Università degli Studi di Napoli Federico II

INFN Istituto Nazionale di Fisica Nucleare

ATLAS Experiment

# Application of a Machine Learning algorithm for the background estimation in the search for New Physics

*Riunione Gruppo 1 - 21/12/2021*

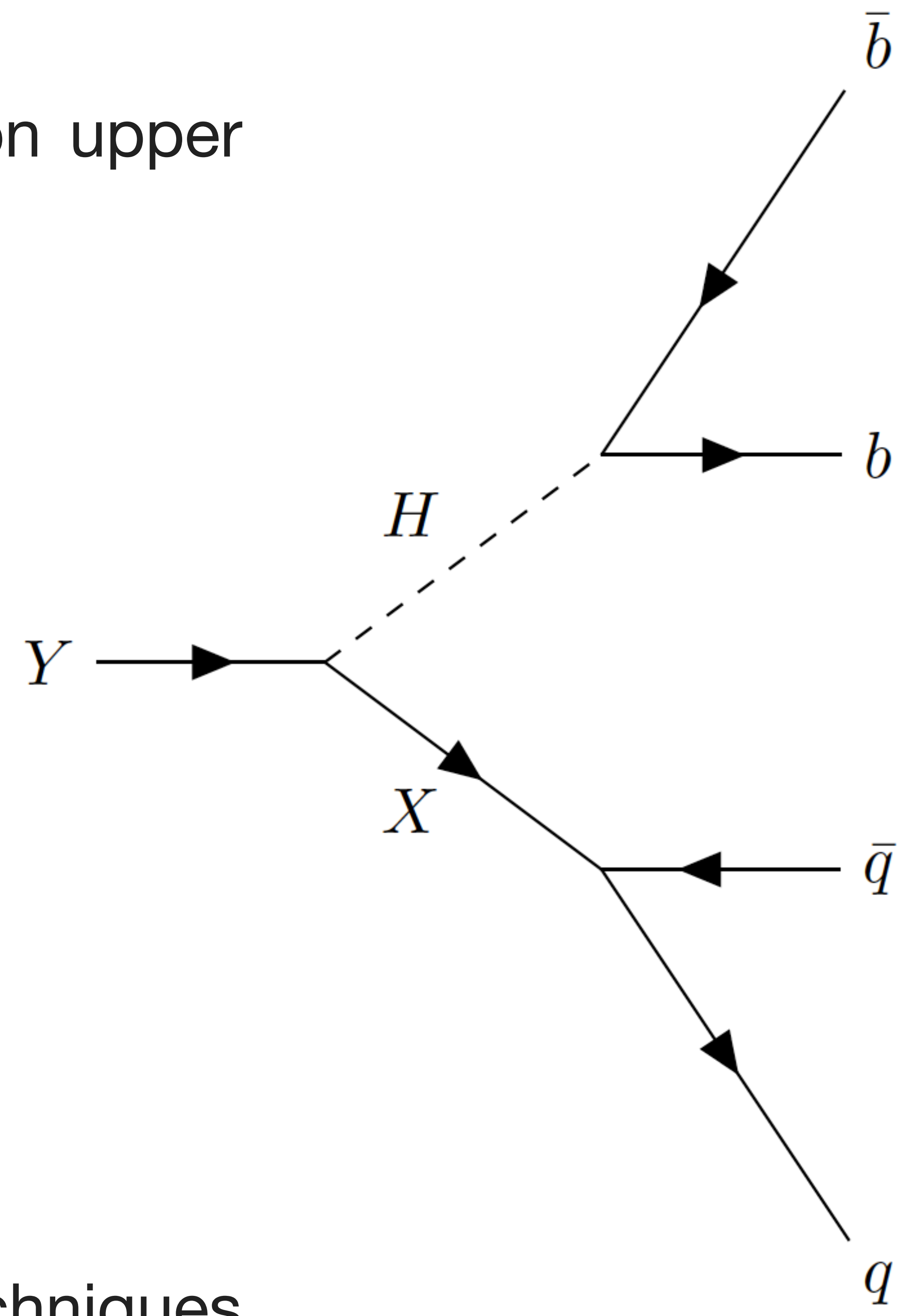
*Silvia Auricchio - PhD student in Physics - 35th cycle*

# Introduction

- Main aspects of **a general search for New Physics** in the fully hadronic final state
- **Problem faced:** the need of a background estimation from data
- **Method used:** a Machine Learning algorithm trained on data
- **Obtained results**

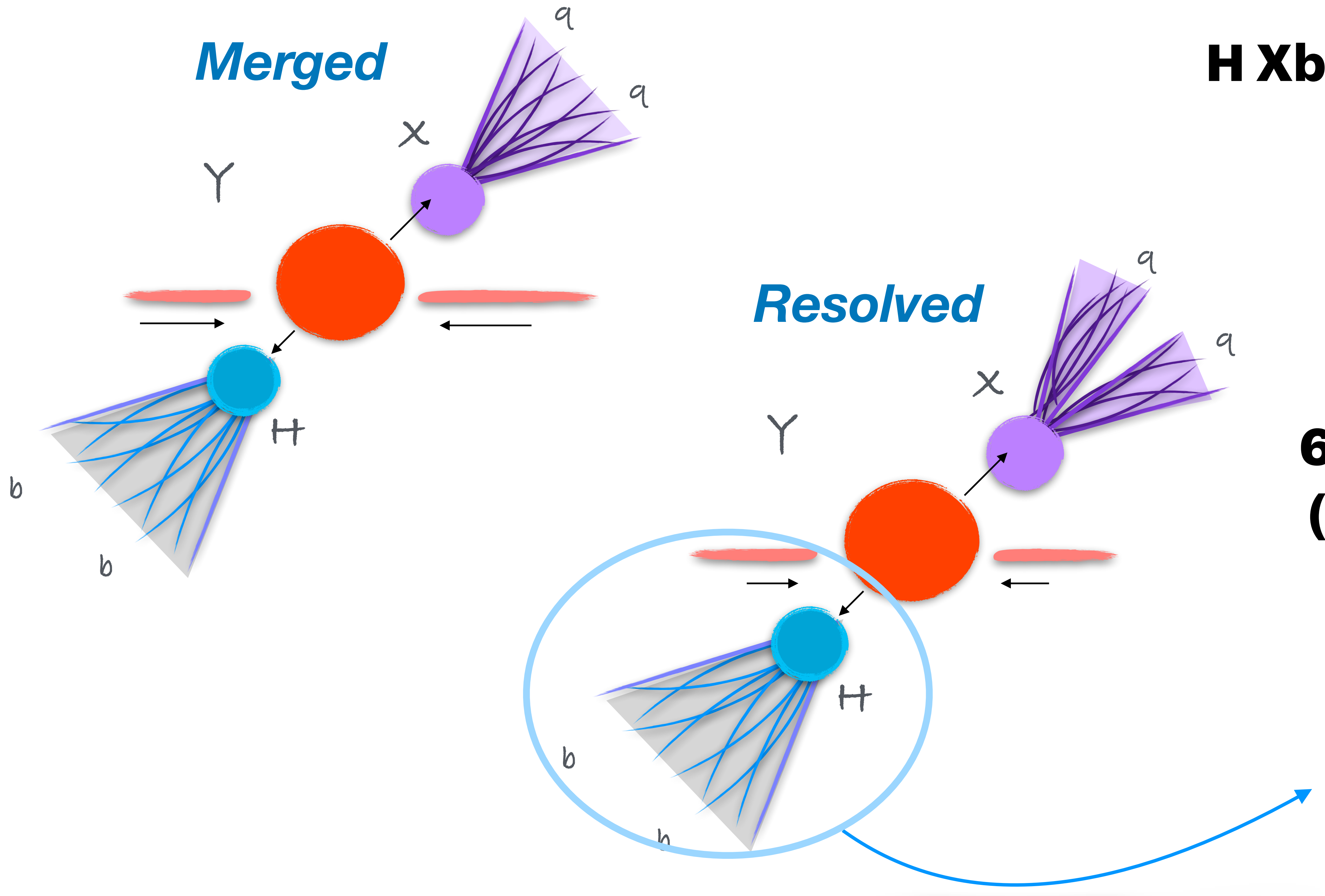
# Y->XH->qqbb analysis

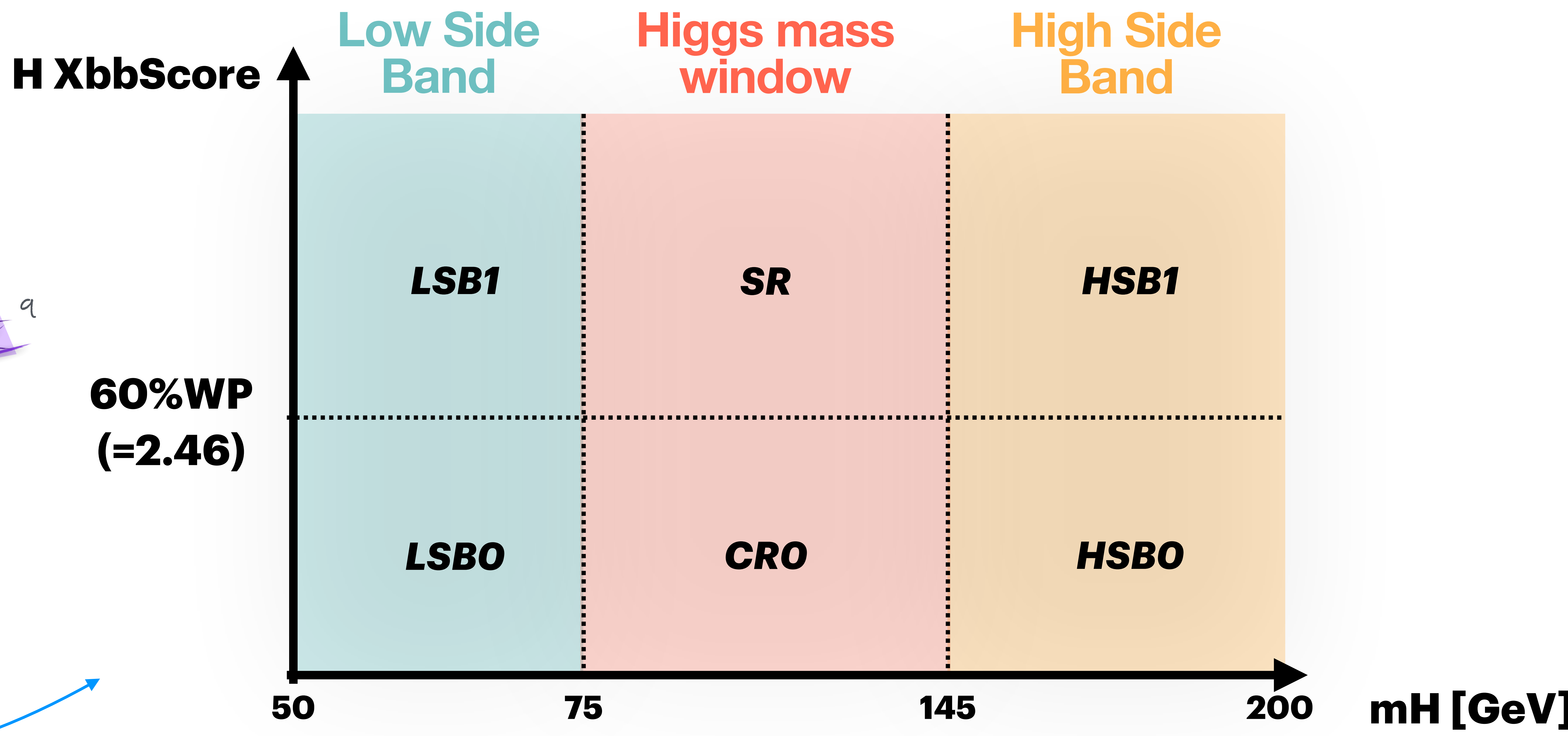
- Search for a heavy resonance (Y) decaying into a SM Higgs (H) and a new particle (X) in a fully hadronic final state
- Model independent search, Heavy Vector Triplet model as a benchmark for cross section upper limits
- High Y mass (>1 TeV), Higgs selected from $b\bar{b}$ decay
- X and H reconstructed as:
  - 2 large-R jets (merged regime)
  - 1 large-R jet (H) and two small-R jets (X) (resolved regime) **New!**
- Background composition: ~97% QCD di-jet processes, ~3% $t\bar{t}$ and V+jets processes.
- Previous analysis performed on 2015-2016 data $(36.1\ fb^{-1})$
  - Current analysis exploits the full Run 2 datasets $(139\ fb^{-1})$ and adopts new techniques (XbbTagger, DNN-based background estimation, Anomaly Score for model-independent search) **New!**
  - Signal grid extension up to $m_X/m_Y$ ~ 0.5 **New!**



# Event categorization

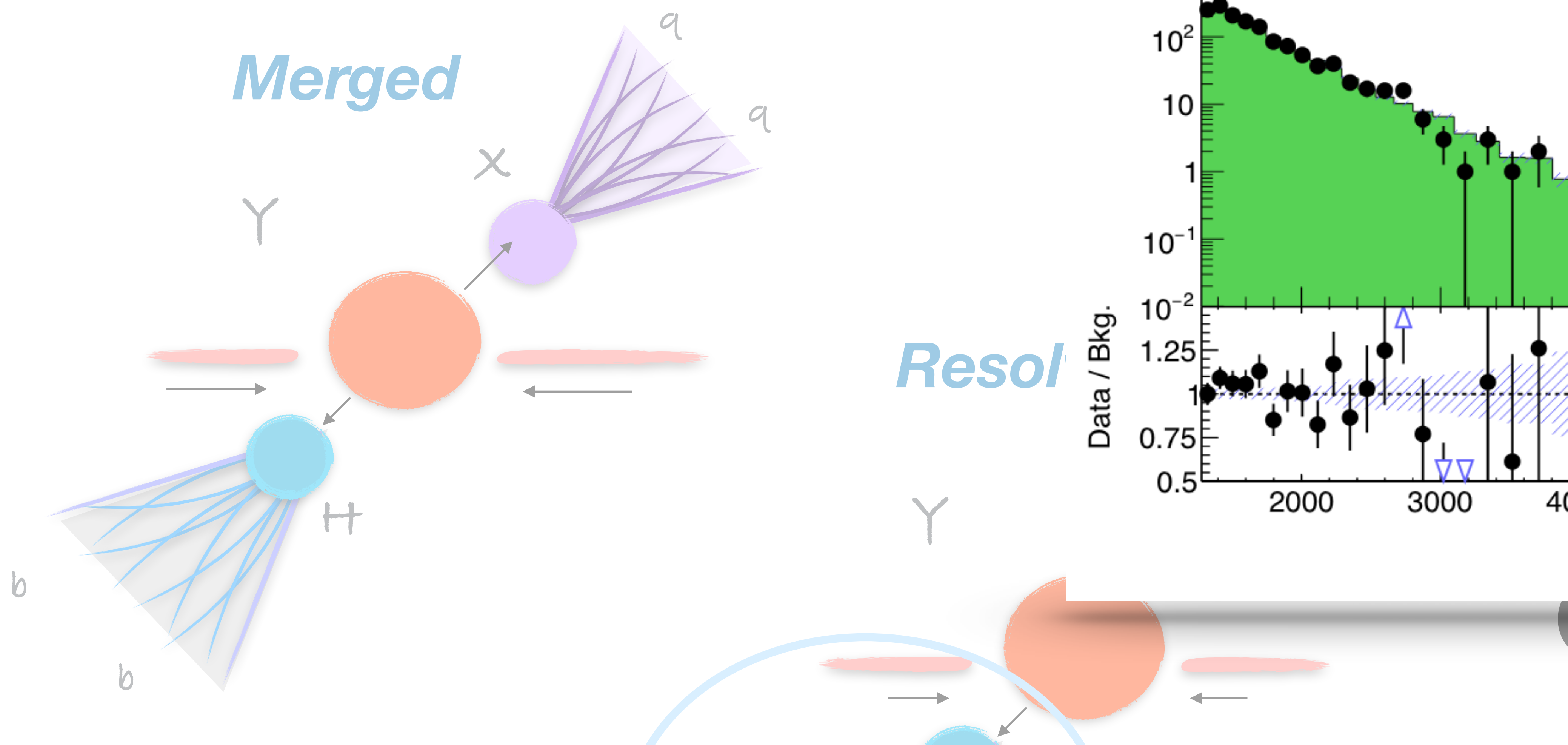
- 2 strategies for X boson reconstruction (with a large-R jet or two small-R jets)
- According to Higgs mass value and Higgs Xbb score (for H->bb tagging), 6 regions are defined

