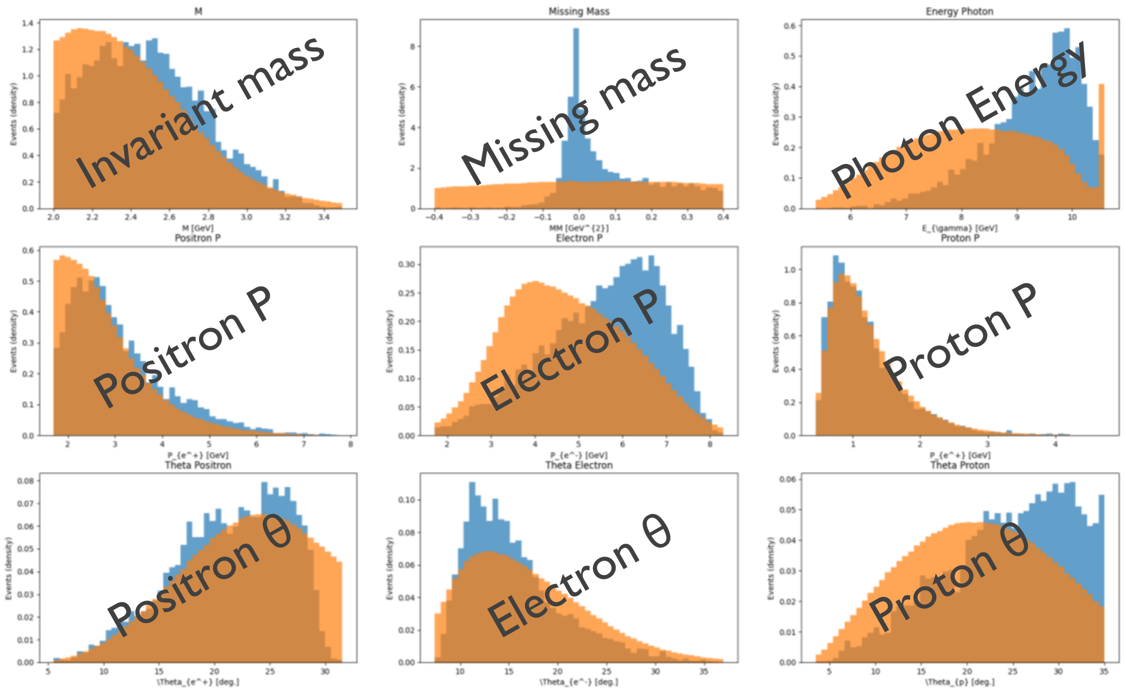
- Use a ML method to compute a weight event-by-event so that source and target distribution match
- Weights are obtained by optimizing a ML algorithm to distinguish target from source:

$$\omega = \frac{f_{target}(\mathbf{x})}{f_{source}(\mathbf{x})} = \frac{p_{target}(\mathbf{x})}{p_{source}(\mathbf{x})}$$

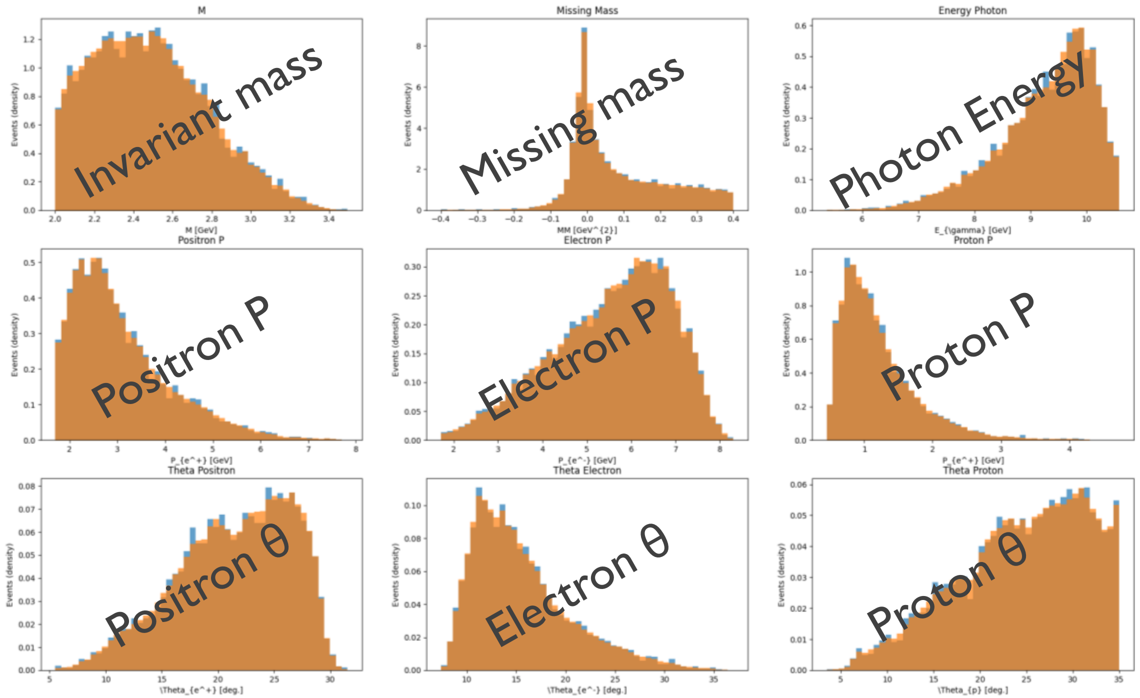
- Using method from Alex Rogozhnikov 2016 *J. Phys.: Conf. Ser.* **762** 012036. Code available here
- Advantages:
  1) As many variables as needed can be matched
  2) No/less of a dimensionality curse
  3) Easy to use, no need to handle complex bin indexing

