Entropy, Geometry and Collider Physics

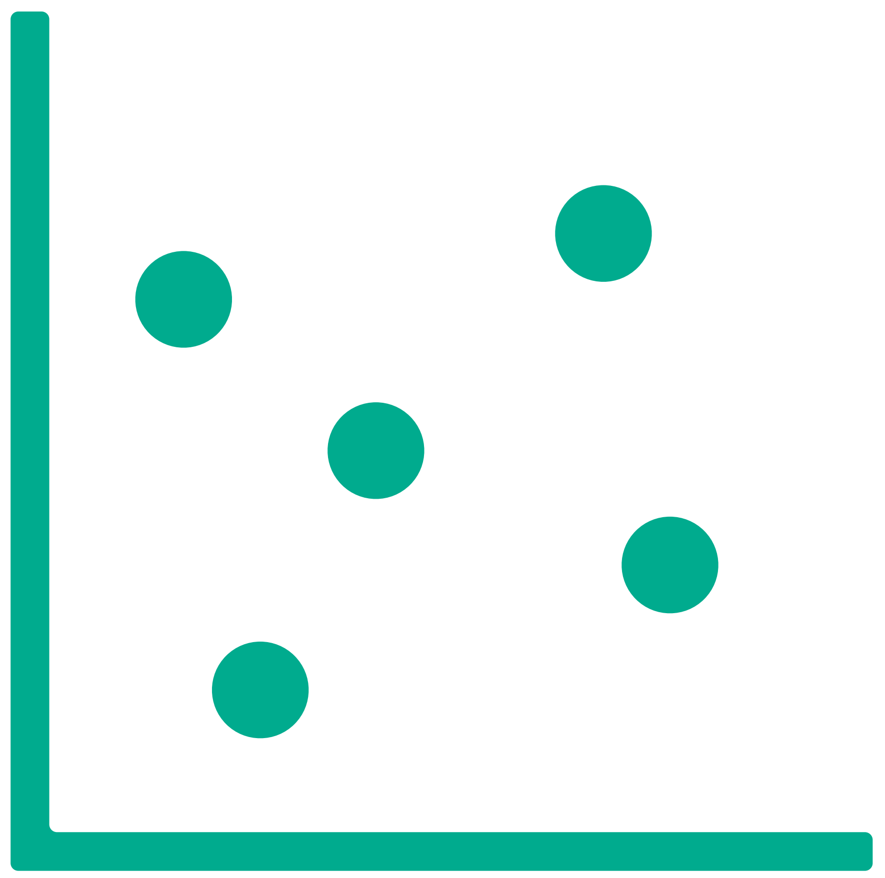
Mutual Information and Machine Unlearning

2023. 2. 21

BSM workshop at CAU, Seoul, Korea

W Cho   S Han   H D Kim

Information Theory
motivation



- How much information is present in HEP data?
- How much information do collider variables share?
  Dependence and Independence
- What is the most relevant and efficient set of input data?
- Is it possible to construct machine learning (ML) models robust to the physical variables of interest?
- Or … robust to the uncertainties introduced by unknown systematics?

Information Theory 101

Information

- An event with the probability 1 has no information
- An event with less probability has more information
- Total information from two independent events should be the sum of each information
- Shannon Information

$$I = \log \frac{1}{p(x)} = -\log p(x)$$

  satisfies all the conditions listed above

Information Theory 101

Differential entropy

- Entropy is the average of the information $H(X) = \sum_i p_i \log \frac{1}{p_i}$
- Differential entropy (or continuous entropy) $H(\Delta X) = \int_{\Delta X} dx\ p(x)\ \log \frac{1}{p(x)}$
- We write it as an expectation value with pdf p $H(X) = E_p[-\log p(x)]$
- Joint entropy $H(X, Y) = E_p[-\log p(x, y)]$

Information Theory 101

Differential entropy

- Entropy is the average of the information $$H(X) = \sum_i p_i \log \frac{1}{p_i}$$
- Differential entropy (or continuous entropy) $$H(\Delta X) = \int_{\Delta X} dx\ p(x)\ \log \frac{1}{p(x)}$$
- Mapping $\bar{p}_{x_i}\Delta \to p_i$, we get $H(\Delta X) = H(X) + \log \Delta$ which can be negative as $\Delta \to 0$
- Differential entropy is not positive definite

Information Theory 101

Entropy

- Cross entropy

$$H(f;g) = E_f[-\log g(x)] = \int dx\, f(x)\ \log \frac{1}{g(x)}$$

- Conditional entropy

$$H(Y|X) = E_p[-\log p(y|x)] = \int dx\ dy\ p(x,y)\ \log \frac{1}{p(y|x)}$$

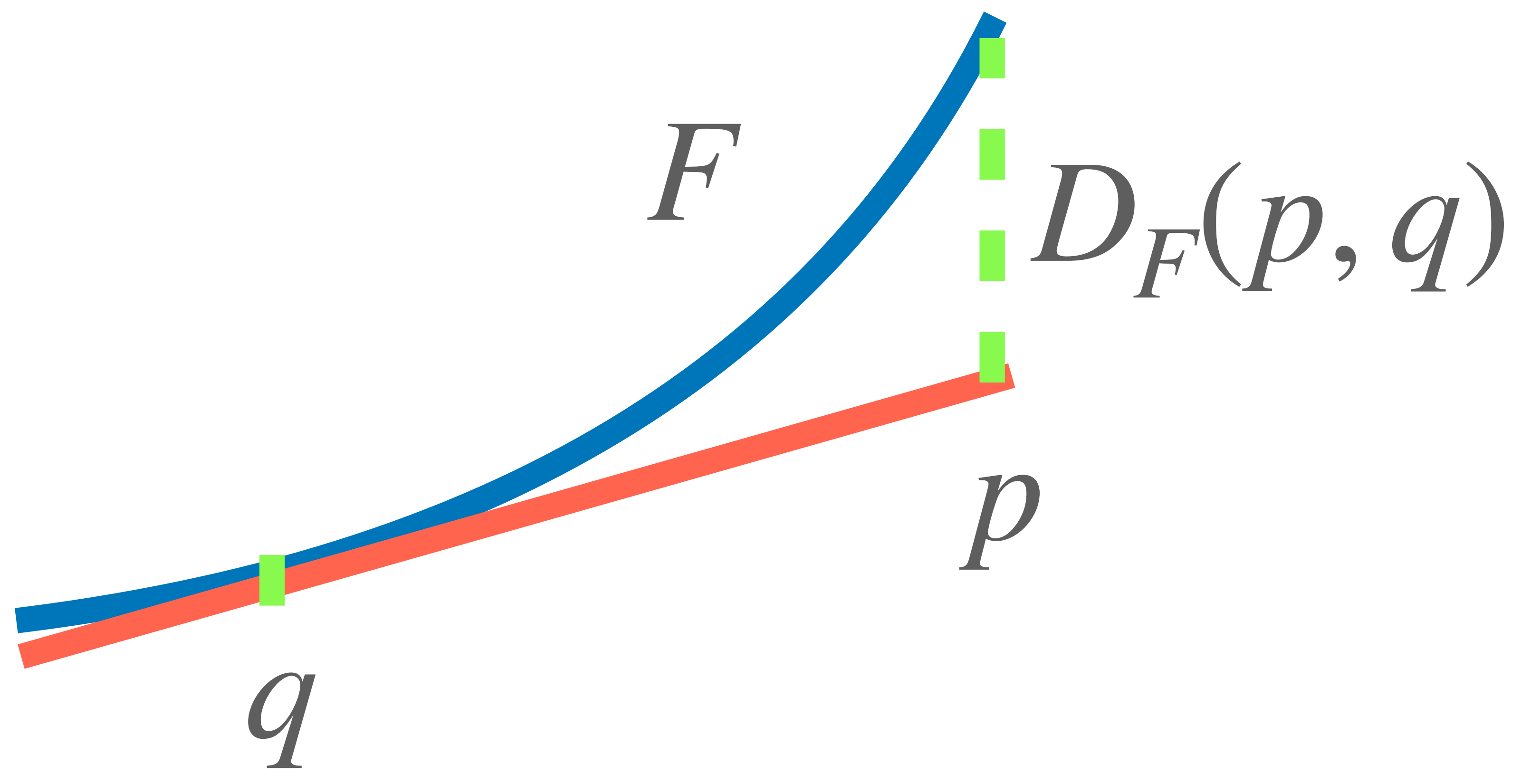
- Joint entropy

$$H(X,Y) = E_p[-\log p(x,y)] = \int dx\ dy\ p(x,y)\ \log \frac{1}{p(x,y)}$$

Information Theory 101

Bregman divergence

- Bregman divergence for a convex function F,
  $D_F(p,q) = F(p) - F(q) - < \nabla F(q), p - q >$
- $D_F(p,q) \geq 0$ for all p, q
- $D_F(p,q) = 0$ iff $p = q$
- Taking $F(p) = \int dx\ p(x)\ \log\ p(x)$, we can define Kullback-Leibler divergence