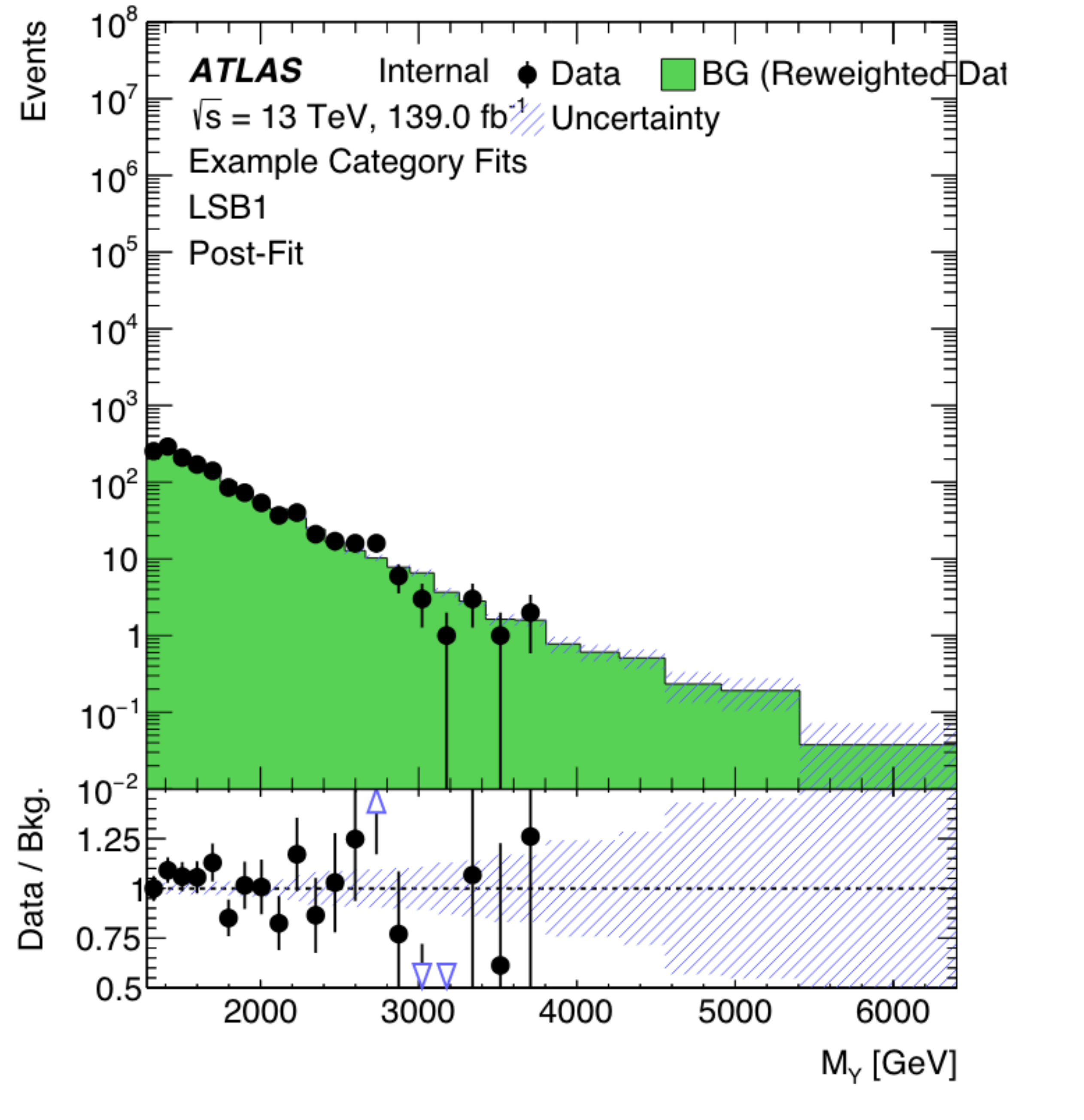


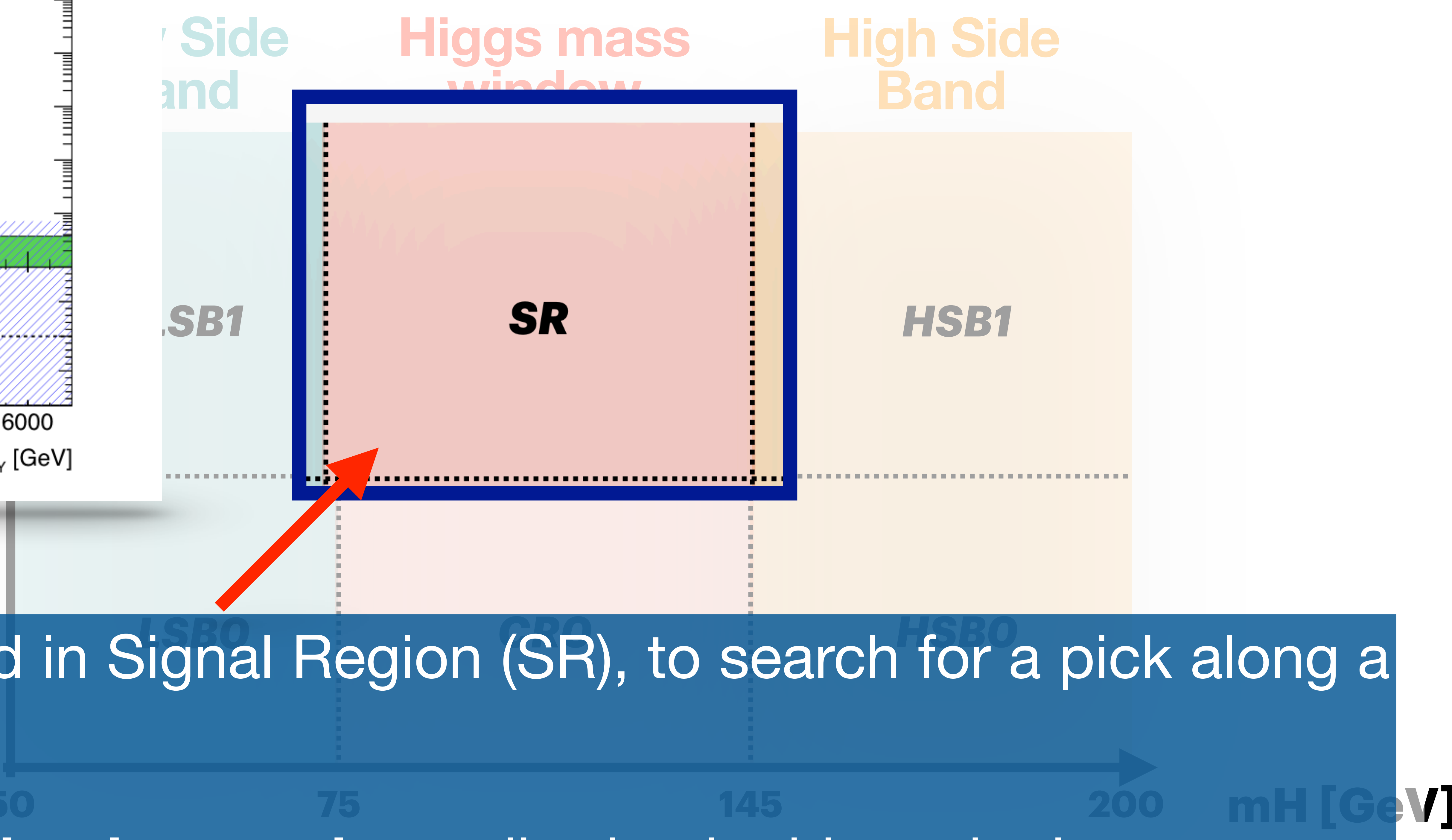
# Event categorization

- 2 kinds of X boson reconstruct ... small-R jets)
- According to Higgs mass value ... ->bb tagging), 6 regions are defined




(=2.46)

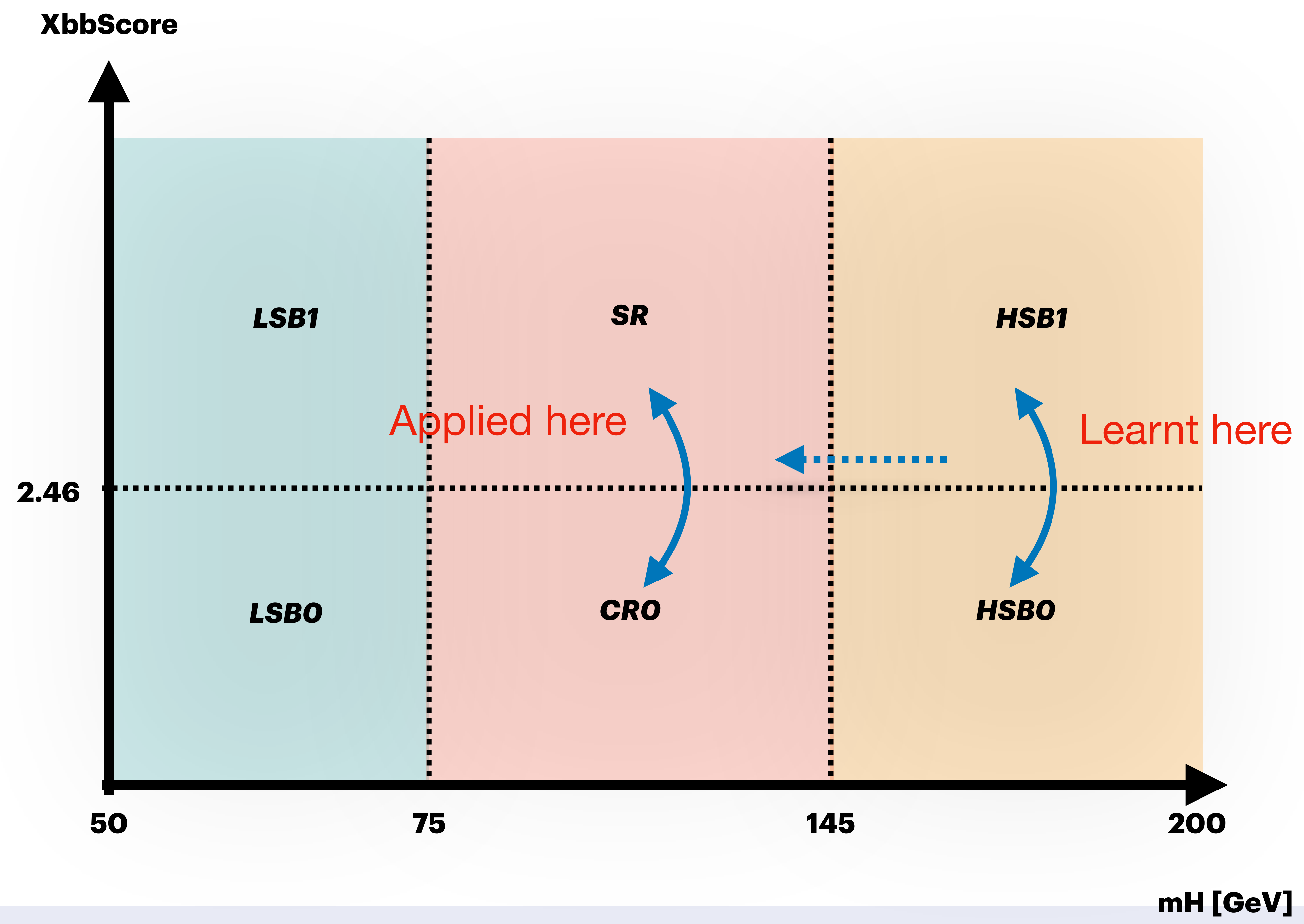



The fit to the invariant mass of XH system is performed in Signal Region (SR), to search for a pick along a smoothly falling background distribution

➡ we need an **estimate of the Standard Model background** contribution in this region!

# Background estimation

- >97% expected background consists in QCD di-jet processes
- Monte Carlo simulations for QCD are not precise enough, therefore we need a data-driven method
- A re-weighting function is needed to map CR0 into SR data
- This function can be learnt in HSB and applied in Higgs mass window
    - Validated assumption: Xbbscore -tagged/-untagged ratio is independent from $m_H$

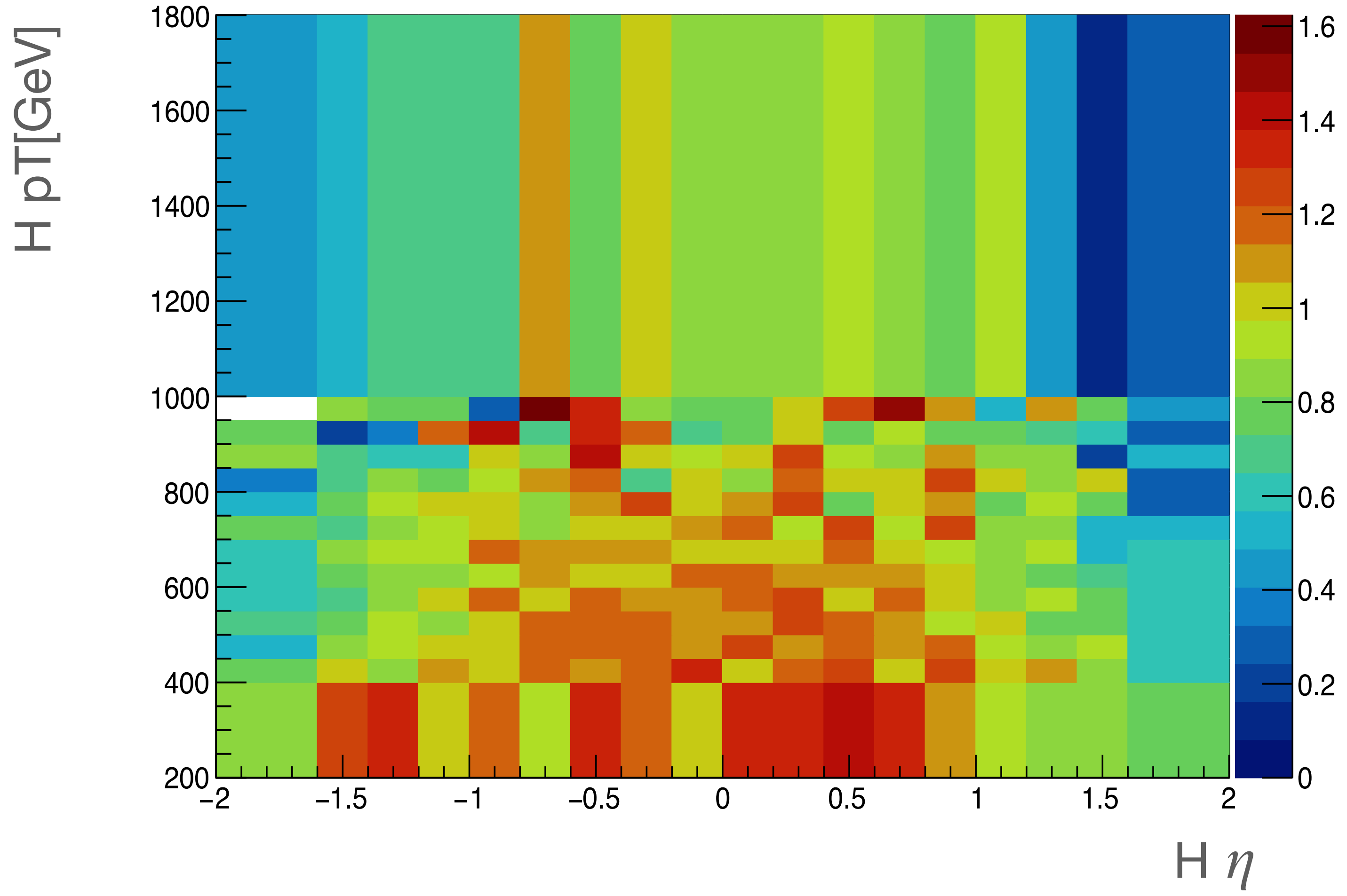


# Histogram-based method

- Simple procedure, adopted in the previous paper:
  - Divide significant histograms between HSB1 and HSB0 to obtain re-weighting factors
  - Apply the same factors to CR0 and obtain the shape in SR

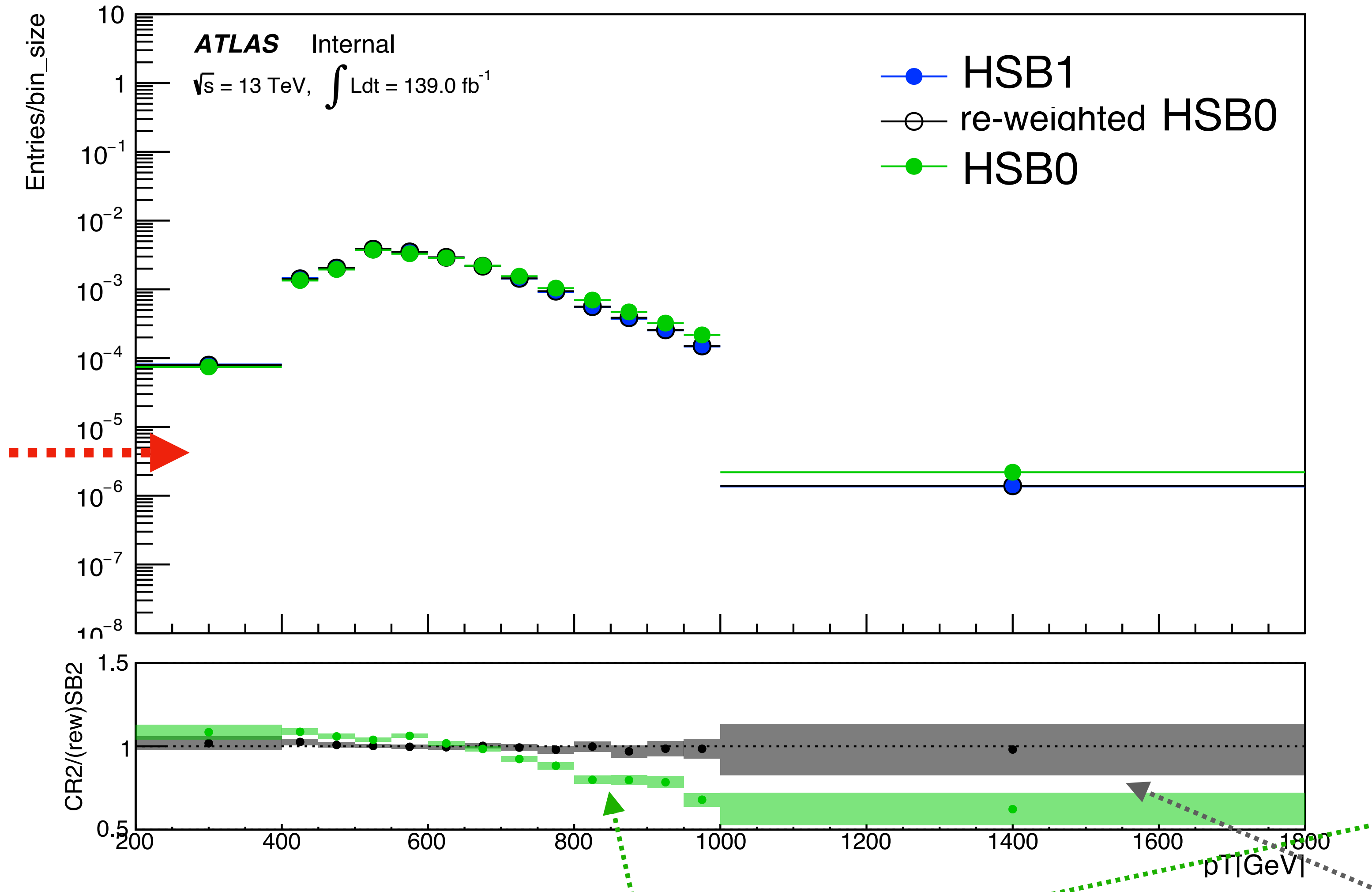
The limit of this method is that only a finite set of variables is well re-weighted

**HSB1/HSB0**

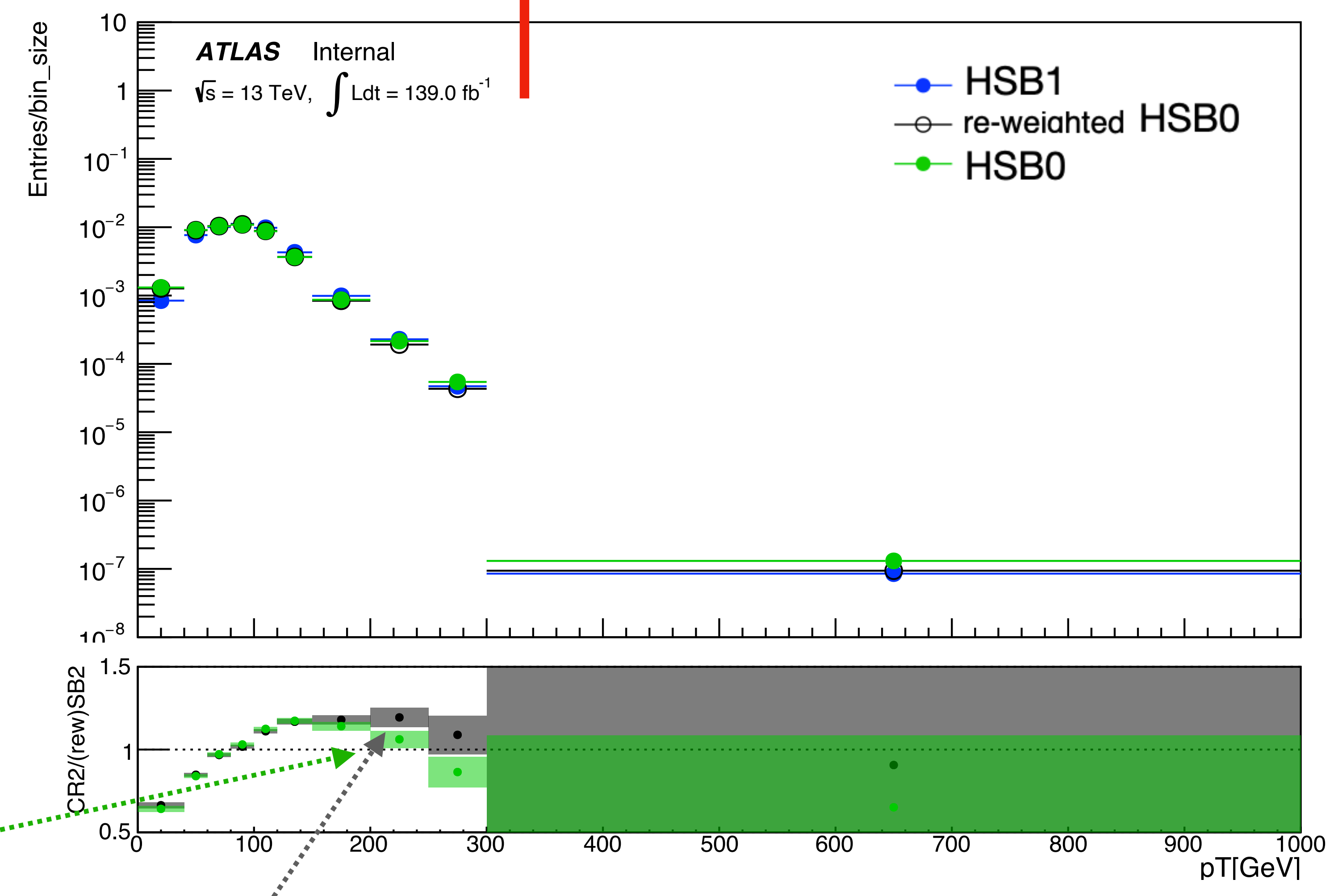



*Higgs candidate pT and η*

H pT




Trackjet 2 pT




*Before reweighing* *After reweighing*

# DNN-based Reweighting

- The re-weighting problem consists in learning some function $w(x)$ between two probability densities $p_0(x)$ and $p_1(x)$:

$$w(x) \cdot p_0(x) = p_1(x) \; -> \; w(x) = \frac{p_1(x)}{p_0(x)}$$

The re-weighting function has the form of a probability density ratio

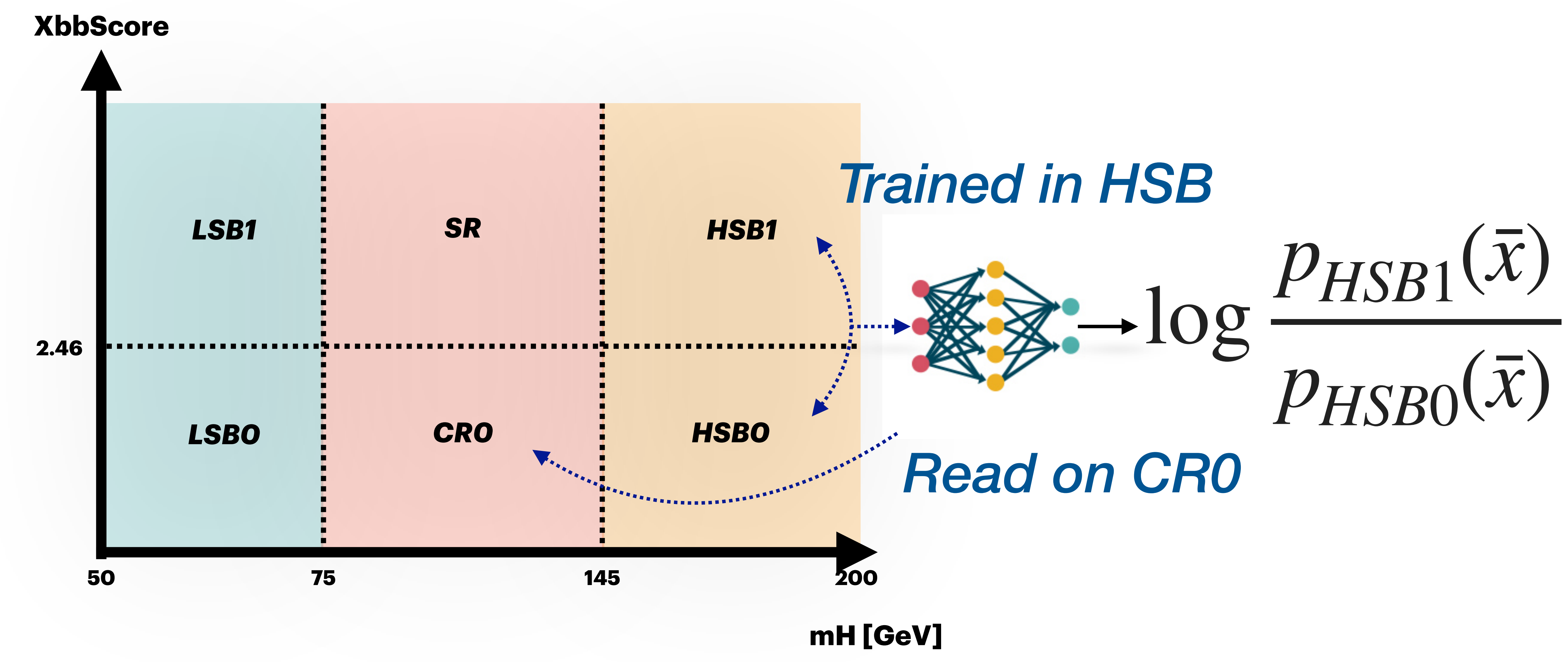
- It can be directly estimated from data with a DNN, by minimizing the loss function:

$$J(\theta) = E_{p_0}\sqrt{e^{u(\bar{x},\Theta)}} + E_{p_1}\frac{1}{\sqrt{e^{u(\bar{x},\Theta))}}}$$

- The DNN prediction has the form

$$u_r(\bar{x}, \bar{\Theta}) = \log \frac{p_1(\bar{x})}{p_0(\bar{x})}$$

Intrinsic multi-dimensionality!



# DNN-based Reweighting

- The re-weighting problem consists in learning some function $w(x)$ between two probability densities $p_0(x)$ and $p_1(x)$:

- The likelihood ratio estimation problem is well known in Statistics, for sure not restricted only to this particular case

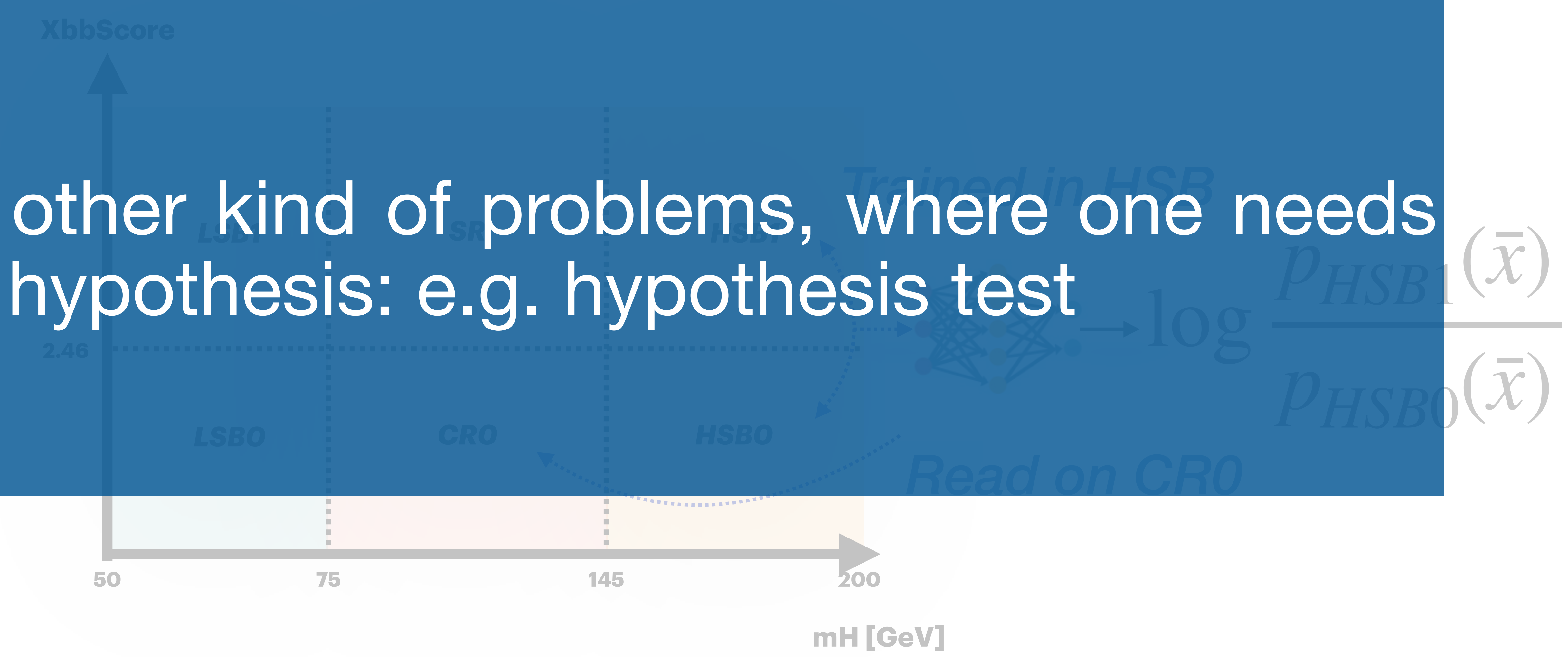
- Applications of the method possible for other kind of problems, where one needs to know the likelihood ratio between two hypothesis: e.g. hypothesis test

- It can be directly estimated from data with a DNN, by minimizing the loss function:

- The DNN prediction has the form

$$u_r(\bar{x}, \bar{\Theta}) = \log \frac{p_1(\bar{x})}{p_0(\bar{x})}$$

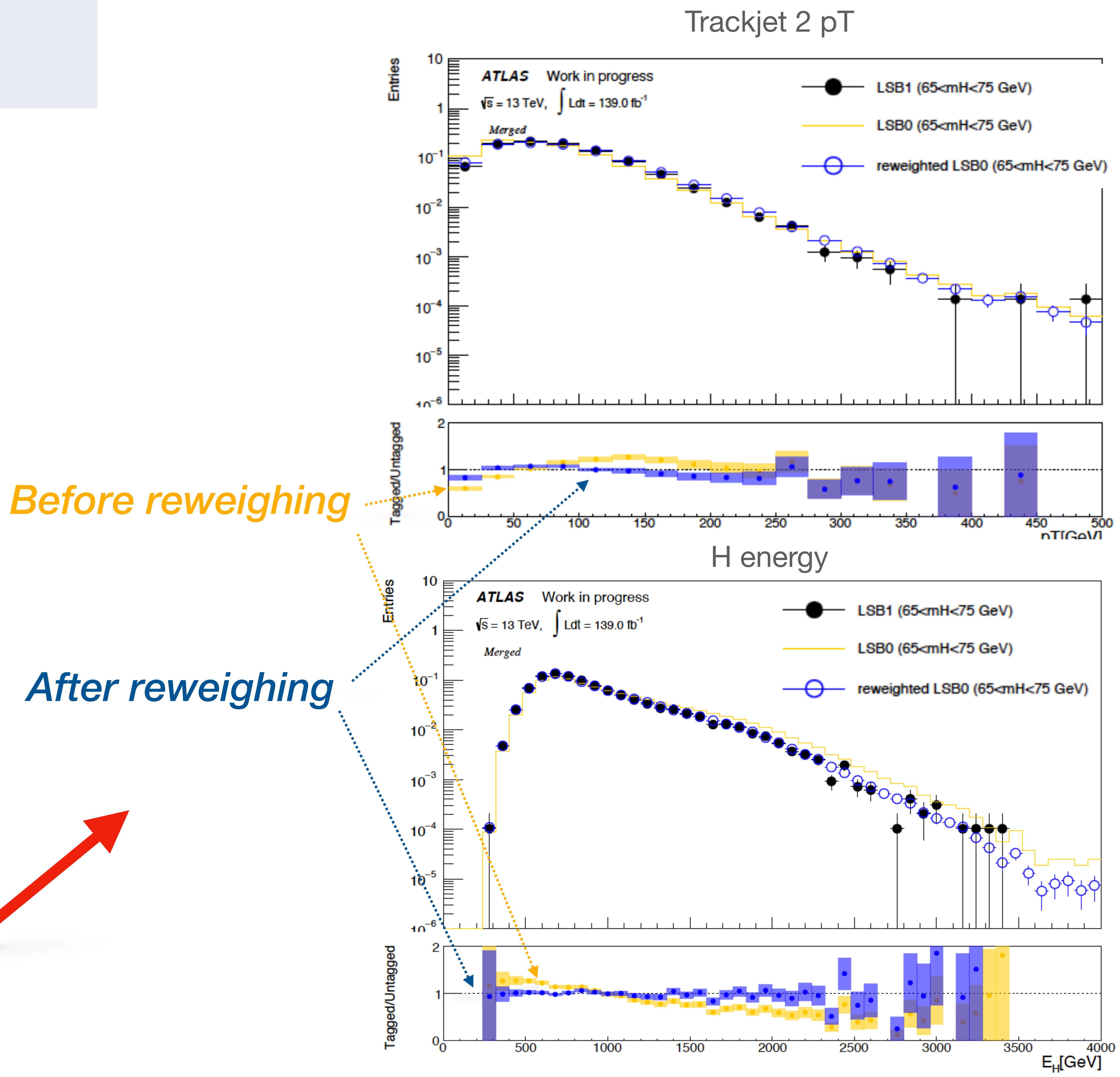
Intrinsic multi-dimensionality!



# Obtained results

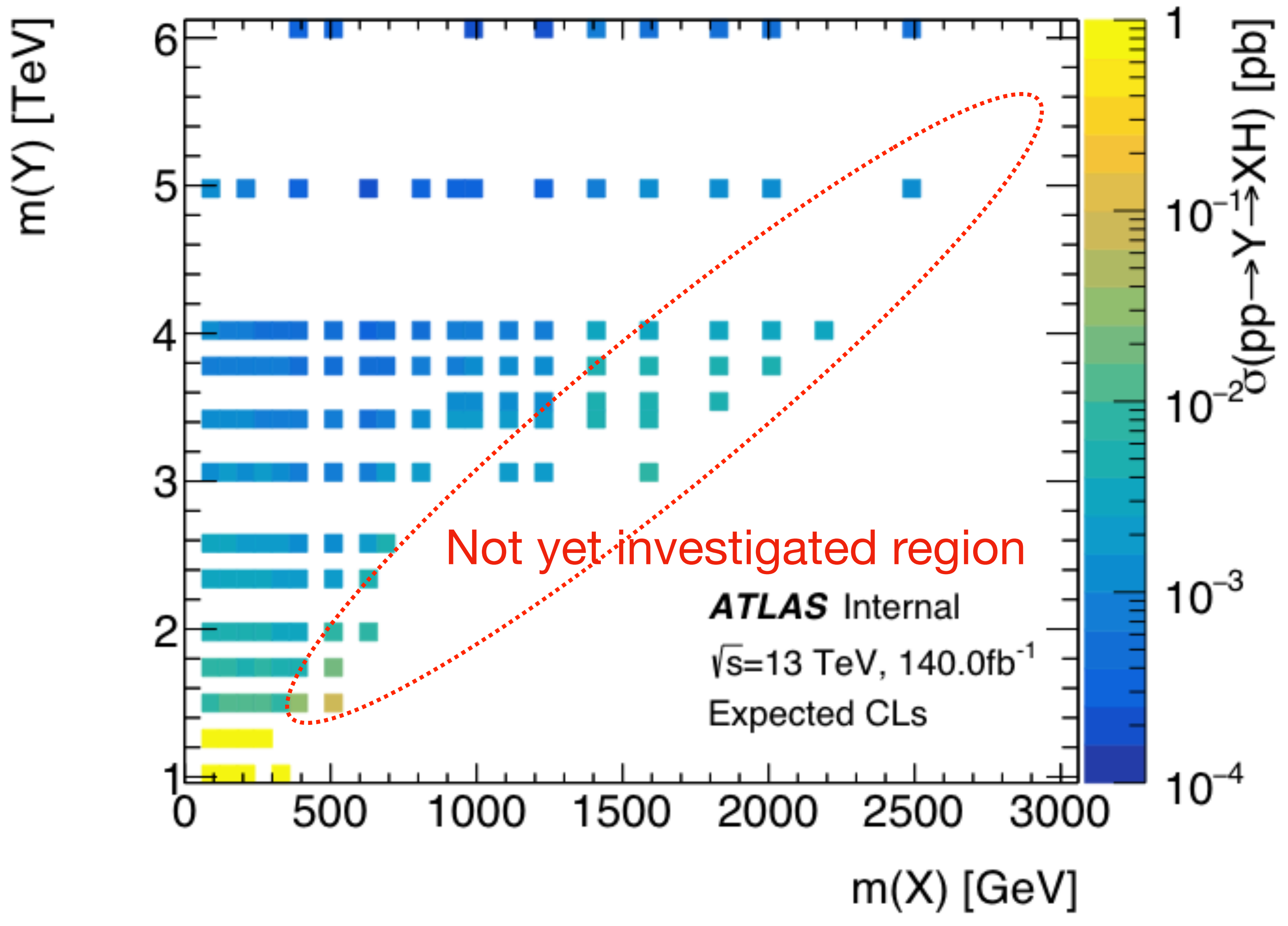
- DNN, with 3 fully-connected inner layers, 20 neurons each, implemented with Keras (and Tensorflow as backend)
- Variables used for training:
  - Higgs candidate 4-momentum and number of track jets associated
  - The first two pT-leading track jets 4-momentum, associated to higgs candidate

Results are very satisfying for all the variables of interest



# Conclusions

**Preliminary expected limits**



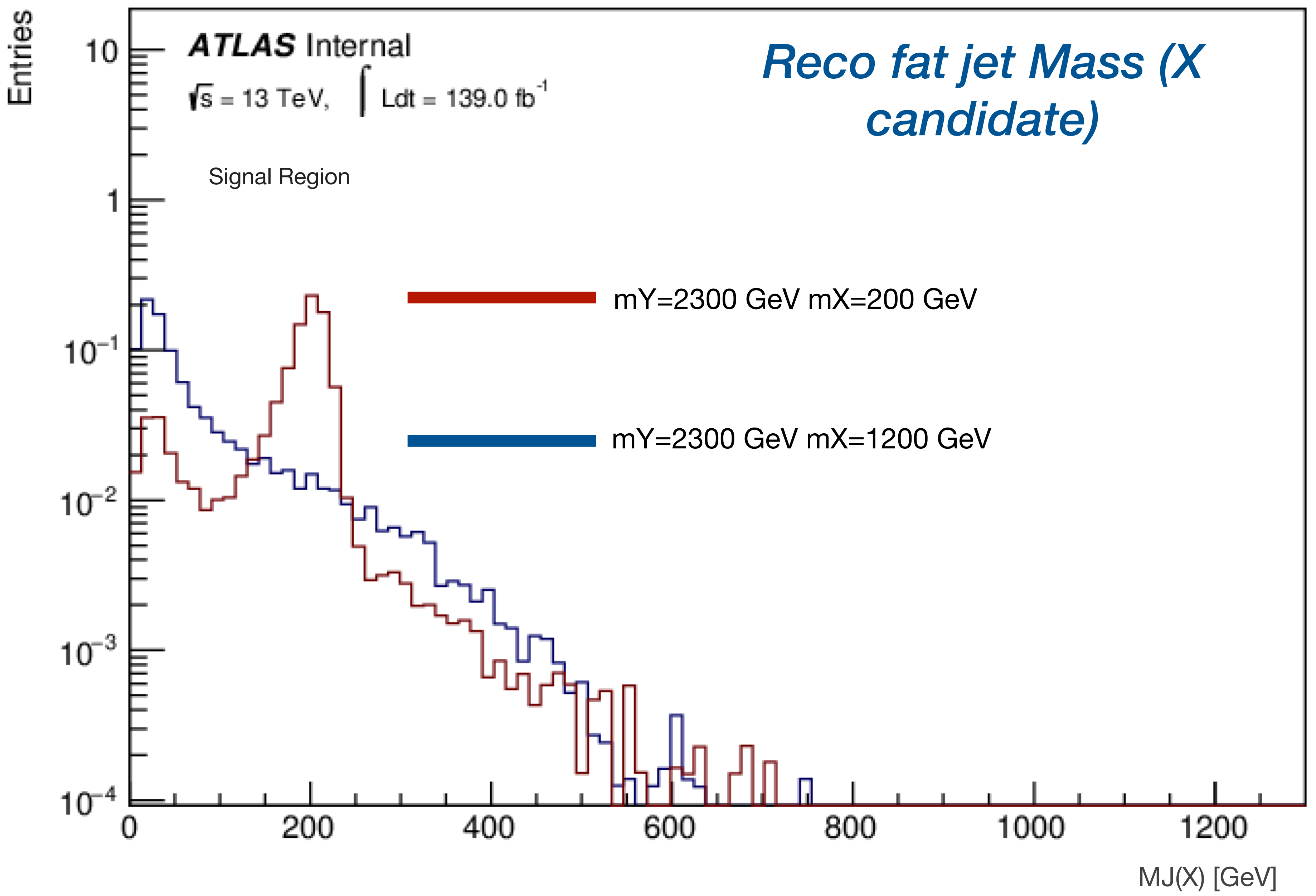

- DNN method for background estimation implemented in the analysis, as well as the systematics associated
  - Computationally expensive (100 network trained), plan to run on GPU
- The strategy of the analysis is ~99% finalized
  - Preliminary expected limits on cross section show improvements in sensitivity wrt to the previous paper results
  - 12th of January 2022 subgroup pre-approval meeting for unblinding approval
- Planning to show results on Moriond 2022

