

# Performance of heavy-flavour jet identification algorithms in boosted topologies at the CMS experiment

**Matej Roguljić (Johns Hopkins University)  
on behalf of the CMS Collaboration**

**DPF-PHENO, Pittsburgh,  
15th May, 2024**

uncertainty of  $\pm 5\%$  on the shape of the  $Wb\bar{b}$  background in the  $H \rightarrow b\bar{b}$  signal region. In conclusion, the extraction of a signal from  $H \rightarrow b\bar{b}$  decays in the  $WH$  channel will be very difficult at the LHC, even under the most optimistic assumptions for the  $b$ -tagging performance and calibration of the shape and magnitude of the various background sources from the data itself.

[SNOWMASS-2001-P111](#)

It is widely considered that, for Higgs boson searches at the Large Hadron Collider,  $WH$  and  $ZH$  production where the Higgs boson decays to  $b\bar{b}$  are poor search channels due to large backgrounds.

We show that at high transverse momenta, employing state-of-the-art jet reconstruction and decomposition techniques, these processes can be recovered as promising search channels for the standard model Higgs boson around 120 GeV in mass.

[Phys.Rev.Lett. 100 \(2008\) 242001](#)

- Presenting results of [CMS-PAS-BTV-22-001](#)
- Calibrating algorithms (taggers) for  $X \rightarrow b\bar{b}(c\bar{c})$  jets used in Run 2 by CMS

- Lorentz-boosted regime has become widely used at the LHC
- Reconstruct merged decay products with a single large-area jet
- Heavy-flavour tagging is used to recognize origins of such jets

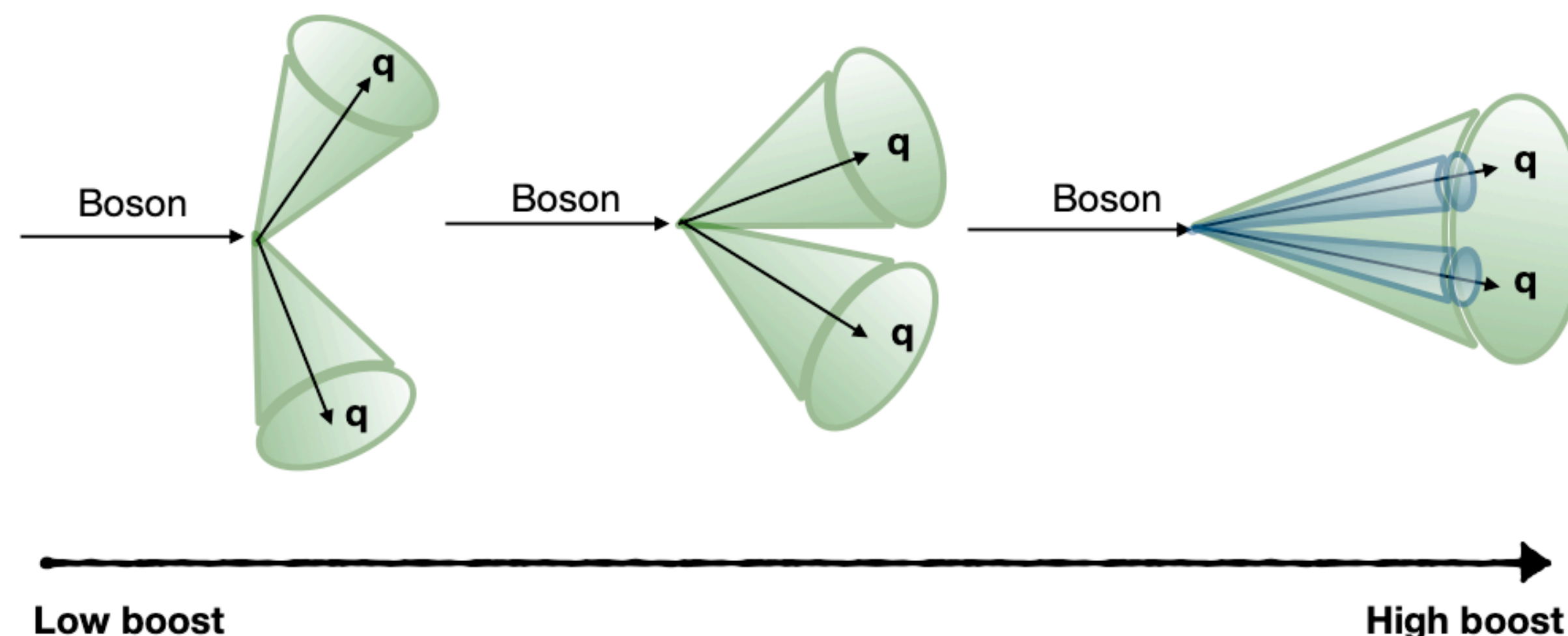


Figure from [CMS-PHO-EVENTS-2022-018](#)

# Heavy-flavour tagging

- Techniques used to distinguish jets originating from b/c quarks from other jets
  - Powerful tool for reducing background
- Distinguishing features such as the presence of secondary vertices due to relatively long b/c-hadron lifetimes
- Boosted jet tagging also exploits the substructure of the jet
- Provide jet properties and/or particle-content to ML algorithms to extract the most information out of the jet