Reweighting method using BDT

Before reweighting



After reweighting



Events in the training region (Target)

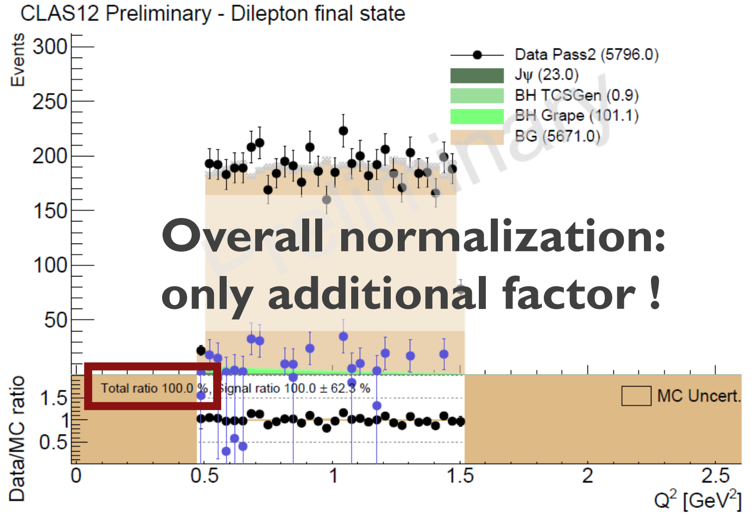
Mixed background (Source)

Jefferson Lab clas

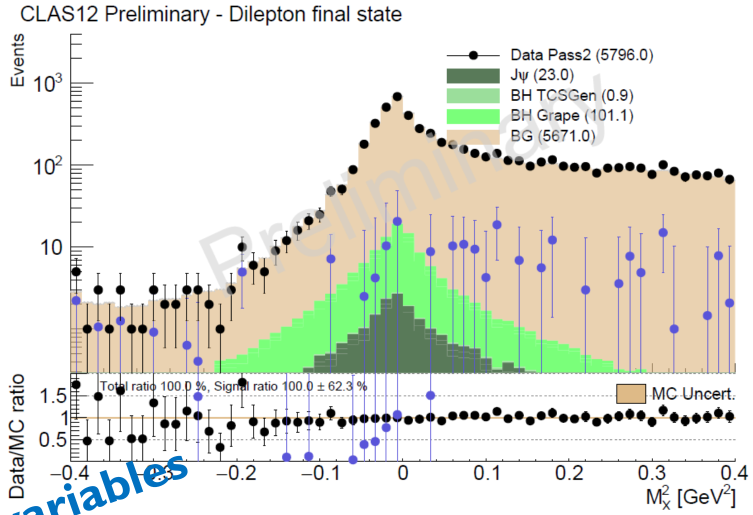
Reweighting training – Fall 2018 inbending