Information Theory 101

Bregman divergence to Kullback-Leibler divergence

- Bregman divergence for convex function F,
  $D_F(p, q) = F(p) - F(q) - < \nabla F(q), p - q >$
- $D_F(p, q) = \int dx \left(p \log p - q \log q - (p - q) \log q - (p - q)\right)$
- $D_{KL}(p, q) = \int dx \; p(x) \; \log \; (\frac{p(x)}{q(x)})$ if $\int dx \; p(x) = \int dx \; q(x) = 1$

Information Theory 101

Kullback-Leibler divergence

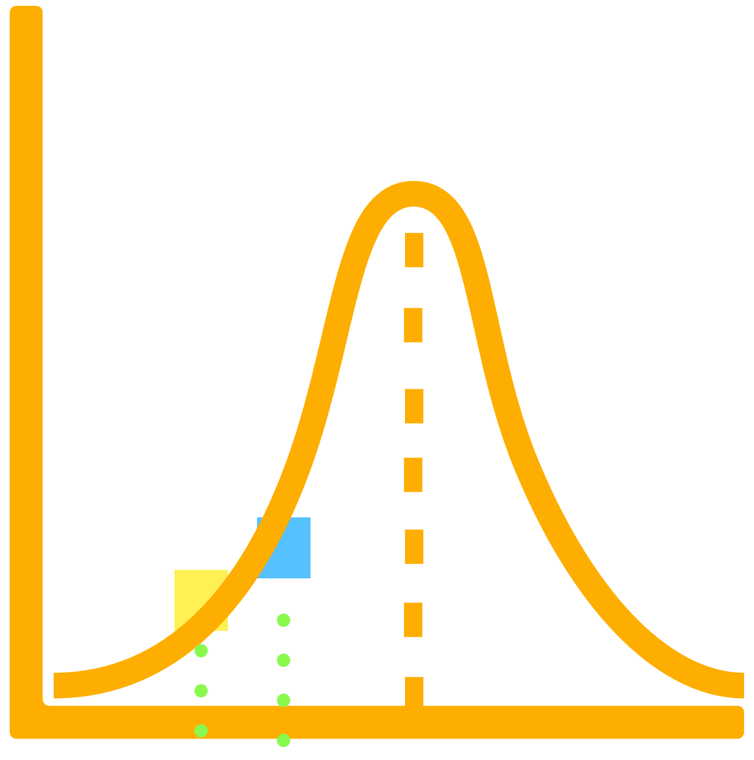
- Kullback-Leibler (KL) divergence $D_{KL}(p\,|\,q) = \int dx\ p(x)\ \log\ \frac{p(x)}{q(x)} = E_f\ [\log \frac{p(x)}{q(x)}]$
- In terms of entropy, it is a combination of the cross entropy and the entropy $D_{KL}(p\,|\,q) = H(p;q) - H(p)$
- Sorry for inconvenience due to $\rightarrow$
  - ; for relative entropy and mutual information
  - $|$ is used for the conditional pdf/entropy but is used for KL divergence
  - , is used for joint pdf/entropy but is used for Bregman divergence

Information Theory 101

Kullback-Leibler (KL) divergence

- Kullback-Leibler (KL) divergence

$$D_{KL}(p\,|\,q) = \int dx\ p(x)\ \log\ \frac{p(x)}{q(x)} = E_f\ [\log \frac{p(x)}{q(x)}] \geq 0$$



- $D_{KL}(p\,|\,q) = 0$ iff $p(x) = q(x)$ for all x

- From $q(x) = p(x)$, varying $q(x) : \Delta q(x_1)\Delta x = -\ \Delta q(x_2)\Delta x$ keeping $\int dx\ q(x) = 1$,

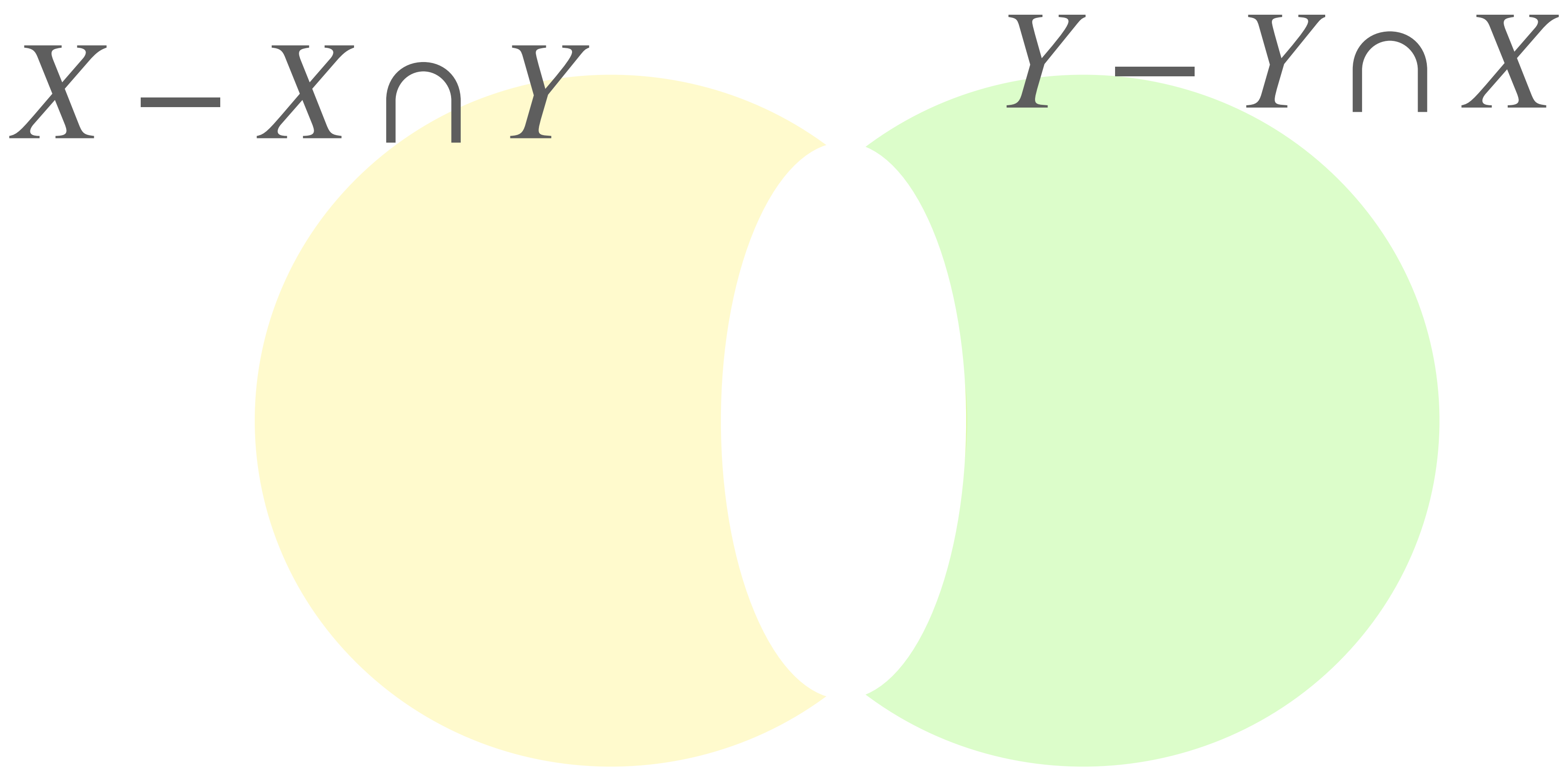
$$\delta D_{KL}(p\,|\,q) = \Delta x \left(-p_1 \log(1 + \frac{\Delta q}{p_1}) - p_2 \log(1 - \frac{\Delta q}{p_2})\right) = \sum_{i=1,2} \frac{\Delta^2}{p_i}\Delta x > 0$$

Information Theory 101

Conditional probability and entropy

- Conditional probability $p(x, y) = p(x \,|\, y)p(y) = p(y \,|\, x)p(x)$
- Conditional entropy : $Y - Y \cap X$

  $$H(Y \,|\, X) = \int dx \; dy \; p(x, y) \; \log \frac{p(x)}{p(x, y)} = \; H(X, Y) - H(X)$$