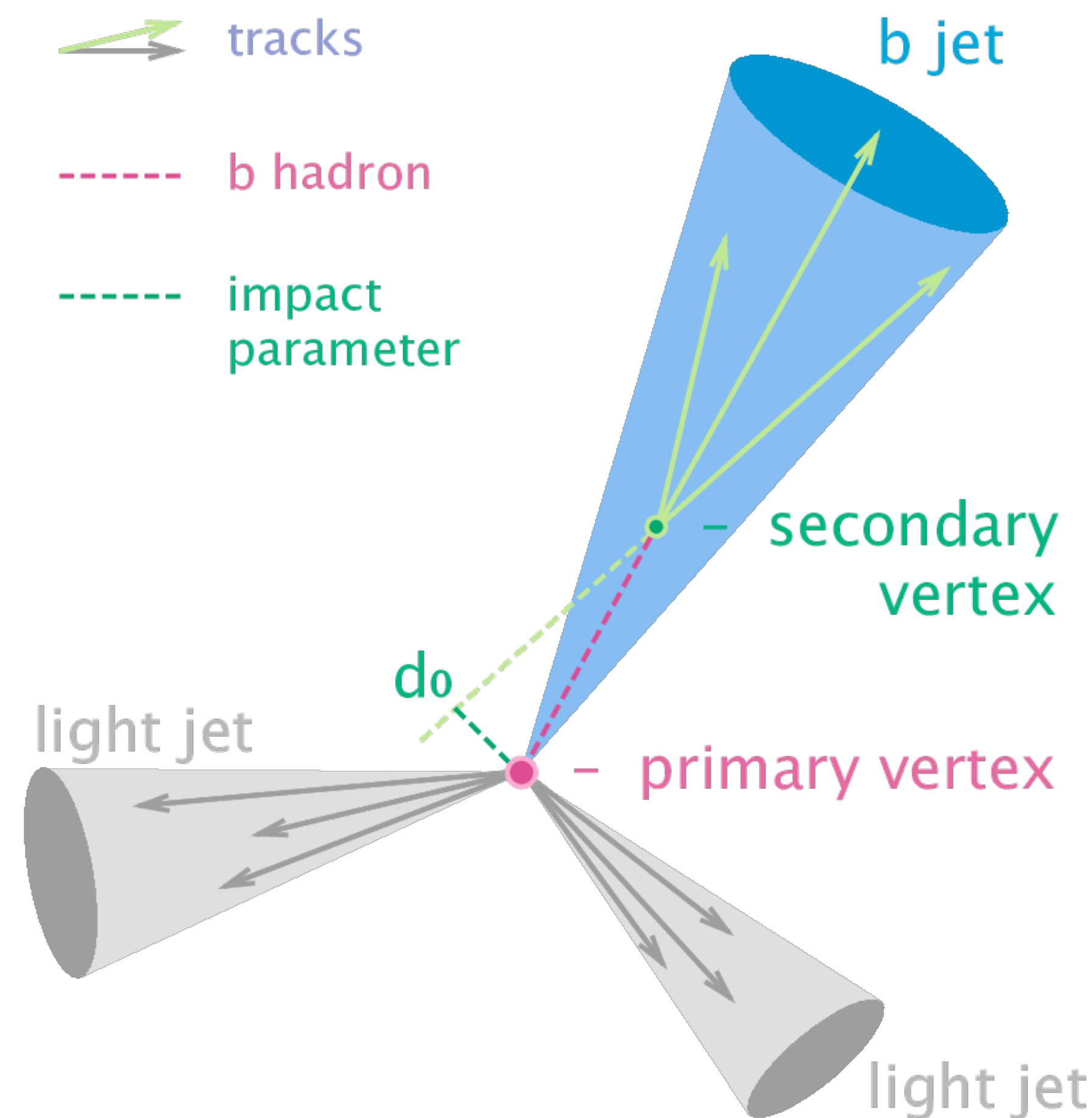
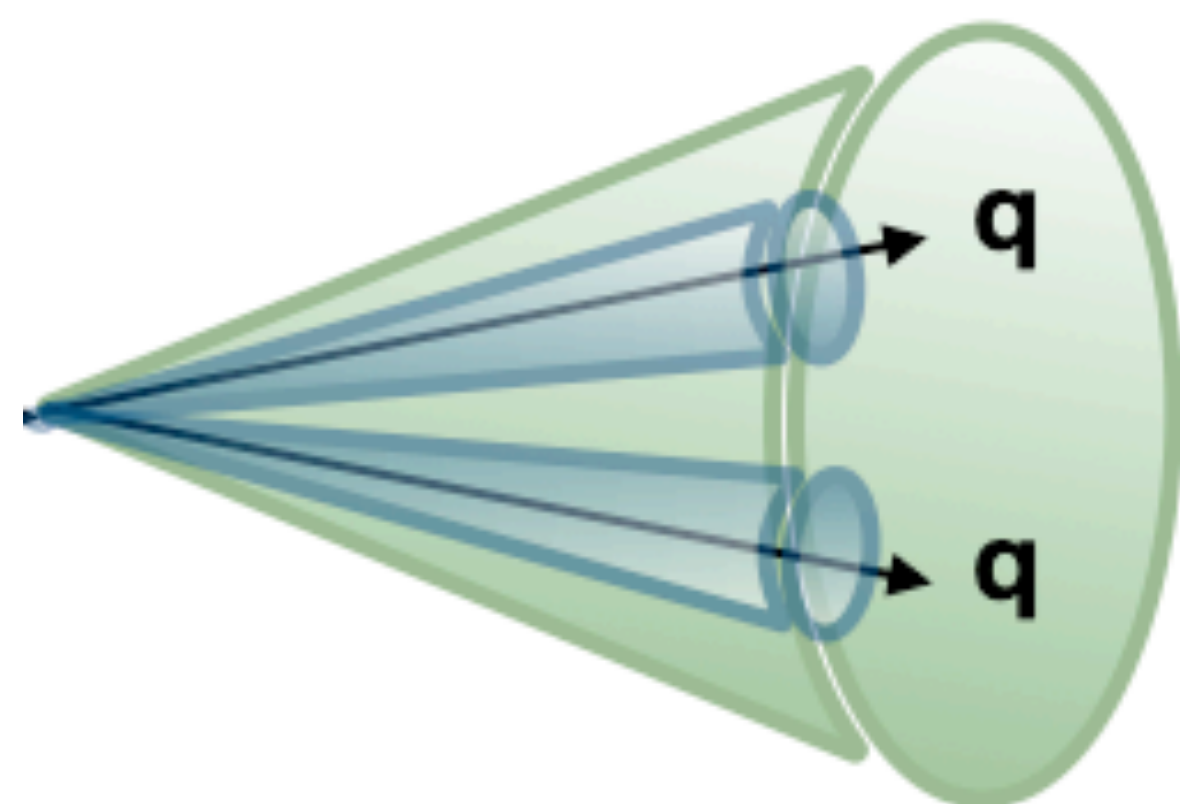
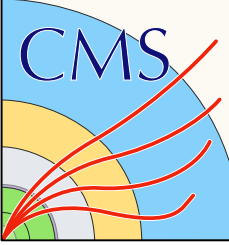
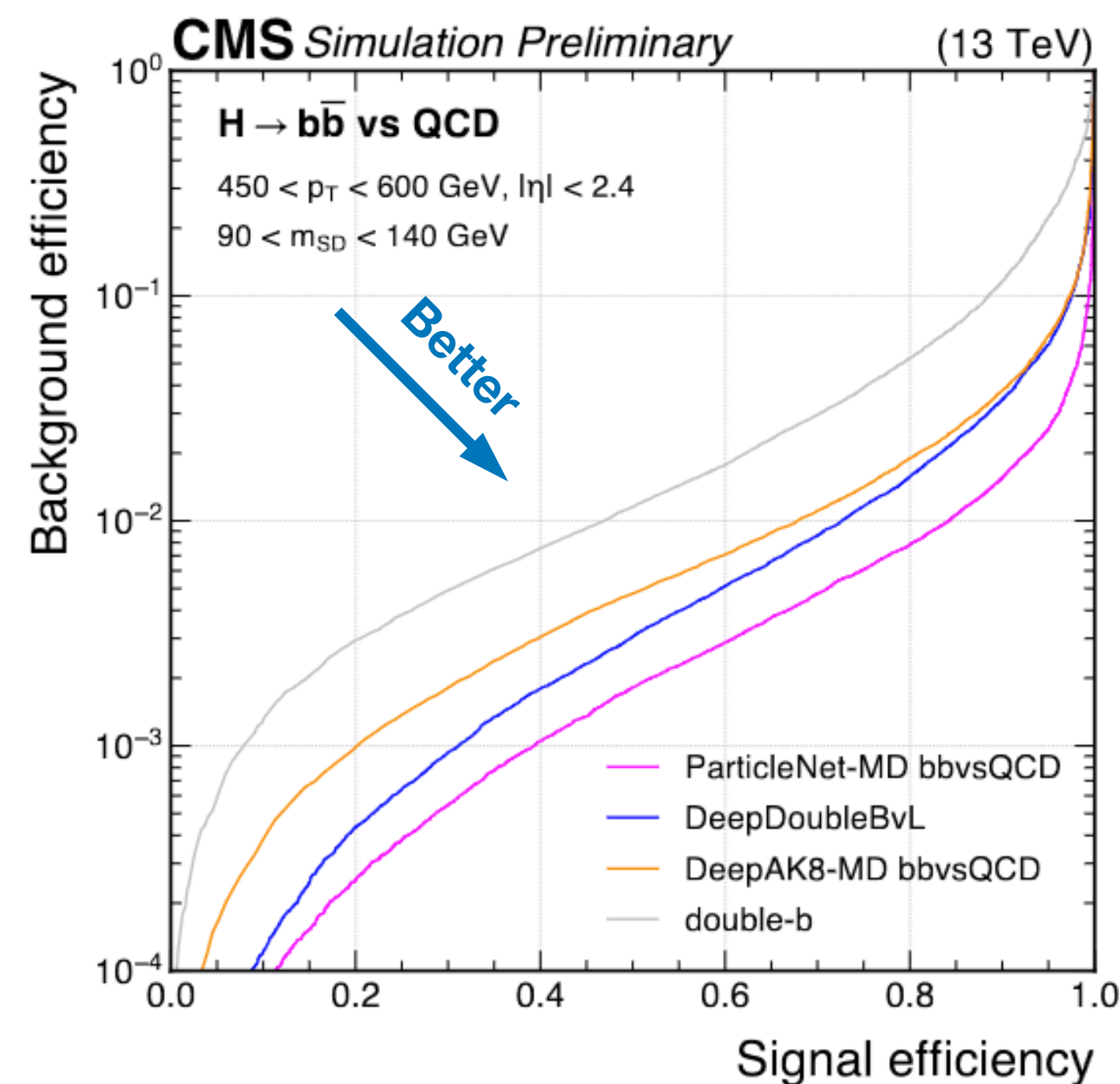
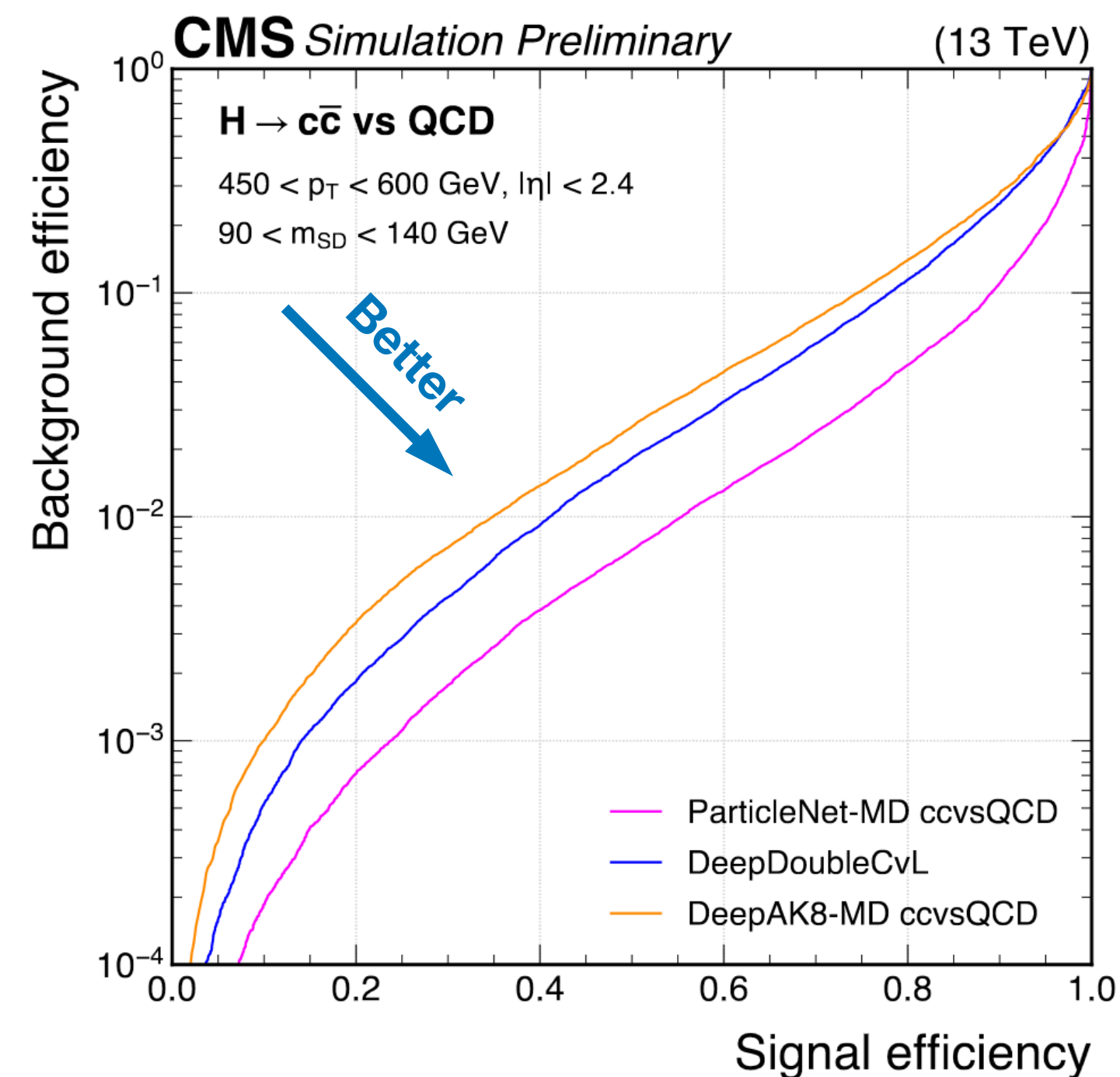


Figure source



# $X \rightarrow b\bar{b}(c\bar{c})$ taggers at CMS

- Huge progress in tagger performance from start to end of Run 2
  - Double-b tagger [JINST 13 \(2018\) P05011](#)
    - 2016  $H \rightarrow b\bar{b}$  ([PRL 120.071802](#))
  - DeepAK8-MD [JINST 15 \(2020\) P06005](#)
    - $HH \rightarrow b\bar{b}\ell^+\ell^-$  ([JHEP05 \(2022\) 005](#))
  - DeepDoubleX [CMS-DP-2022-041](#)
    - Run 2  $H \rightarrow b\bar{b}$  ([CMS-PAS-HIG-21-020](#))
  - ParticleNet-MD [PRD 101.056019](#), [CMS-DP-2020-002](#)
    - $X \rightarrow HY \rightarrow 4b$  ([PLB \(2022\) 137392](#))