# Backup

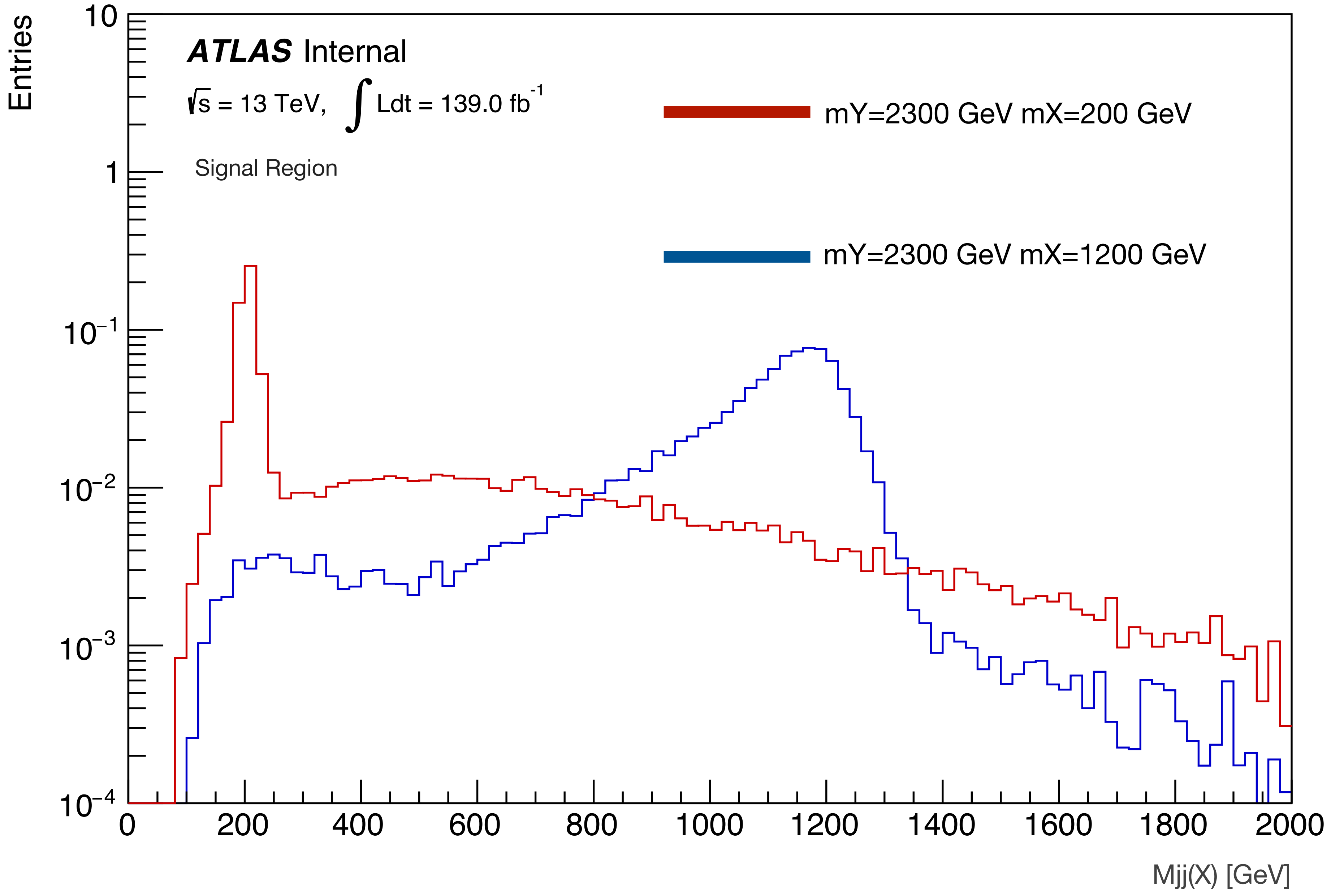
# X mass distribution in merged and resolved regimes

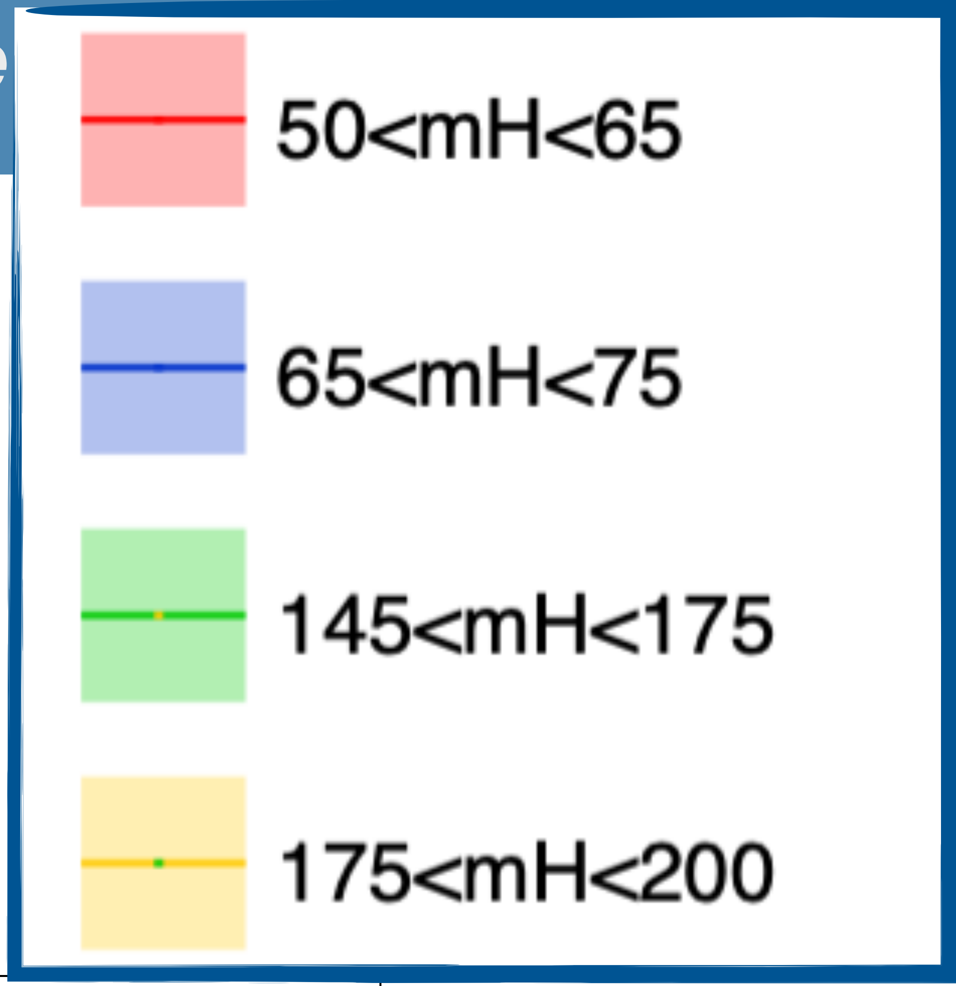
- Signals:
  - mY=2300 GeV and mX = 200 GeV ==> **mX/mY = 0.09**
  - mY = 2300 GeV and mX = 1200 GeV ==> **mX/mY = 0.52**

## 1 large-R jet




## 2 small-R jets

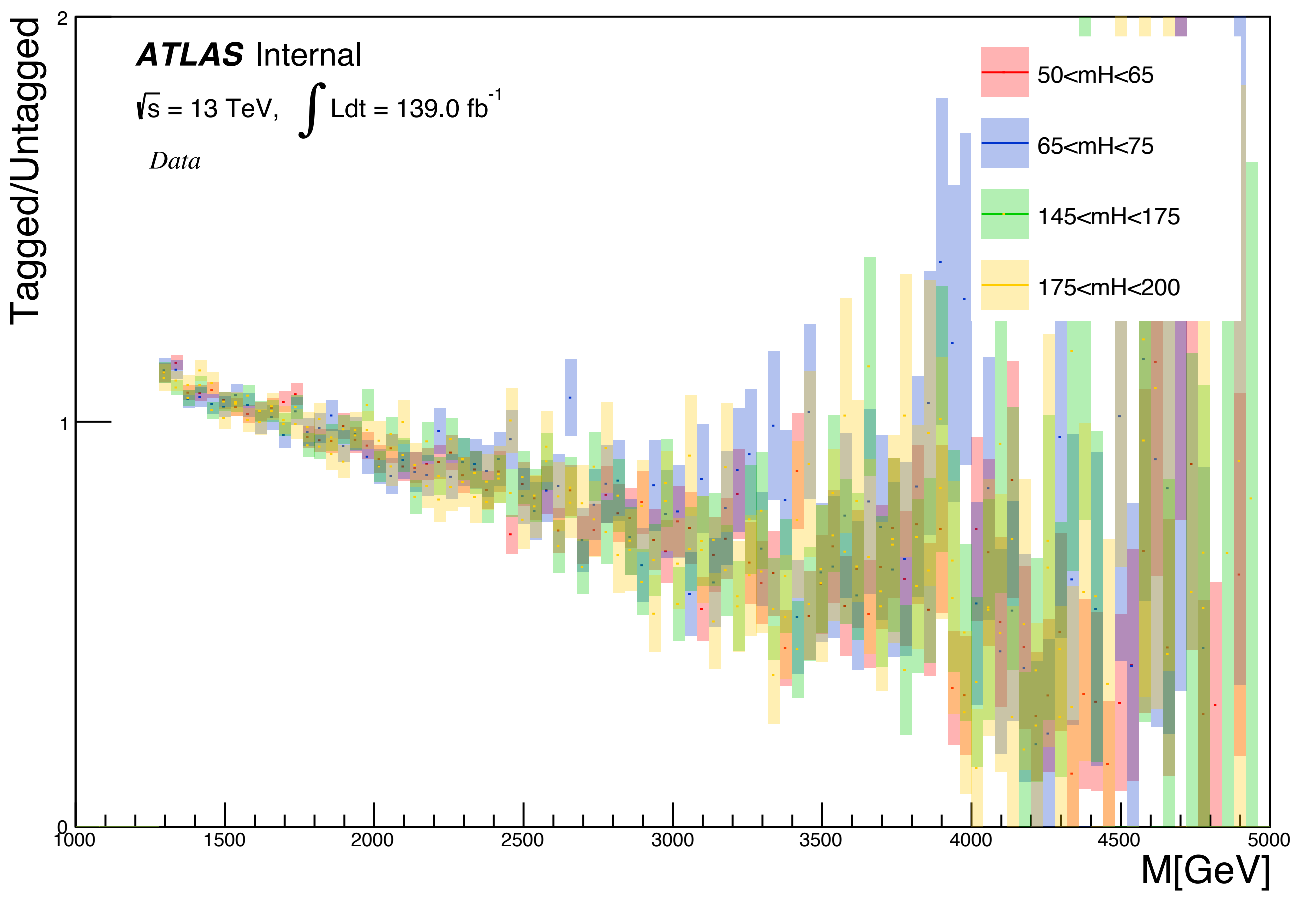




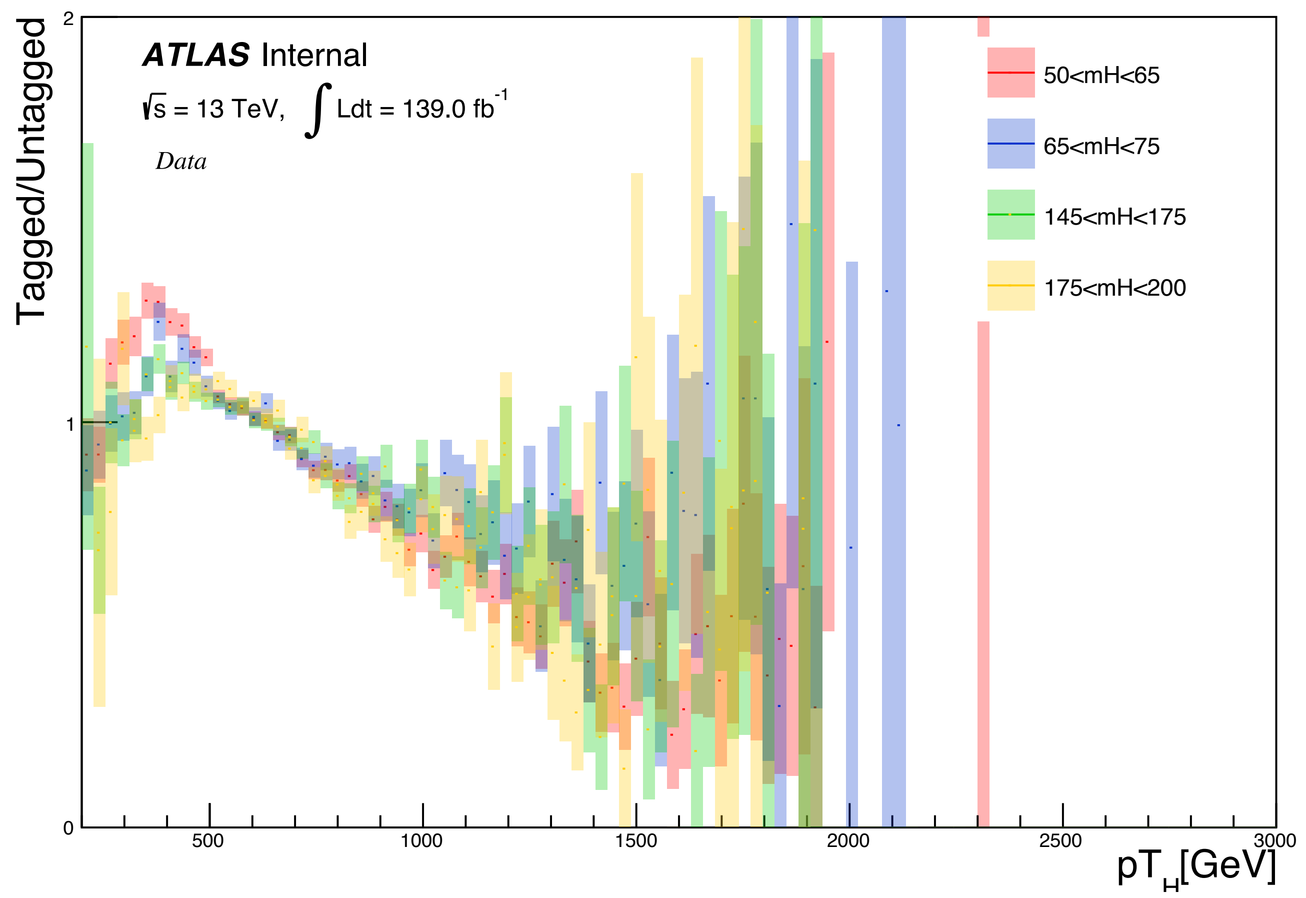

# Independence of Xbbscore from mH

- Xbbscore tagged/untagged ratios, compared among several mH windows (other studies here)

