Applications of reweighting techniques with XGBoost in BESIII analysis

Beijiang Liu, Xiaoyan Shen, Xian Xiong

Institute of High Energy Physics

Outline

- Introduction
- Two use cases
  - Efficiency calculation
  - Background estimation
- Summary

Introduction

- Efficiency calculation
  - Monte-Carlo(MC) consistent with Real data(RD) is needed
- Background estimation
  - MC: describe Background shape accurately in real data

- Structures in data
- PHSP MC can not describe the data in most of the cases
- PWA results may not be available

- How to get data-like MC?

Typical method for reweighting

- Easy to correct one distribution of 1D or 2D to another distribution
- Reweighting with bins:

$$weight\ factor_{bin} = \frac{\omega_{bin,RD}}{\omega_{bin,MC}}$$

- Difficult for a high-dimension reweighting
  - Higher statistics are needed
  - correlations among variables

Reweighting with Machine Learning

- Density ratio in reweighting [1]:

$$\frac{f_{RD}(x)}{f_{MC}(x)}$$

- Classifier to distinguish Data and MC provide probabilities $p_{RD}(x)$ and $p_{MC}(x)$

$$weight\ factor\ w(x) = \frac{f_{RD}(x)}{f_{MC}(x)} \sim \frac{p_{RD}(x)}{p_{MC}(x)}$$

- We utilize this approach with XGBOOST algorithm

- Define symmetrized χ^2 by binning the multi-dimensional space[2]:

$$\chi^2 = \Sigma_{bin} \frac{\left(w_{bin,original} - w_{bin,target}\right)^2}{w_{bin,original} + w_{bin,target}}$$

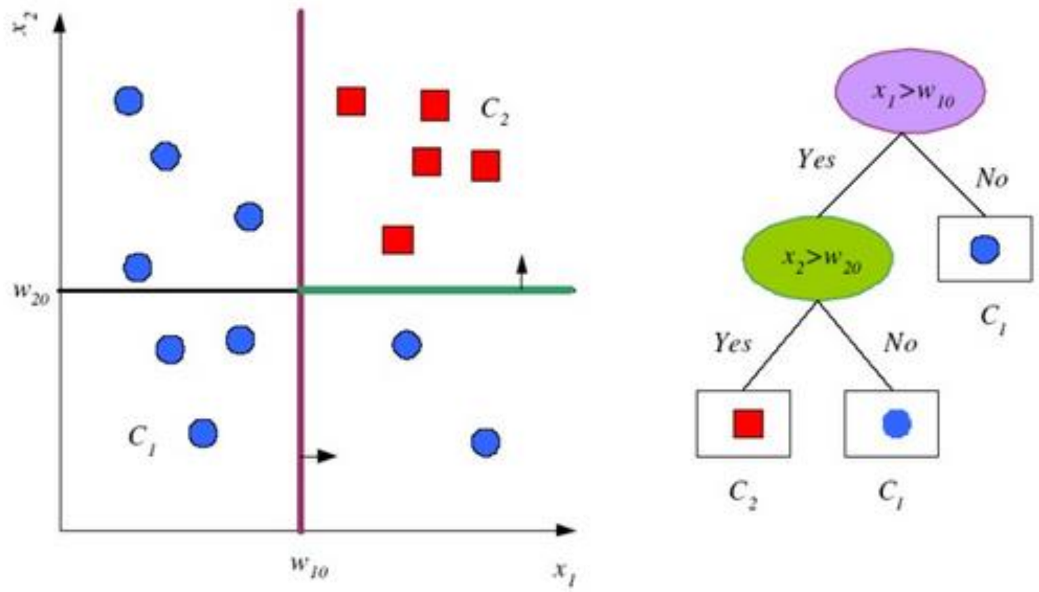
- Finding regions with high difference by maximizing χ^2 and give the weights directly

*applications in LHCb, e.g.[3]

[1] Martschei, D., Feindt, M., Honc, S., & Wagner-Kuhr, J. (2012). Advanced event reweighting using multivariate analysis. *Journal of Physics: Conference Series*, *368*(1), 012028. https://doi.org/10.1088/1742-6596/368/1/012028
[2]Rogozhnikov, A. (2016). Reweighting with Boosted Decision Trees. *Journal of Physics: Conference Series*, *762*(1), 012036. https://doi.org/10.1088/1742-6596/762/1/012036
[3] LHCb Collaboration(Roel Aaij(CERN) et al.), JHEP 1703 (2017) 001

Introduction for ML

- Machine learning:
  - Algorithms: learn from and make predictions on data
  - Supervised learning: Classification, Regression, …
  - Unsupervised learning: Clustering, …
- Decision Tree:
  - Continuously split according to a certain parameter to form a Decision Tree
  - Ability to predict in the unseen data



Introduction for XGBoost

- XGBoost : eXtreme Gradient Boosting
  - Boosting: additive training of Decision Tree
  - Gradient: first and second derivative
  - High flexibility: DIY loss function, multi-variables input
  - More robust: L1 and L2 penalty factors
  - Faster: multi-thread parrallel

  - Adaptive splitting: $Gain = \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma$

G,H: first and second derivative
λ, γ:para reflect the complexity of tree

Input:
High dimension

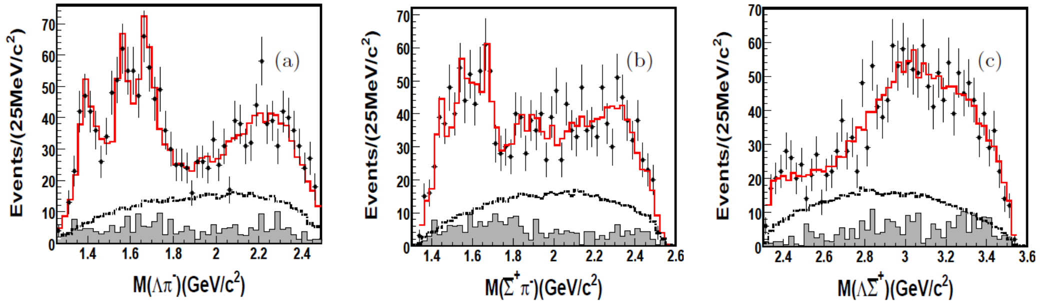
→

Output:
low dimension

For both classification and regression problem

Use case 1: Efficiency calculation

Observation of $\psi' \to \Lambda\bar{\Sigma}^{\pm}\pi^{\mp} + c.c.$



- Many structures in the data
- PHSP MC cannot describe the data

To get the efficiency:

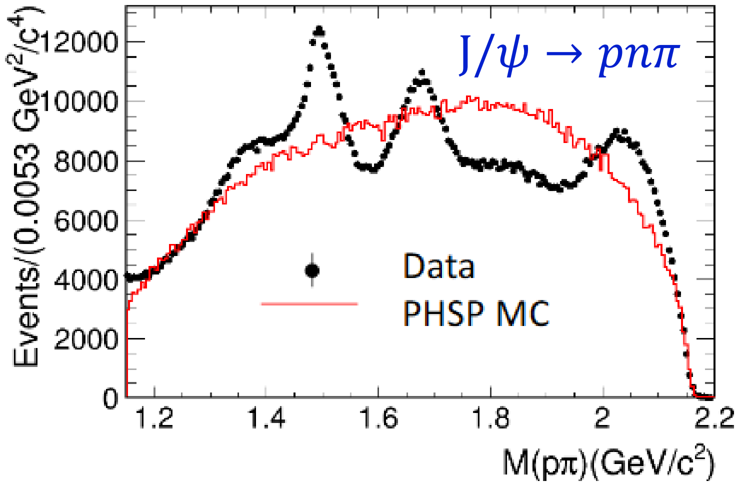
- Solution
  - do a “quick” PWA, e.g. BESIII Phys.Rev. D88, 112007 (2013)
- New solution
  - reweighting

A test with $J/\psi \to pn\pi$

- Input: 'Data'(DIY MC) and PHSP MC sample
- Use a XGBoost classifier to get the probabilities of $p_{RD}(x)$ and $p_{MC}(x)$
- weight factor: $w(x) = \frac{p_{RD}(x)}{p_{MC}(x)}$
- Efficiency with reweighting:

$$\epsilon = \frac{\sum_{i=1}^{N_{phsp}} w_i^{phsp}}{\sum_{i=1}^{N_{truth}} w_i^{truth}}$$

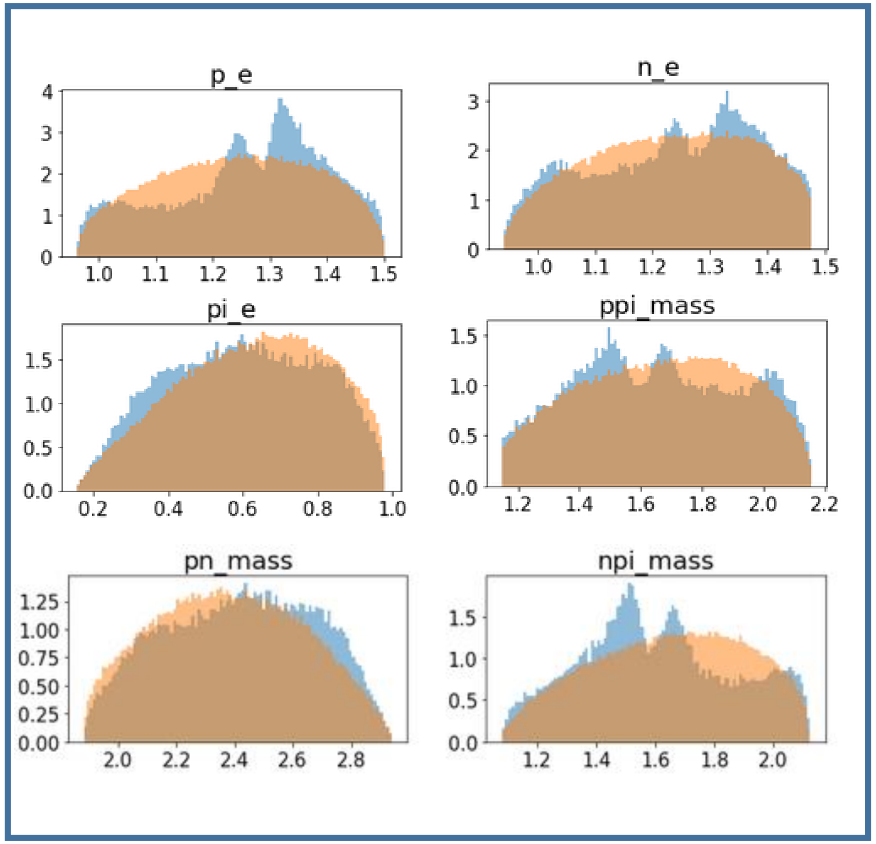
Yanjia Xiao



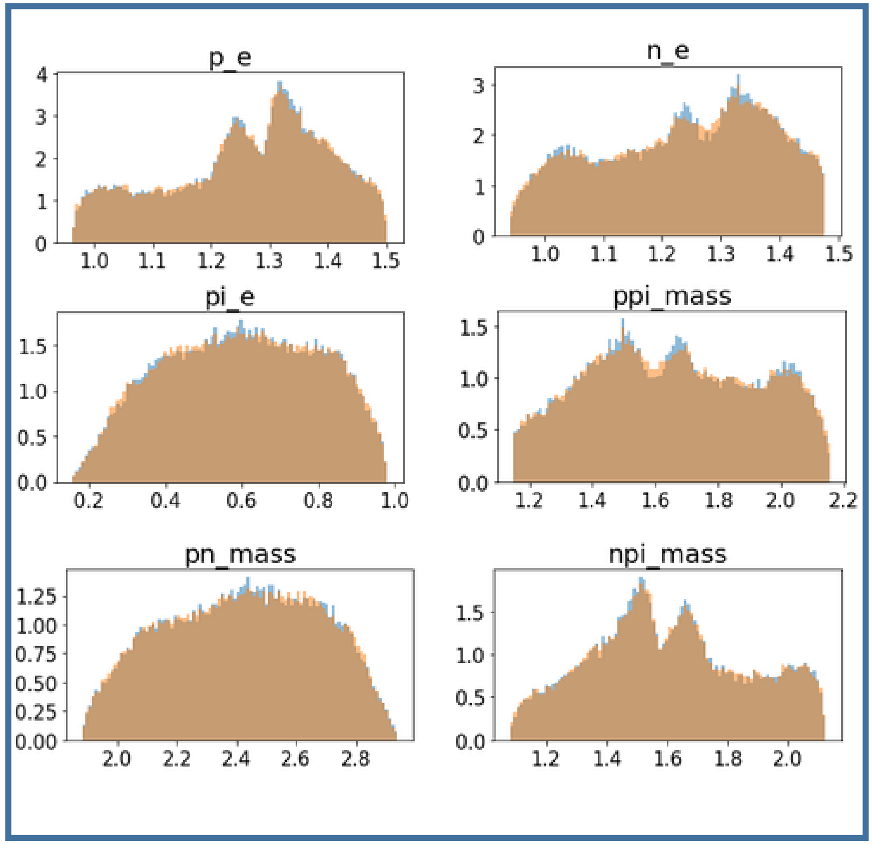
- Number of d.o.f : 4
- Input Variables:
  - Four-momentum of p, n and π
  - $M(pn), M(p\pi), M(n\pi)$