$Q^2 \in [0.5, 1.5] GeV^2$

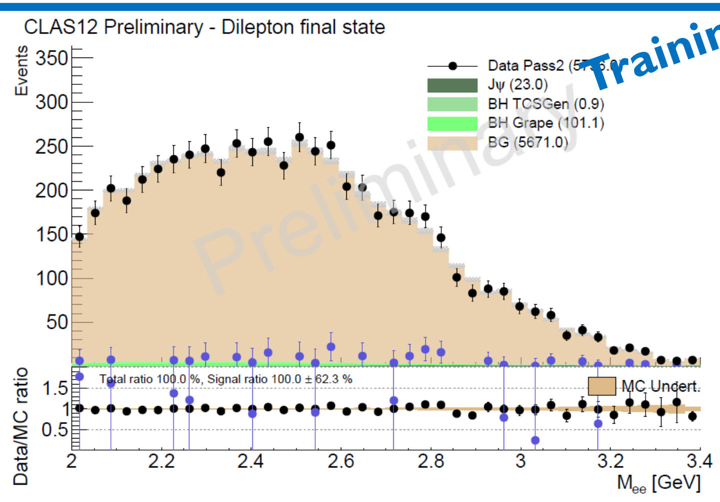
Training region

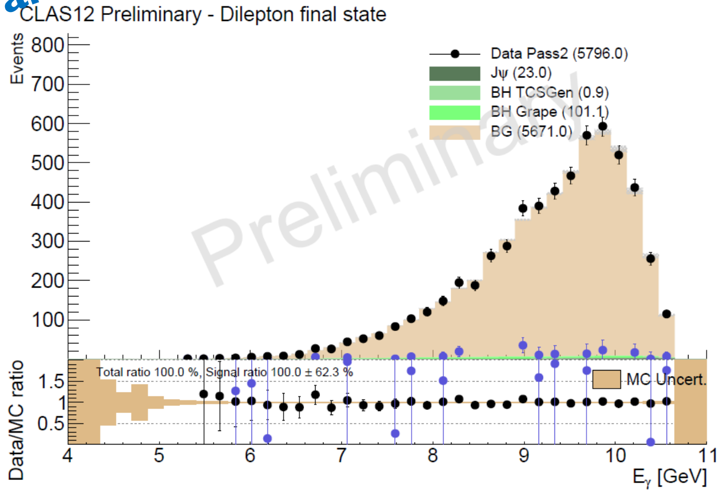


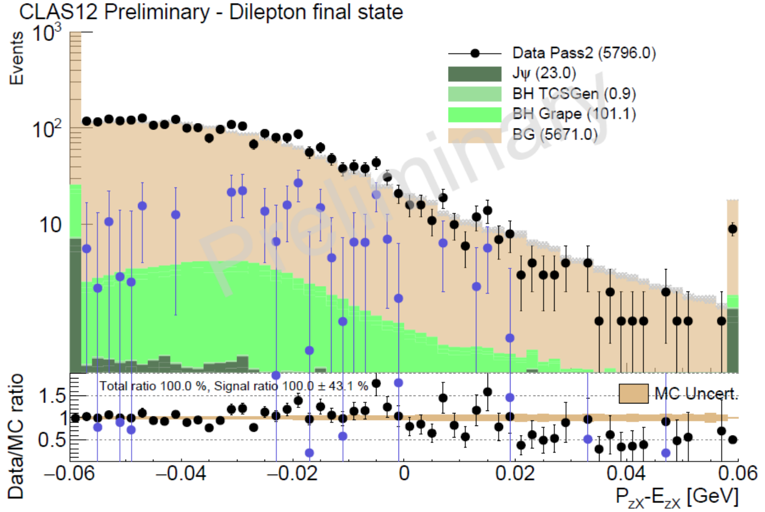
Overall normalization: only additional factor !