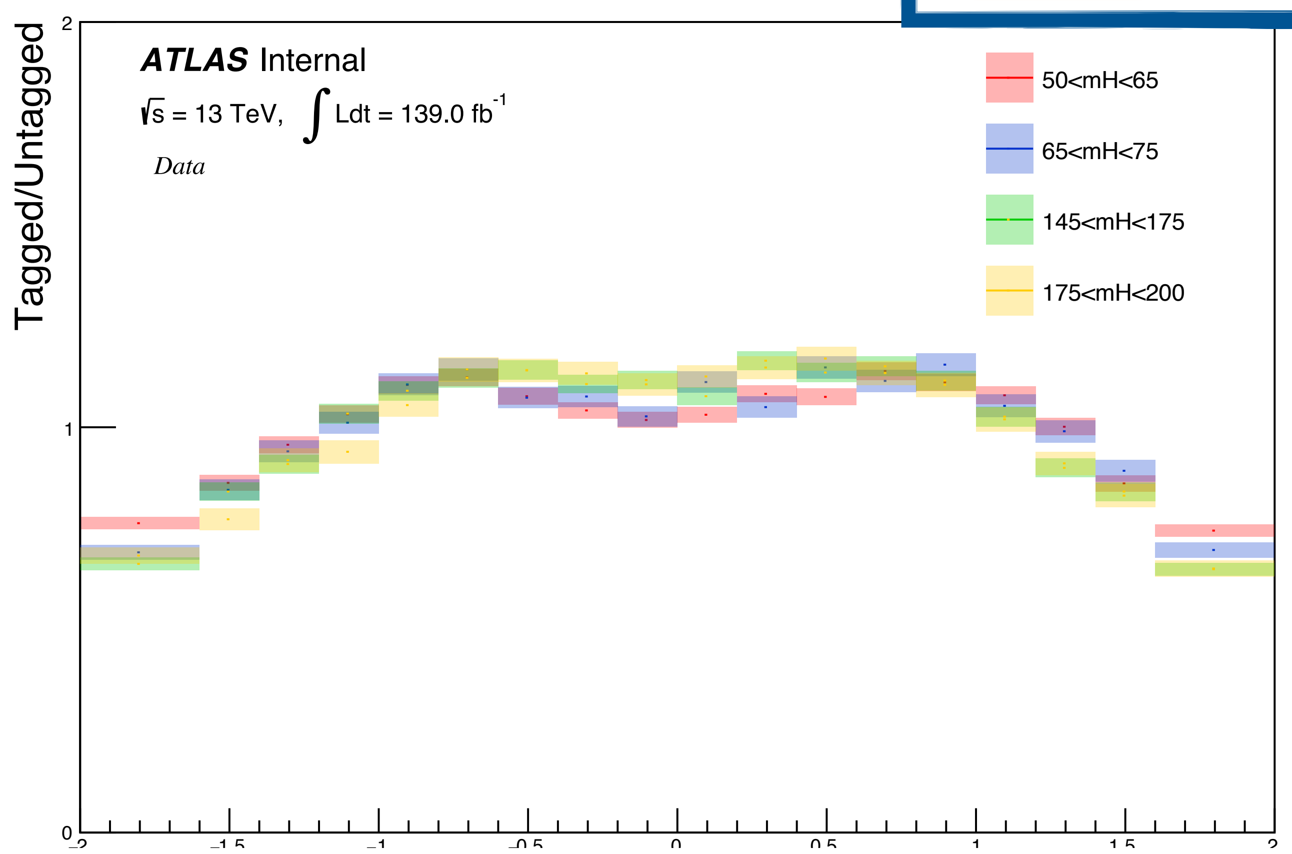
mVH

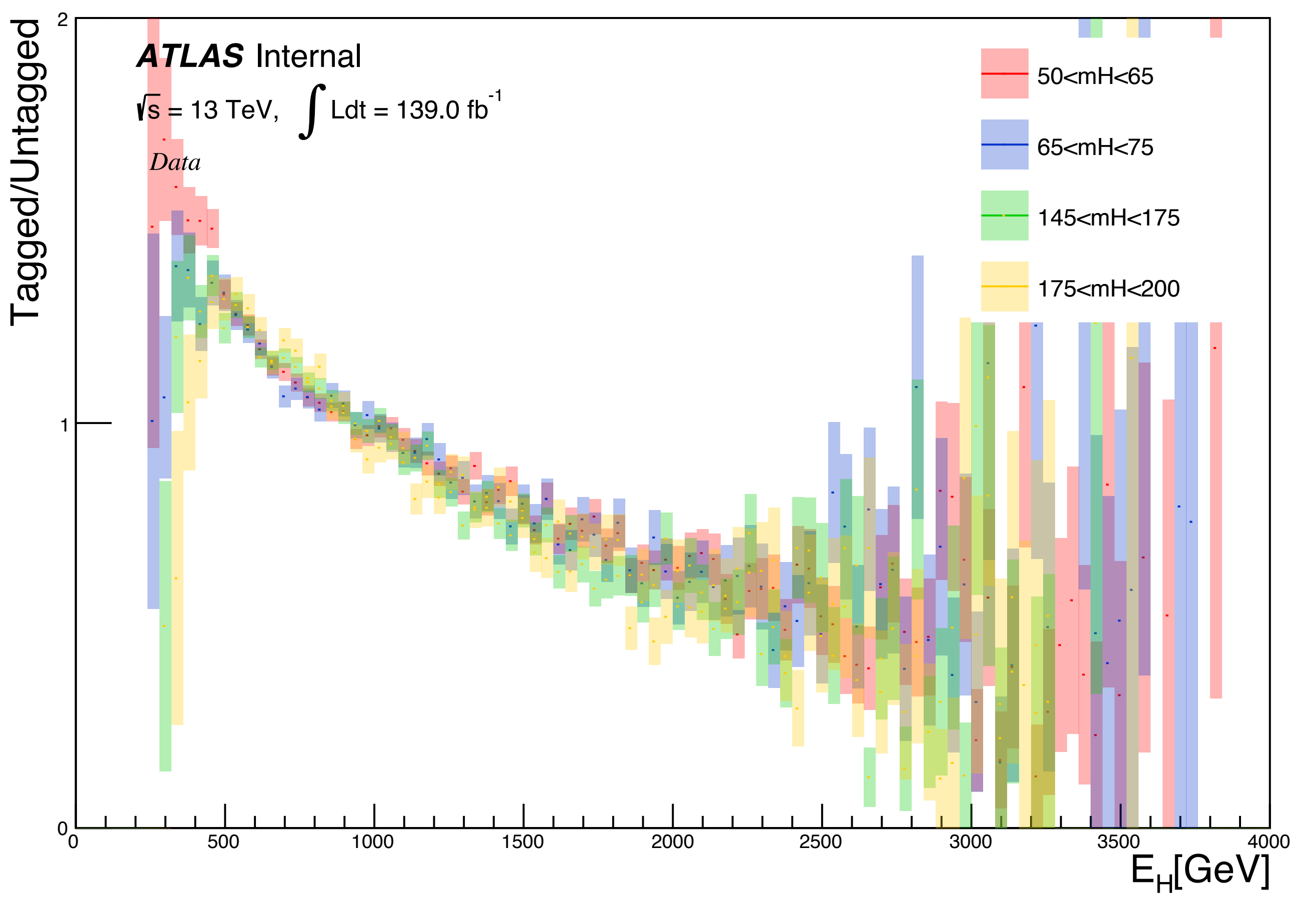



pTH

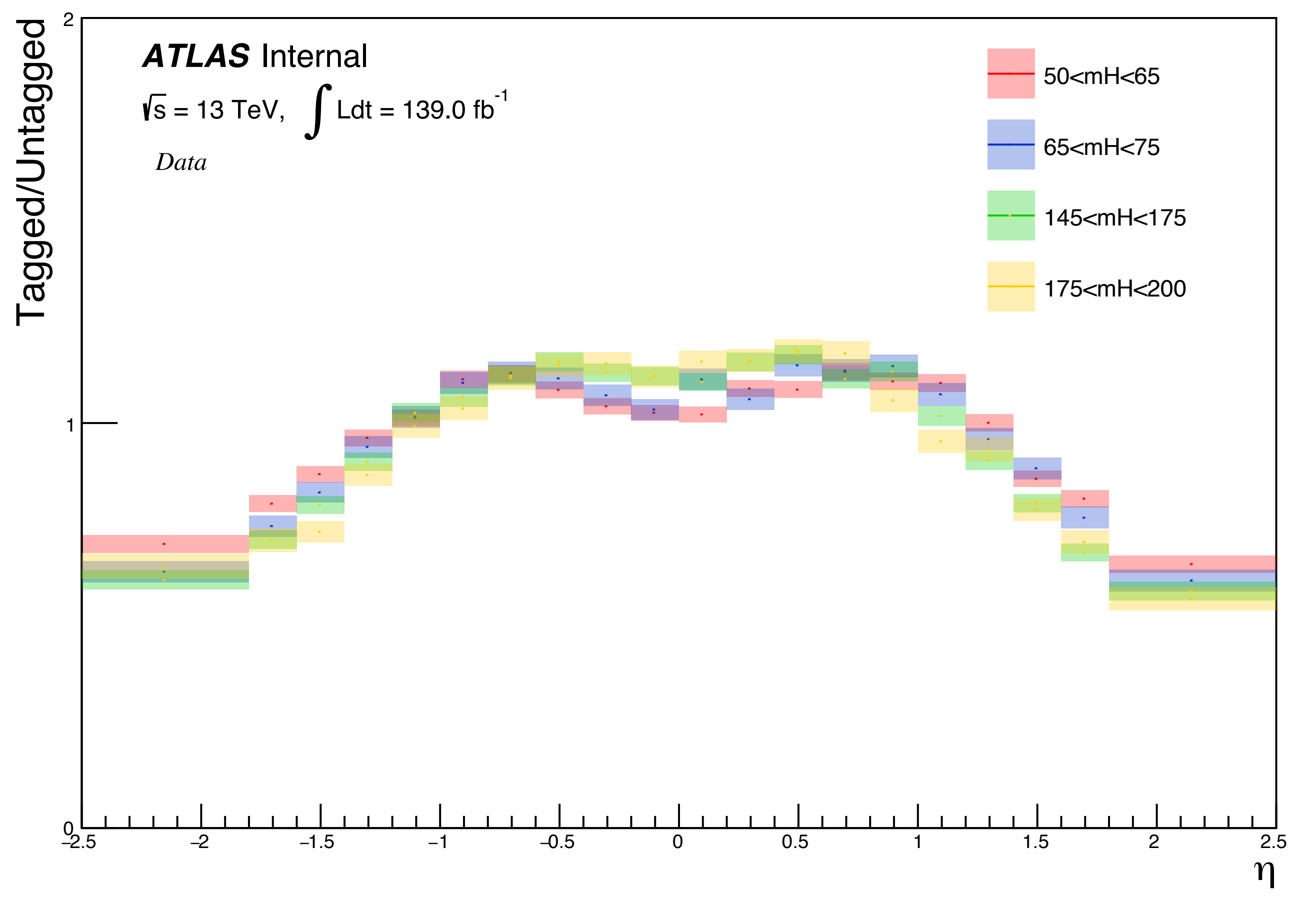



etaH

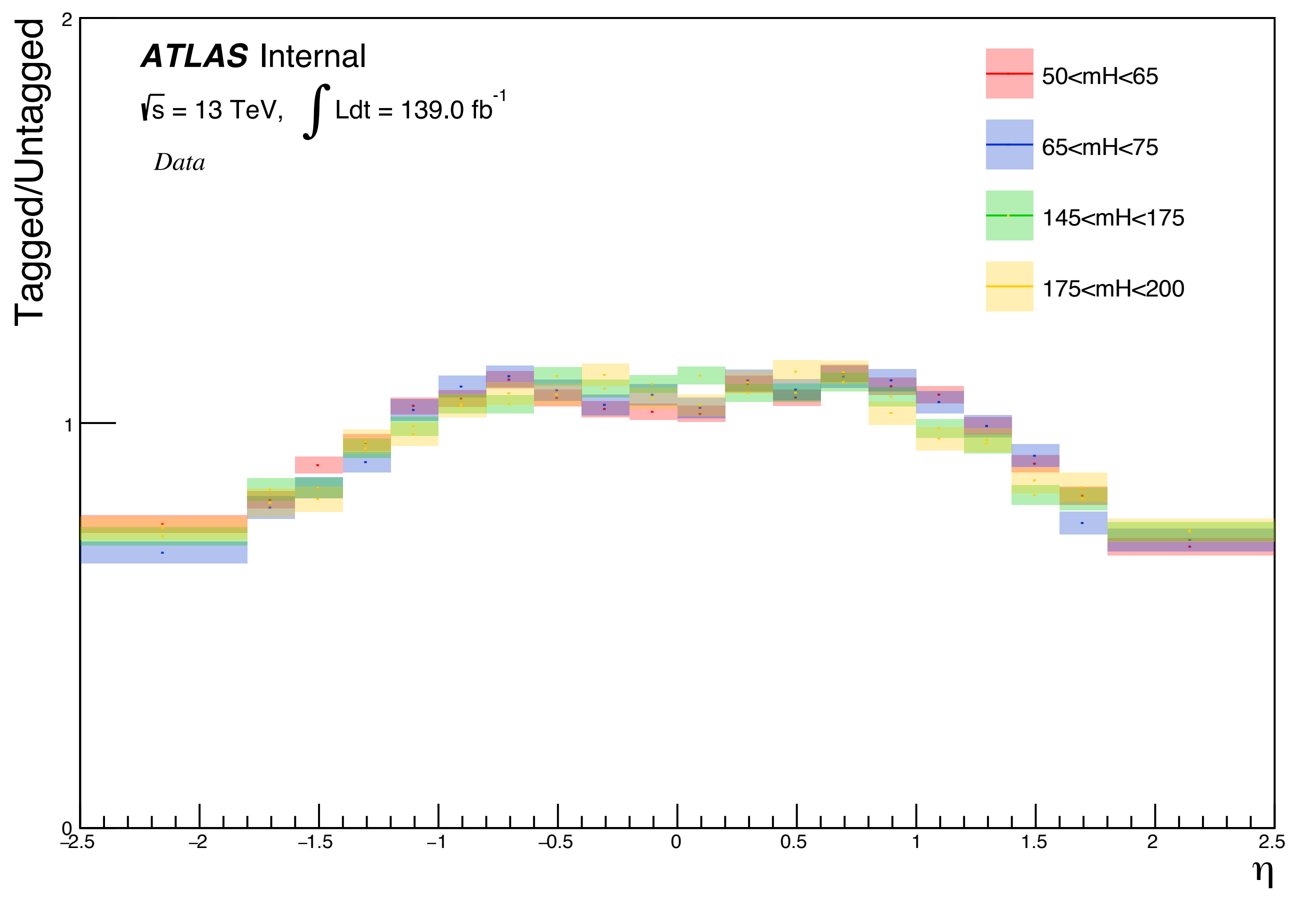



trackjet1_eta



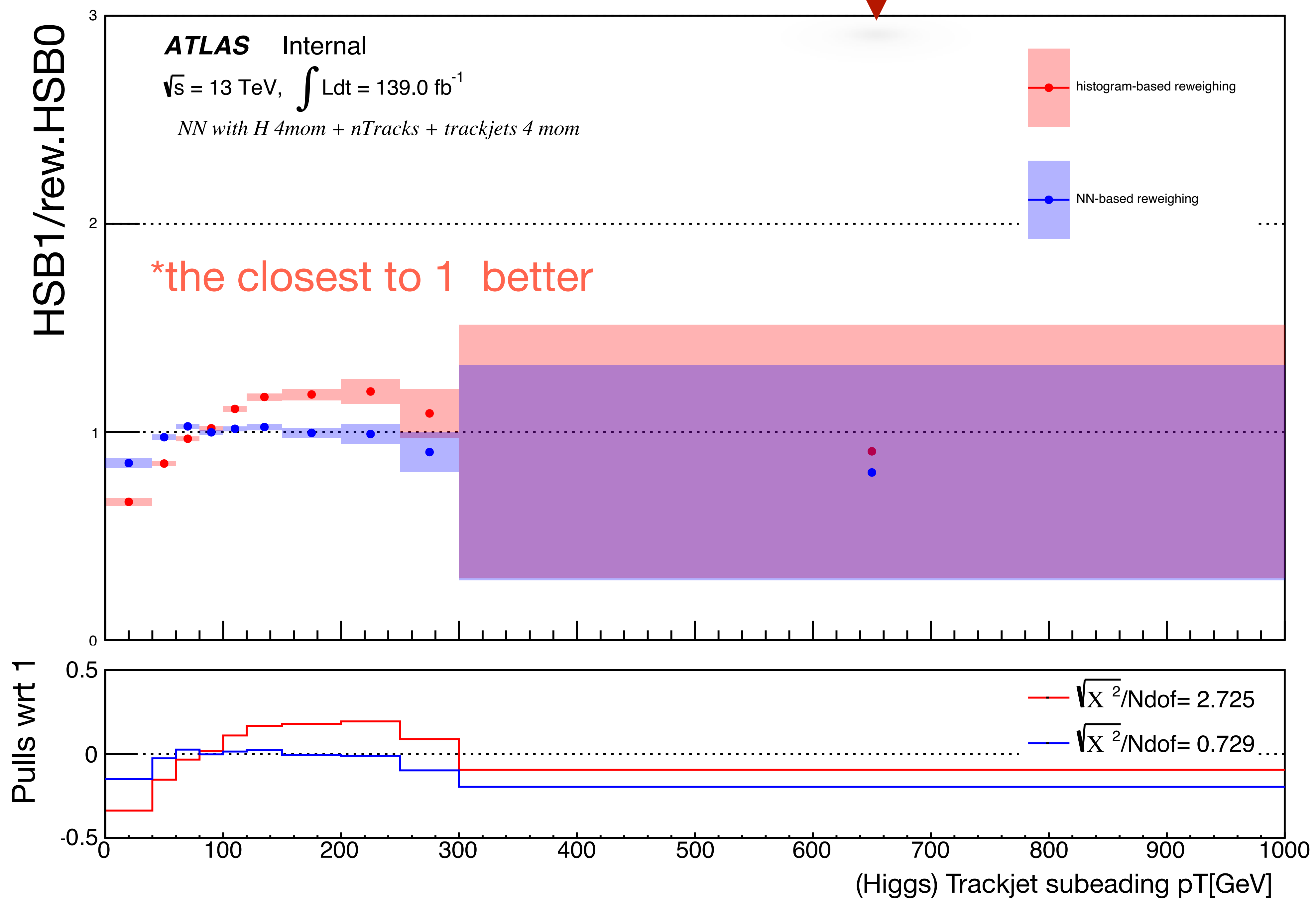

trackjet2_eta

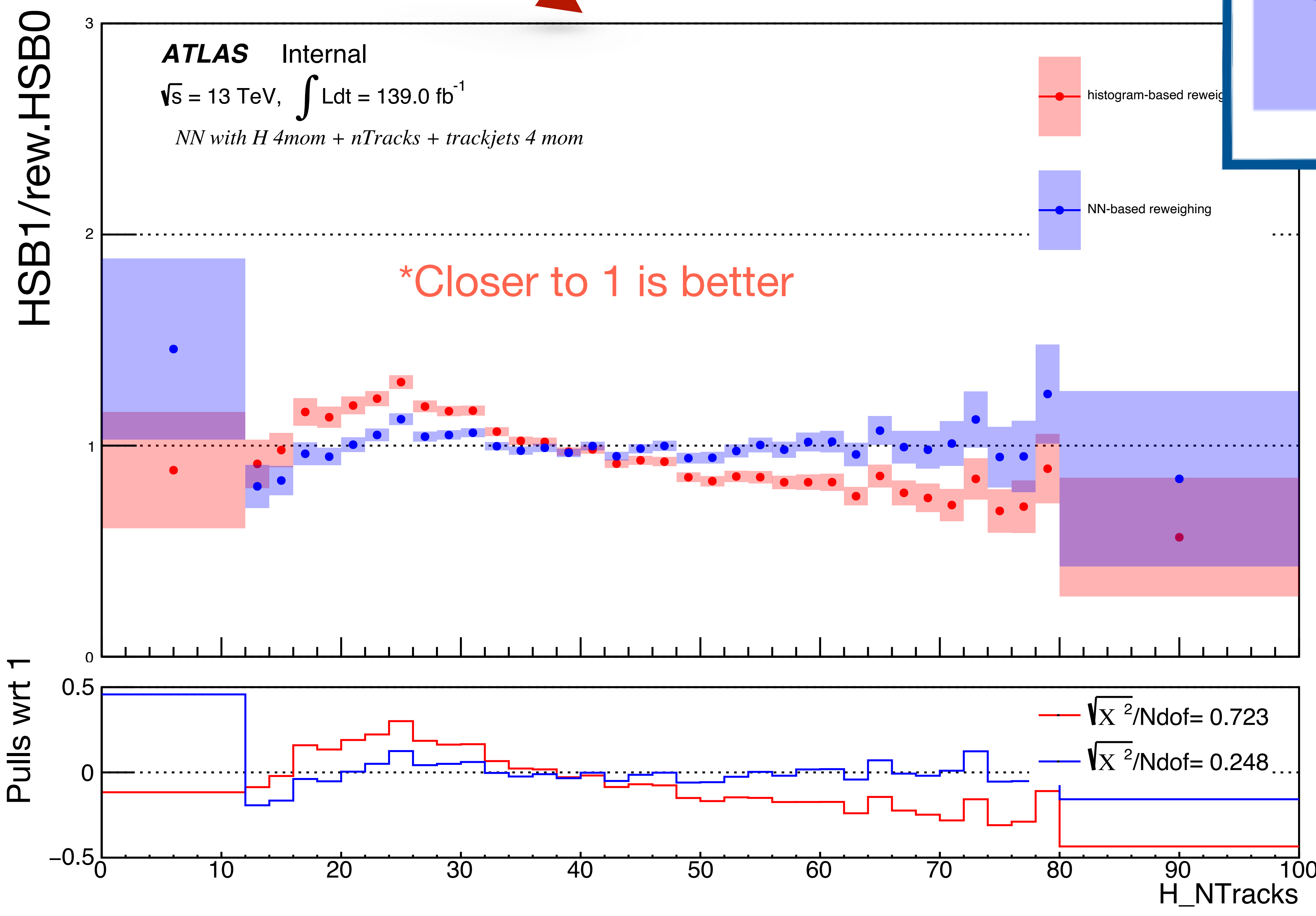


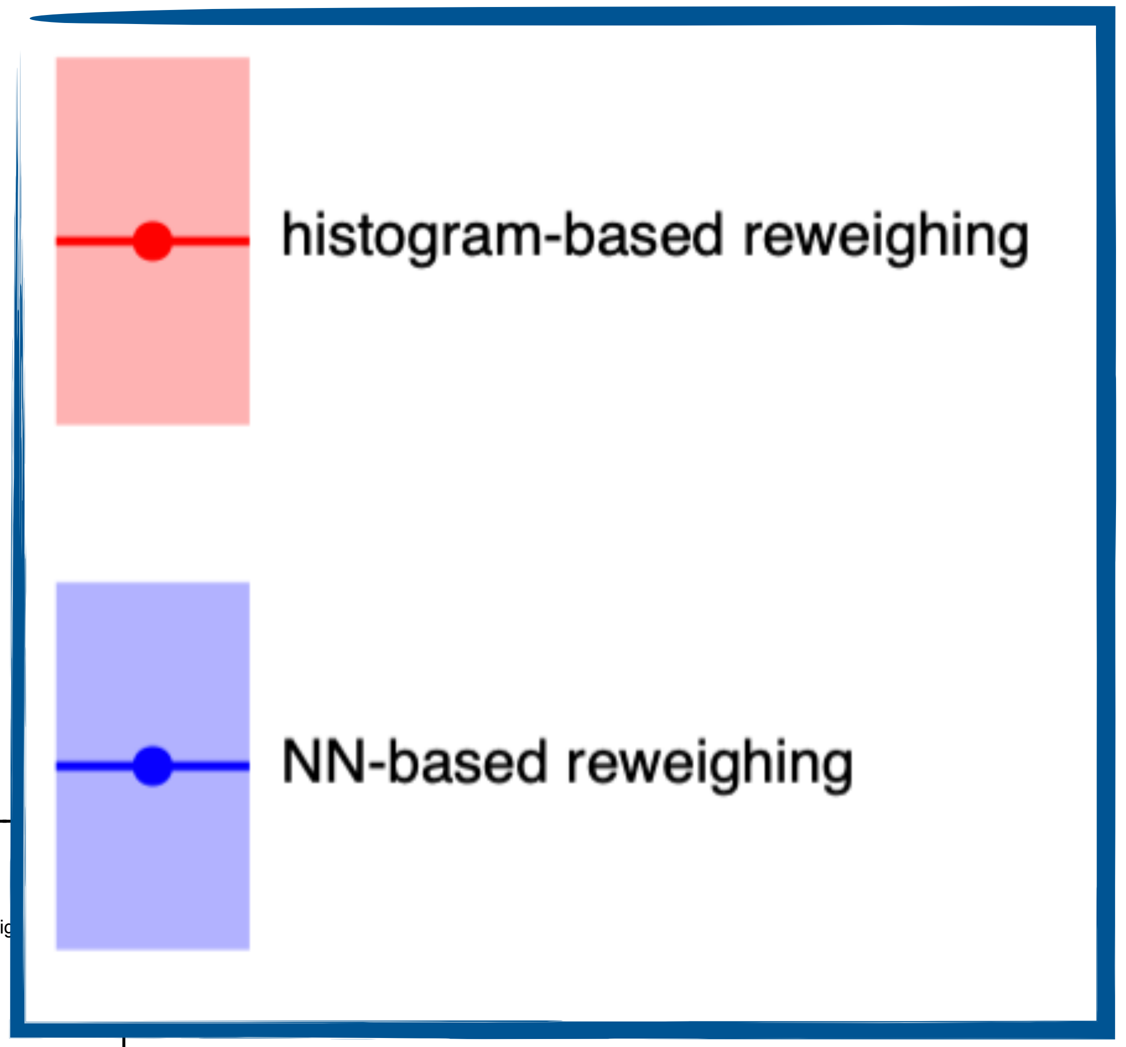
# Comparisons with the histogram method

- Good reweighing on variables not well reweighted from the histograms method

## *Ratio HSB1/reweighted HSB0*



# DNN-based method

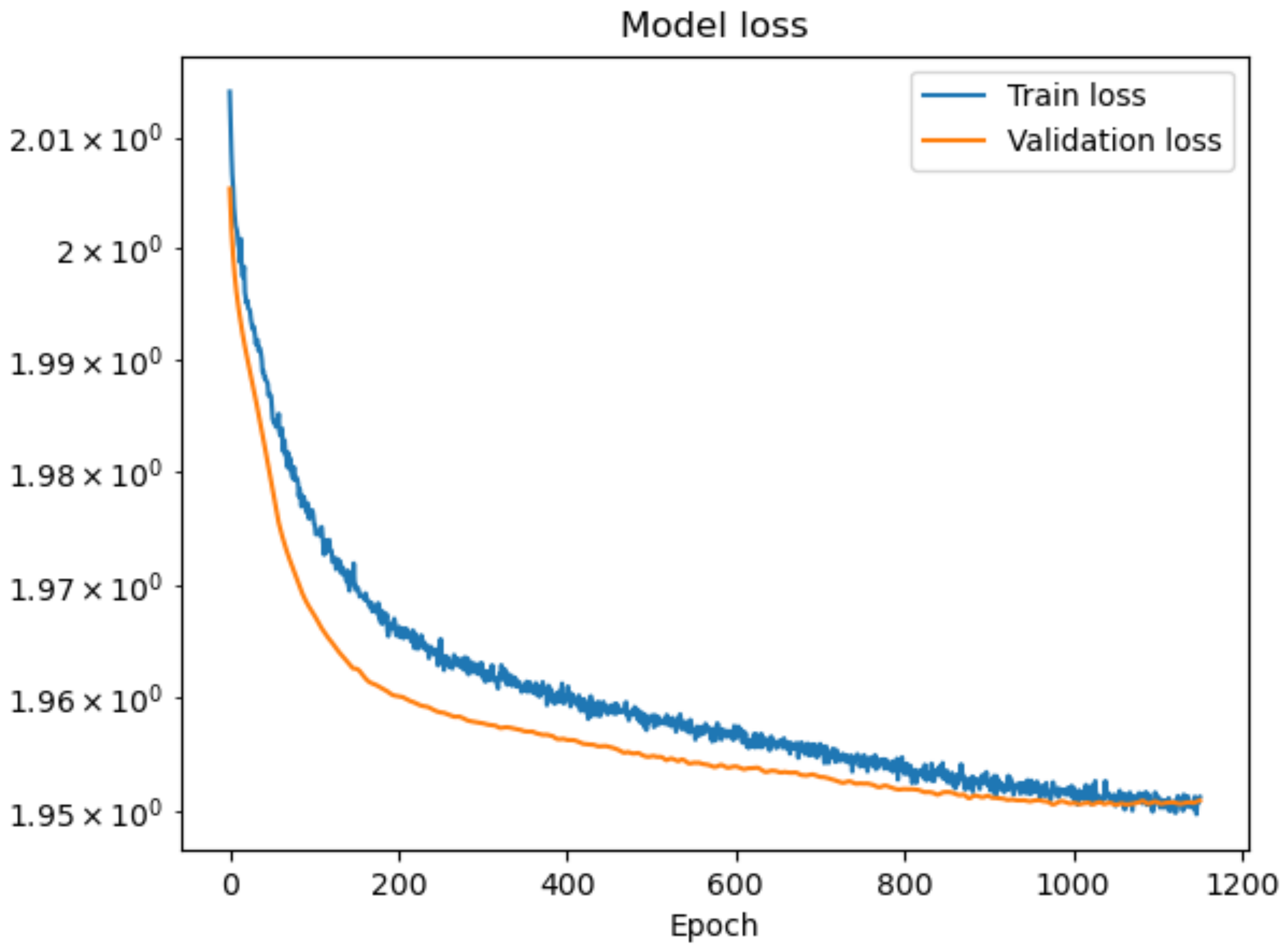
- Keras DNN model with:
  - 20 neurons per inner layer
  - 3 fully-connected inner layers
  - 0.1 dropout
- Training parameters:
  - Max 1600 epochs, with early stopping at 100 epochs