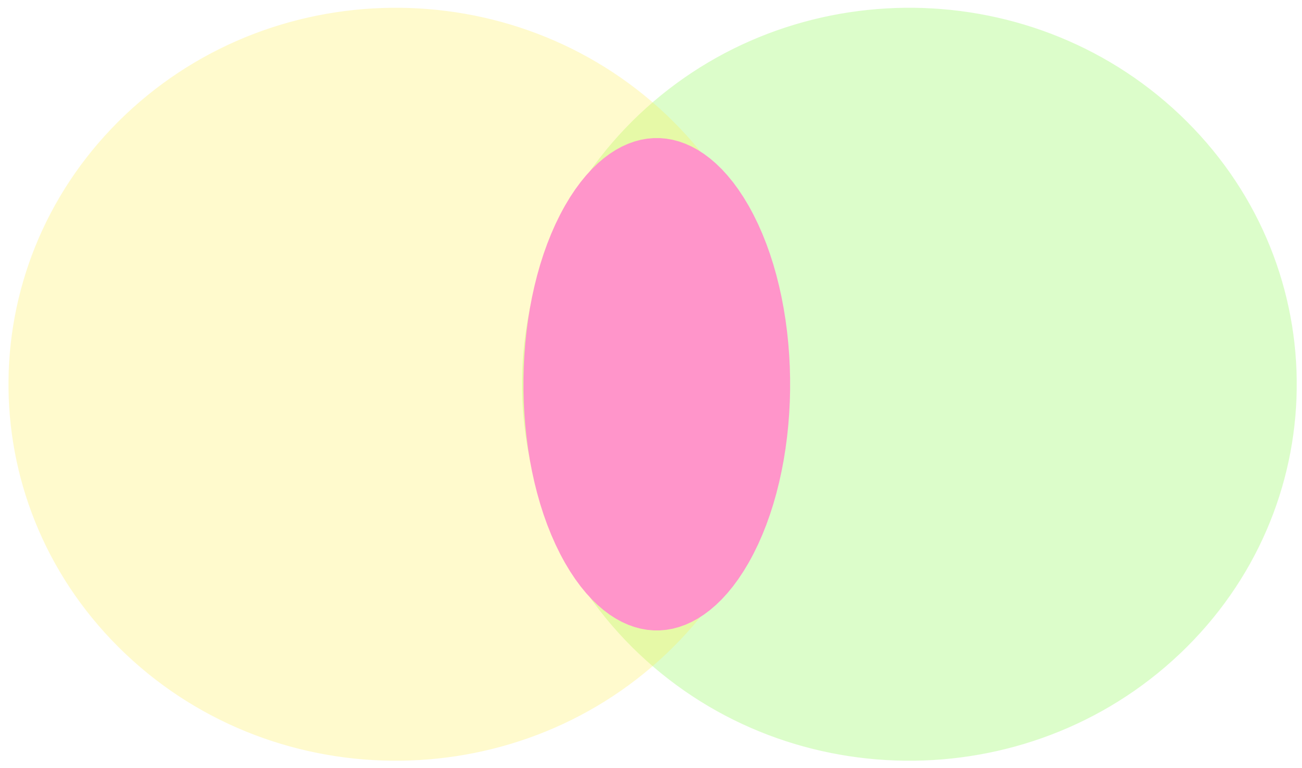


- Similarly, $H(X \,|\, Y) = H(X, Y) - H(Y)$
- $H(X, Y) = H(X \,|\, Y) + H(Y) = H(Y \,|\, X) + H(X)$
- $2H(X, Y) = H(X \,|\, Y) + H(Y \,|\, X) + H(X) + H(Y)$

Information Theory 101

Mutual information

- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$
- Mutual information : entropy of $X \cap Y$
  $$I(X;Y) = \int dxdyp(x,y)\ \log \frac{p(x,y)}{p(x)p(y)}$$
- Mutual information from KL divergence
  $$I(X;Y) = D_{KL}(p(x,y)|p(x)p(y))$$
- MI = diff. ent of X + diff. ent. of Y - diff. ent. of $X \cup Y$



$X \cap Y$

Information Theory 101

Mutual information

- Mutual information vs Pearson's correlation coefficient
- $I(X;Y) = H(X) + H(Y) - H(X,Y) = \int dxdy\ p(x,y)\ \log \frac{p(x,y)}{p(x)p(y)}$
- $I(\Delta X;\Delta Y) = -\frac{1}{2}\log(1 - r_\Delta^2)$ for the correlation r_Δ (bivariate normal pdf)
- $I(X;Y) \geq 0$ implies that $I(X;Y) = 0$ is achieved only when X and Y are independent, i.e., $p(x,y) = p(x)p(y)$ for all x and y.

Information Theory 101

MI vs Pearson's correlation coefficient

- Covariance can be computed as a sum
- Zero covariance does not guarantee the independence
- If they have positive correlation in some parts and negative correlation in other parts, the total correlation can be zero (or small)
- On the other hand $I(X;Y) = 0$ guarantees the independence

Mutual Information for Machine Unlearning

Independence of the variables

- When the integrand is positive definite, $f(x) \geq 0$, the vanishing integral provides a strong condition to the integrand,
  $$F = \int dx\, f(x) = 0 \qquad \rightarrow \qquad f(x) = 0 \quad \text{for all } x$$
- Similarly, for MI as it is defined using KL divergence
  $$I = D_{KL}(p(x, y)\,|\,p(x)p(y)) = 0 \qquad \rightarrow \qquad p(x, y) = p(x)p(y) \quad \text{for all } x, y$$
- The variables are independent if $I = 0$

Pointwise Mutual Information (PMI)
(*Language model : PMI^k)

- Pointwise Mutual Information x, y $$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$
- Positive PMI $$PPMI(x, y) = \max\left(\log \frac{p(x, y)}{p(x)p(y)},\ 0 \right)$$
- Normalized PMI $$NPMI(x, y) = \frac{\log \frac{p(x, y)}{p(x)p(y)}}{\log \frac{1}{p(x, y)}}$$
- $-1 \leq NPMI \leq 1$ (1:correlated, 0:independent, -1:exclusive)

Metric
between pairs of points

- Metric $d(X,Y) = H(X,Y) - I(X;Y)$ satisfies triangle inequality, non-negativity, indiscernability $0 \le d(X,Y) \le H(X,Y)$
- Normalized metric $D(X,Y) = \dfrac{d(X,Y)}{H(X,Y)} = 1 - \dfrac{I(X;Y)}{H(X,Y)}, \quad 0 \le D(X,Y) \le 1$

Information Quality Ratio (IQR)

Redundancy or uncertainty (ref: correlation from variance and covariance $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$)

- Redundancy $$R = \frac{I(X;Y)}{H(X) + H(Y)}$$
- Symmetric uncertainty $$U = 2R = \frac{2I(X;Y)}{H(X) + H(Y)}$$
- Information Quality Ratio $$IQR(X,Y) = \frac{I(X;Y)}{H(X,Y)}$$
- Normalized mutual information $$NMI = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

How to estimate MI?

Neural Estimators based on variational representation of D_{KL}

- MI is hard to compute unless the exact pdf is known
- In most cases, the underlying pdf is not known a priori
- Thus MI is hard to estimate, with finite data samples, in a non-parametric way without any assumptions on the pdf
- MINE: Mutual Information Neutral Estimation arXiv:1801.04062
- Finding tractable representation of MI to obtain a relevant gradient flow from a MI loss function to train down to the input connected models would be important for deep learning models

Mutual Information Neural Estimation (MINE)

Donsker-Varadhan Representation of KL

- Donsker-Varadhan (DV) representation $D_{KL}(p\,|\,q) = \sup_{T:\Omega\to\mathbb{R}} E_p[T] - \log(E_q[e^T])$
- $D_{KL}(p\,|\,q) \geq E_p[T] - \log(E_q[e^T])$
- (proof) $Z = E_q[e^T] = \int dx q(x) e^{T(x)}$ and $g(x) = \frac{1}{Z} e^{T(x)} q(x)$,
  $E_p[T] - \log E_q[e^T] = E_p(T - \log E_q[E^T]) = E_p(\log \frac{e^T q(x)}{E_q[e^T]} \frac{1}{q(x)}) = E_p(\log \frac{g(x)}{q(x)})$
  therefore, $E_p(\log \frac{p(x)}{q(x)} - \log \frac{g(x)}{q(x)}) = E_p(\log \frac{p(x)}{g(x)}) = D_{KL}(p\,|\,g) \geq 0$
- Equality holds for $g(x) = p(x)$

Mutual Information using Neural Estimation

f-Divergence Representation of KL

- f-divergence representation $D_{KL}(p\,|\,q) \geq \sup_{T:\Omega\to\mathbb{R}} E_p[T] - (E_q[e^{T-1}])$

- $E_q[e^{T-1}] \geq \log(E_q[e^T])$ from $\dfrac{x}{e} \geq \log x$

- The bound is weaker but is easy to compute and can be useful practically

Mutual Information with DV

DV Representation

- $I(X; Y) \geq I_\Theta(X; Y) = \sup_{\theta \in \Theta} \left(E_{p(x,y)}[T_\theta] - \log(E_{p(x)p(y)}[e^{T_\theta}]) \right)$ where $T_\Theta : X \times Y \to \mathbb{R}$ is a neural network
- Using n samples of X, we can estimate MI which converges well
- The estimated gradient of θ in the network is $\hat{G}_B = E_B[\nabla_\theta T_\theta] - \dfrac{E_B[\nabla_\theta T_\theta e^{T_\theta}]}{E_B[e^{T_\theta}]}$ where B is a batch of data
- The first (second) term from the joint (marginal) distribution