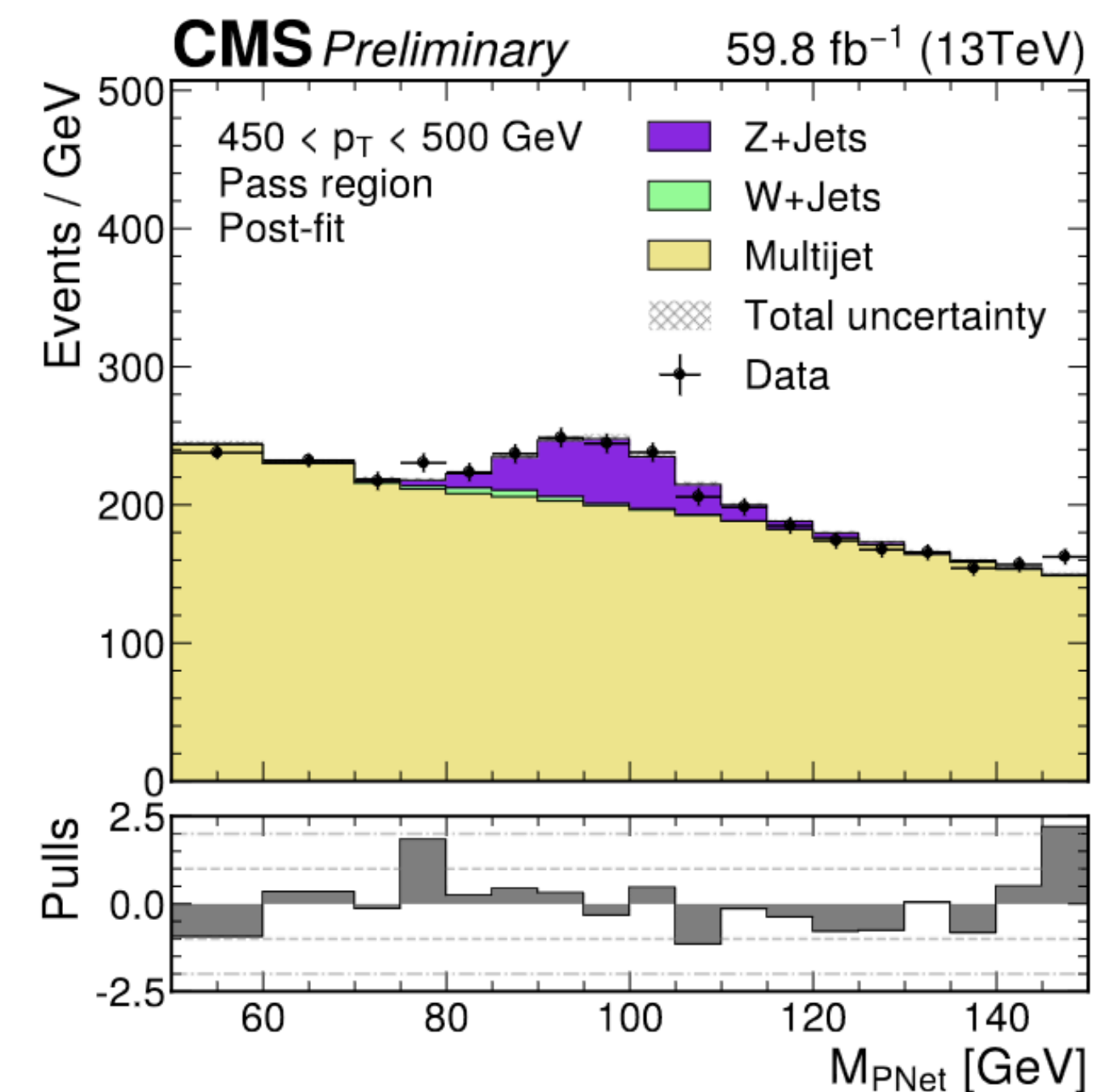
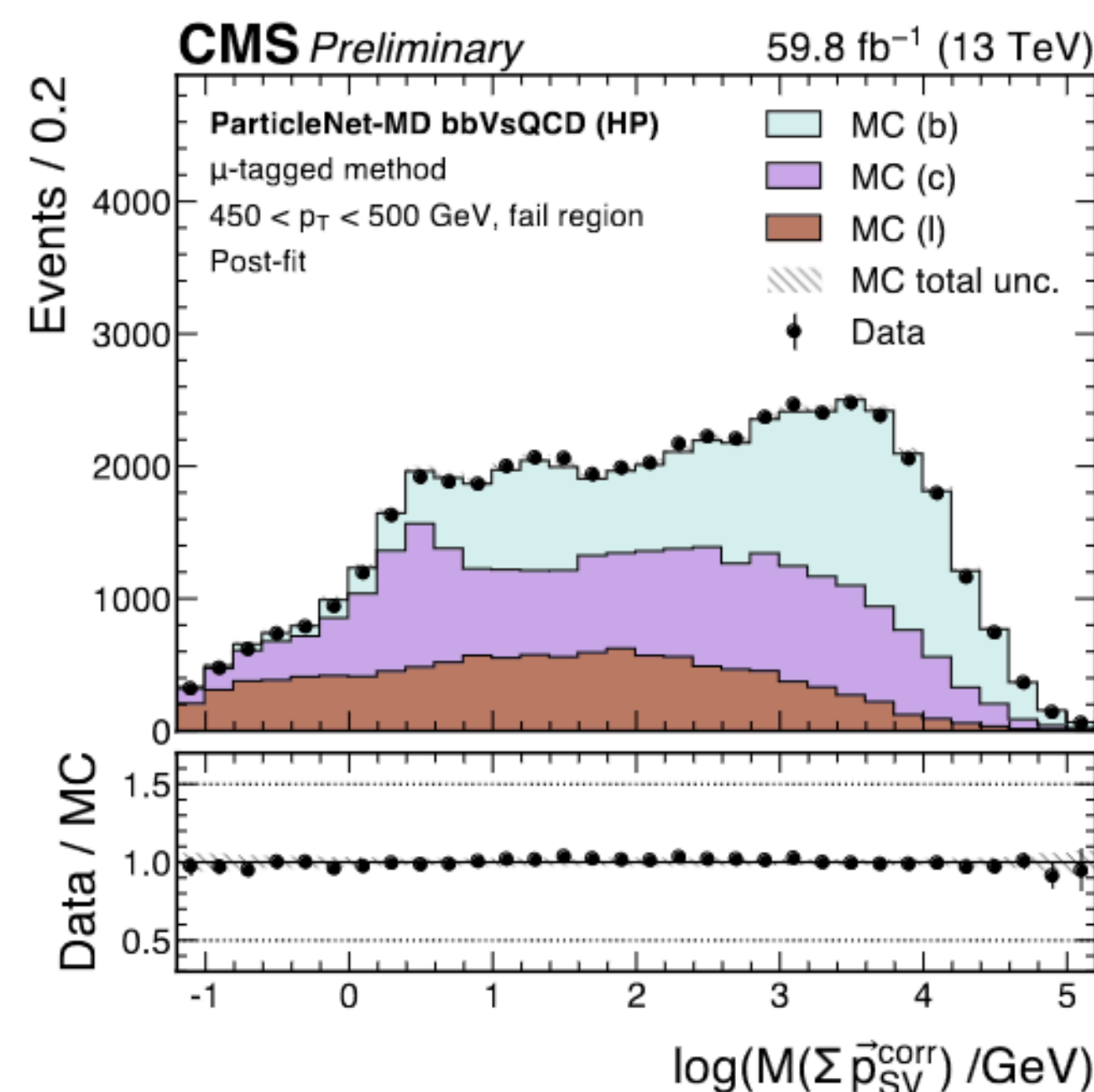
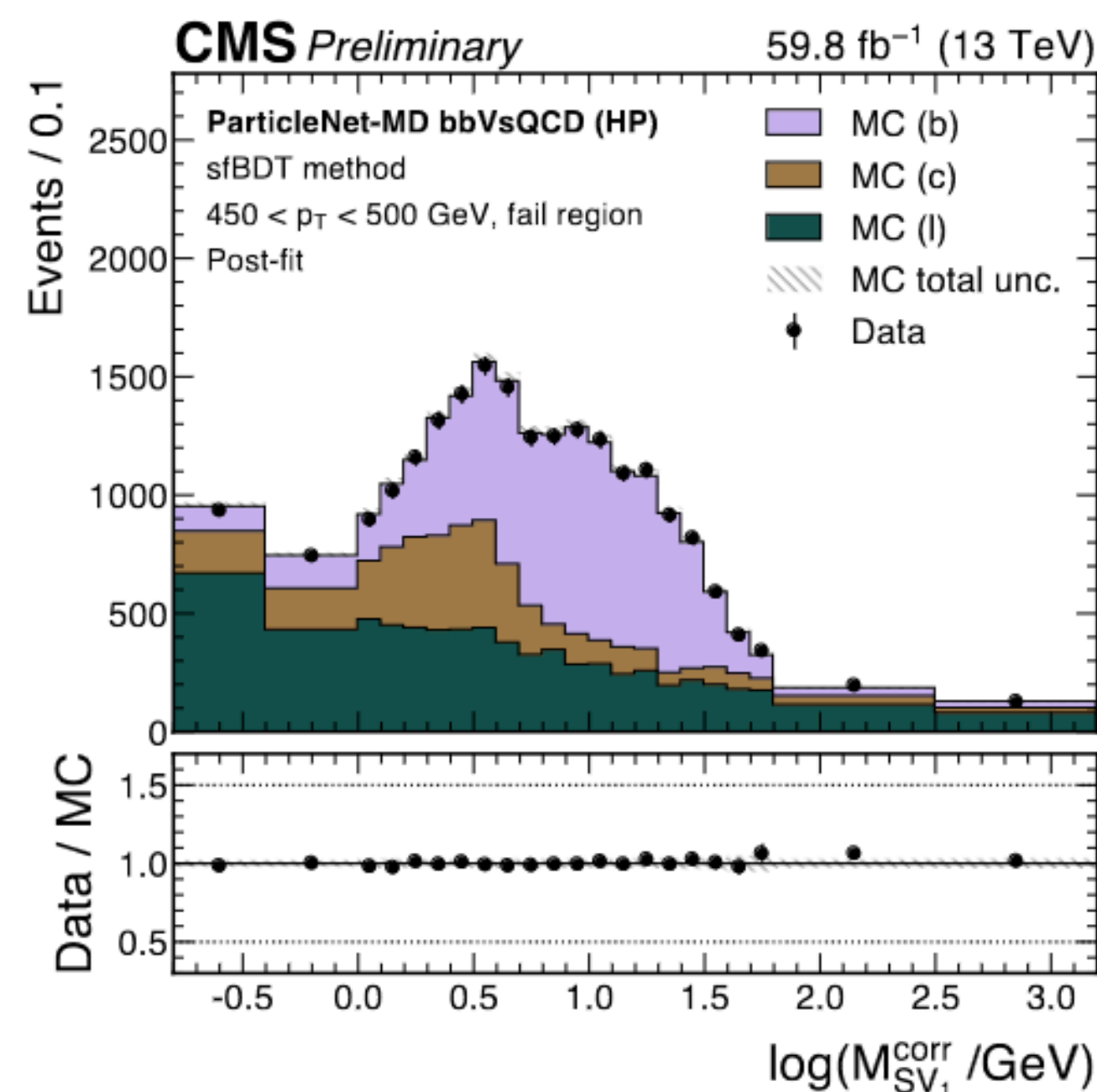
These and rest of the figures are from [CMS-PAS-BTV-22-001](#)

# Calibrating Xbb tagging algorithms

- How can we know the performance of our tagger on signal,  $H \rightarrow b\bar{b}$ , jets?
  - Ideally, we would isolate a pure sample of  $H \rightarrow b\bar{b}$  jets and measure it, but it is not possible
  - We select an ensemble of signal-like (proxy) jets and perform measurements on them
  - Assume similar performance between signal and proxy jets
- [CMS-PAS-BTV-22-001](#) presents three calibration methods, differing in the selection of proxy jets

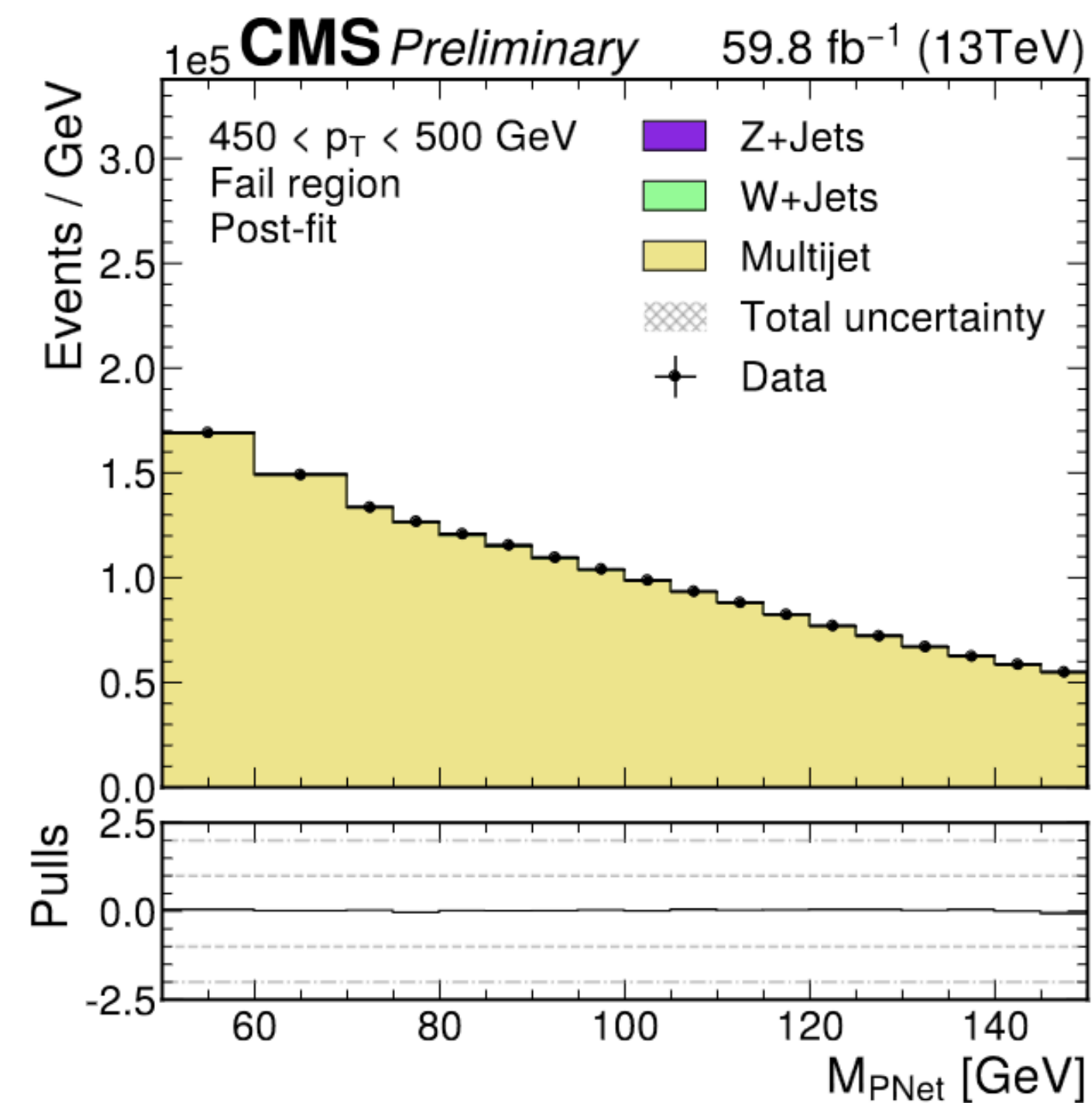
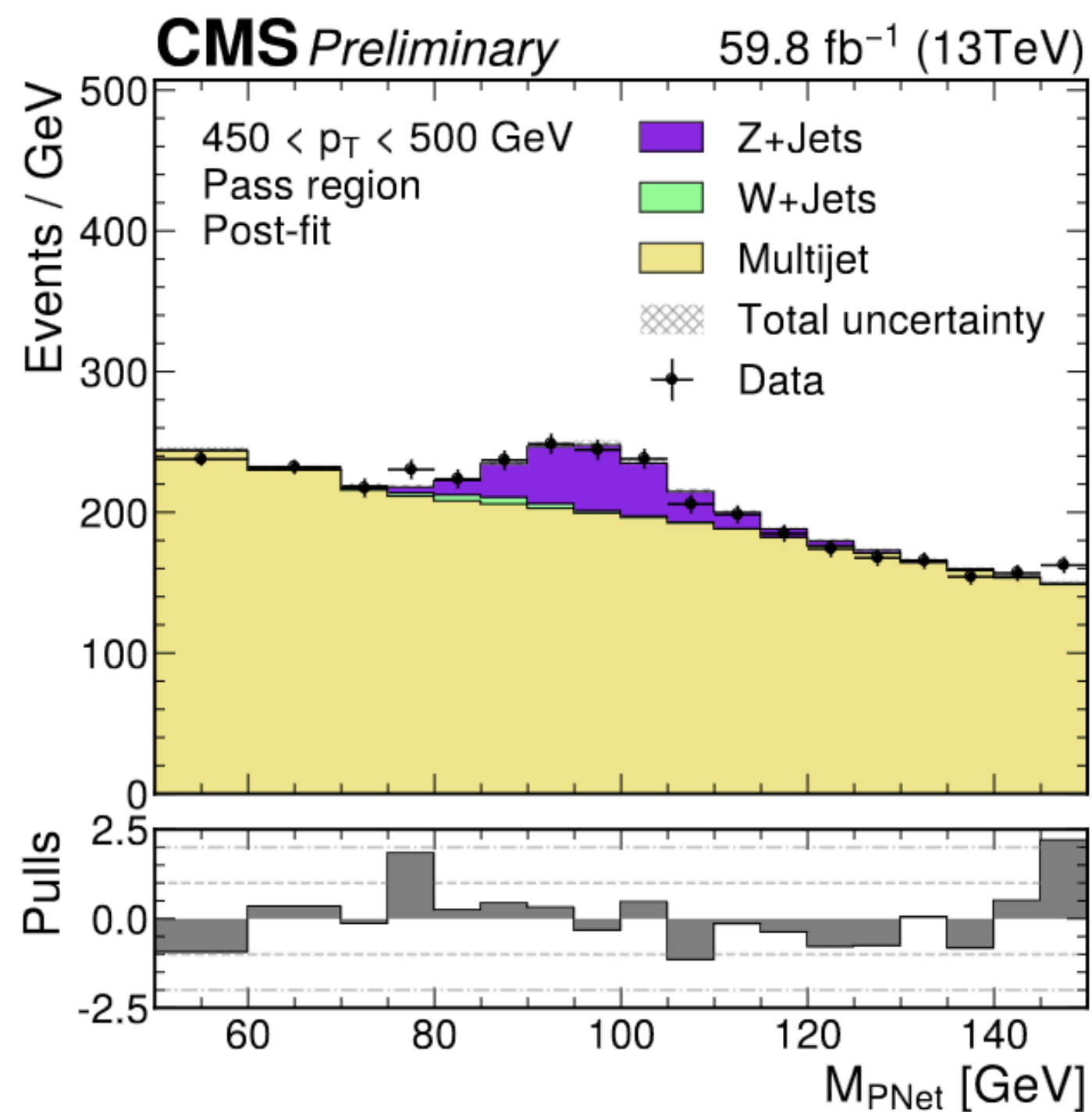
This is what we want to measure in data