Results

Original distribution



After reweighting



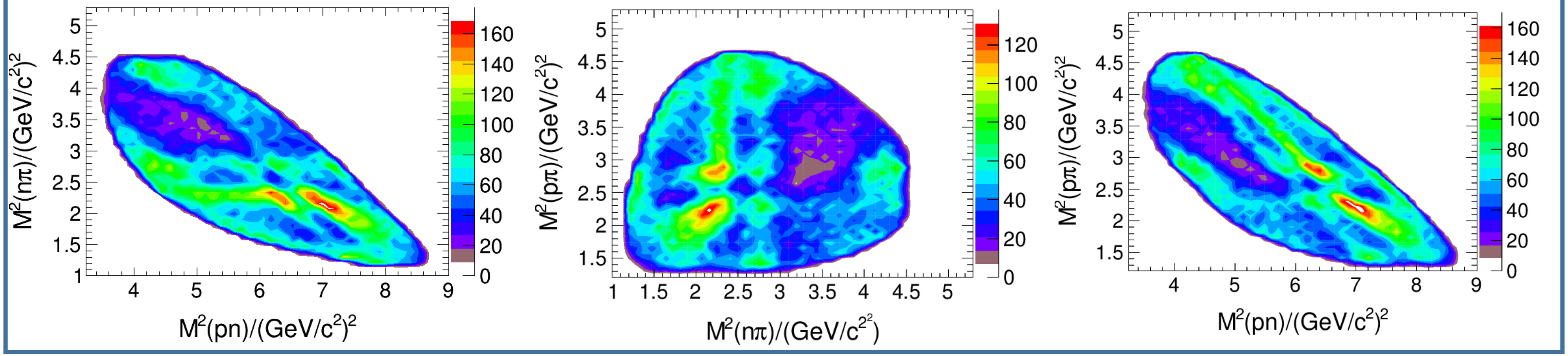
- ‘Data’ (DIY MC)(blue)
- PHSP(brown)

	Efficiency
‘Data’(DIY MC)	$(68.42\pm0.16)\%$
PHSP MC	$(75.71\pm0.08)\%$
Reweighting	67.94%

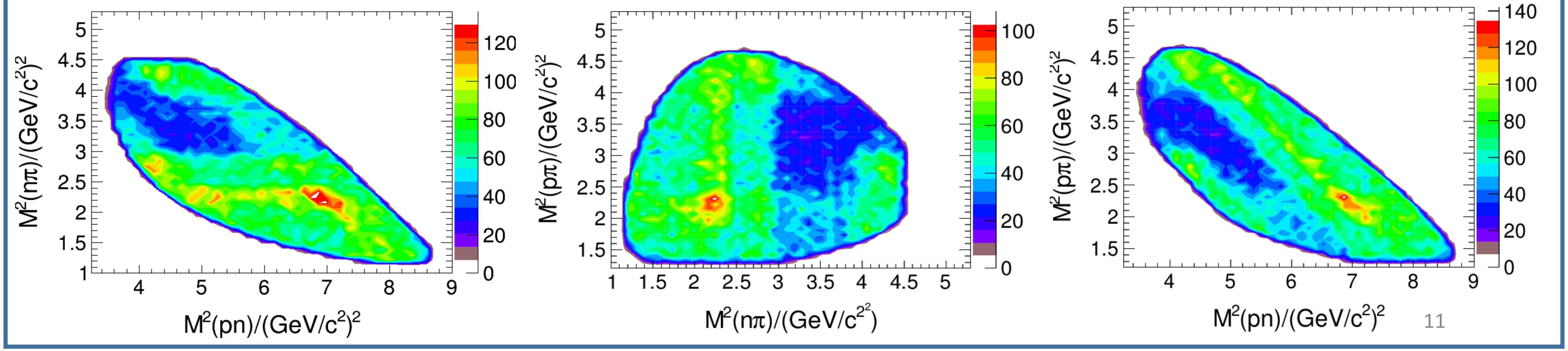
- Reweighted PHSP MC can describe the ‘data’ (DIY MC as pseudo data)
- Efficiency of reweighted PHSP MC is consistent with the input

Results

Target (DIY MC):