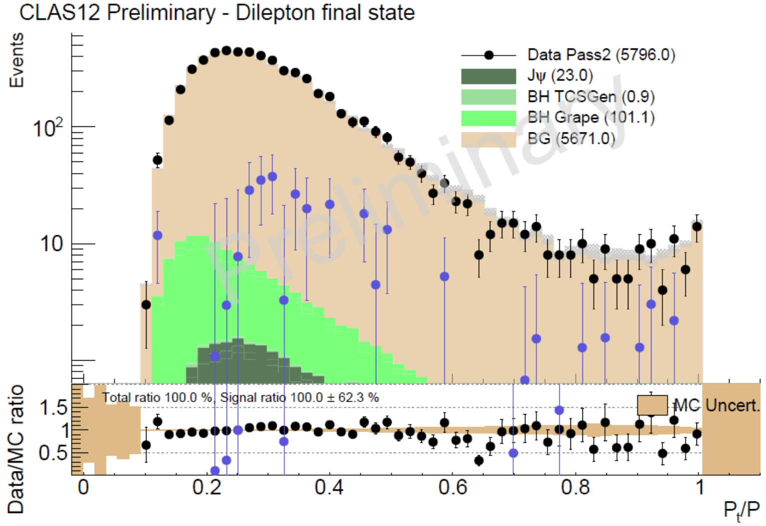


Training variables





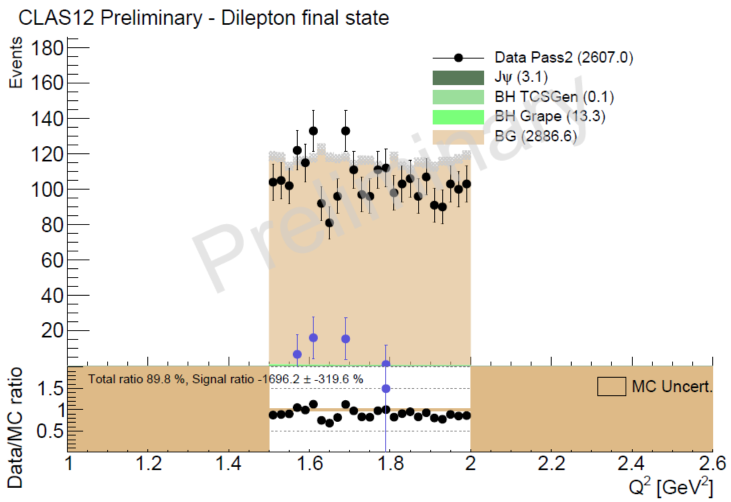


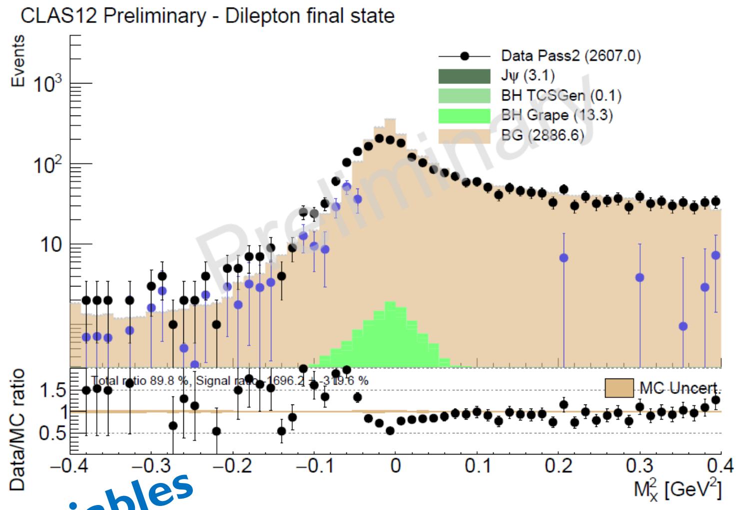


Reweighting validation – Fall 2018 inbending

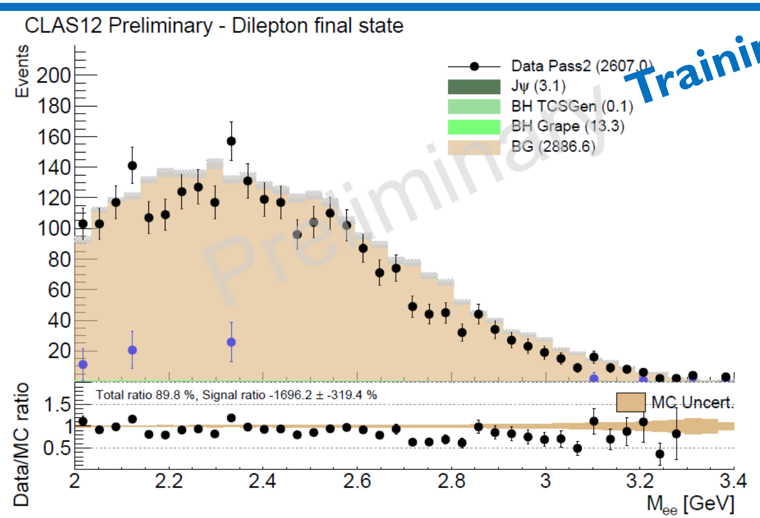