$$\frac{N_{jet}^{Pass}}{N_{jet}^{Fail} + N_{jet}^{Pass}}$$

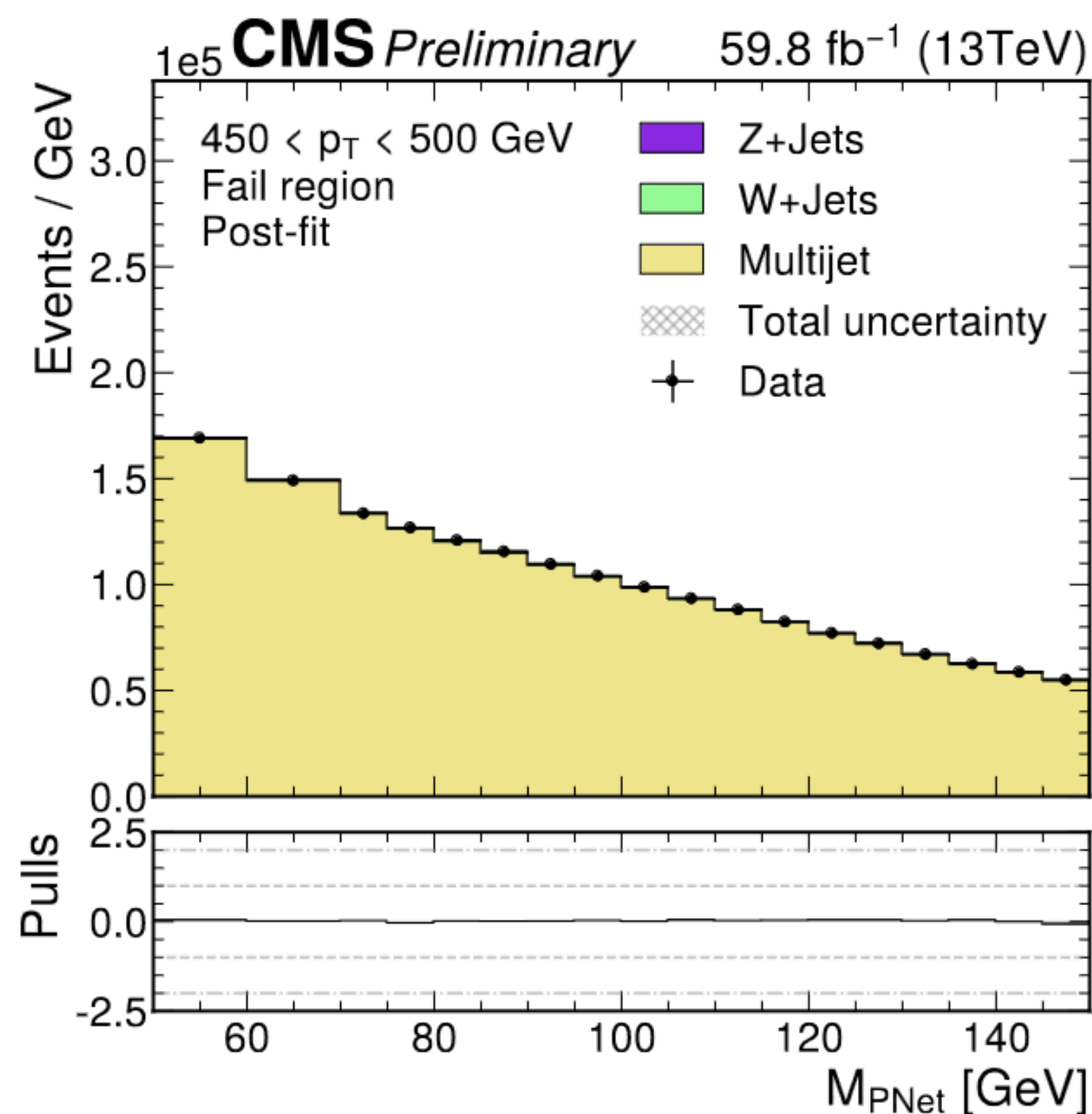


# Method 1: Z boson proxy

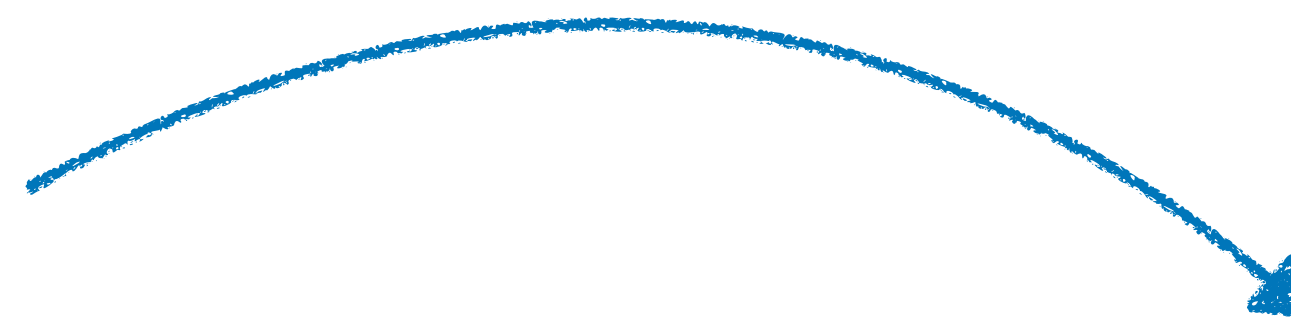
- Uses jets originating from  $Z \rightarrow b\bar{b}$  jets as an excellent proxy
  - The only difference between the proxy and the signal is the mass: 90 vs 125 GeV
  - Taggers are intentionally trained not to have a dependency on the mass
    - Keeps the shape of the mass distribution between regions which fail and pass the taggers the same
- Downside: very large multijet background makes the measurement challenging



# Method 1: Background estimate

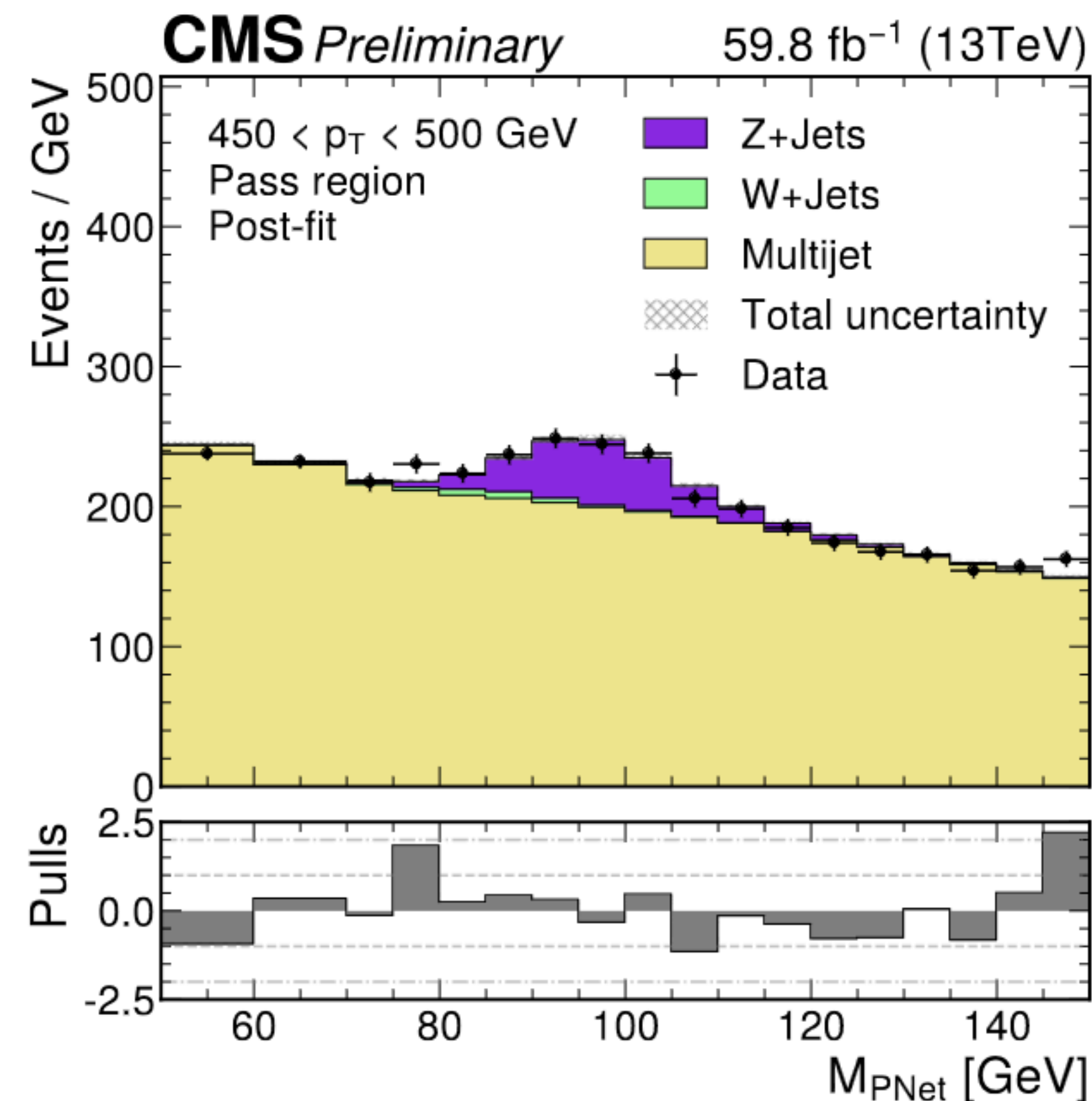


Assume that the shape of the bkg. in pass is similar to that in fail



Multiply shape in fail by  $R_{P/F}(M)$  to allow some small, smooth shape change