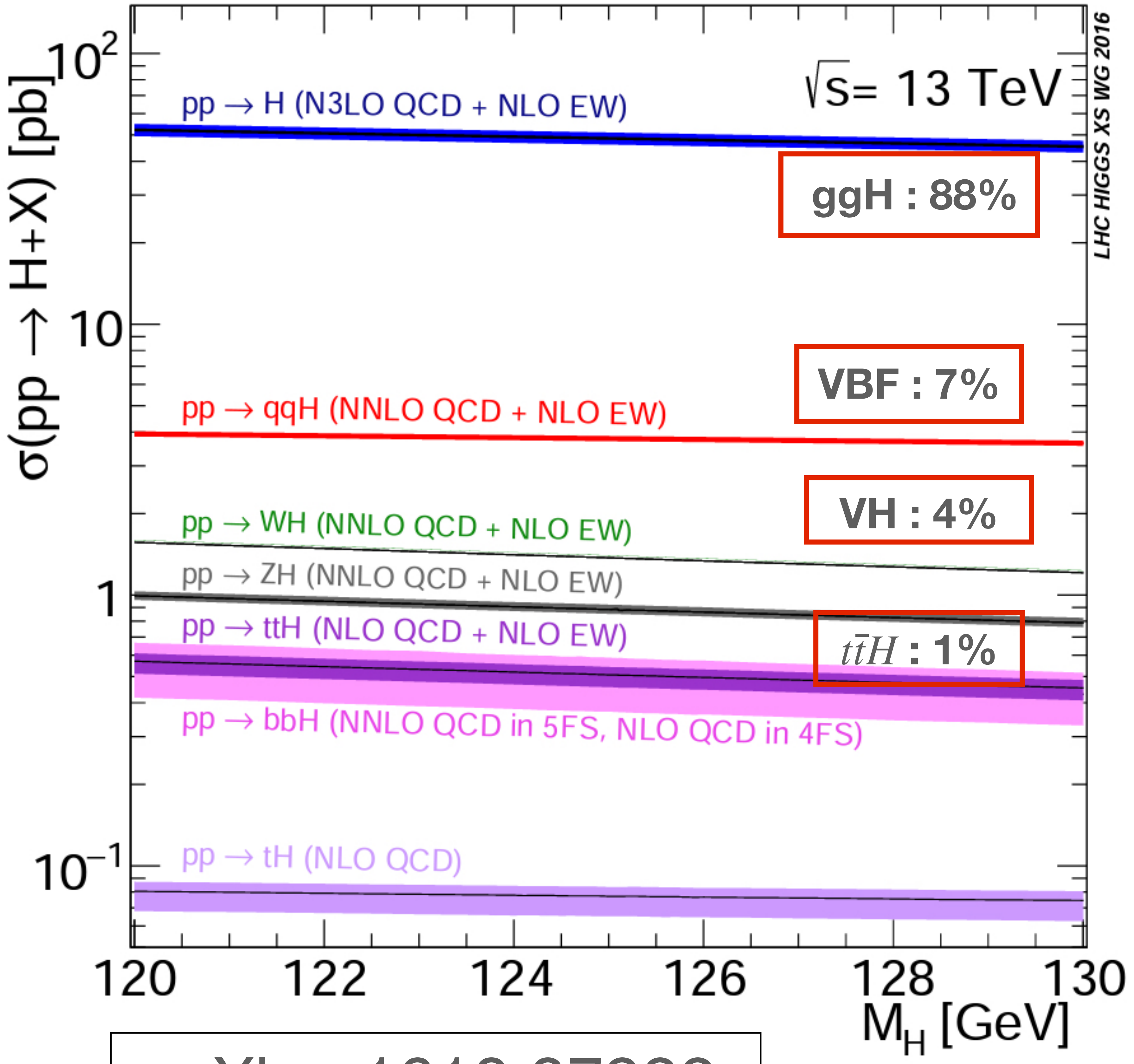
Application : Higgs to dimuon

Machine not to learn invariant mass of dimuon

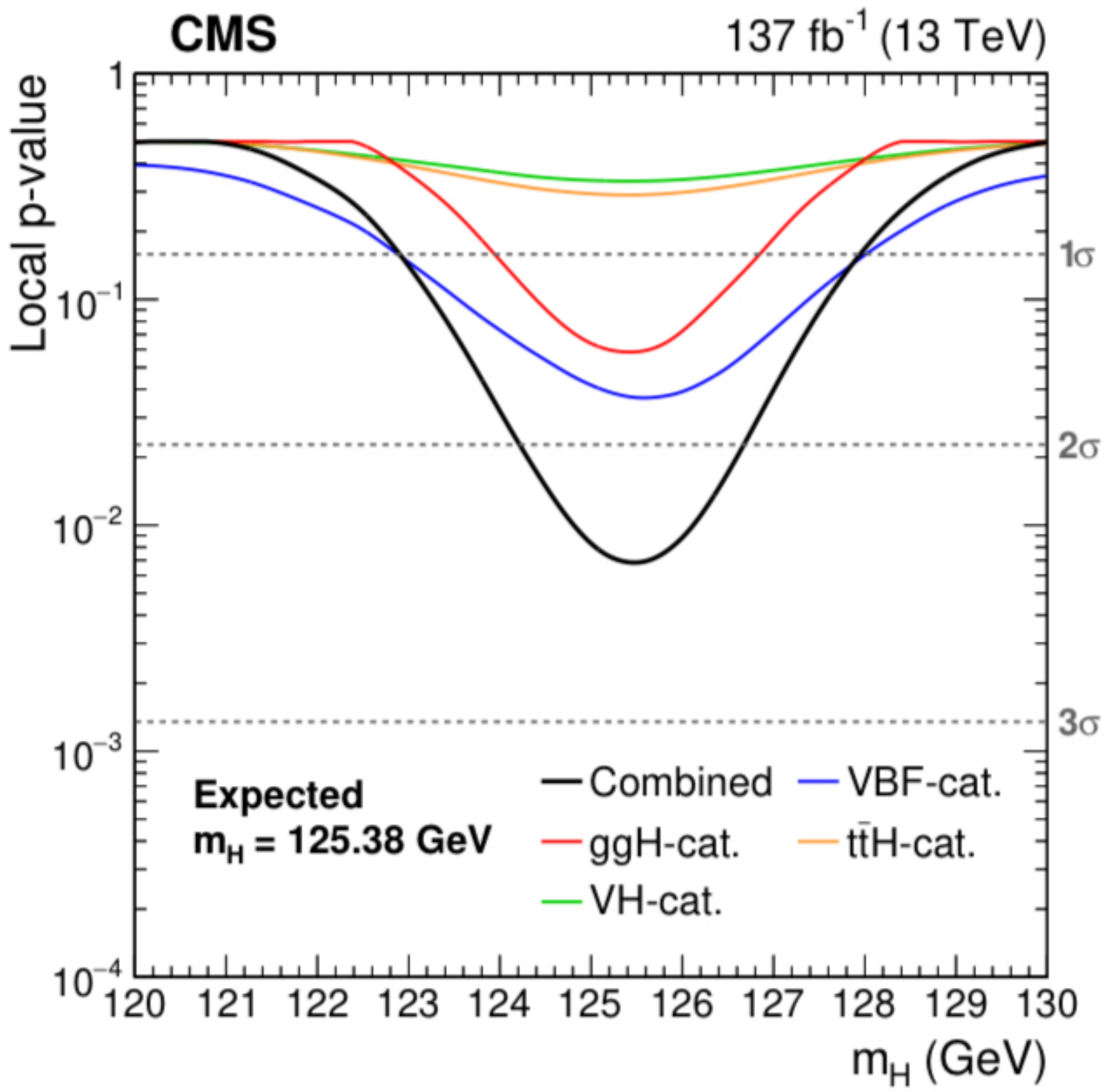
- VBF gives the best sensitivity while ggF has the larger cross section
- Mainly due to the background form DY for ggF
- VBF has a relatively clean background
- Initial state radiation (ISR) can be used to enhance signal to background ratio for ggF channel Higgs since ggF ISR is gluon rich while DY ISR is quark rich
- Process dependent discrimination would work using quark/gluon jet discrimination from deep learning
- However, invariant mass distribution is distorted by categorization

Motivation

from SB Han, talk at KPS 2022



arXiv : 1610.07922



JHEP 01(2021) 148

- CMS : 3.0σ excess (Expected : 2.5σ; VBF ~ 1.8σ / ggH ~ 1.6σ)

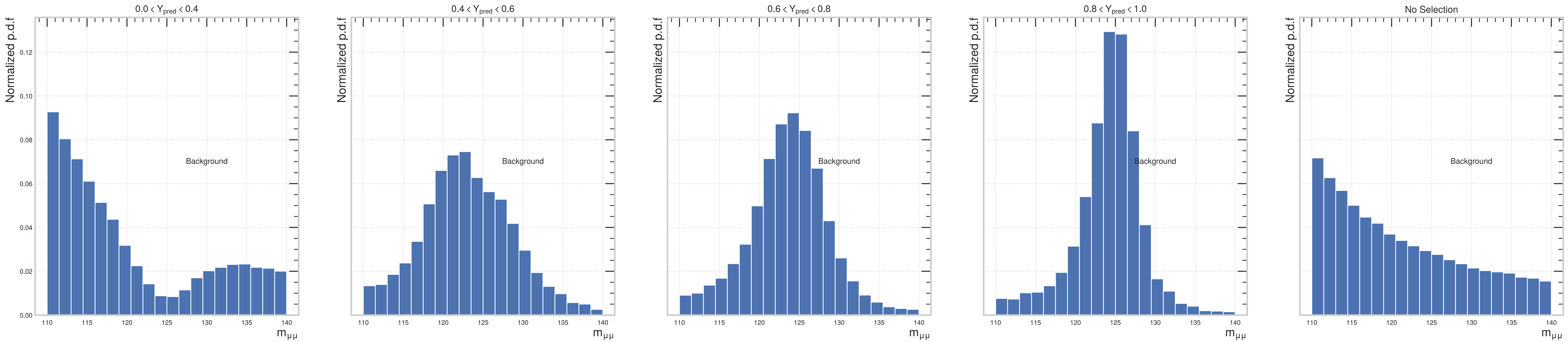
Before

Invariant mass distribution

signal

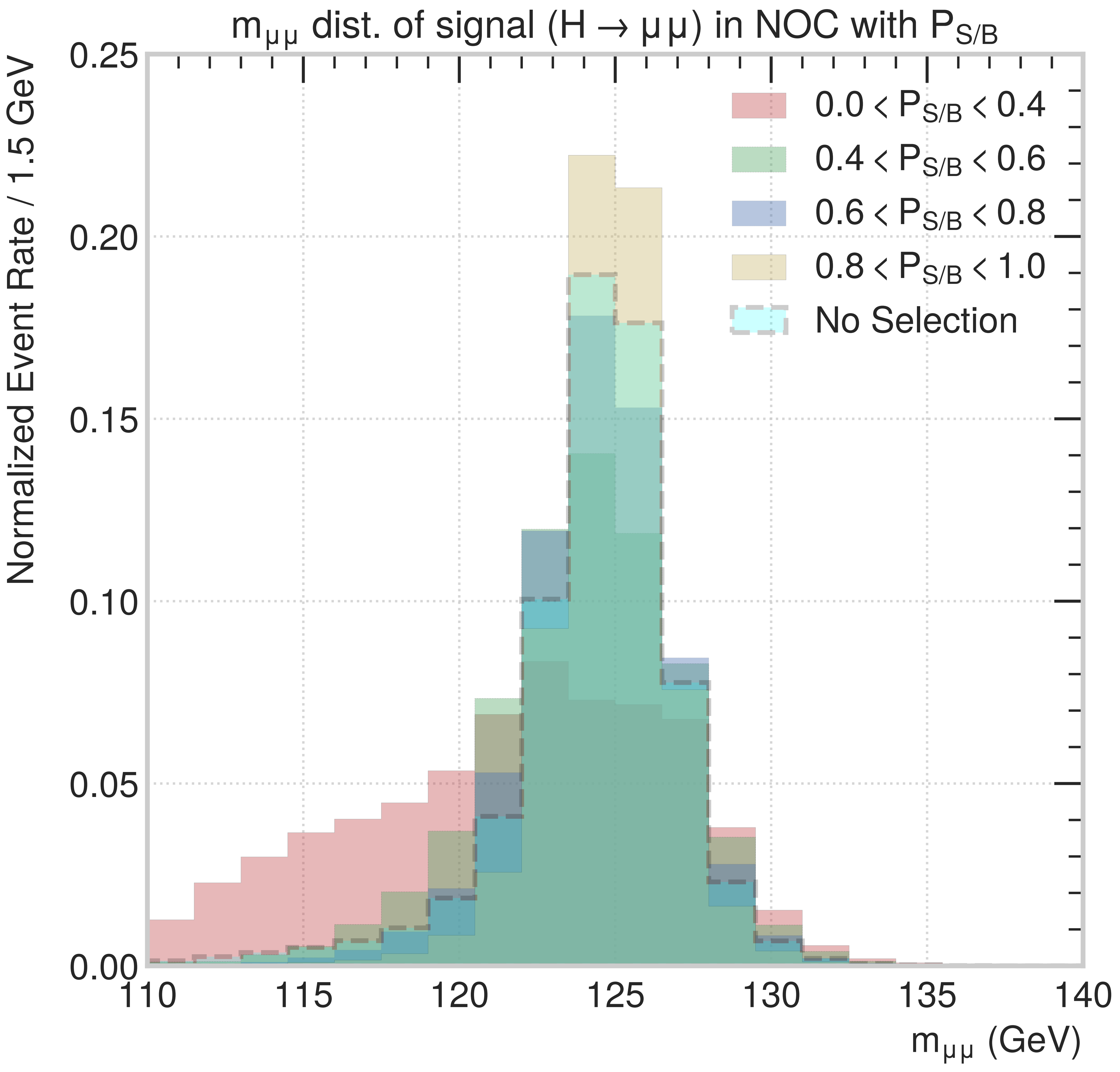
Invariant mass distribution

background

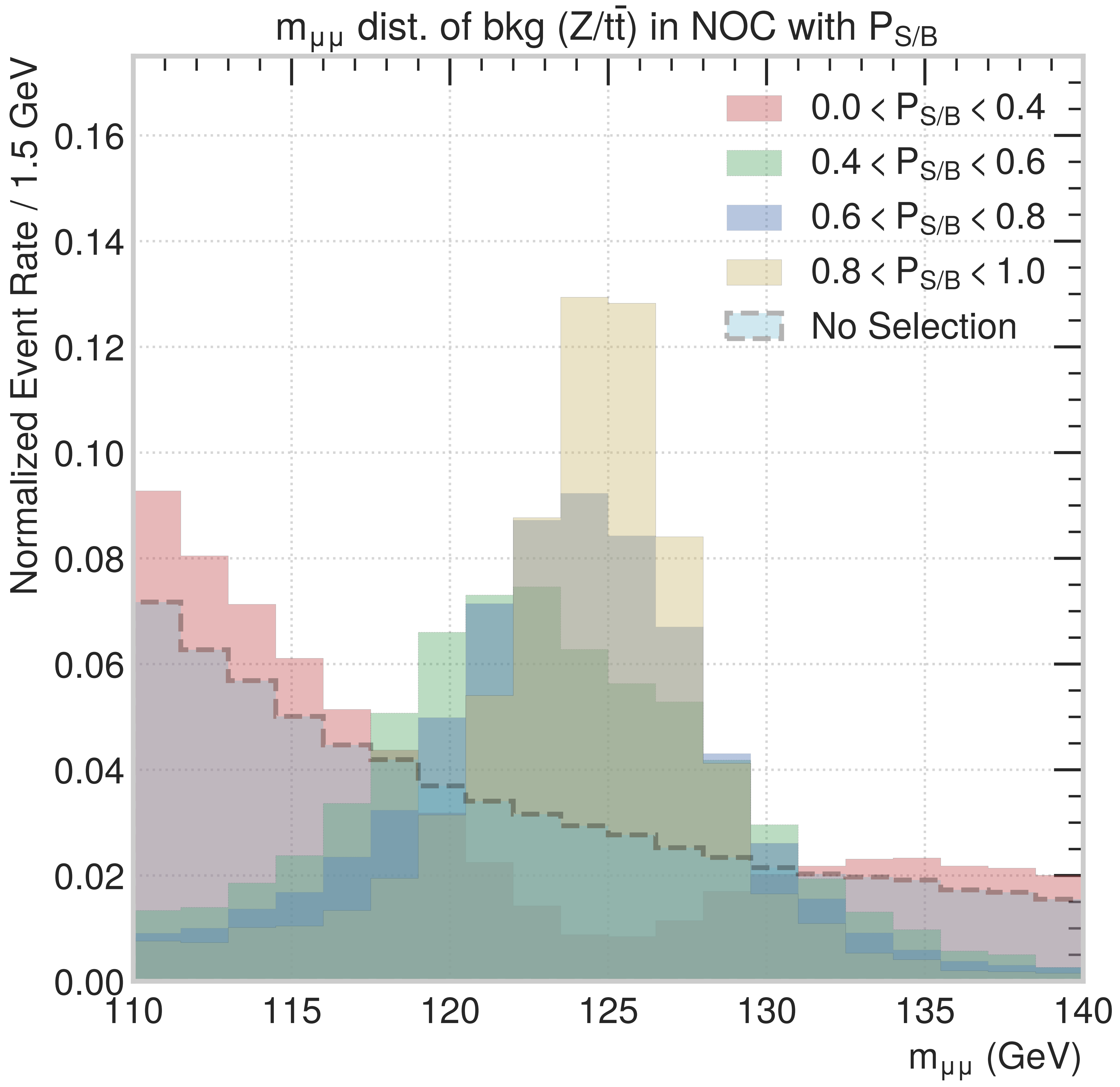


Invariant mass distribution

signal

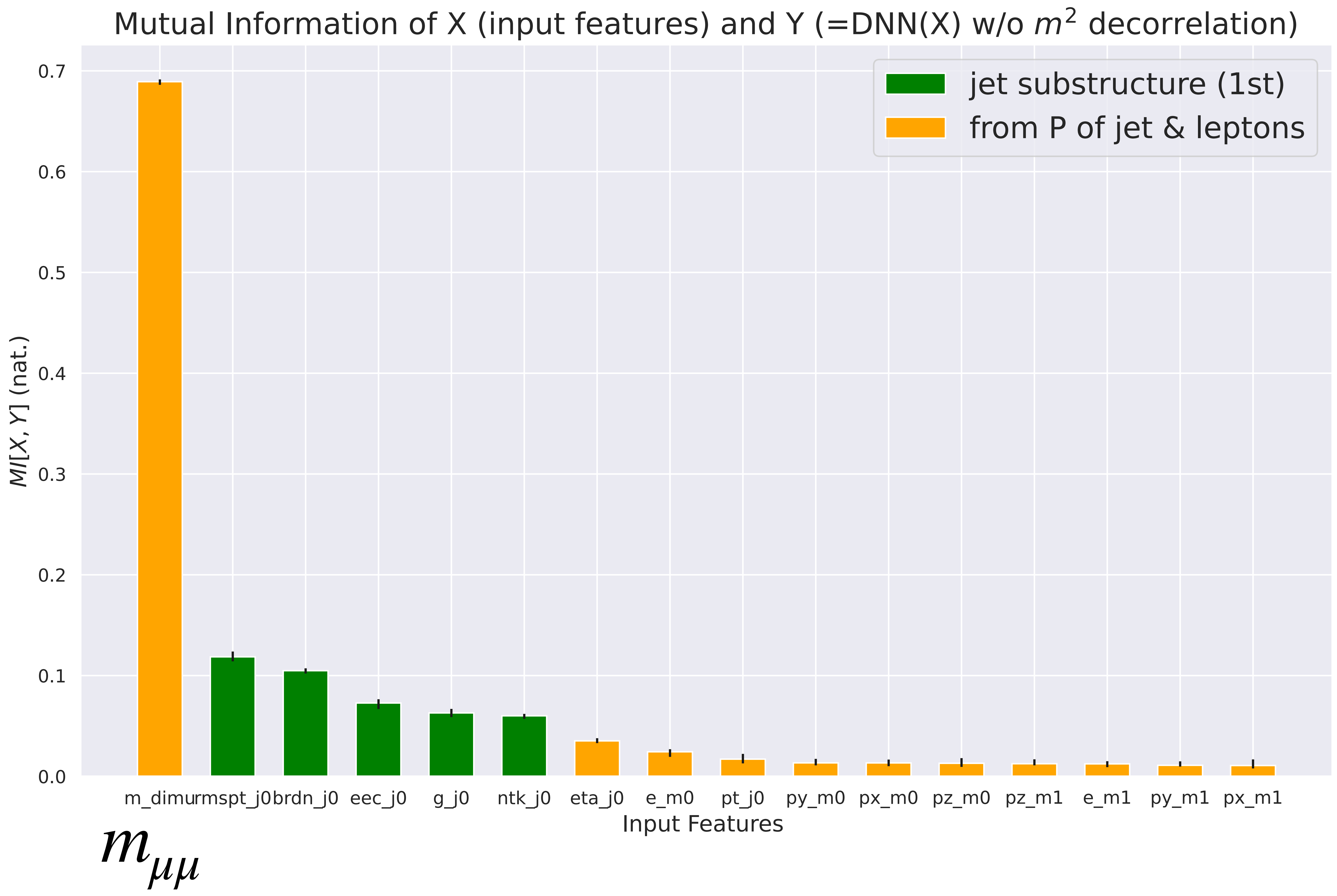


background



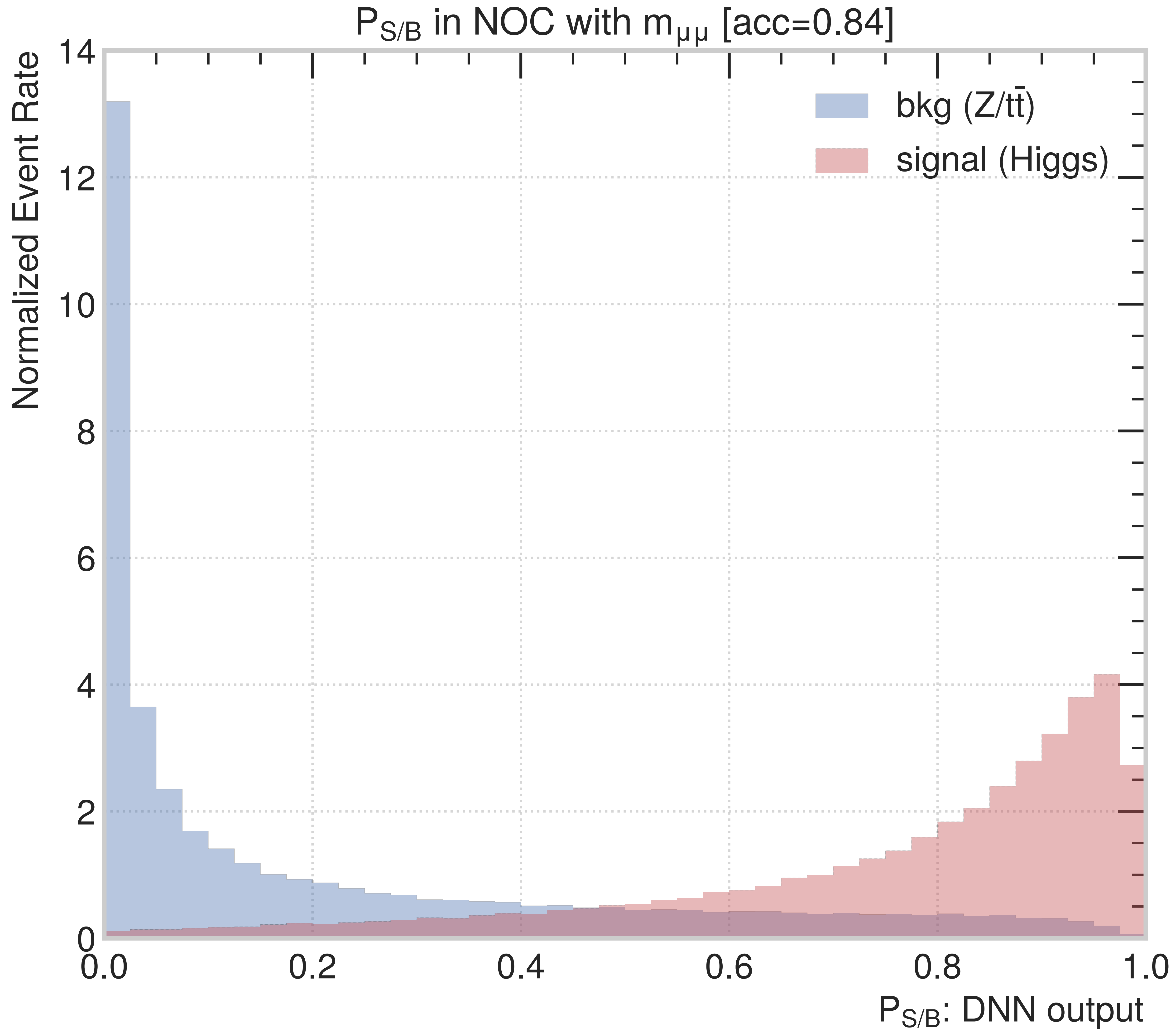
Mutual Information