$Q^2 \in [1.5, 2.0] GeV^2$

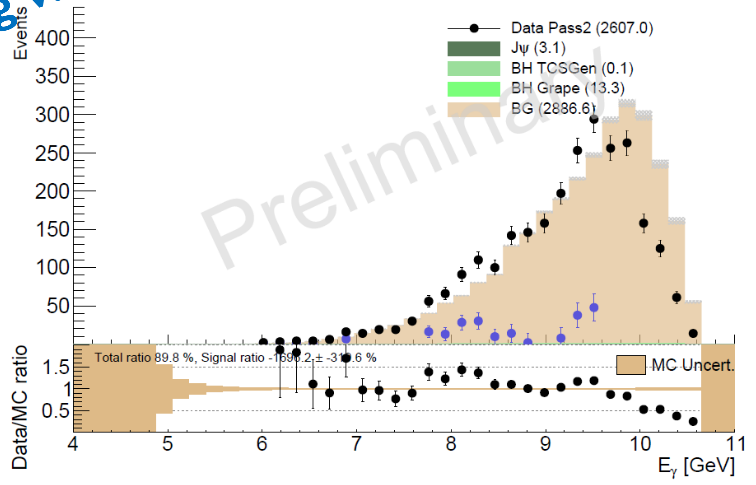
Validation region

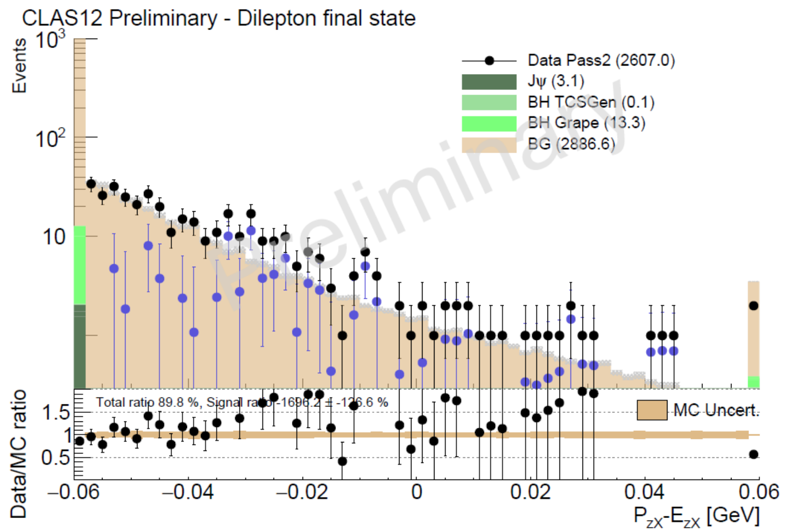


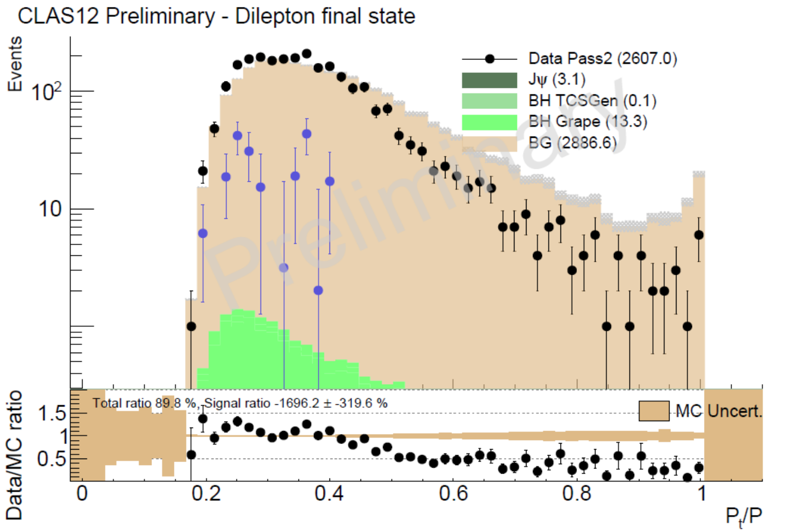


Training variables





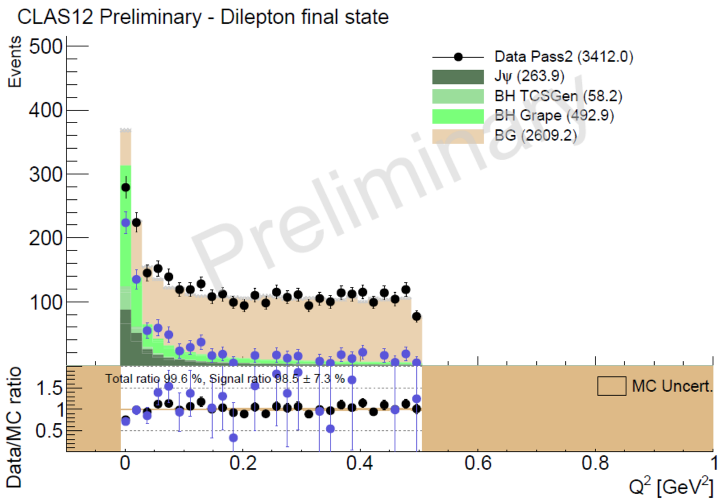


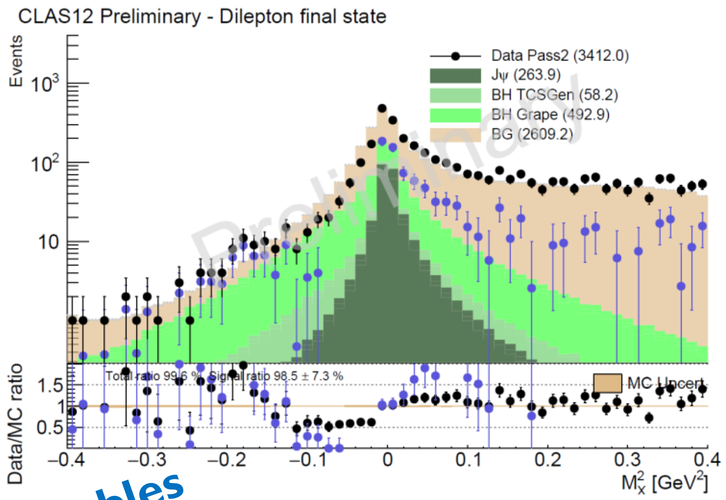


Full comparison data/MC – Fall 2018 inbending

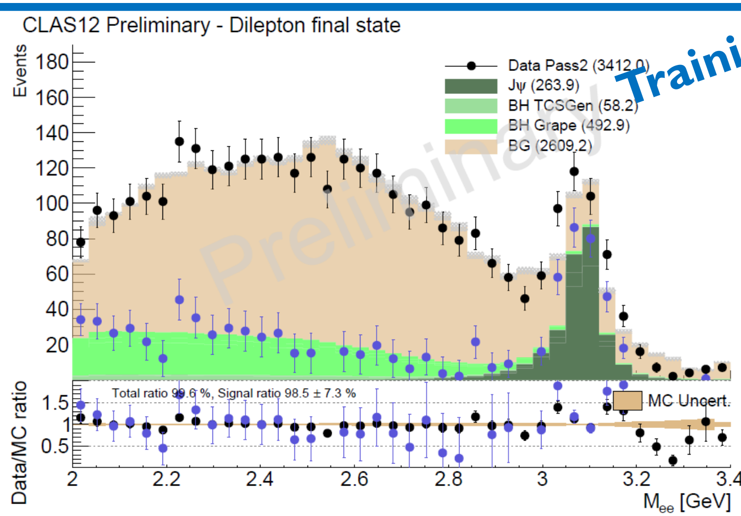
$Q^2 \in [0.0, 0.5] GeV^2$

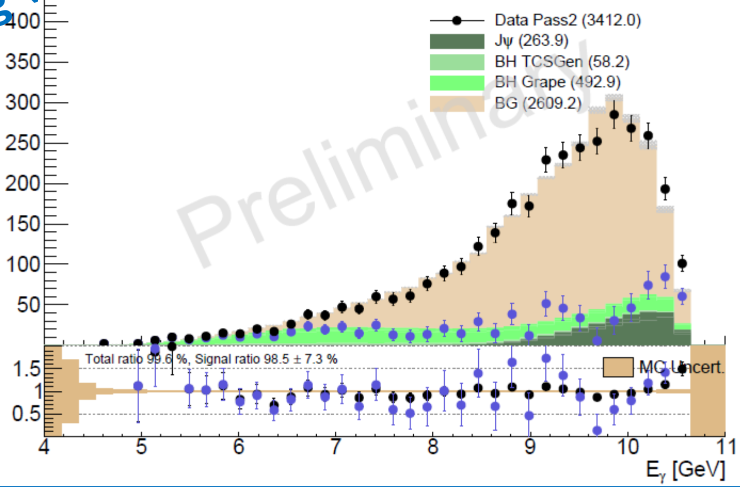
Signal region

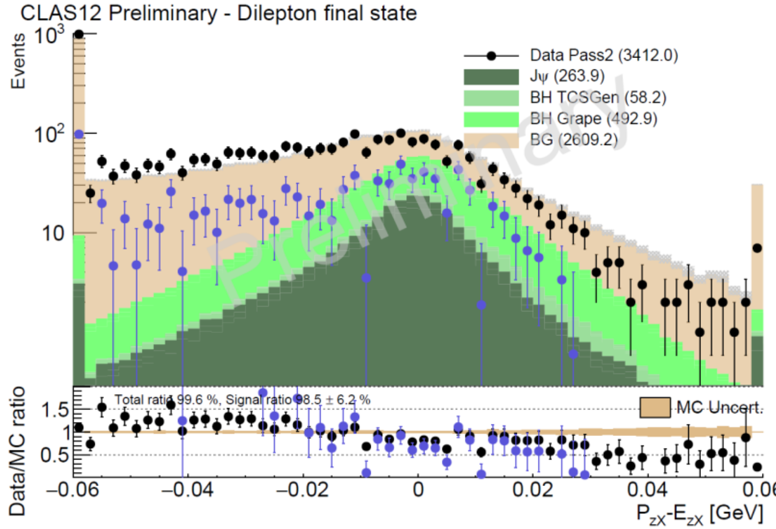