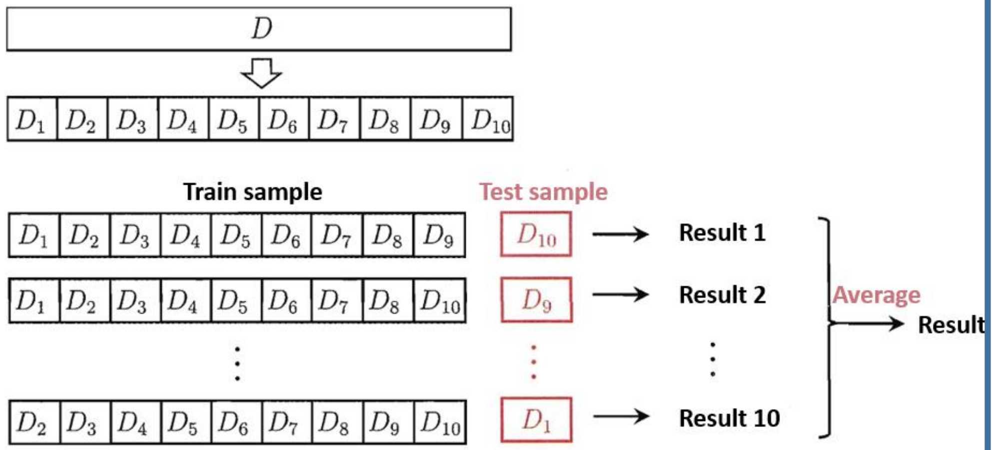


Reweighted MC



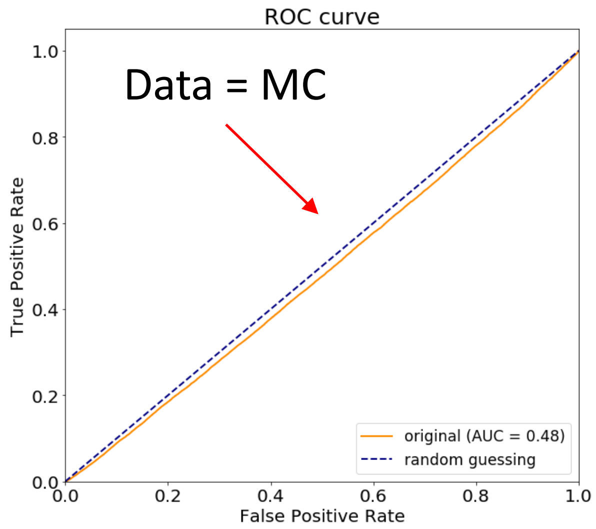
Performance

- 10-fold cross validation：
- Decrease the uncertainties caused by sample division



Performance

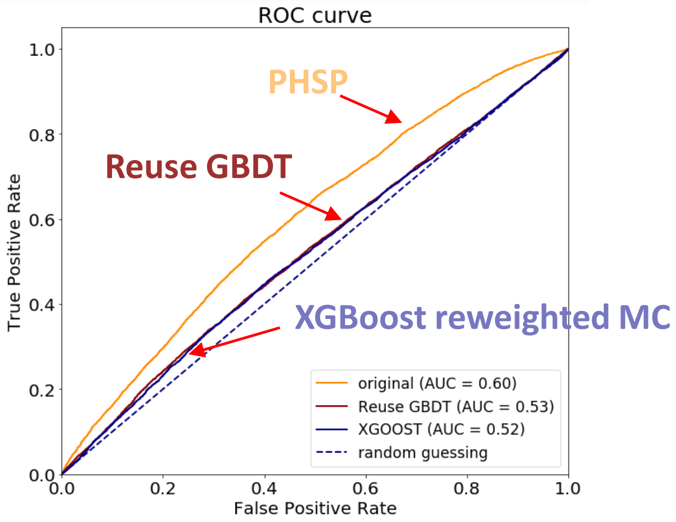
- ROC curve: a metric to evaluate the goodness of a classifier
- If two samples are same:
  - Classifier can't discriminate
  - ROC curve will close to dash line



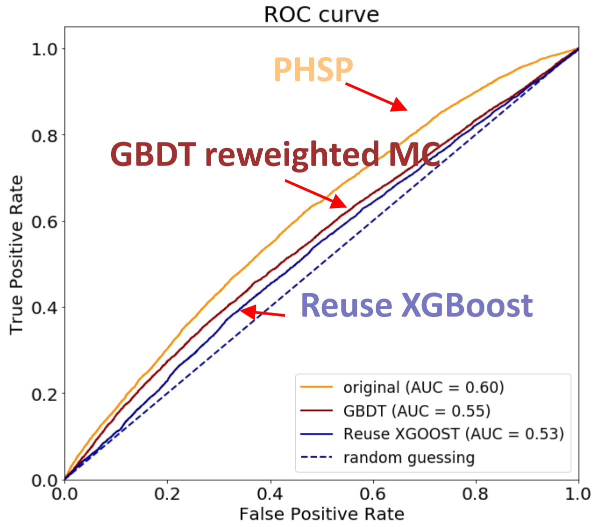
Performance

- For comparison, the symmetrized χ^2 approach with GBDT is also performed
- The performance has been checked
  - The discrepancy of PHSP and data can be significantly reduced with both approach
  - It seems that the results of GBDT can be slightly improved by the our implementation

Reuse GBDT



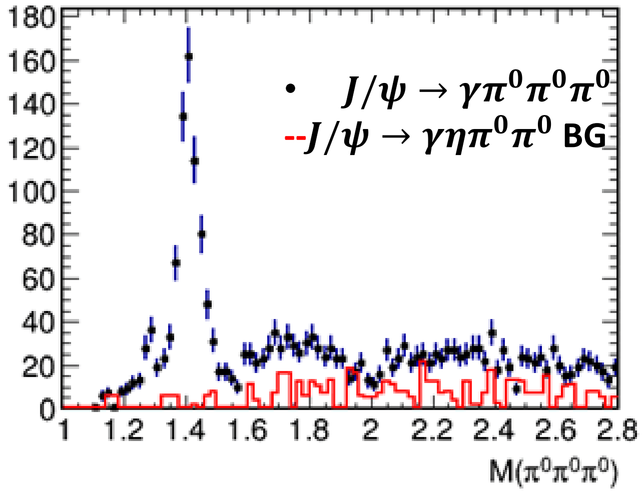
Reuse XGBoost



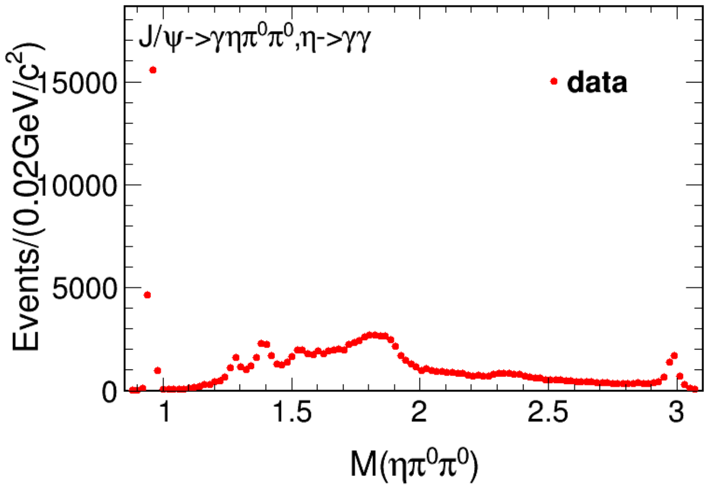
Use case 2: Background estimation

By Qiao zhou

- BG estimation is important, especially in PWA
- In the analysis of $J/\psi \to \gamma\pi^0\pi^0\pi^0$, there's a dominate irreducible BG $J/\psi \to \gamma\eta\pi^0\pi^0$
  - rich structures
  - can't be subtracted with sideband
- How to model an irreducible BG?
  - Perform another PWA to $J/\psi \to \gamma\eta\pi^0\pi^0$ first?