Figure out the most important observables in deep learning



Scoreboard

signal vs background



After

Invariant mass distribution

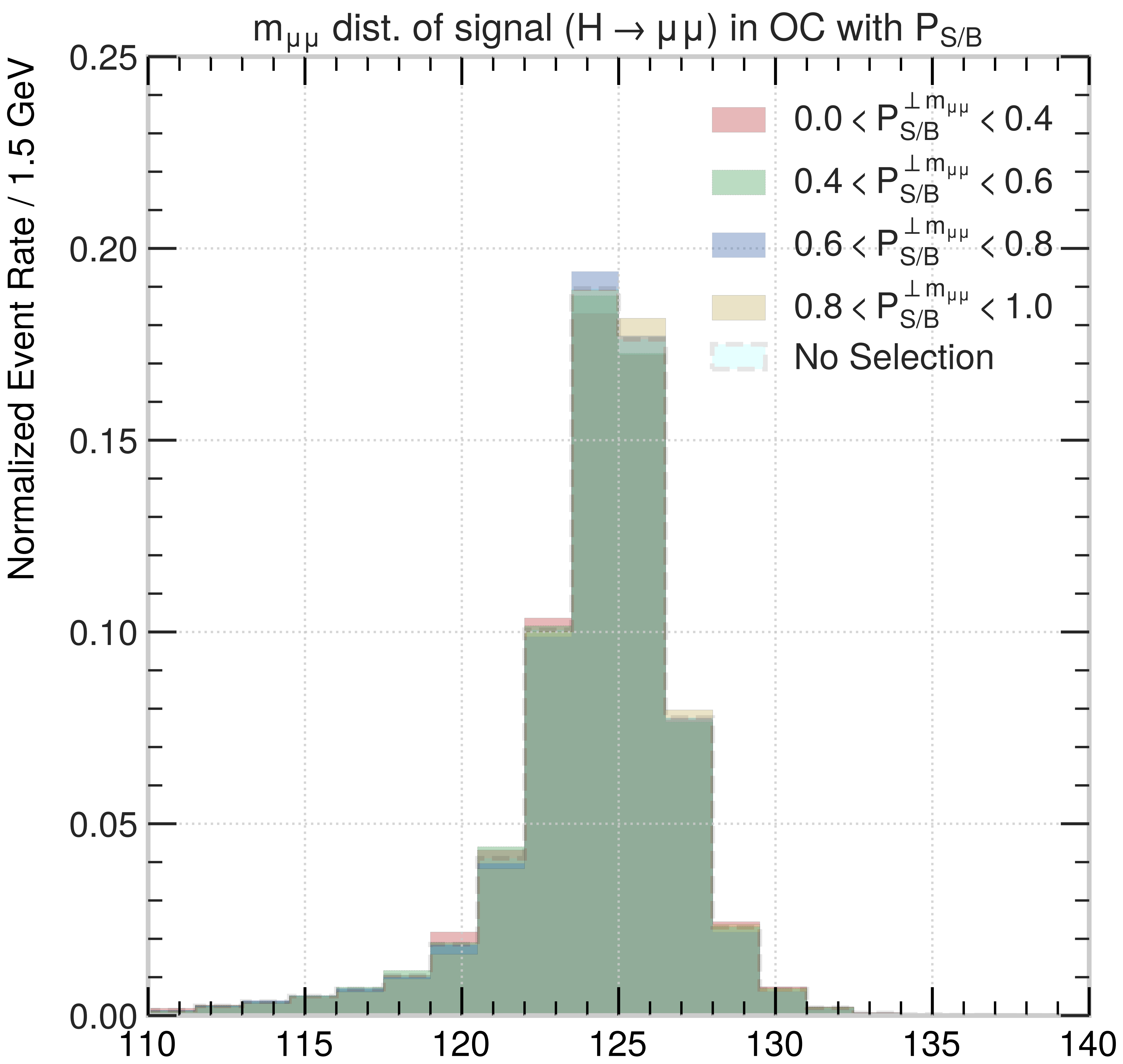
signal

Invariant mass distribution

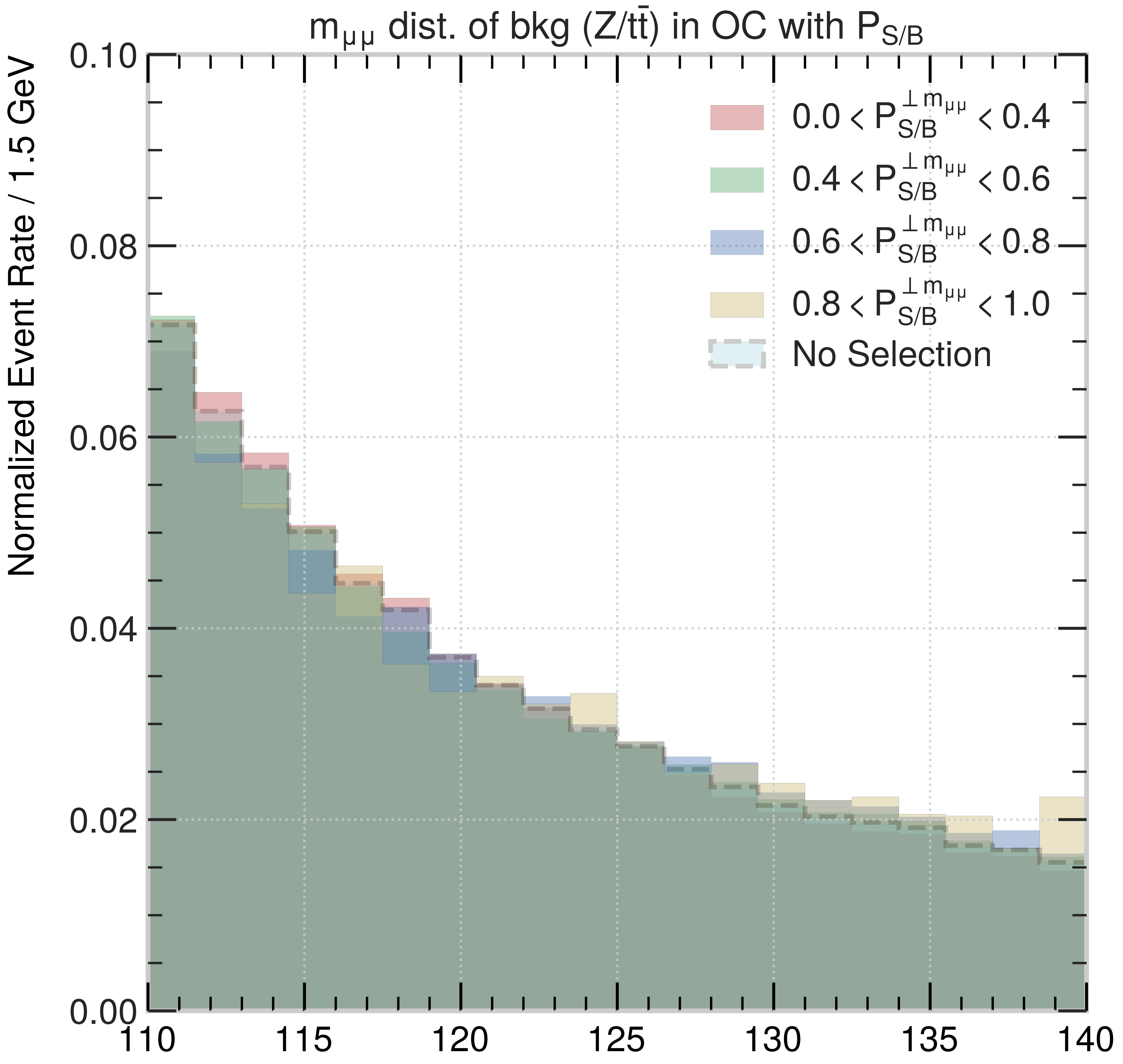
background

Invariant mass distribution

signal

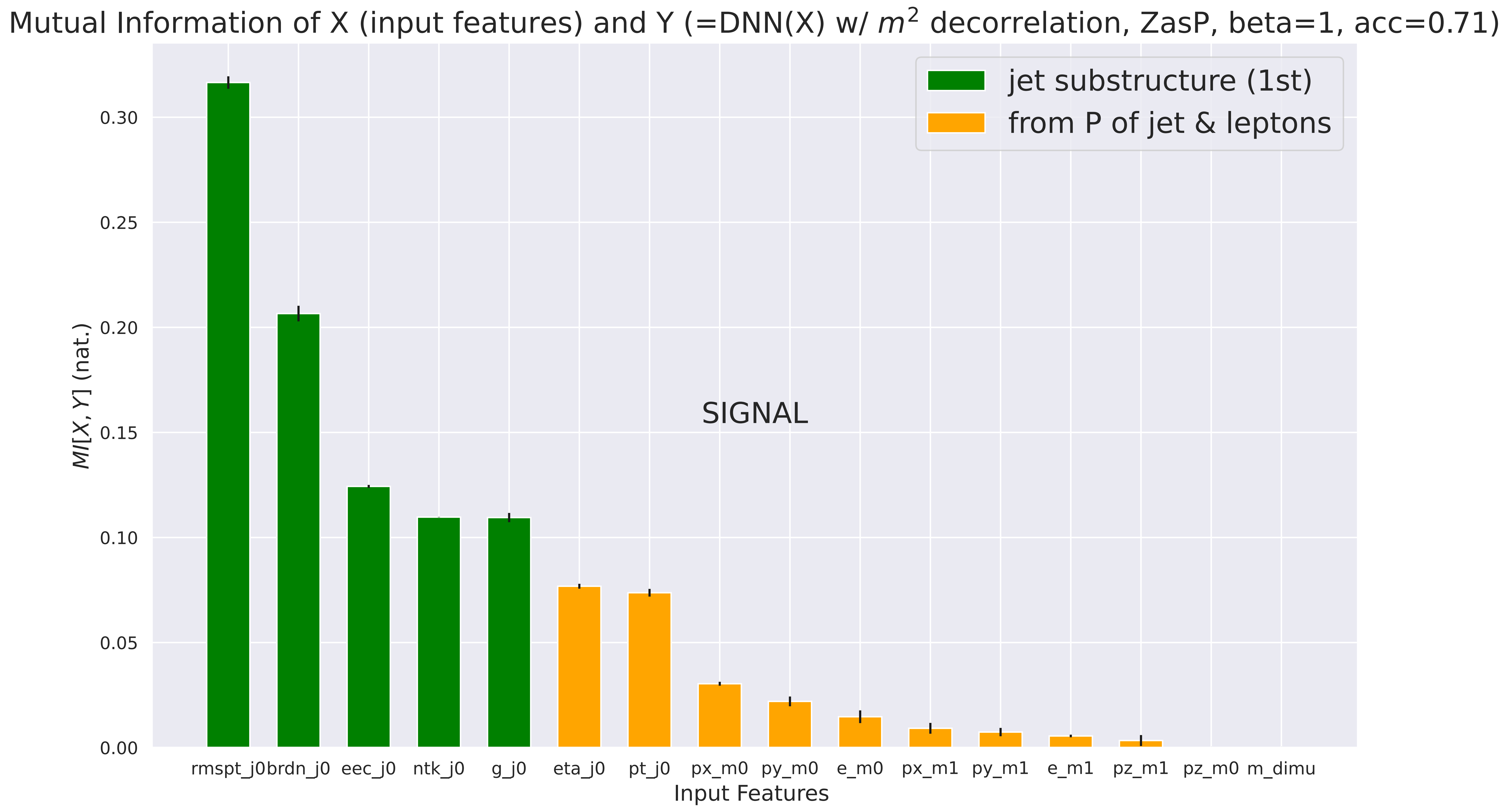


background