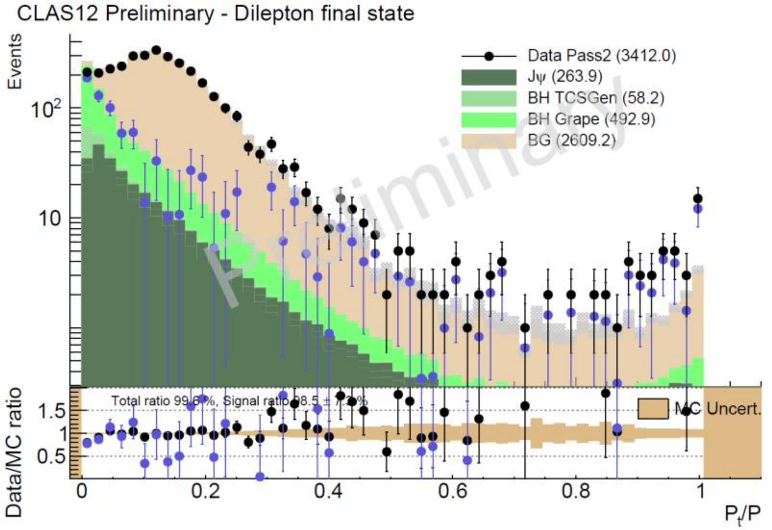


Training variables









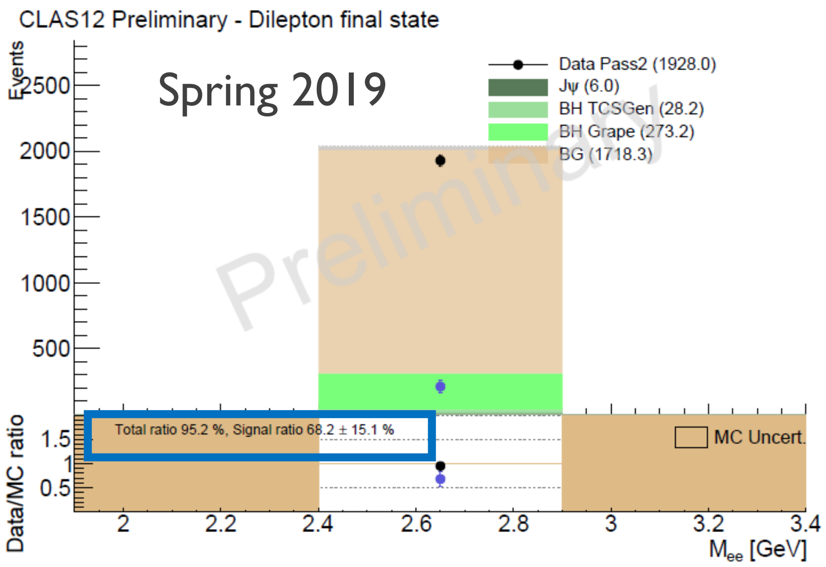
Application: normalization factor

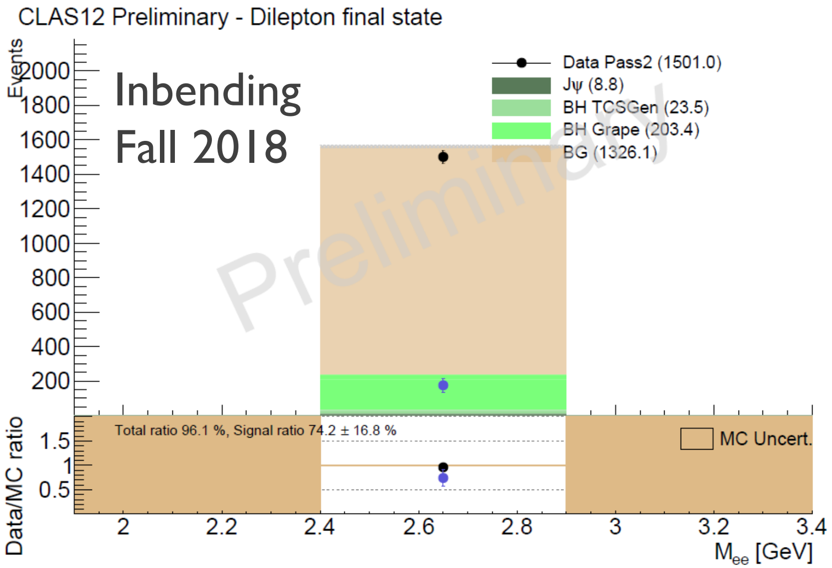
- Normalization factor can be computed as:

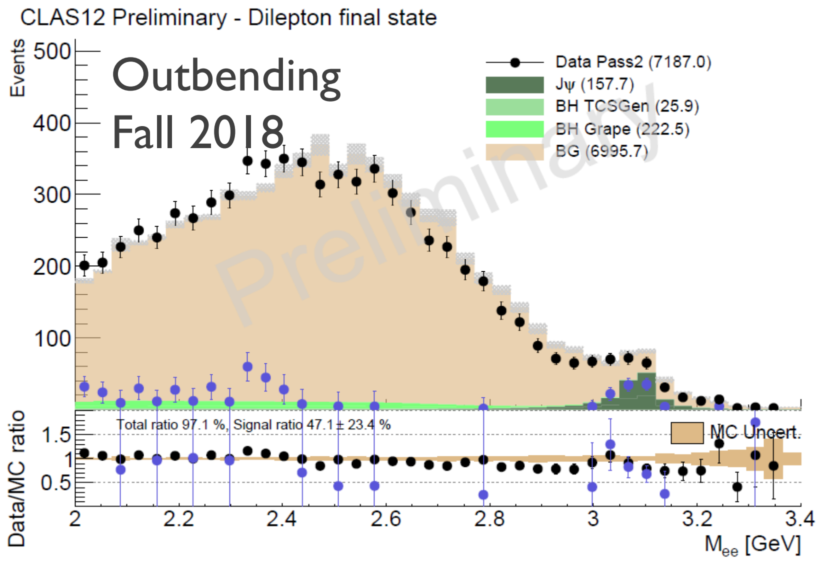
$$\omega_c = \frac{N_{Data} - N_{BG}}{N_{SIM\ \ BH}}$$

- Results:
  - Fall 2018 inbending – 74%
  - Spring 2019 inbending – 69%
  - Fall 2018 outbending – to be continued

Normalization factor: 0.7







Take-aways

I The JPsi analysis needs a good understanding of the background to extract the overall normalization.

II We have developed a method based on reweighted - event mixing.

III We have shown that this method allows to model the background of the Jpsi final state.

V **Systematics uncertainties will be studied (and reduced), as the overall normalization is crucial for the cross-section extraction**

Back-up

Data and MC samples

- Analysis on Pass 2 data. All *main* Fall 18 (Inbending and outbending) and Spring 19 runs are processed.
- Simulations are processed through OSG with pass 2 configuration
- The QADB tool is used to clean-up data and retrieve the accumulated charge per DST files
- The RCDB interface of clas12root is used to retrieve the beam current for each run
- Accumulated charge is computed per beam current for each configuration

	Config / Beam currents / Charge					
	Fall 18 In.			Fall 18 Out.		Sp. 19
Generator	45 nA 26.312 mC	50 nA 4.000 mC	55 nA 5.355 mC	40 nA 11.831 mC	50 nA 20.620 mC	50 nA 45.994 mC
Grape	8.2M each					6.7 M
TCSGen	2M each					1.5 M
JPsiGen	2M each					
JPsiGen (No rad.)	3M each					
Total of 24 MC samples and 3 Data samples						

Data/MC normalization

- Each event is weighted by:

$$\omega = \frac{\mathcal{L} \cdot \sigma_{tot}}{nb_{GEN}}$$ for generator providing integrated CS, $$\omega = \frac{\mathcal{L} \cdot w_{GEN}}{nb_{GEN}}$$ for weighted generator.

- Where the luminosity is obtained from target specification:

$$\mathcal{L} = \frac{l \cdot \rho \cdot N_A \cdot C \cdot Q}{e} = 1316.875 \cdot Q(\text{in mC})$$

Length of the target l = 5 cm
Density of the target $\rho = 0.07$ g/cm^3
Avogadro constant $N_A = 6.02 \times 10^{23}$ mol^{-1}
Unit charge $e = 1.6 \times 10^{-19}$ C
Conversion to pb $C = 10^{-36}$

Note on normalization method

Jefferson Lab clas

Full comparison data/MC – Fall 2018 inbending

$Q^2 \in [0.0, 0.5] GeV^2$ Region C (Signal) - Final state particle kinematics