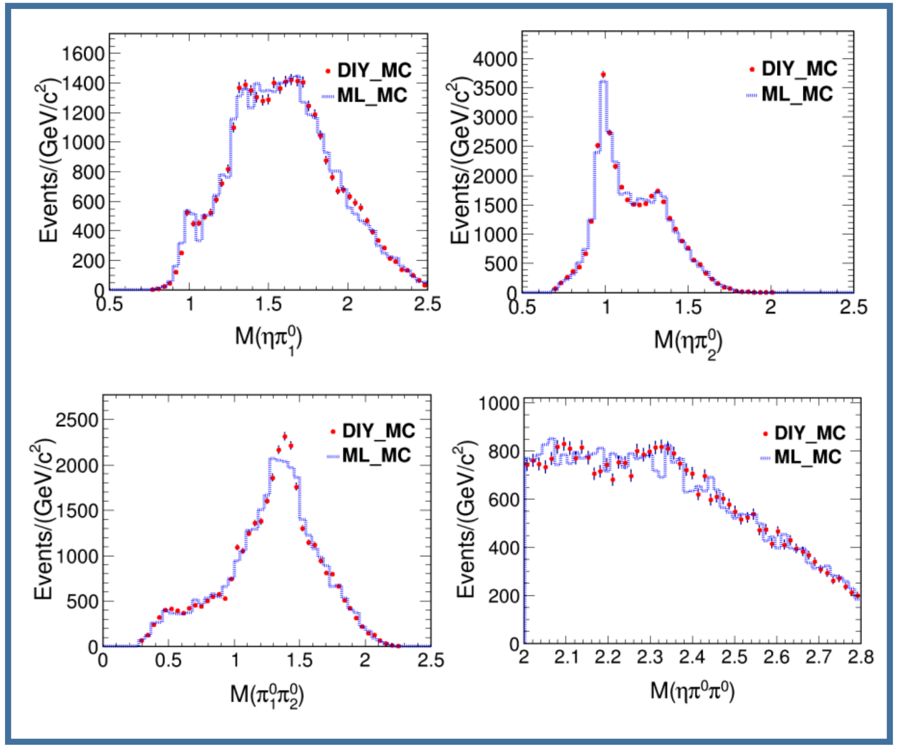
Can we get the contribution of $J/\psi \to \gamma\eta\pi^0\pi^0$ to $J/\psi \to \gamma\pi^0\pi^0\pi^0$ from the selected $J/\psi \to \gamma\eta\pi^0\pi^0$ events ?



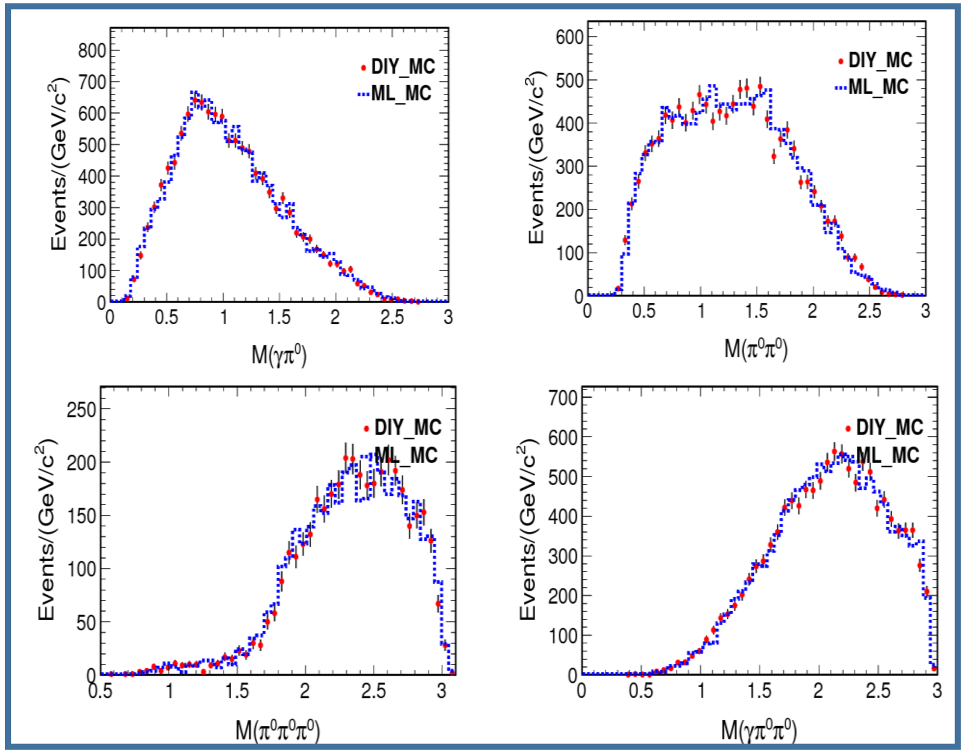
Background estimation in $J/\psi \to \gamma\pi^0\pi^0\pi^0$

- Reweight $J/\psi \to \gamma\eta\pi^0\pi^0$ PHSP MC to obtain a data-like MC, N.dof=7

$J/\psi \to \gamma\eta\pi^0\pi^0$ events after $J/\psi \to \gamma\eta\pi^0\pi^0$ event selection