Mutual Information

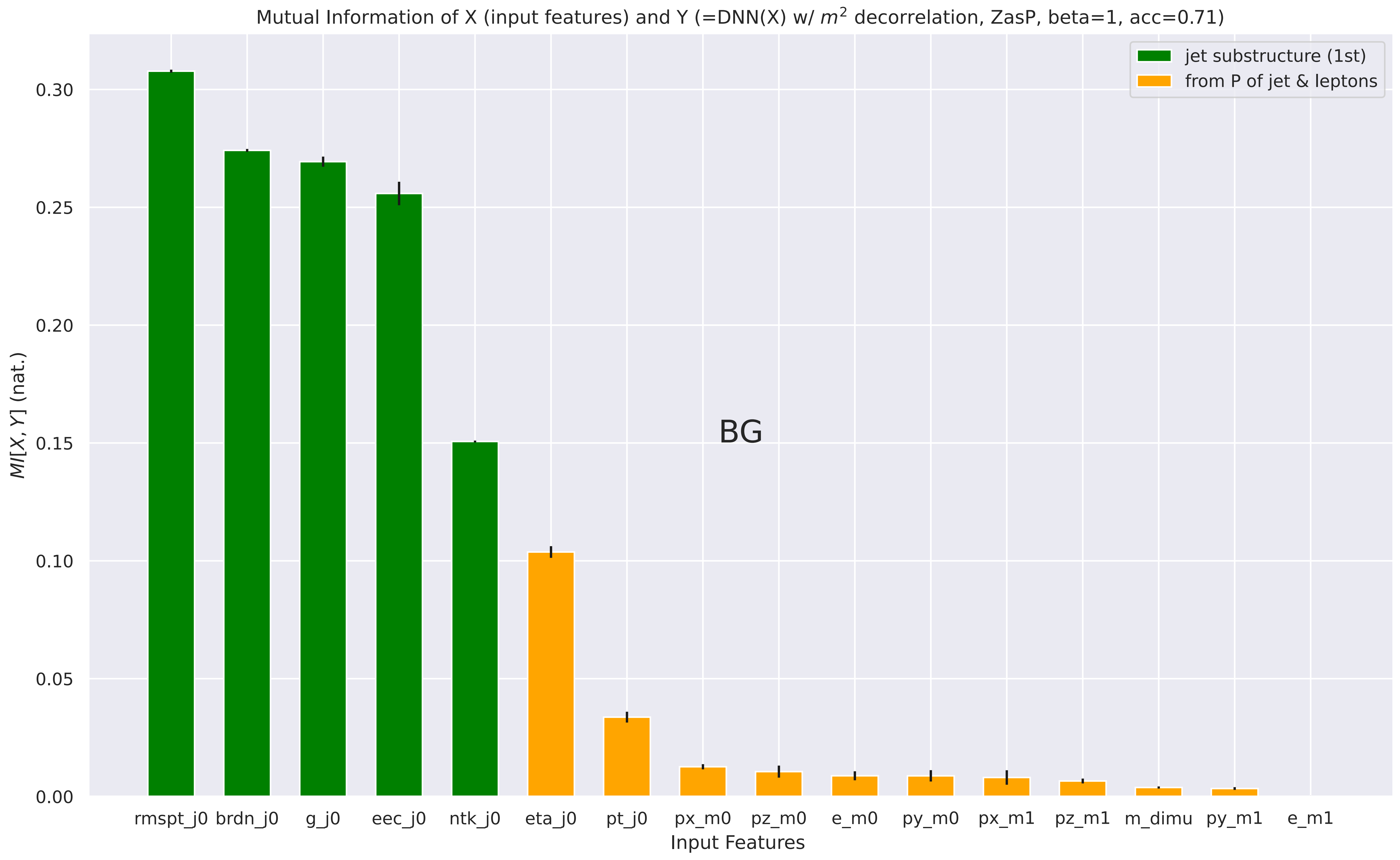
signal



$m_{\mu\mu}$

Mutual Information

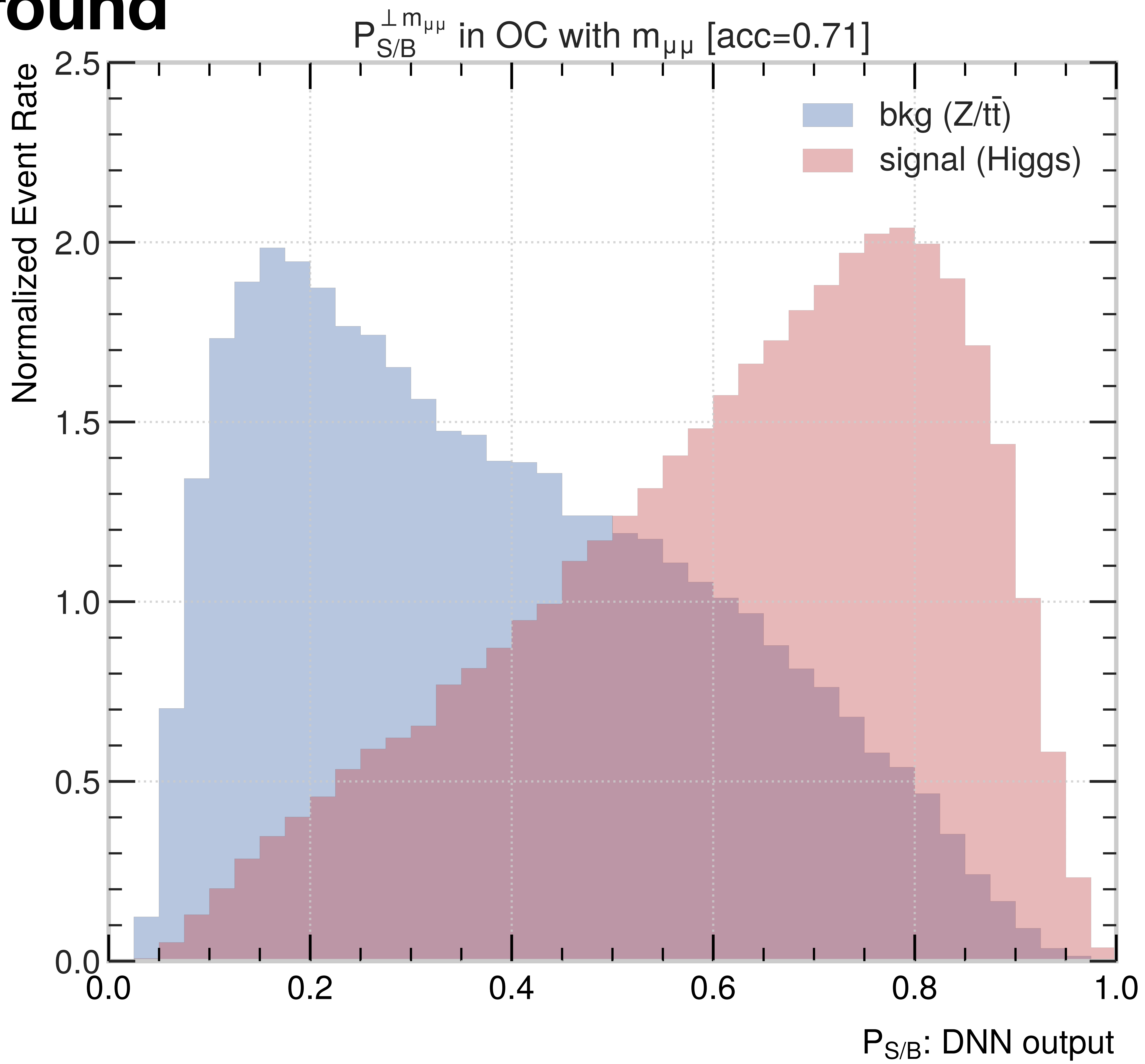
background



$m_{\mu\mu}$

Scoreboard

signal vs background



Summary

Mutual Information as a tool for machine unlearning

- Deep learning is very helpful in many examples including jet substructure studies for signal and background discrimination
- Often it distorts the very nice invariant mass distribution of the signal
- Precision measurement is possible if the nice features are preserved
- Machine can unlearn certain input (dimuon invariant mass in the example) by minimizing MI of the output with certain input
- MI=0 guarantees the independence of two variables

Thank you!