$R_{P/F}(M)$  is a polynomial, parameters are determined during the fit

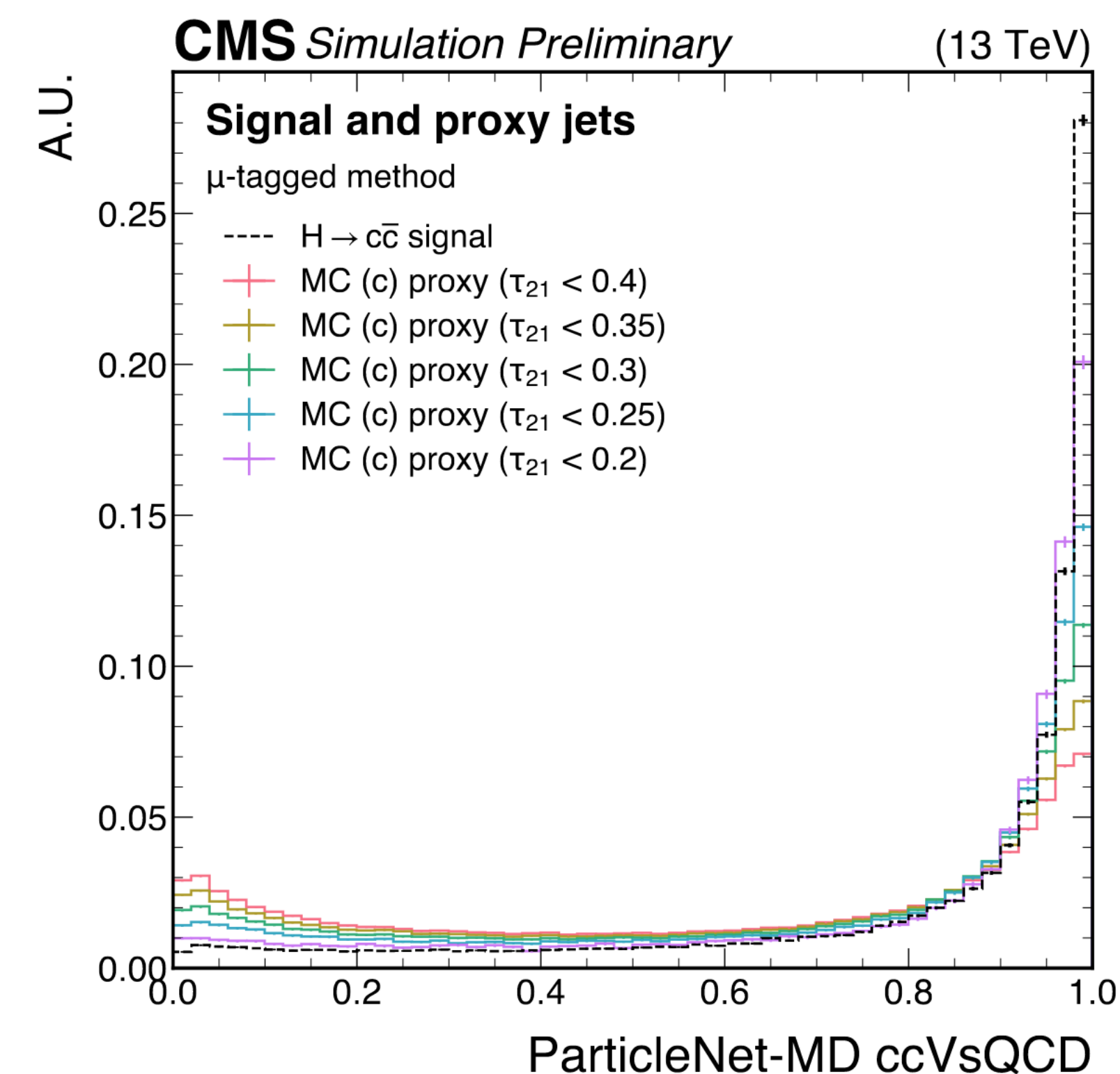
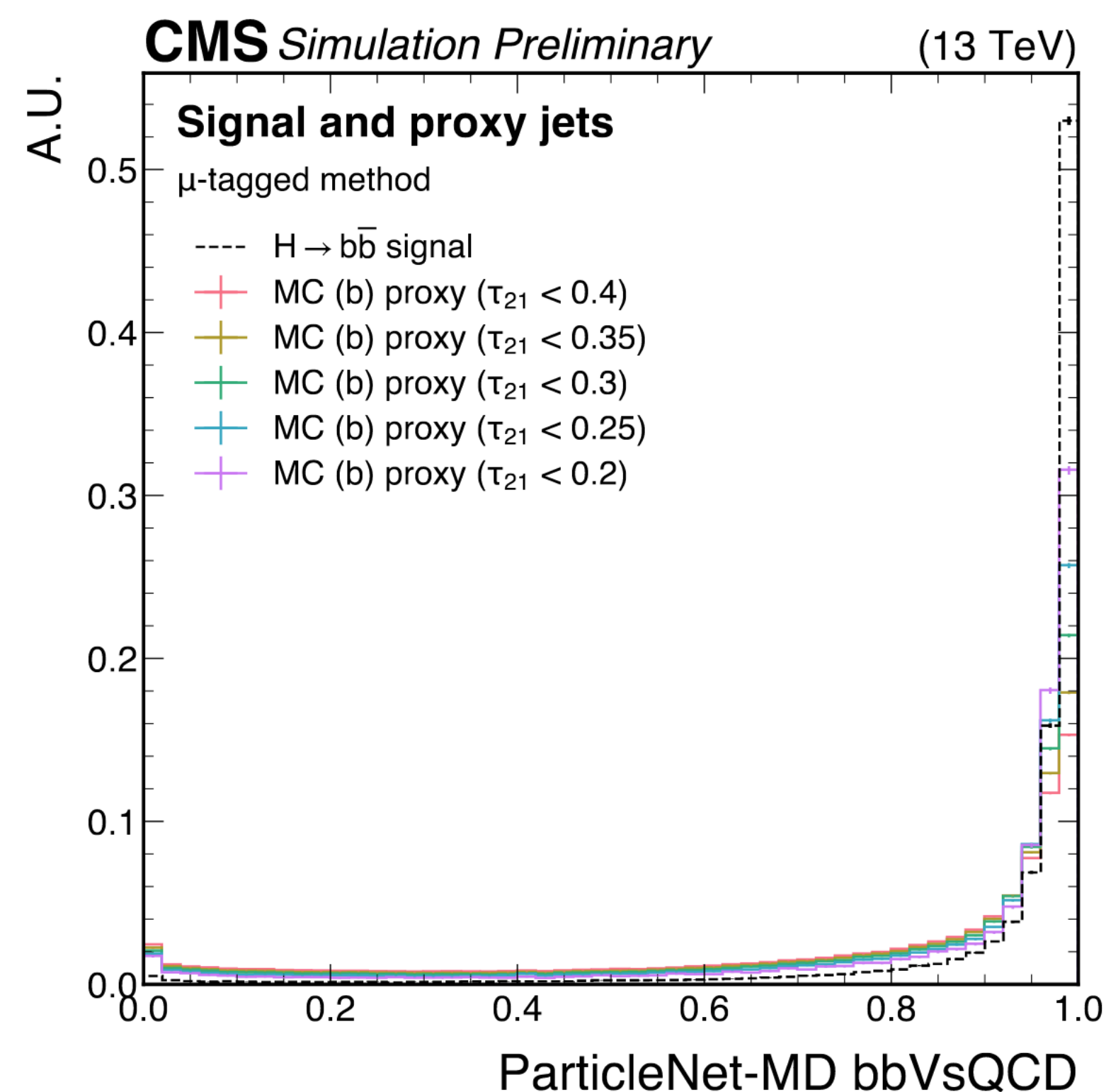


Fail region is dominated by multijet bkg.  
 Whatever we measure in data gives us the shape of this bkg. (modulo small contribution from  $V + jets$ )

# Method 2: Muon jets proxy

- Uses jets originating from gluon splitting,  $g \rightarrow b\bar{b}$ , that contain a muon within their cones
  - Enriches the jets with b and c jets
  - Abundant in data
  - Assumes that  $g \rightarrow b\bar{b}$  jets have similar performance as  $H \rightarrow b\bar{b}$ 
    - Applies N-subjettiness,  $\tau_{21}$ , cut to bring proxy jets closer to signal jets

By applying  $\tau_{21}$  cuts, the tagger distribution of the proxy jets starts to resemble that of the signal jets