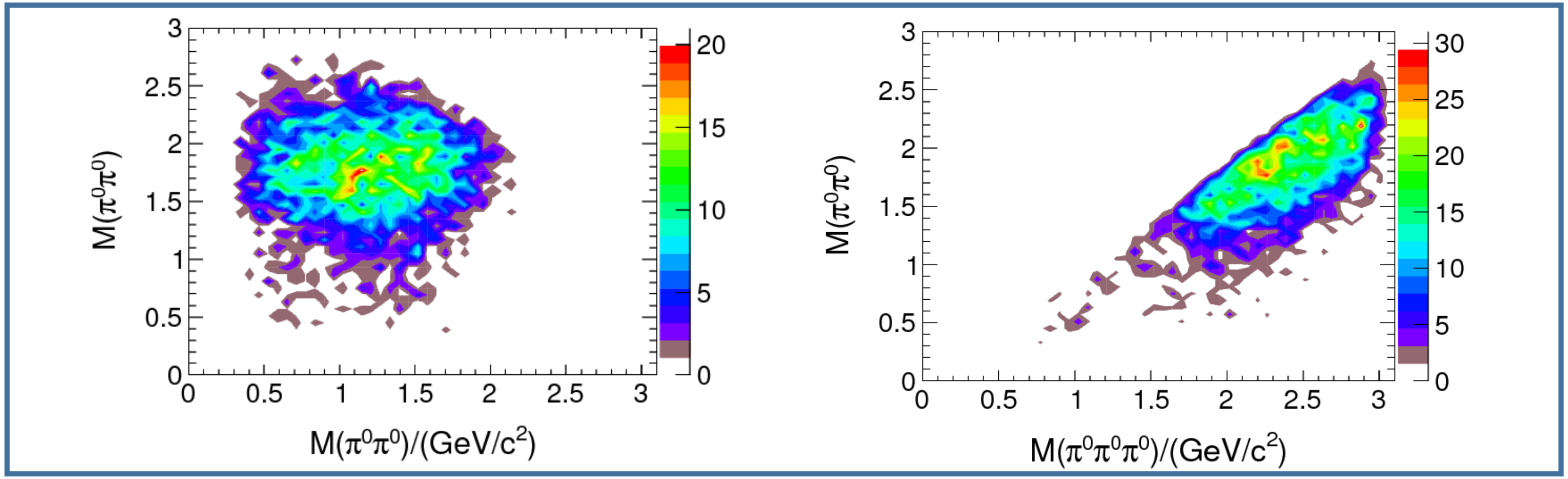
$J/\psi \to \gamma\eta\pi^0\pi^0$ events after $J/\psi \to \gamma\pi^0\pi^0\pi^0$ event selection



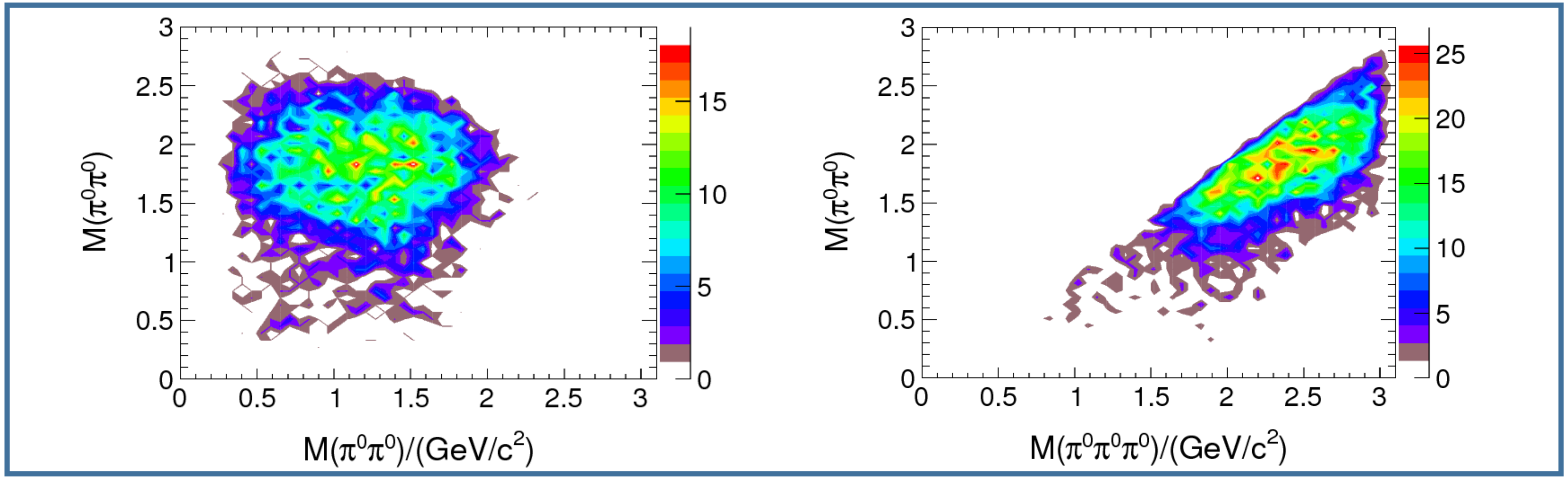
- The reweighted MC can describe the “data” (DIY MC) well