  - Batch size = full dataset

*Customized loss function*

$$J(\theta) = E_{p_0}\sqrt{e^{u(\bar{x},\Theta)}} + E_{p_1}\frac{1}{\sqrt{e^{u(\bar{x},\Theta))}}}$$

*minimized on*

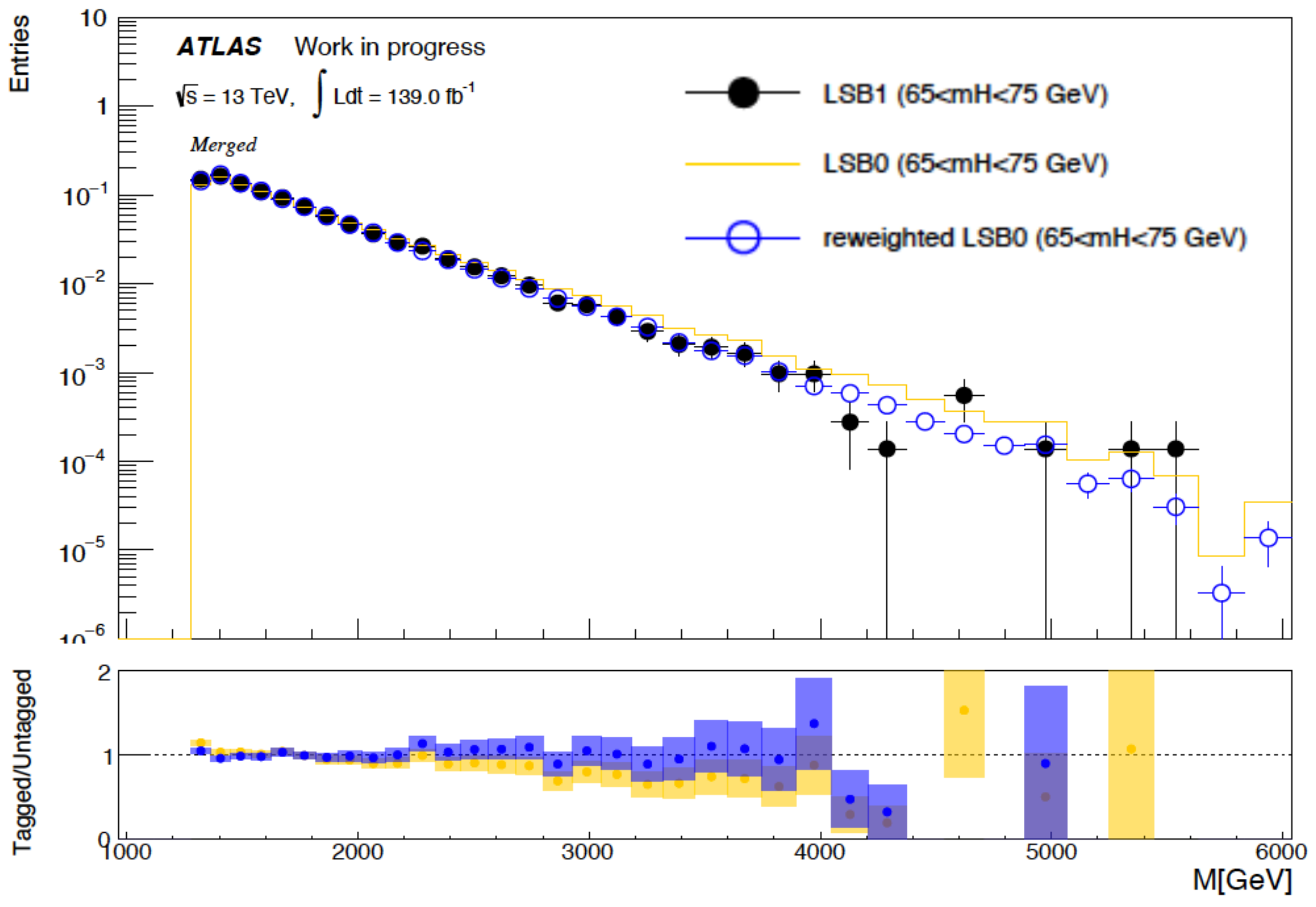
$$\log \frac{p_{HSB1}(\bar{x})}{p_{HSB0}(\bar{x})}$$



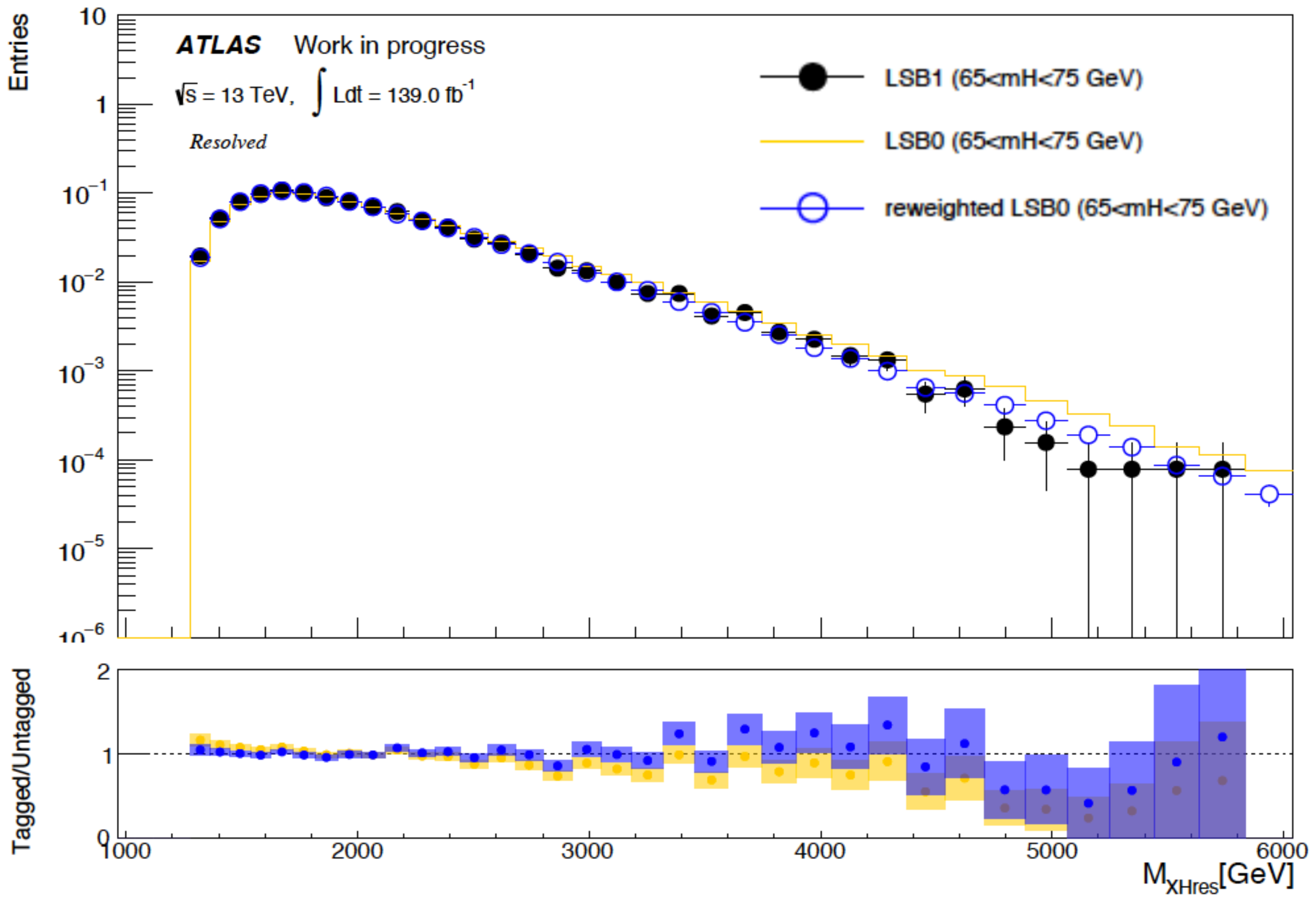
# Background Templates

- Final discriminant are well described!

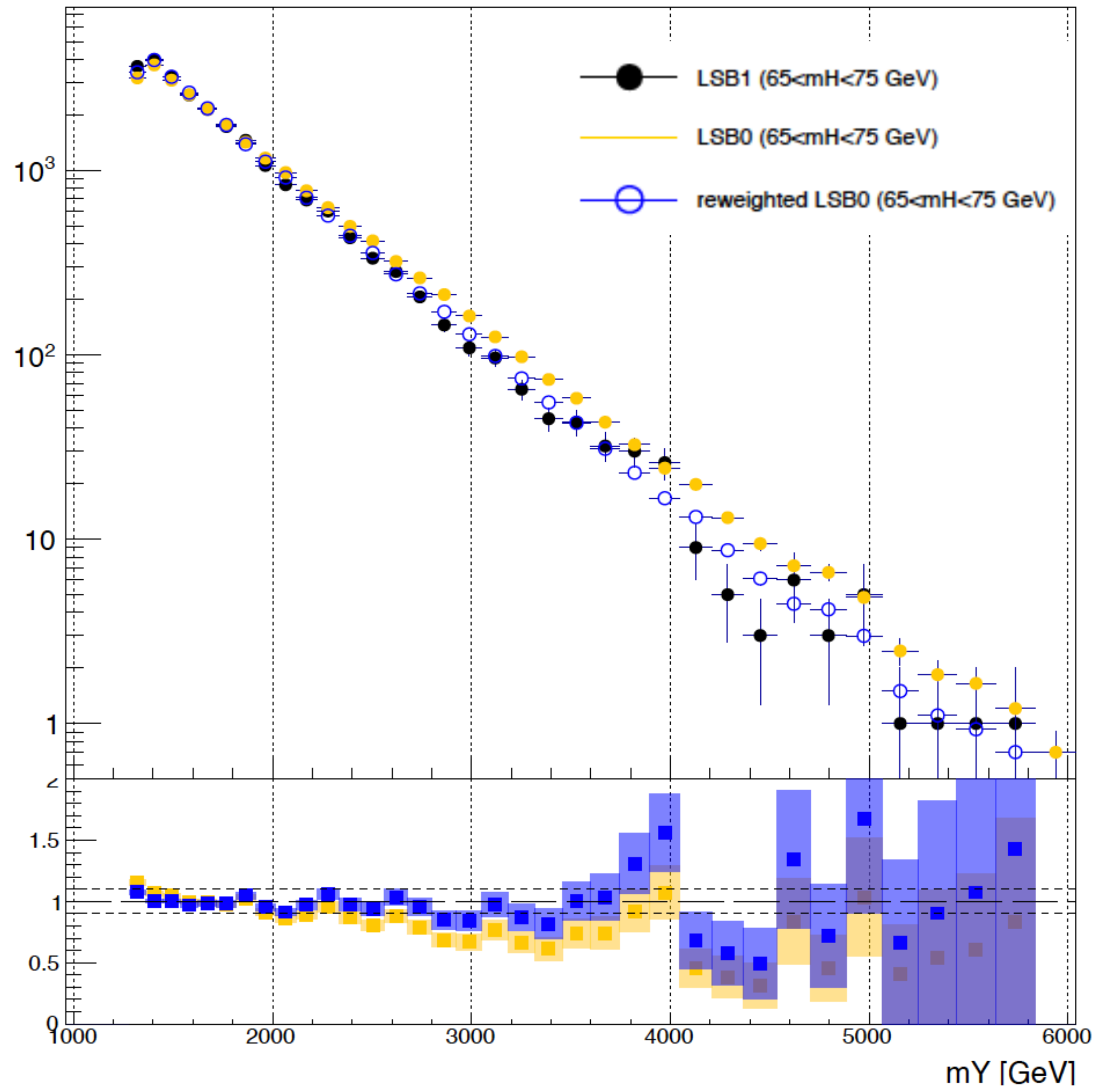
Merged exclusion region




Resolved exclusion region




AS-based discovery region

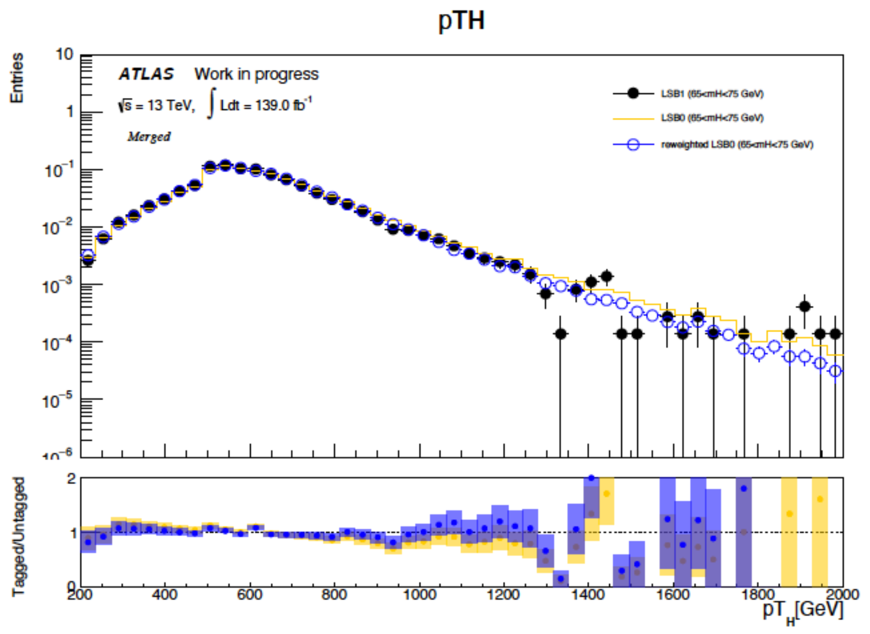


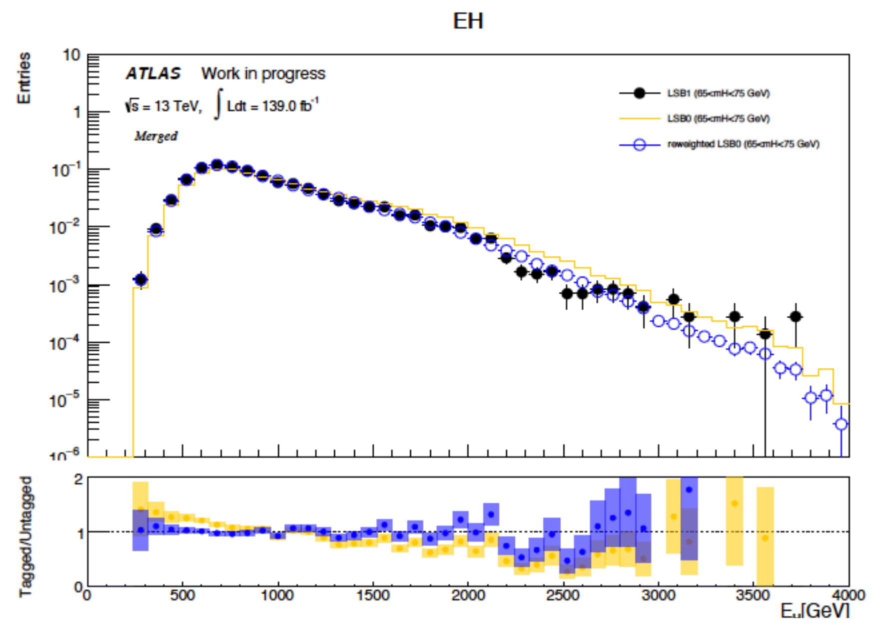

**Ratio legend:**
**Befor reweigthing**
**After reweighting**

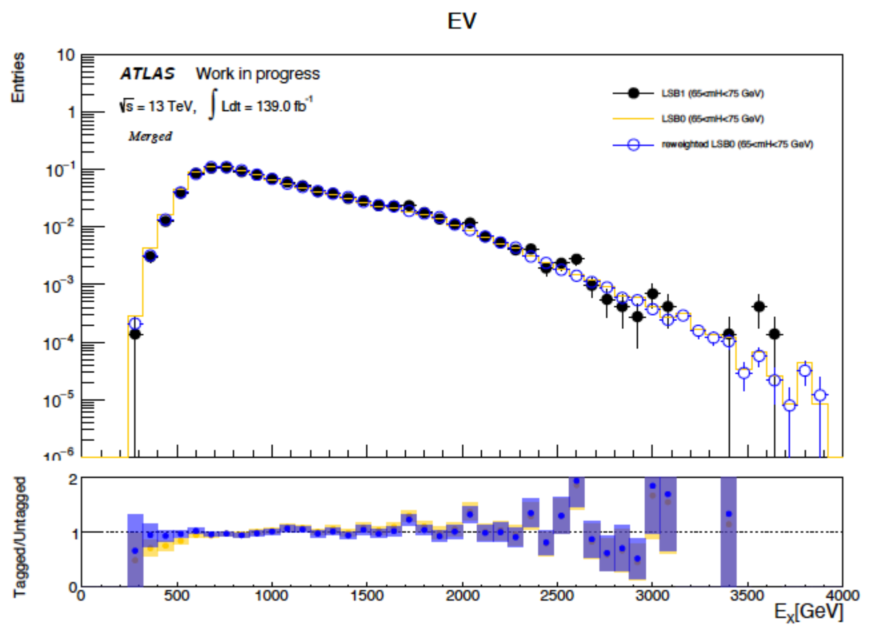
# Reweigthing in LSB (validation region)