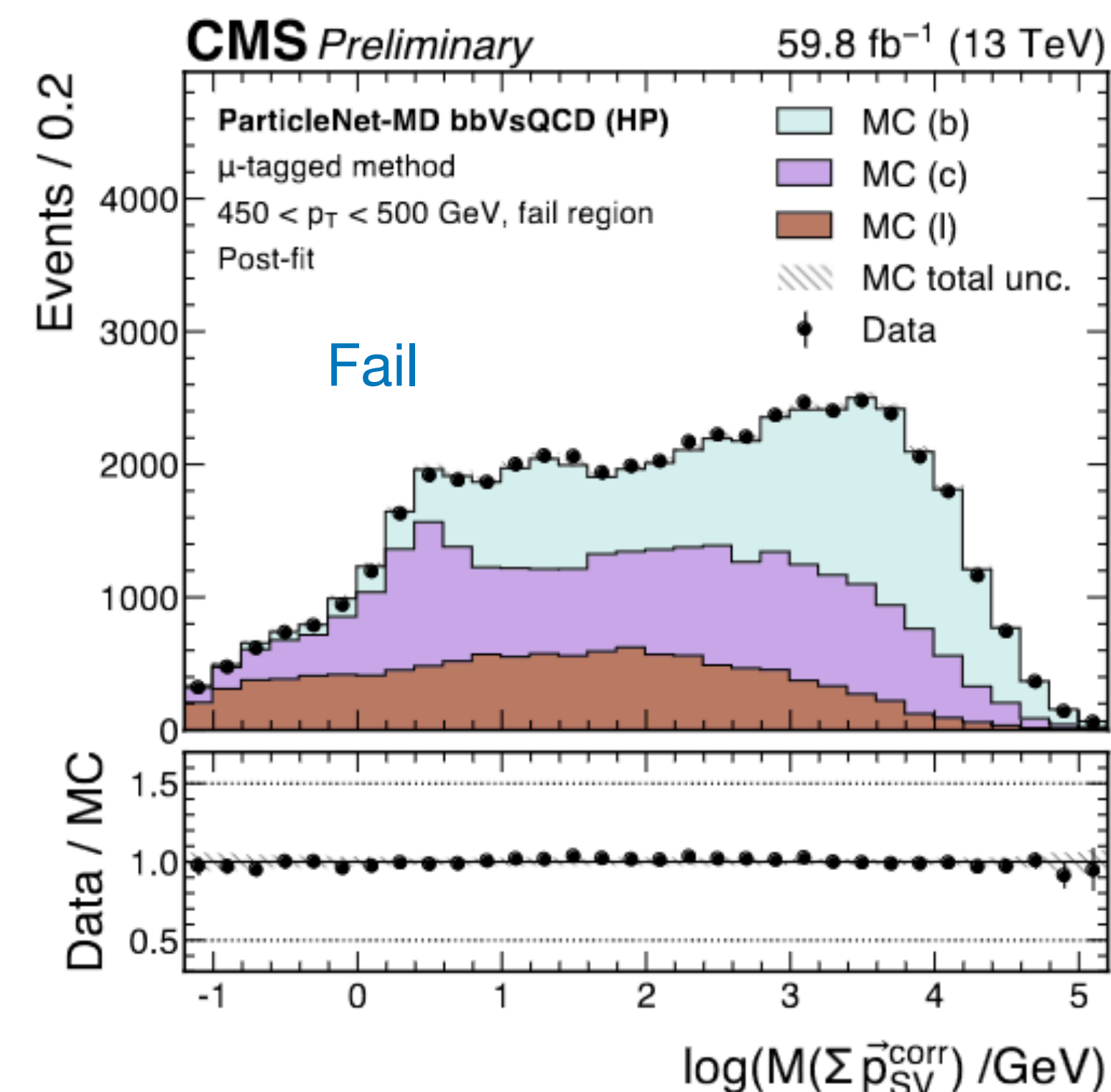
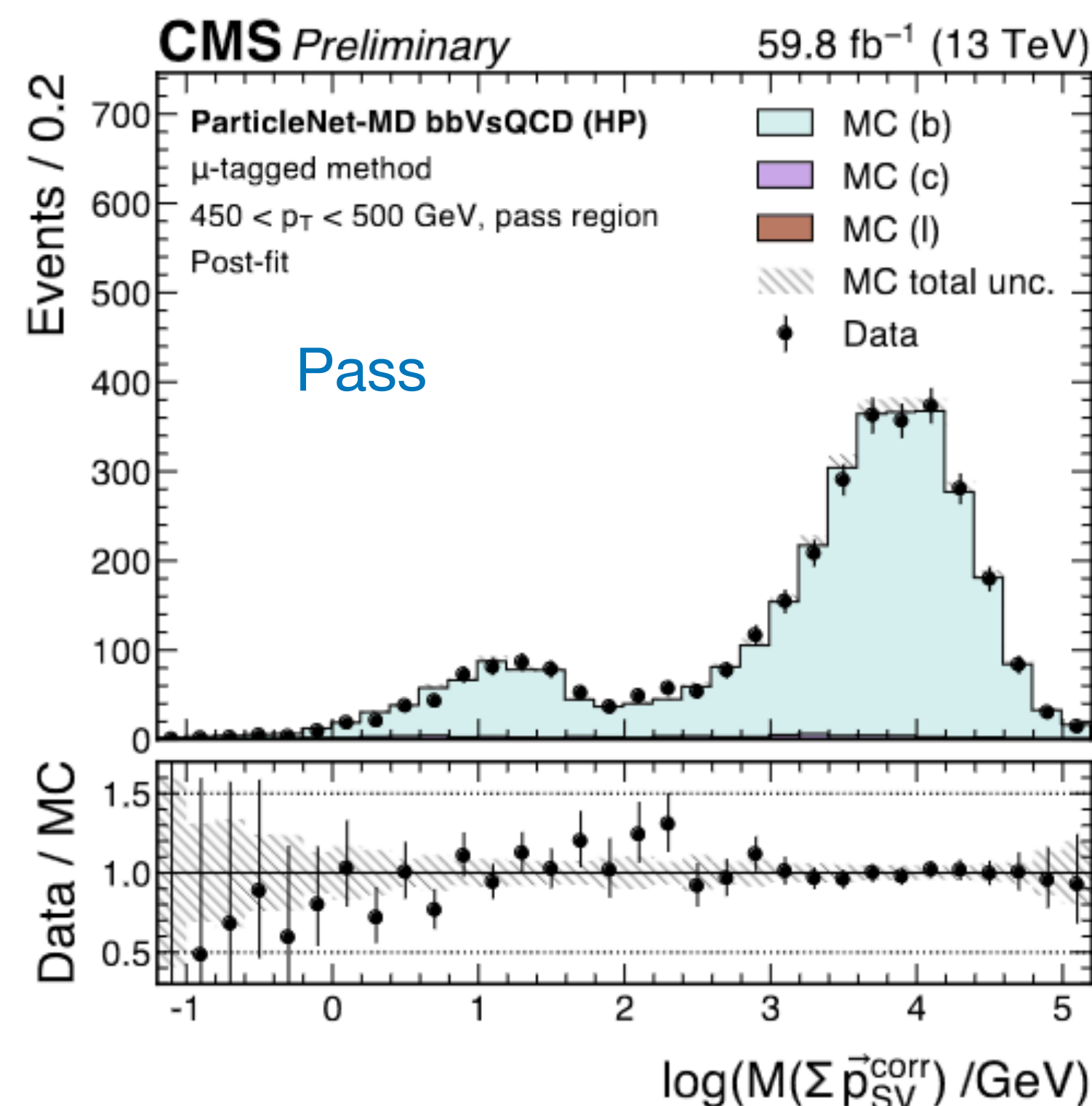


# Method 2: Muon jets proxy

- Uses jets originating from gluon splitting,  $g \rightarrow b\bar{b}$ , that contain a muon within their cones
  - Enriches the jets with b and c jets
  - Abundant in data
  - Assumes that  $g \rightarrow b\bar{b}$  jets have similar performance as  $H \rightarrow b\bar{b}$ 
    - Applies N-subjettiness,  $\tau_{21}$ , cut to bring proxy jets closer to signal jets

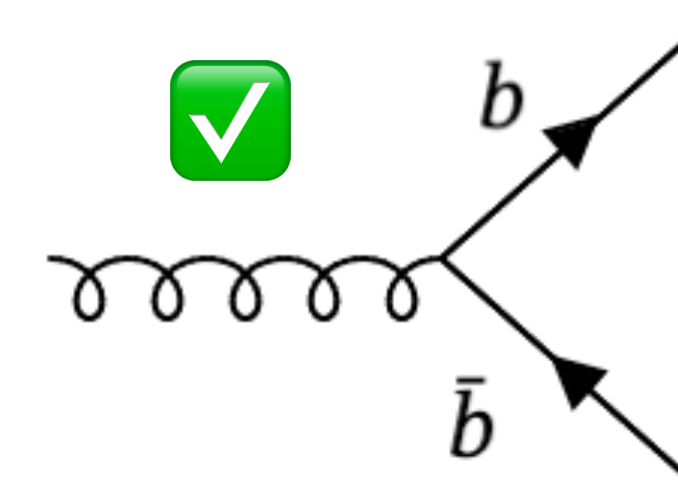
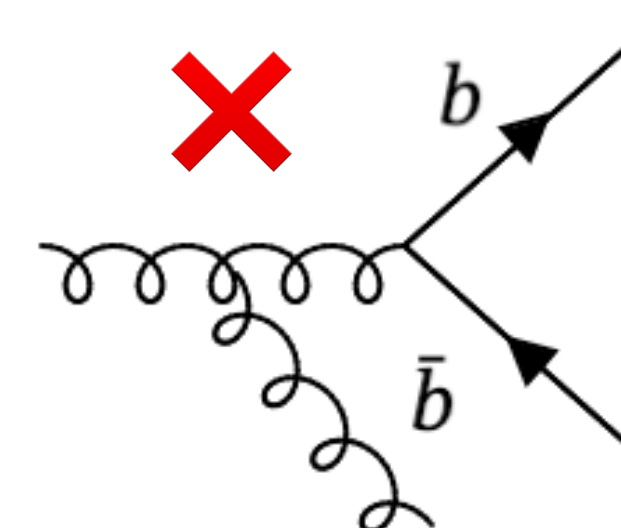
The application of ParticleNet tagger completely removes contributions from other flavours.

Fit variable is the invariant mass of the vector sum of all SV 4-momenta associated with the jet



# Method 3: “sfBDT” method

- Uses jets originating from gluon splitting,  $g \rightarrow b\bar{b}$ , selected by BDT
  - $g \rightarrow b\bar{b}$  jets are more likely to be contaminated by additional gluons
    - BDT is employed to filter-out such jets
- How to select jets for training BDT?
  - Based on generator-level variables



$$\kappa_g = \frac{\sum_{i \in \{g\}} p_{T,i}}{\sum_{i \in \{g, q\}} p_{T,i}}$$

OR

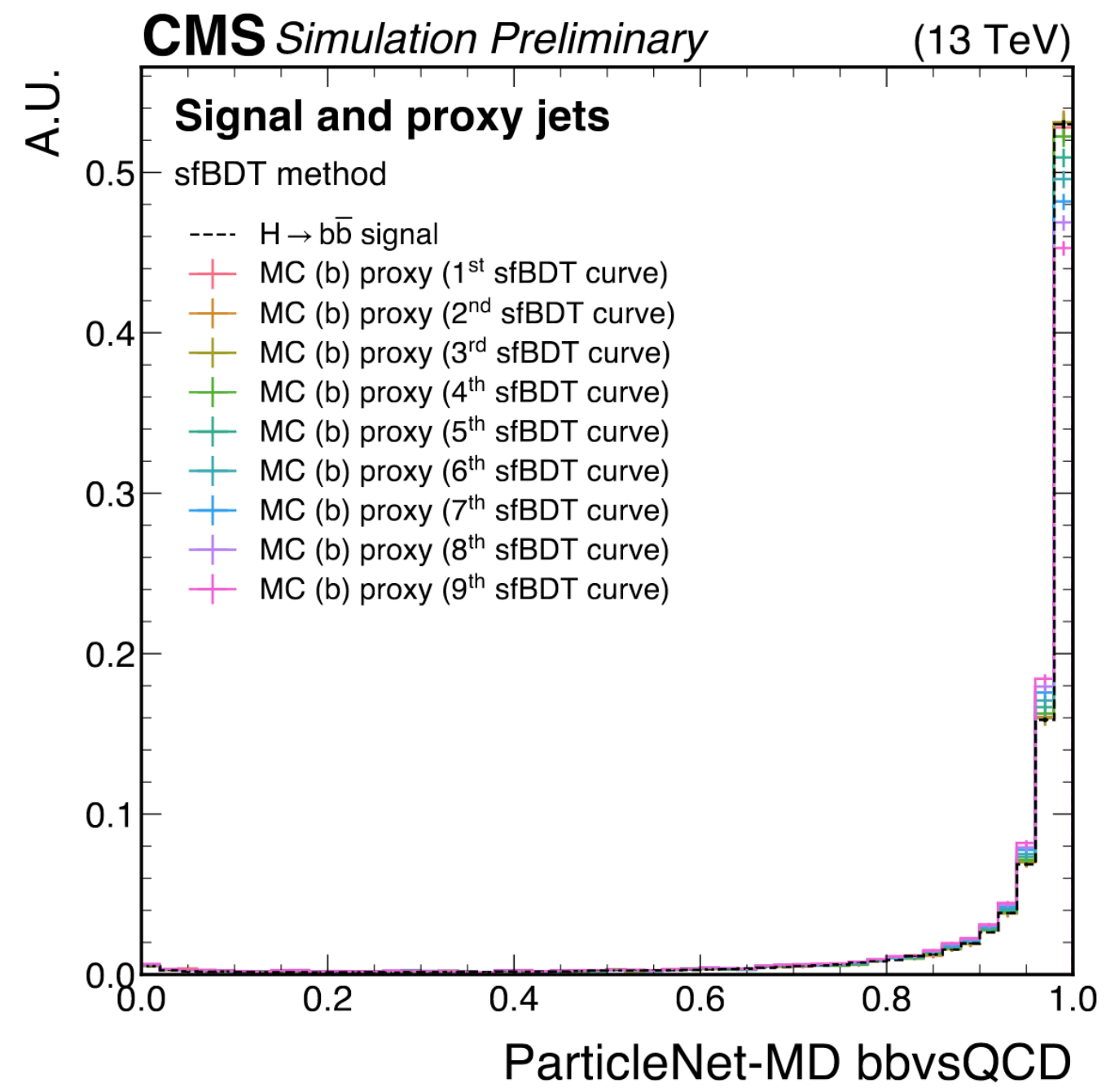
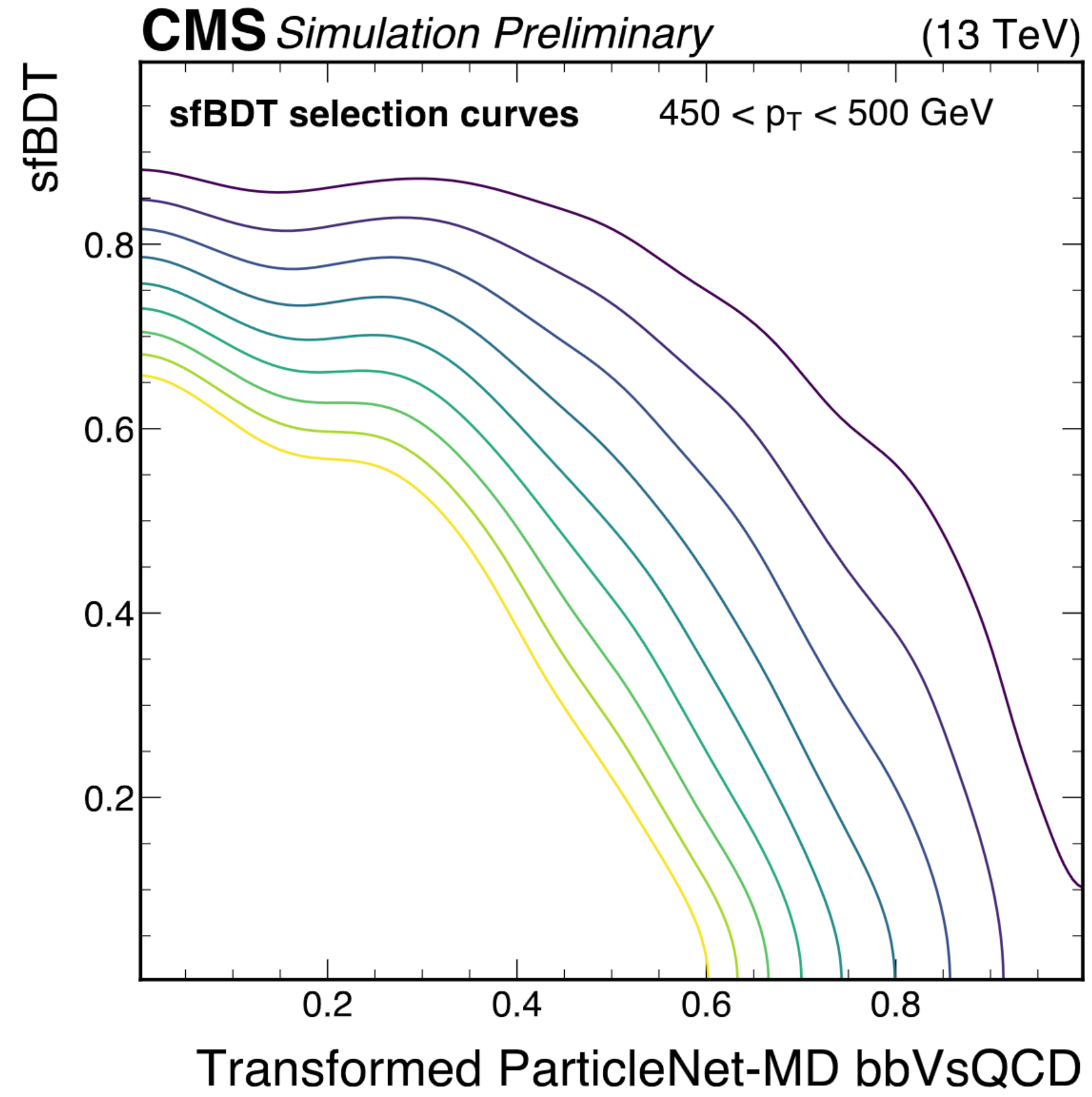
$$\tau_{31}^h = \frac{\sum_{i \in \{\text{had.}\}} p_{T,i} \min[\Delta R_{i, \hat{n}_{3,1}}, \Delta R_{i, \hat{n}_{3,2}}, \Delta R_{i, \hat{n}_{3,3}}]}{\sum_{i \in \{\text{had.}\}} p_{T,i} \Delta R_{i, \hat{n}_{1,1}}}$$

Fraction of momentum carried by gluons  
Train BDT to select jets with low score

N-subjettines  
Train BDT to select jets that do not look like three-prong

- Transform the discriminant so that  $X > X_0$  corresponds to selection efficiency of  $1 - X_0$  on a sample of signal jets
- BDT cut depends on the tagger value of the jets

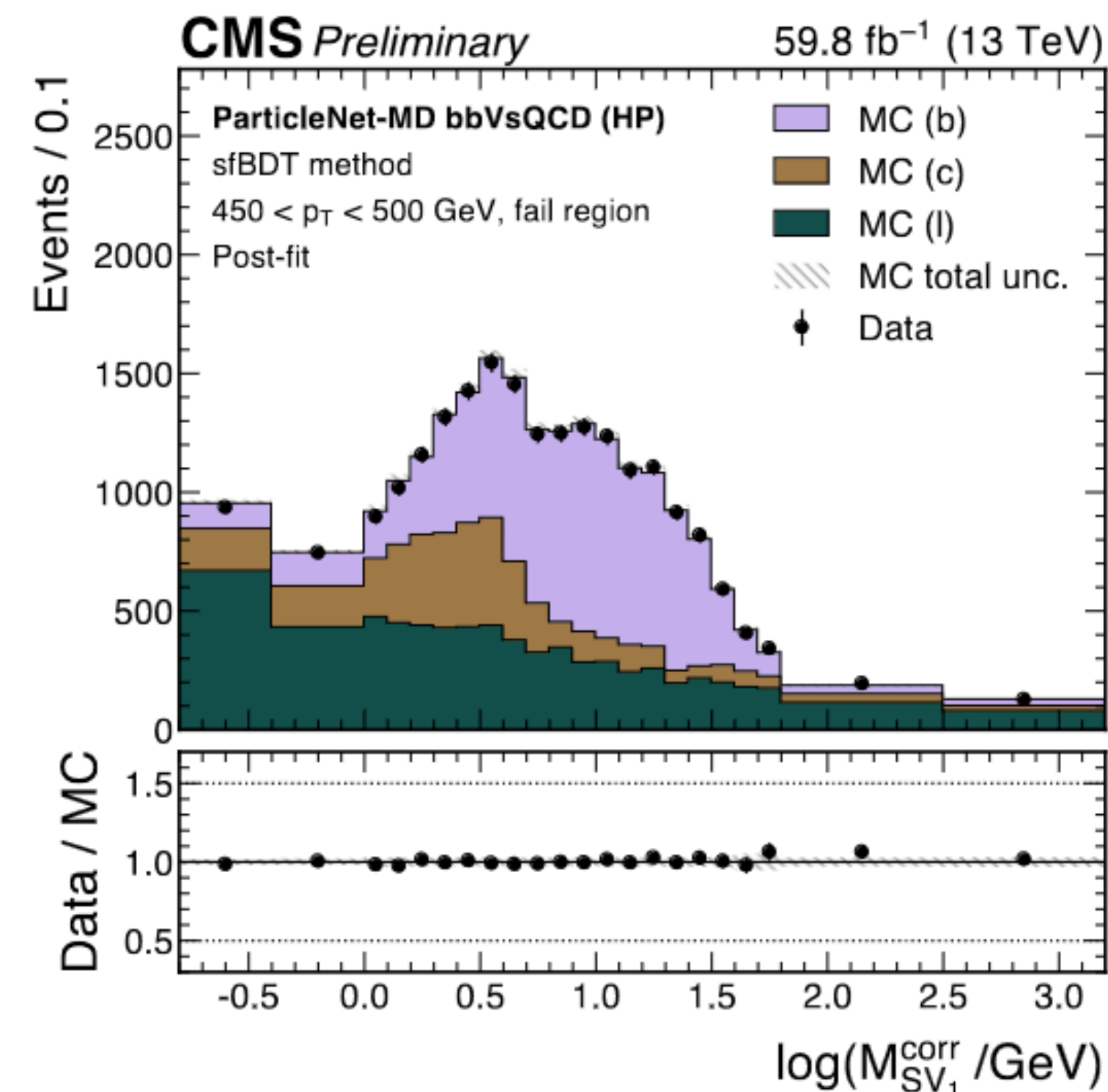
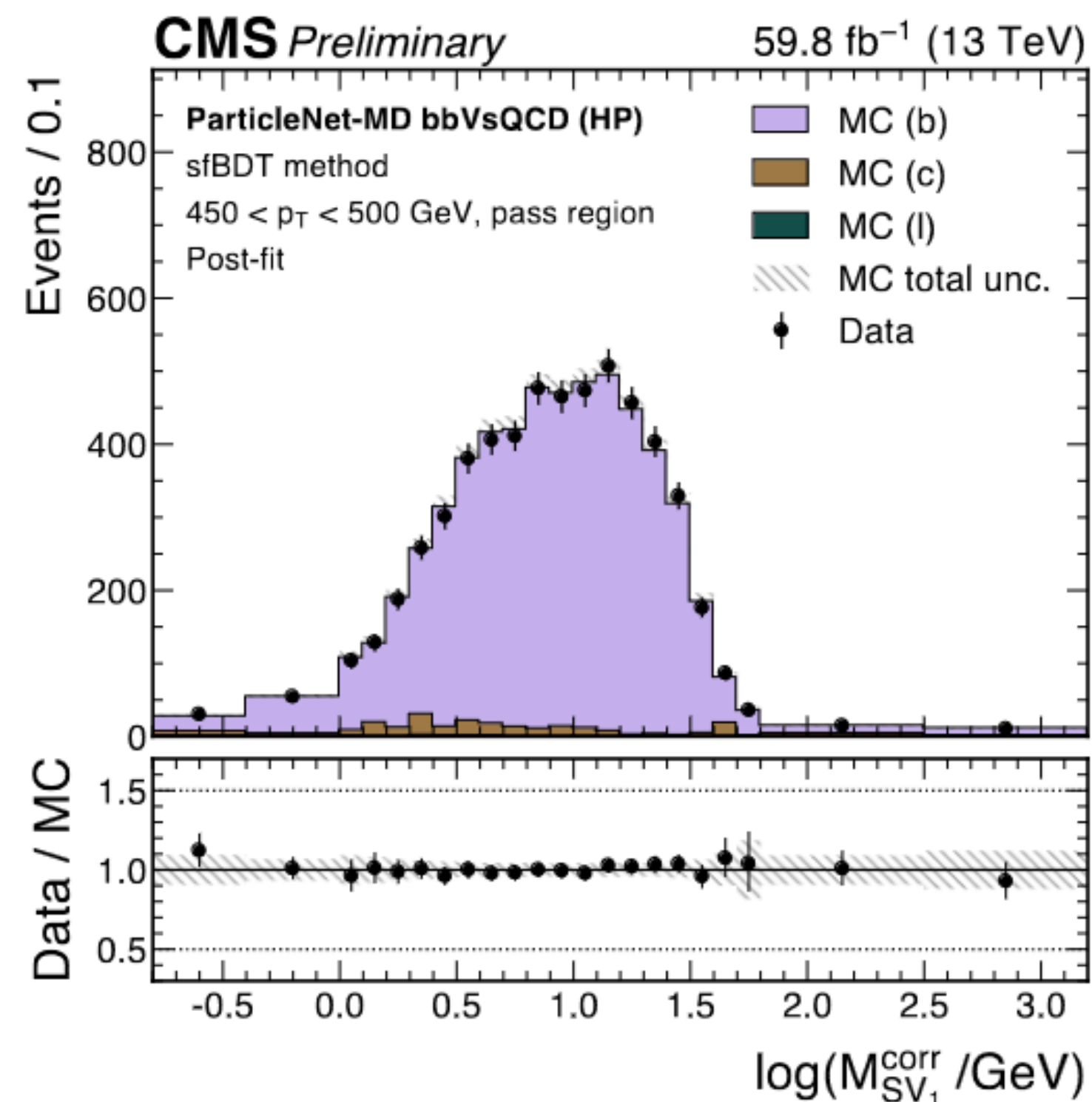
- Given by contours:  $F(x, y) = \int_y^1 f(x, y') dy'$
- $f(x, y)$  is a 2D PDF for the simulated proxy jets



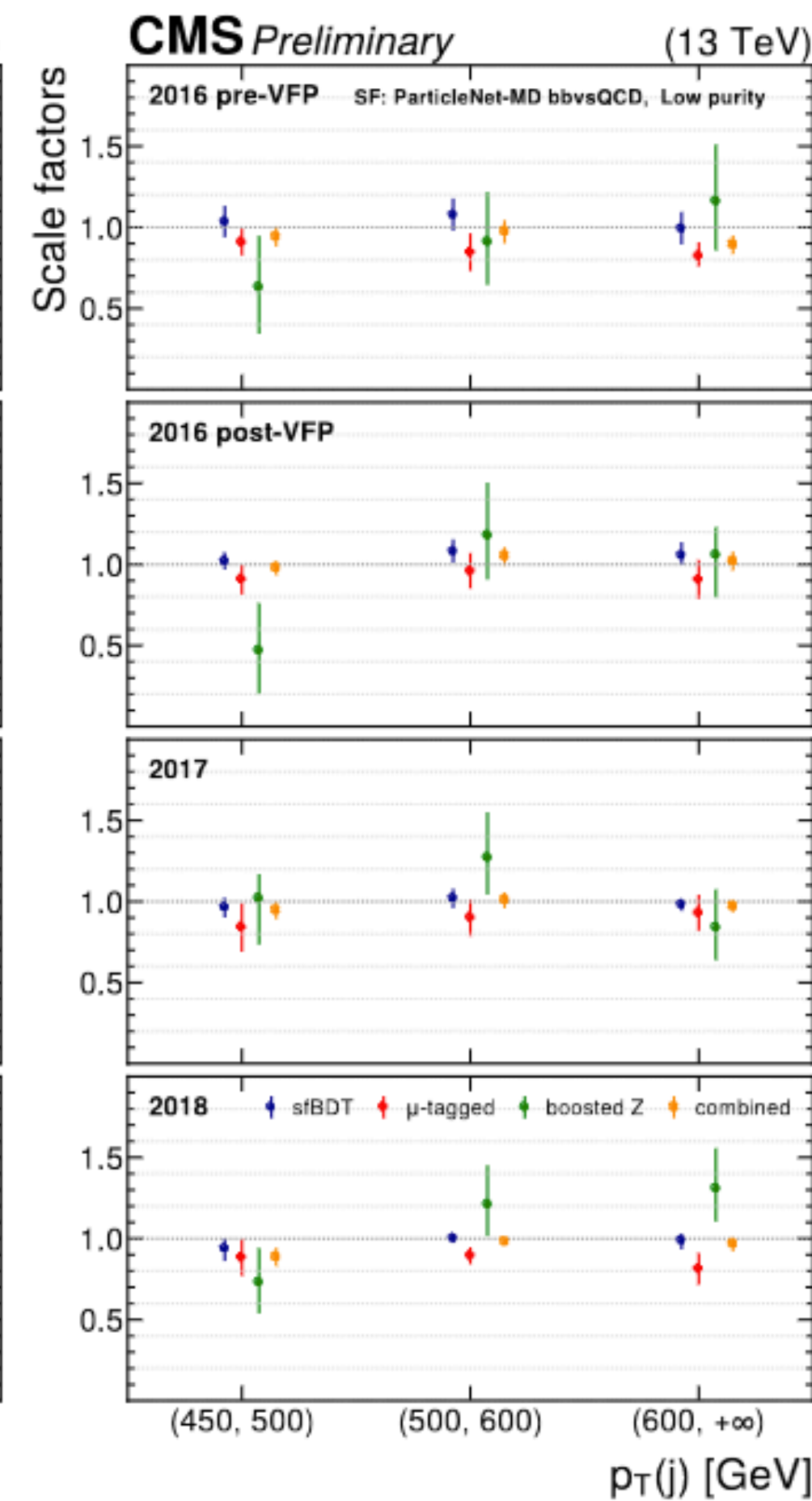
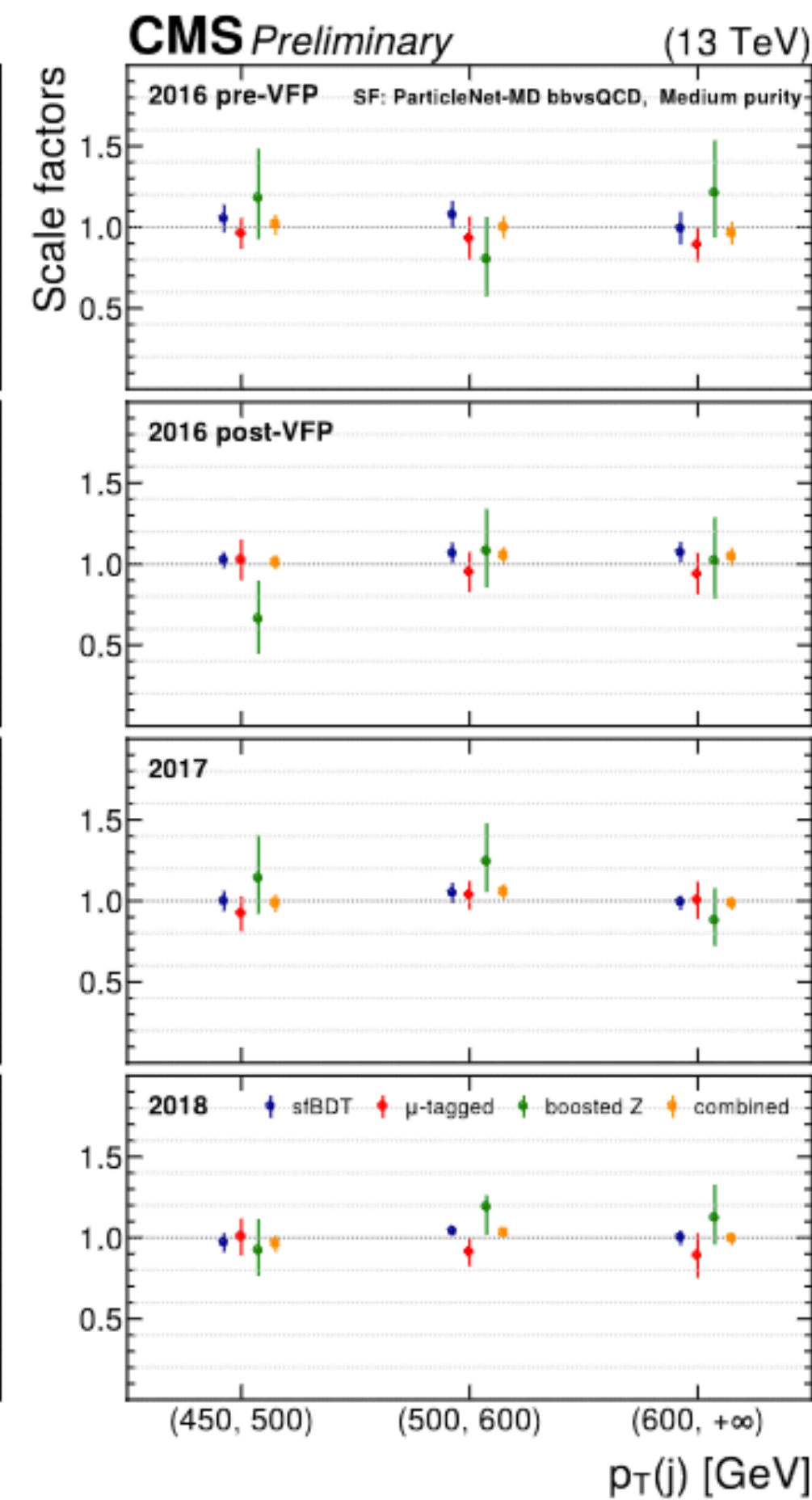
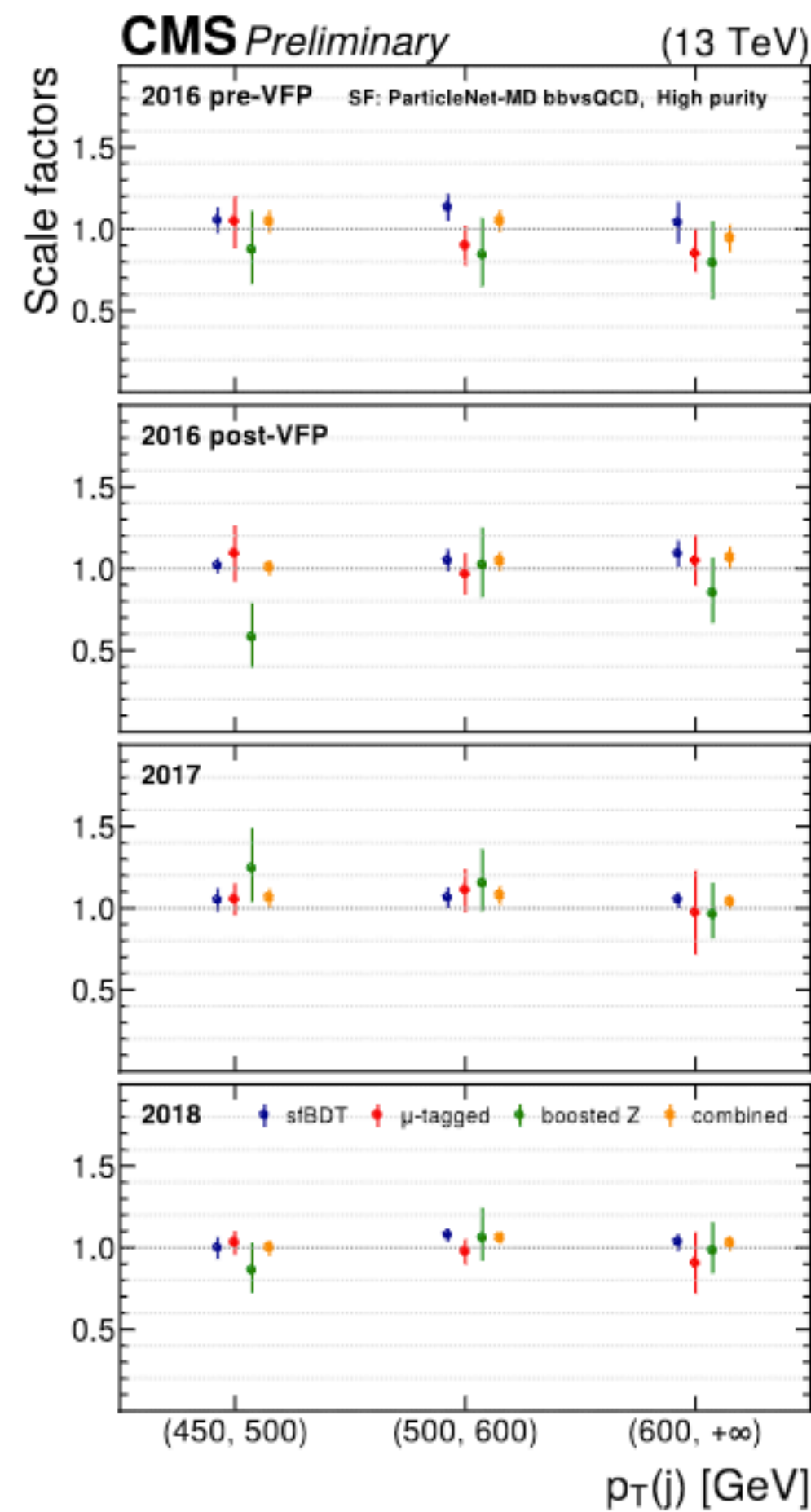
- The application of the BDT cuts brings the tagger discriminant distribution of proxy jets close to that of the signal
- Ensures similarity between signal and proxy jets

BDT is used to select proxy jets  
The efficiency is measured by fitting the number of jets in pass and fail regions

- Maximum likelihood fit
- Parameters of interest are “**scale factors**” for different flavours
  - Ratio of efficiency measured in data and in simulation
  - Move jets from fail to pass and vice-versa, keeping the total number of jets the same
- Multiple source of systematic uncertainties taken into account
  - BDT selection
  - Jet energy scale uncertainties
  - Fraction of jet flavours
  - ...



- Scale factors are mostly consistent with unity
  - Indicates good modelling in simulation
  - Agreement between methods gives confidence in applicability of the calibrations
- Results combined with Best Linear Unbiased Estimator (BLUE) method



# Summary

- Identifying heavy-flavour boosted jets is an important tool for performing analyses at the LHC
- Great progress in tagging performance during Run 2
- Multiple methods employed to validate the performance of the tagging algorithms
  - More details in [CMS-PAS-BTV-22-001](#)

Event display from [CMS physics briefing](#)