Background estimation in $J/\psi \to \gamma\pi^0\pi^0\pi^0$

DIY_MC



ML_MC



Summary

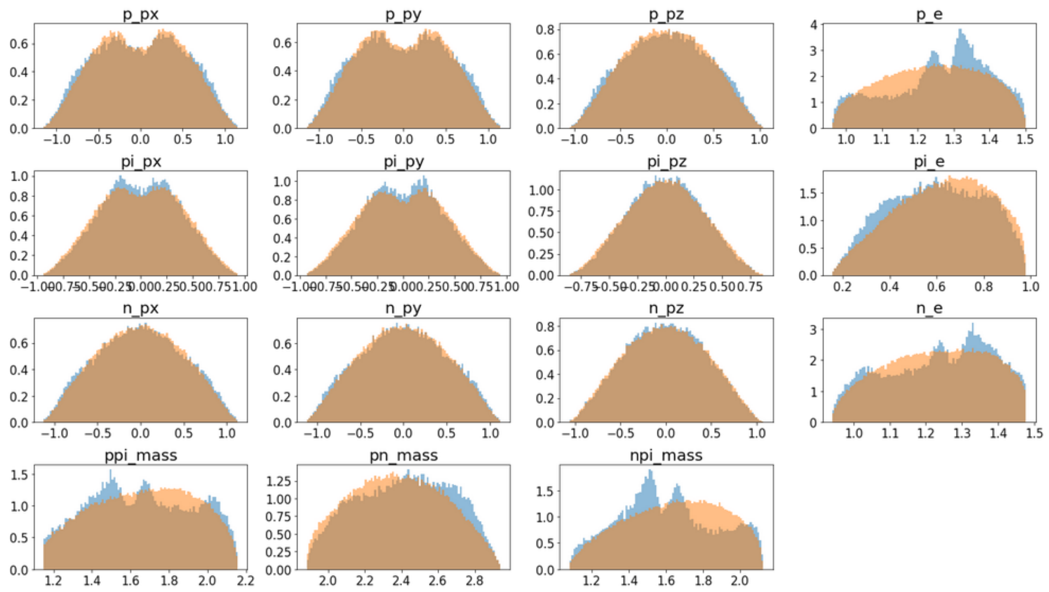
- High dimensional rewighting problem can be solved with ML methods
- We implement a reweighter with XGBoost:
  - Works fine in high dimensional use cases of BESIII analysis:
    - Efficiency calculation
    - Background estimation

- Available on Github: https://github.com/rhineryan/rewighting/blob/master/DoReweight_final.ipynb