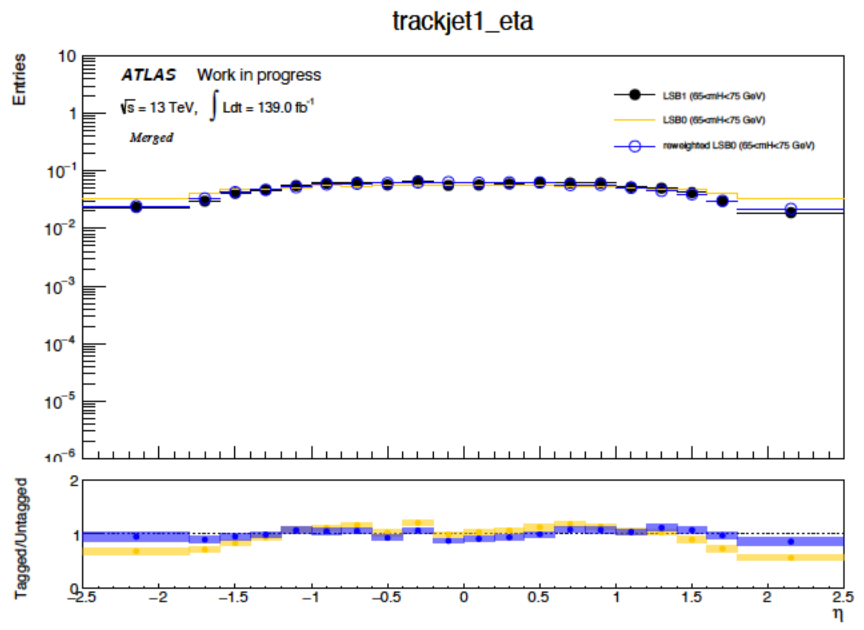
**Ratio legend:**
Befor reweigthing
After reweighting

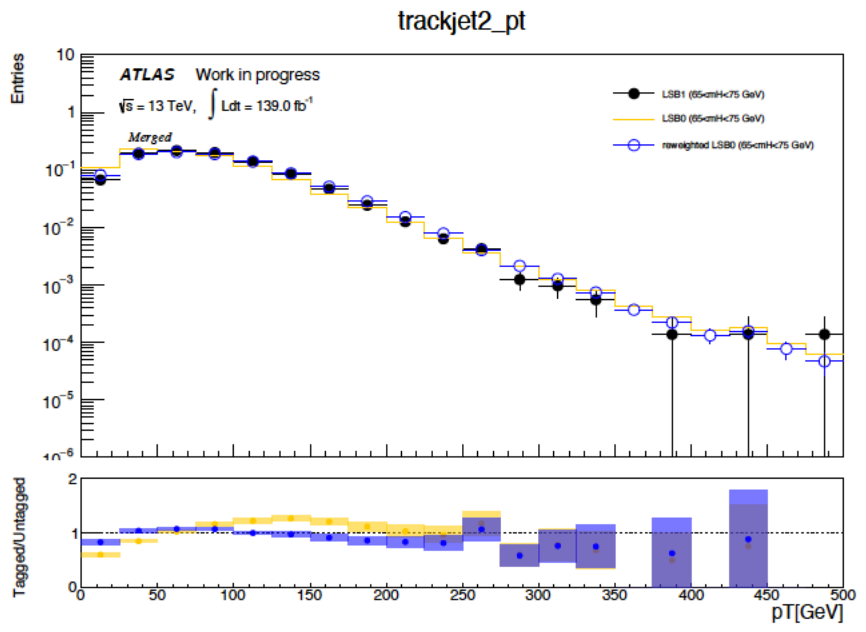
## Merged

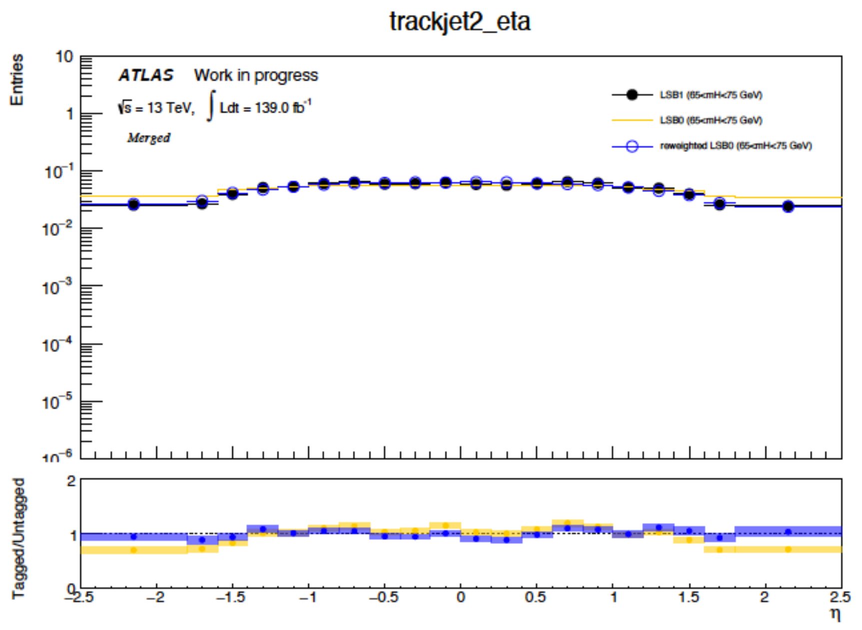


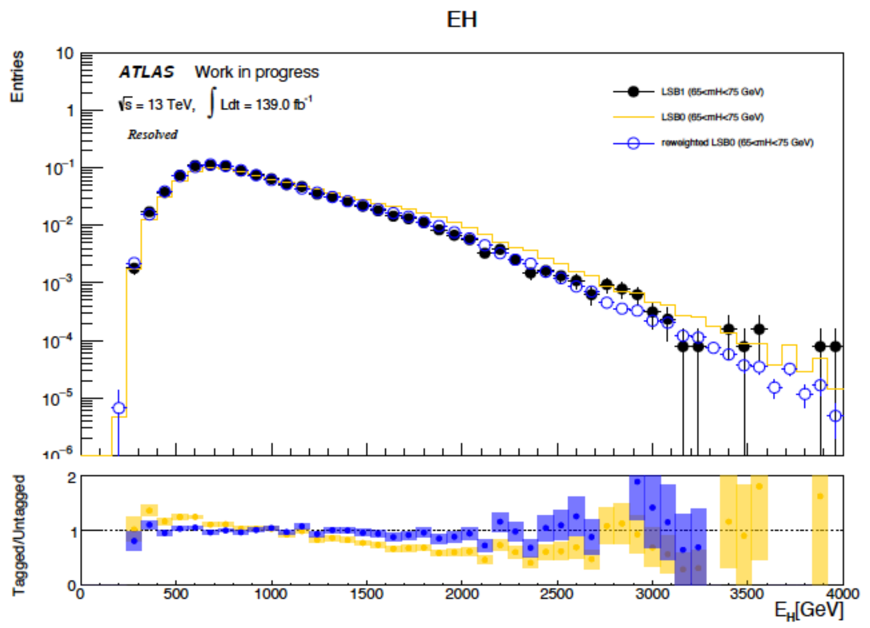



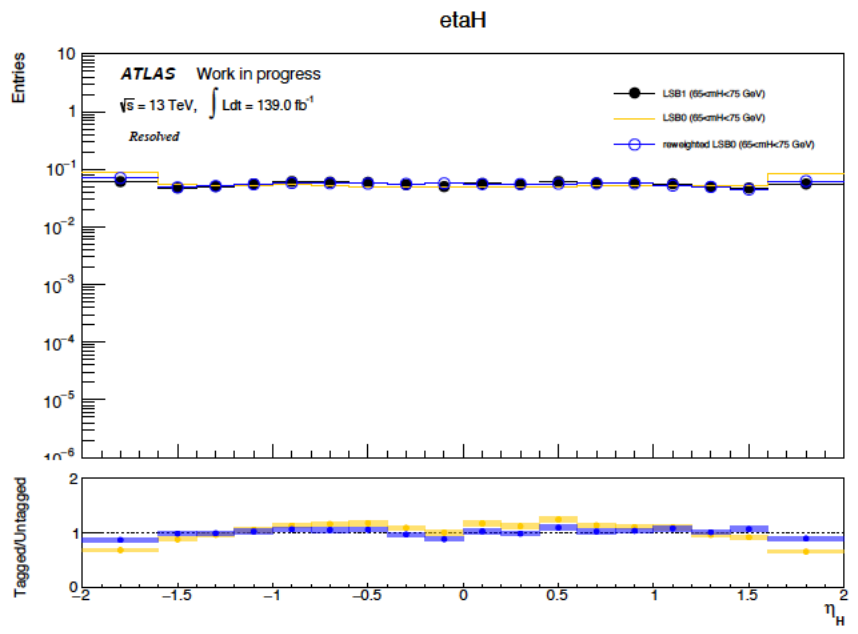



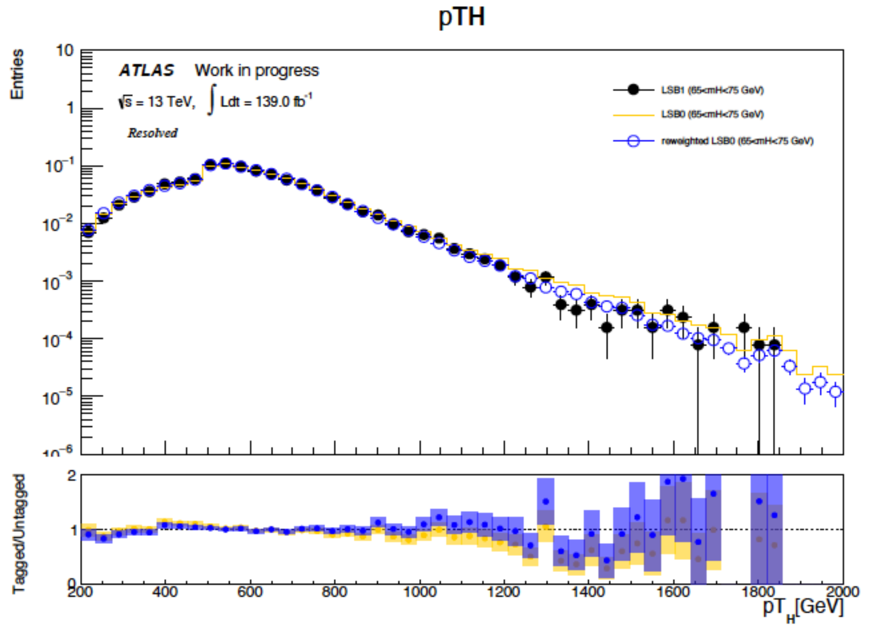



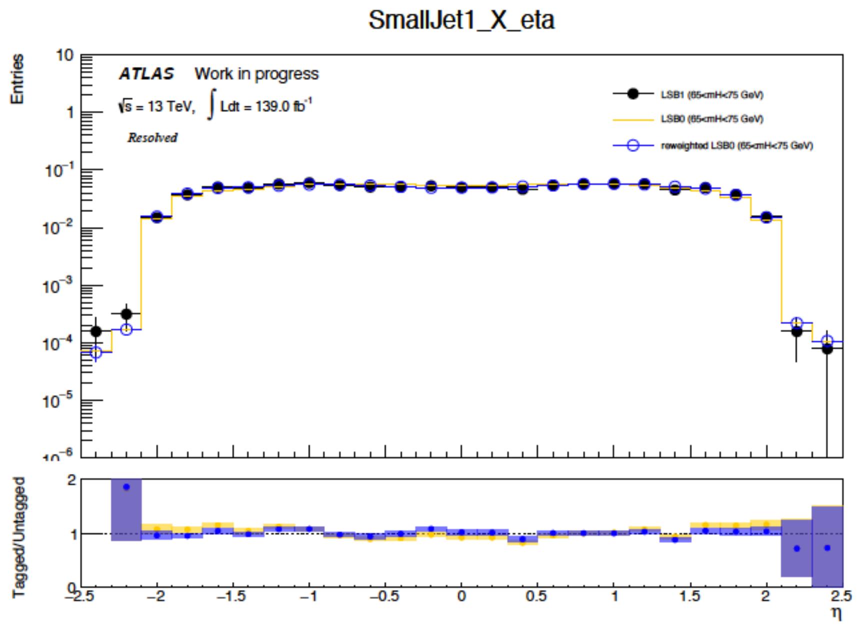



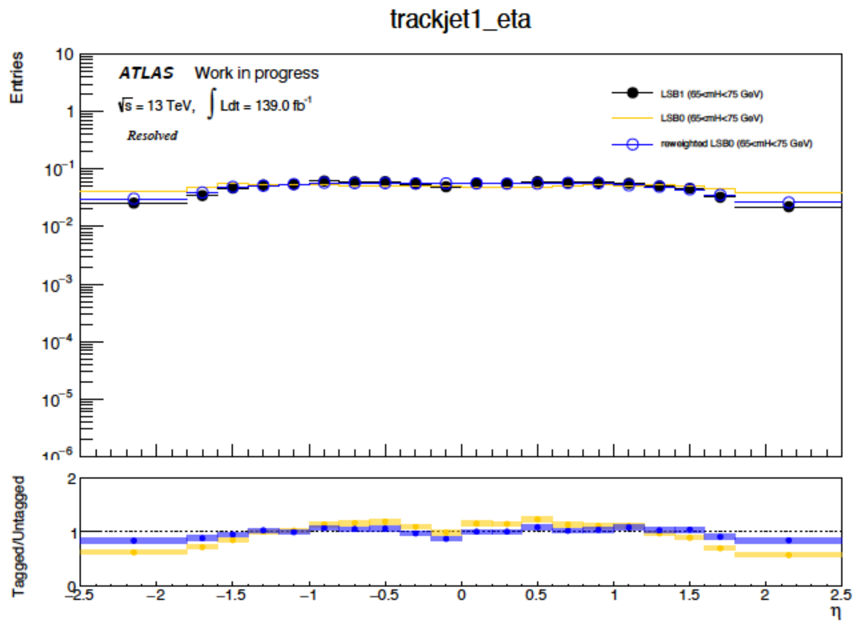

## Resolved

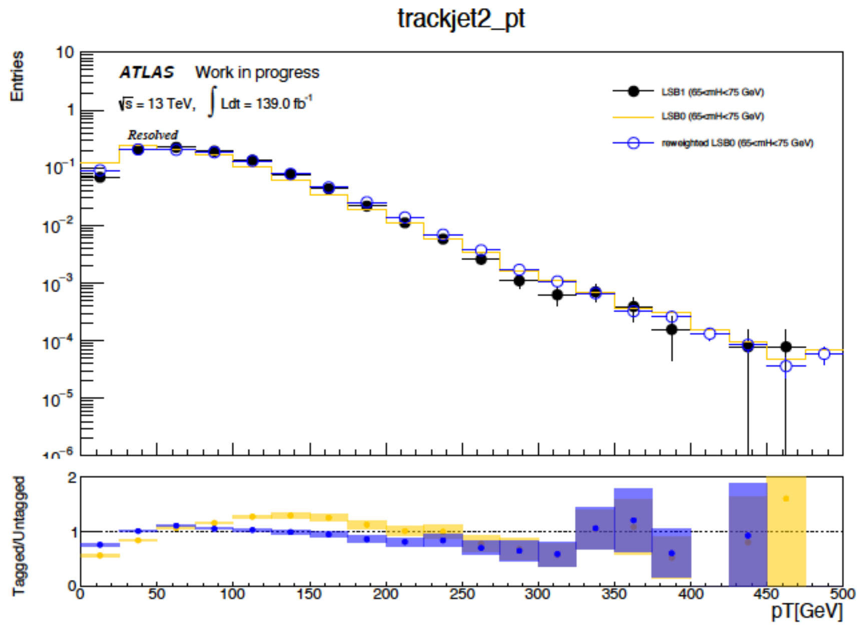


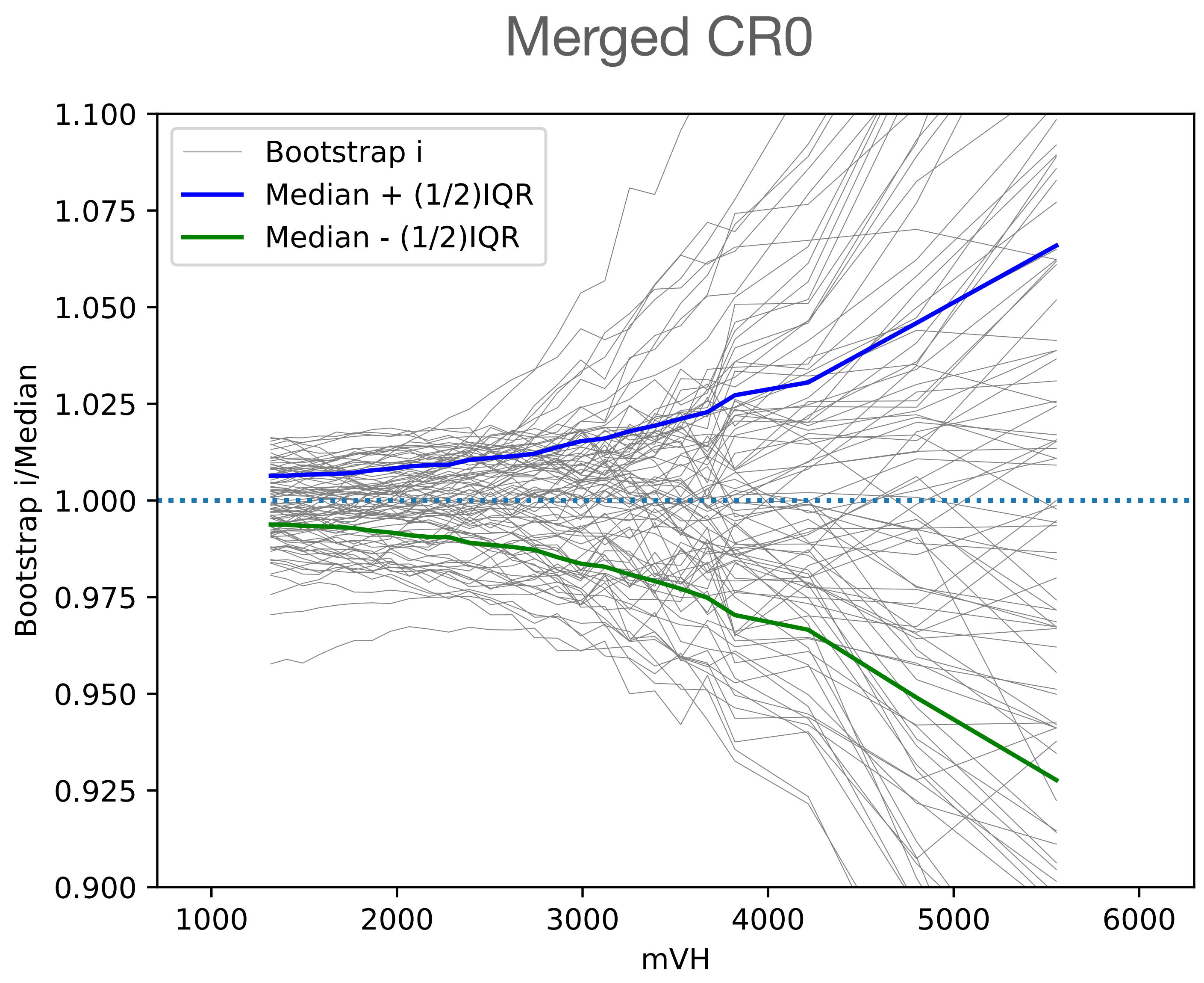
# Background Uncertainties

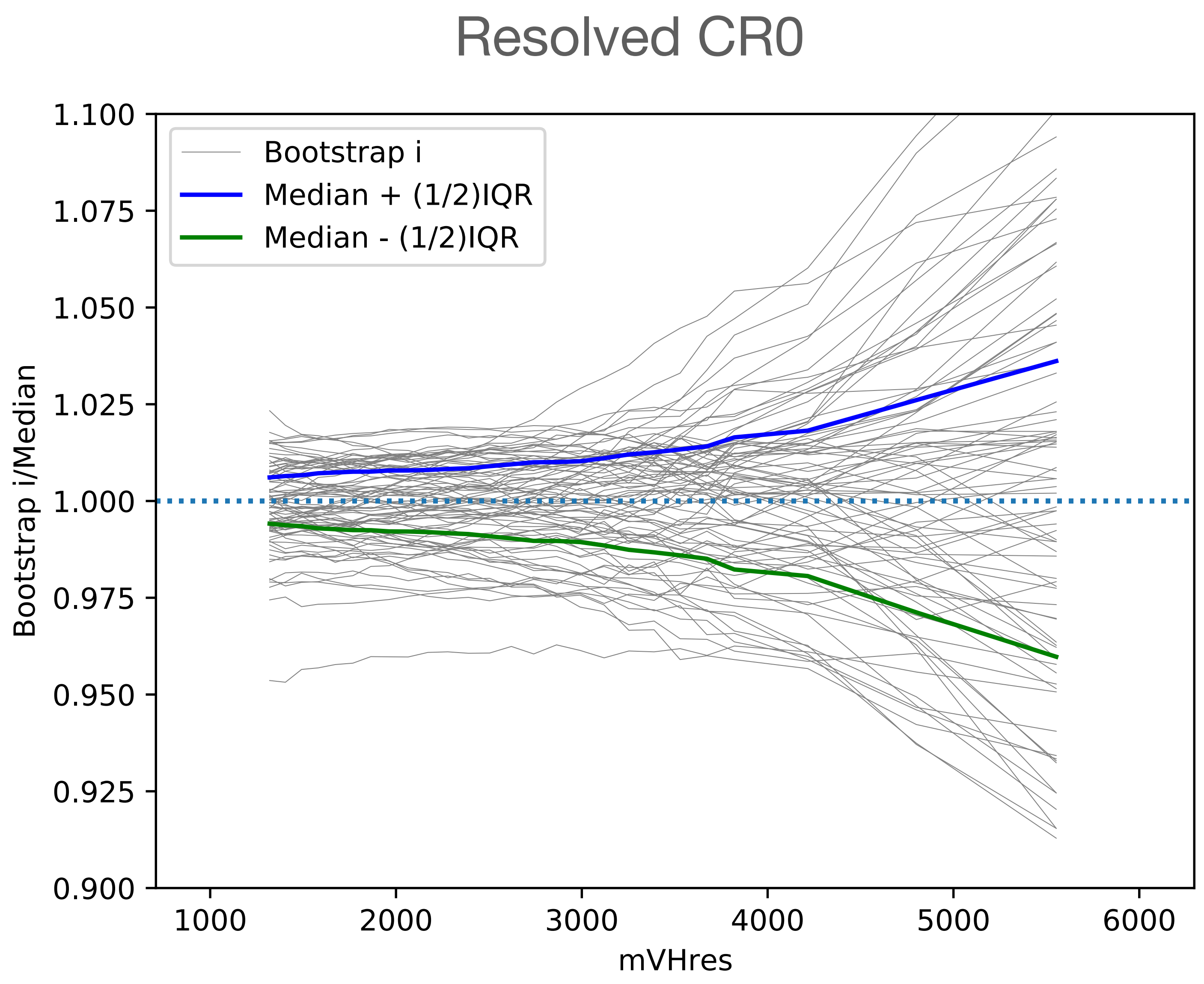
- No standard CP or MC-based systematics on background, since it is fully data-driven
- Three kinds of uncertainties considered:
  - Statistical, intrinsically related to the training procedure (<10%). Summed in quadrature with the Poissonian error in each bin
  - Systematic, on the choice of the training region (~5-10%)
  - Systematic, on the extrapolation of predictions across $m_H$ bins (~10%)
- All estimated inclusively in $m_X$, then applied on $m_Y$ shape in each $m_X$ windows
  - From further studies uncertainties in $m_X$ windows are in good agreement with those inclusively estimated
- Current strategy on correlations: all background uncertainties considered as shape variations, so bin-to-bin correlated.

*All background systematics already incorporated in the fit

# Bootstrap for Statistical Error

- The DNN is trained on a sample with a finite number of events and the weights of the network are randomly initialized at the beginning of the training
- Uncertainty estimated repeating the training N=100 times, randomly sampling the training dataset
  - Nominal histogram reweighed with the median of weights distribution for each event and normalized with the median of the normalization factors
  - Up/down variations obtained taking the median +- half the interquartile range (IQR) of weights distribution for each event and normalized with the median +-IQR/2 of the normalization factors



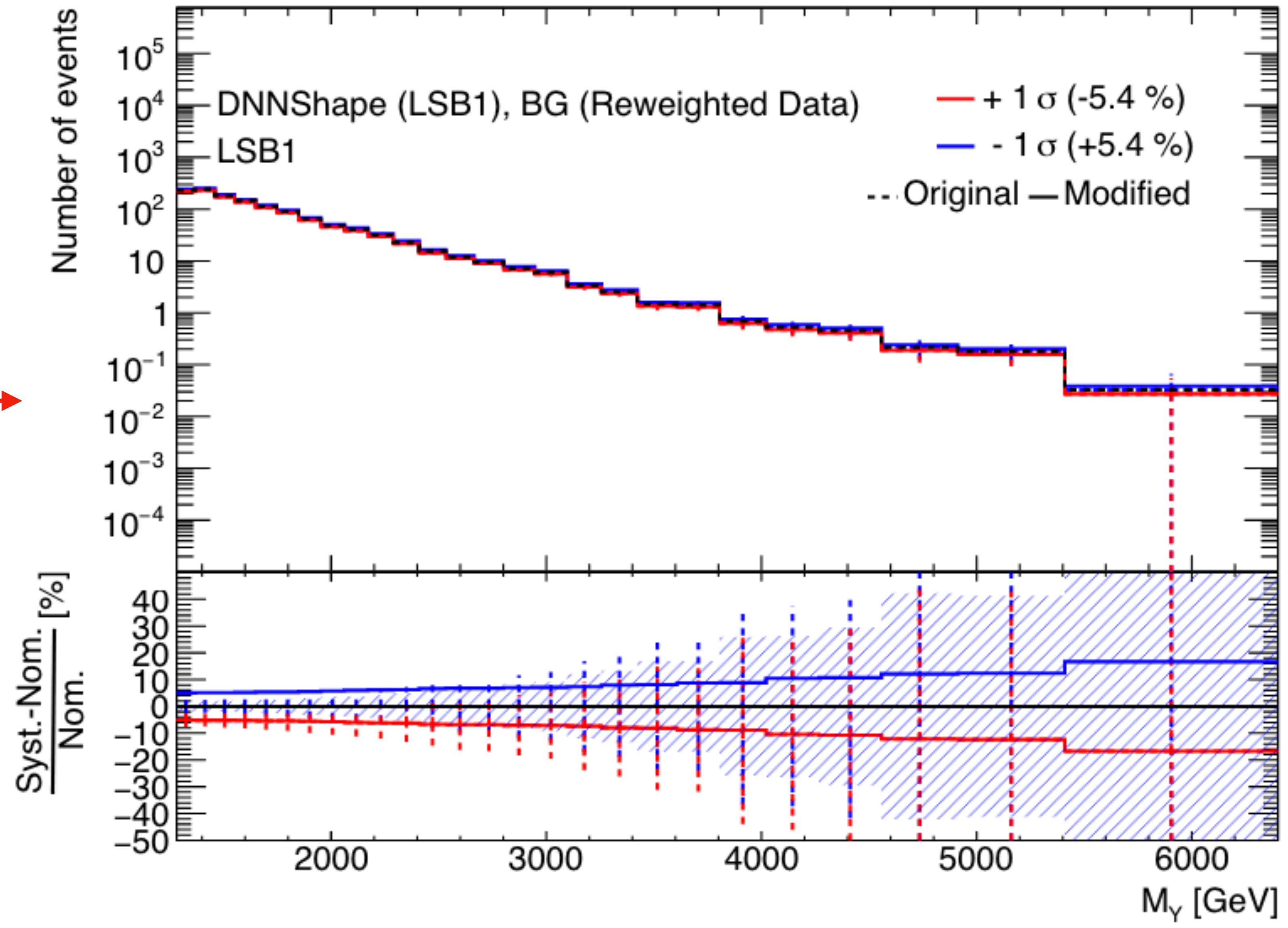



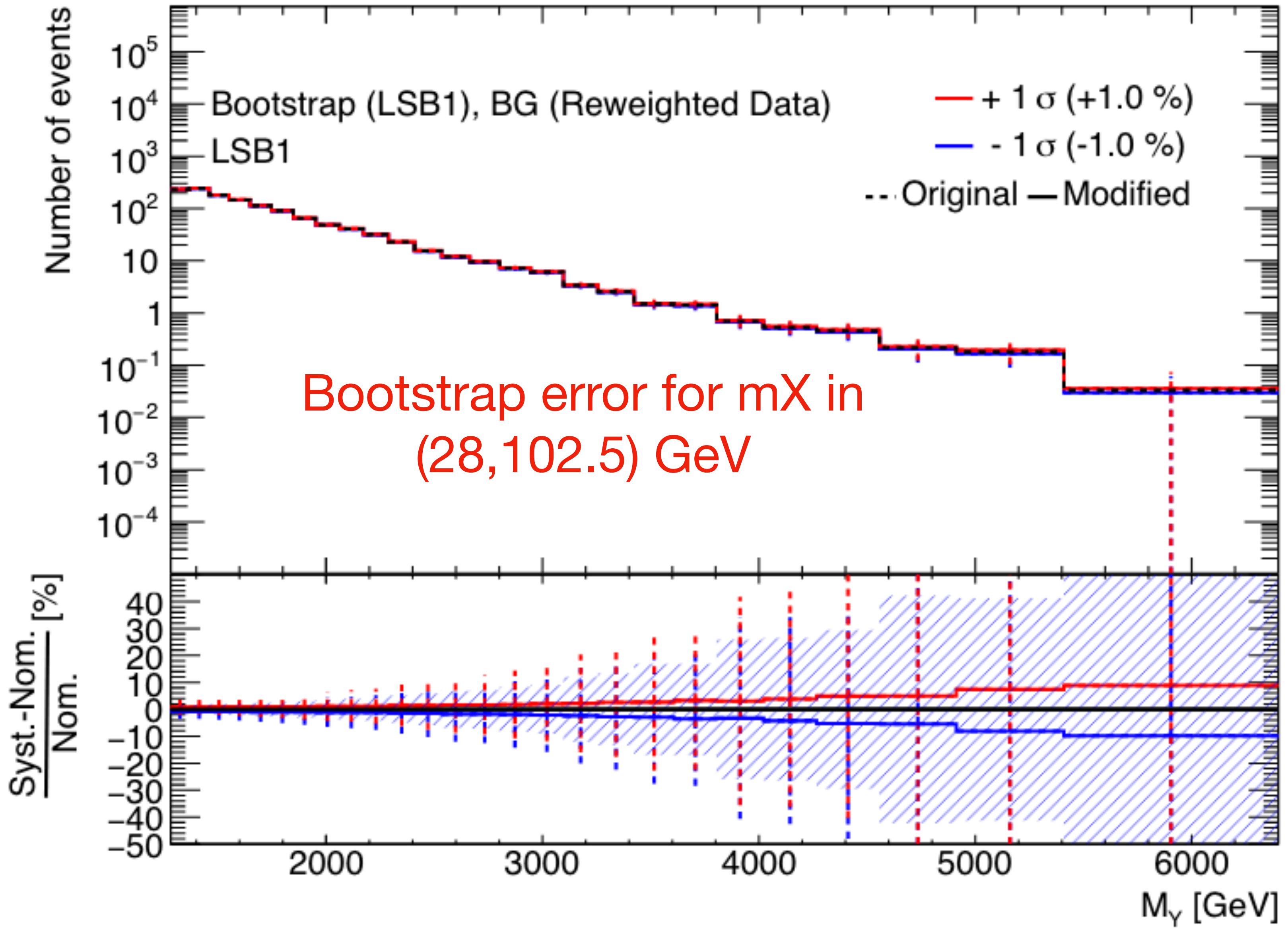

*more details in backup

# Shape Uncertainties - Training Region

- Predictions in the SR may be different if the region used for training changes
  - To quantify this mismodelling, an additional kinematic region ($m_H$ in [165, 200] GeV) is used to train an alternative model (totally identical to the nominal one, only changing the training region)
  - The ratio of the alternative shape to the nominal shape is determined as the NN modeling shape uncertainty

mX in (28,102.5) GeV

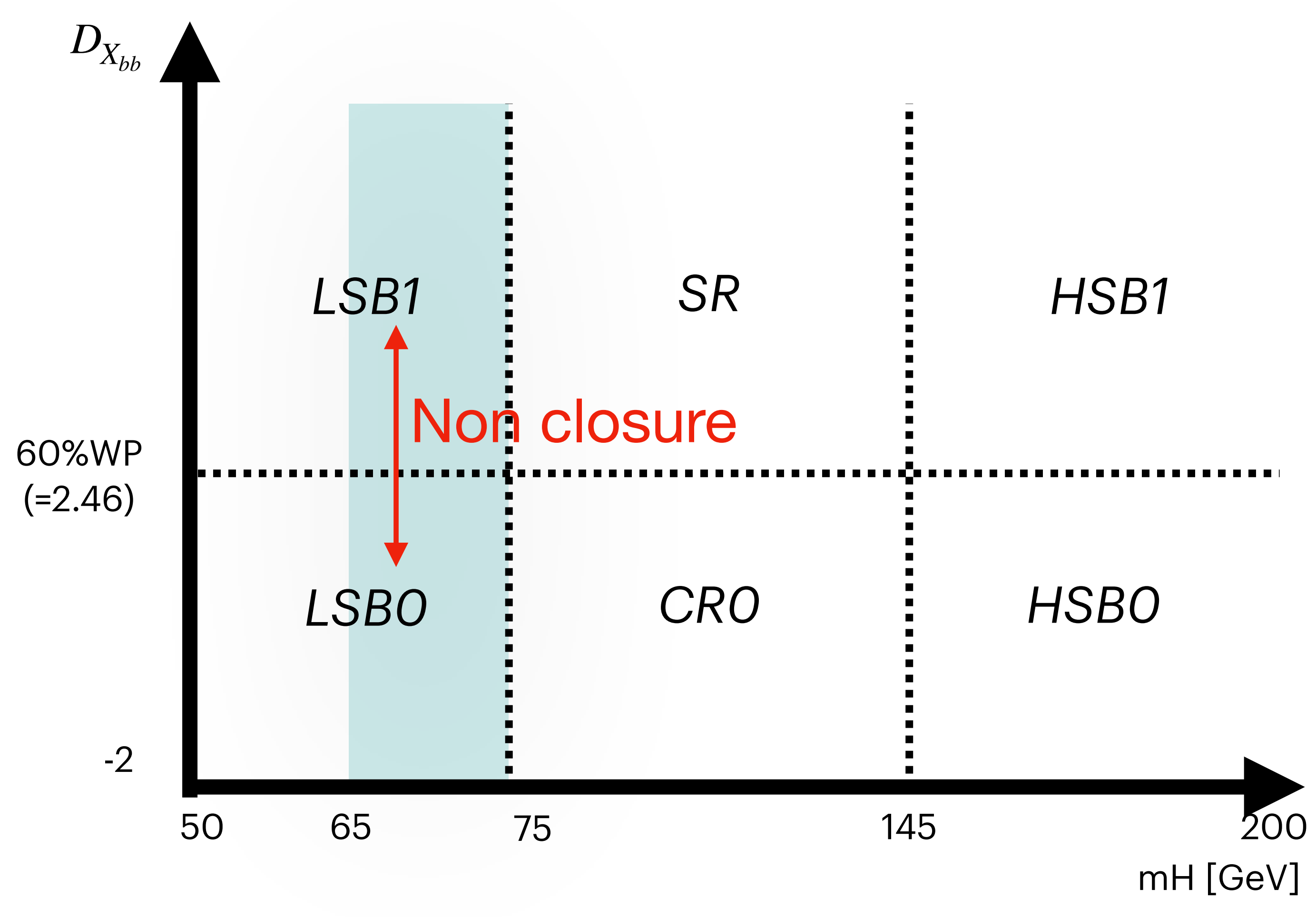


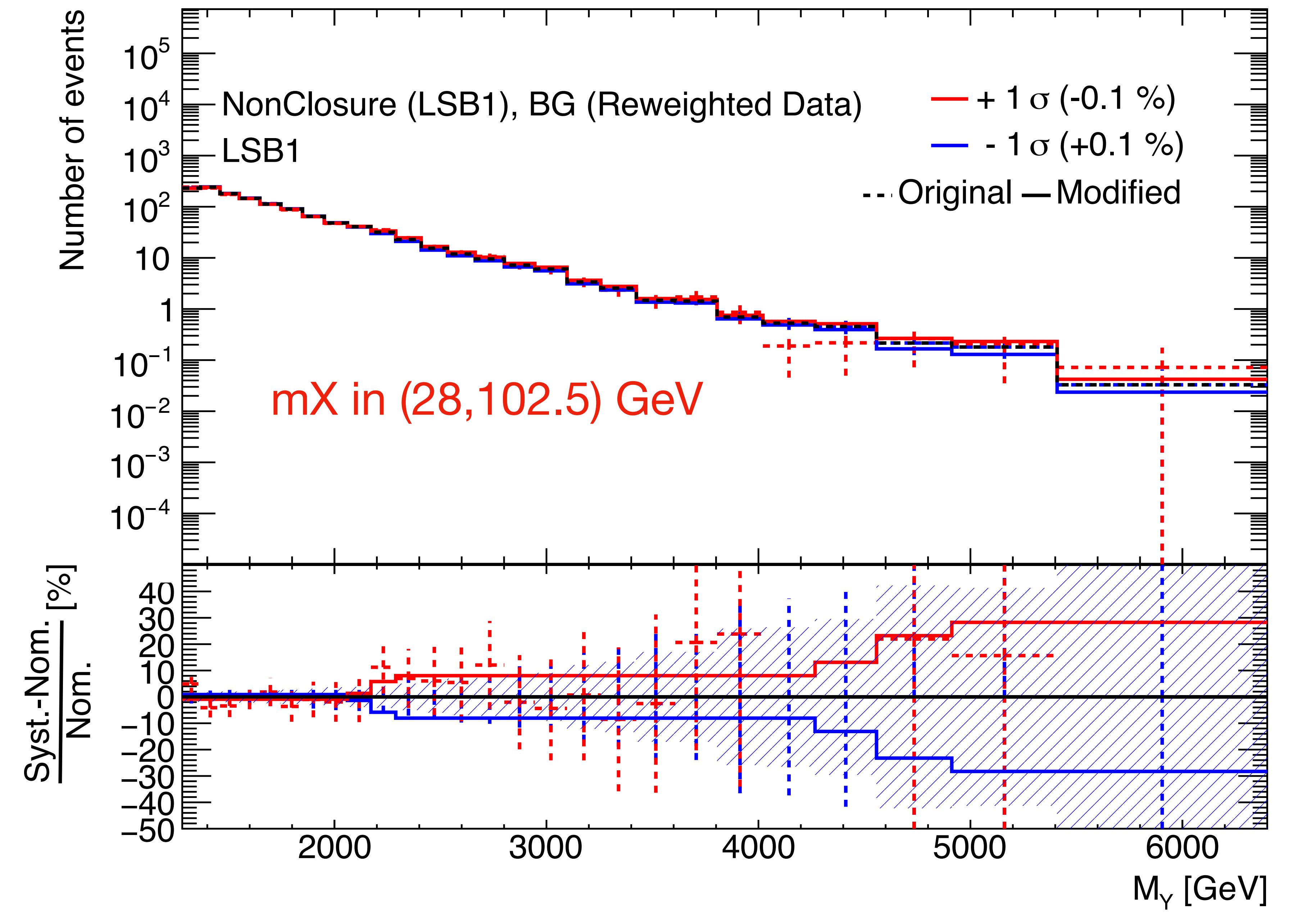




  - Error band smoothed using "*TTBARRESONANCE*" option in TRExFitter

# Shape Uncertainties - Non Closure

- Weights extrapolation process from the training region to the SR may be an additional source of mismodelling.
- Since it is not possible to directly estimate the discrepancy between reweighed data and the target distribution in SR, it is determined by looking at the ratio of data to estimated background in LSB (LSB1 over reweighed LSB0)
  - LSB0 reweighted histograms are scaled to match LSB1 yields




- Error band smoothed using "*TTBARRESONANCE*" option in TRExFitter