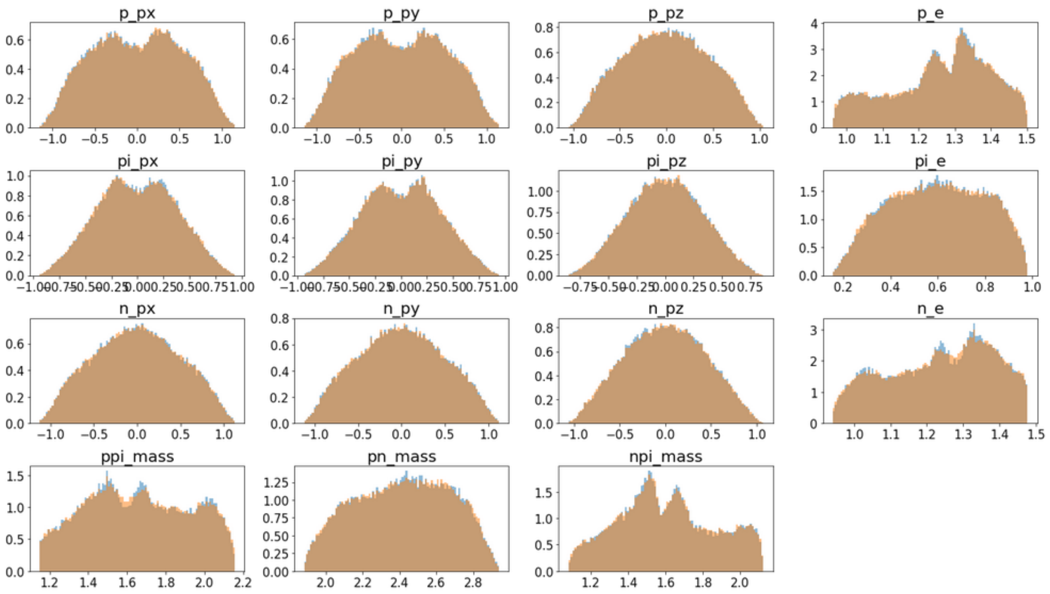
Backup

Efficiency correction with $J/\psi \rightarrow pn\pi$ three body decay

Original distribution



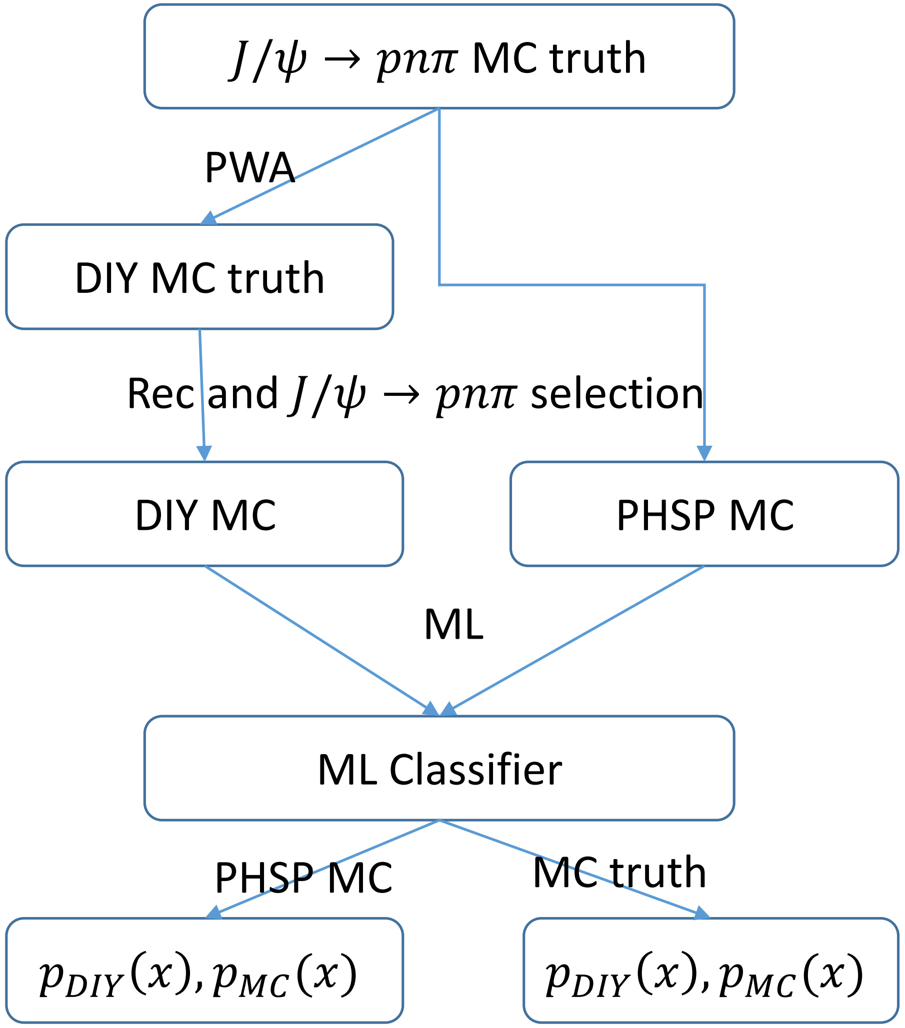
After reweighting



Efficiency correction with $J/\psi \to pn\pi$ three body decay

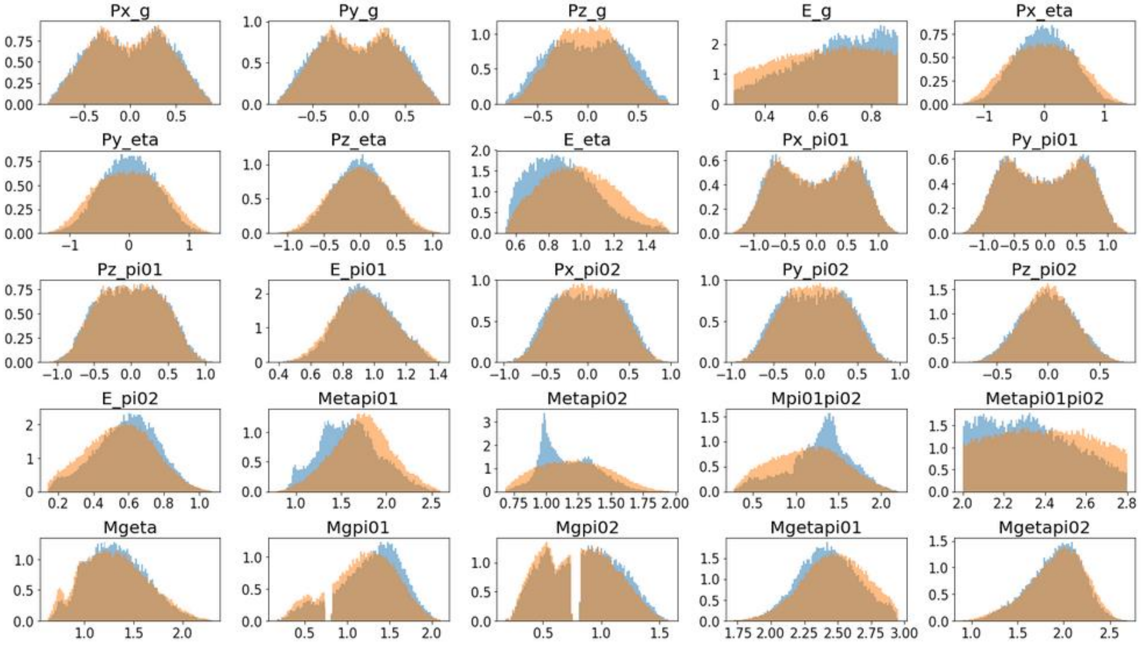
- Input sample: DIY MC and PHSP MC
- Output : probabilities belongs to DIY or PHSP
- weight factor: $w(x) = \frac{p_{DIY}(x)}{p_{MC}(x)}$
- Efficiency after reweighting:

$$\epsilon = \frac{\sum_{i=1}^{N_{phsp}} w_i^{phsp}}{\sum_{i=1}^{N_{truth}} w_i^{truth}}$$

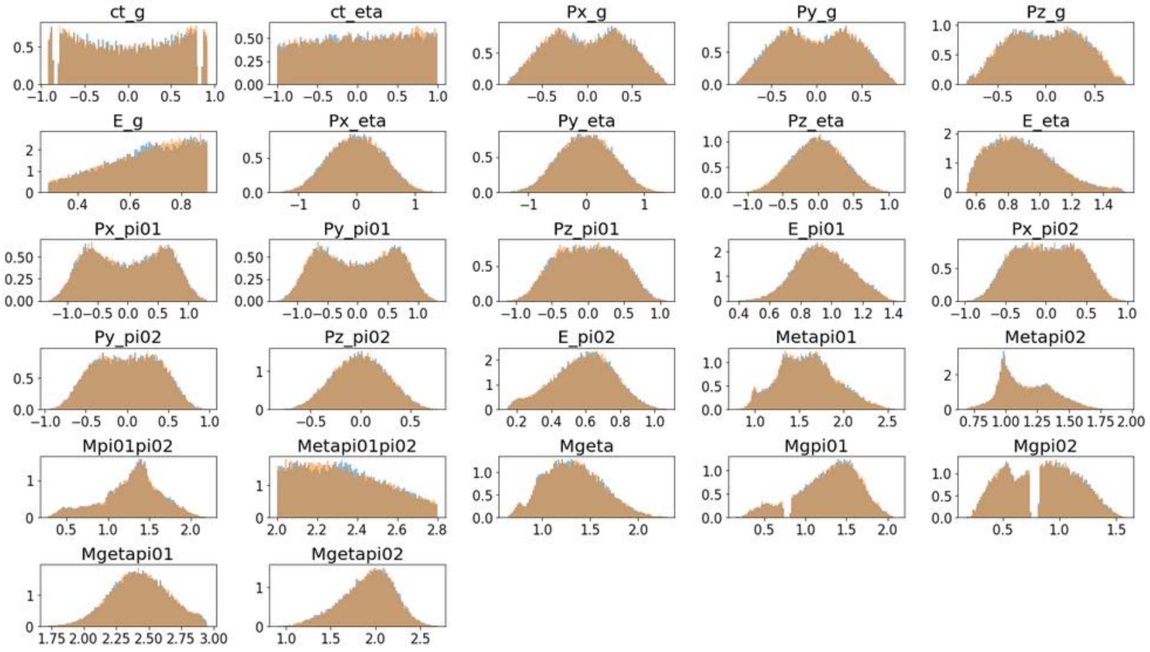


Efficiency correction with $J/\psi \to \gamma\eta\pi^0\pi^0$ four body decay

Original



After reweighting



Background estimation in $J/\psi \to \gamma\pi^0\pi^0\pi^0$

- $J/\psi \to \gamma\eta\pi^0\pi^0$ events contribute to the $J/\psi \to \gamma\pi^0\pi^0\pi^0$ analysis
- Strategy: