Available on CMS information server

CMS DP -2022/005

The Compact Muon Solenoid Experiment

CMS Performance Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

11 March 2022

Calibration of the mass-decorrelated ParticleNet tagger for boosted $\mathrm{b\bar{b}}$ and $\mathrm{c\bar{c}}$ jets using LHC Run 2 data

CMS Collaboration

Abstract

The calibration of the new generation jet tagging algorithms exploiting advanced machine learning techniques becomes a challenging task. This note presents a novel approach for the calibration of the mass-decorrelated ParticleNet (ParticleNet-MD) boosted jet flavour tagging algorithm, focusing on the $\mathrm{X} \rightarrow \mathrm{b\bar{b}}$ and $\mathrm{X} \rightarrow \mathrm{c\bar{c}}$ mode. The approach builds upon the already established method used for the calibration of the previous generation boosted jet taggers, i.e., using proxy jets from gluon splitting to a pair of bottom or charm quarks. However, new techniques have been introduced to improve the similarity between proxy and signal jets and control the systematic uncertainties associated with the corrections. Data-to-simulation scale factors are derived for the three data taking years of Run 2 with the CMS experiment, based on different working points.

Abstract

The calibration of the new generation jet tagging algorithms exploiting advanced machine learning techniques becomes a challenging task. This note presents a new approach for the calibration of the mass-decorrelated ParticleNet boosted jet flavour tagging algorithm, focusing on the $X \to b\bar{b}$ and $X \to c\bar{c}$ mode. The approach builds upon the already established method used for the calibration of the previous generation boosted jet taggers, i.e., using proxy jets from gluon splitting to a pair of bottom or charm quarks. However, new techniques have been introduced to improve the similarity between proxy and signal jets and control the systematic uncertainties associated with the corrections. Data-to-simulation scale factors are derived for the three data taking years of the LHC Run 2 with the CMS experiment, based on different working points.

Glossary

- **AK8/AK15 jets**: Jets clustered with the anti-k_T algorithm [1] with a distance parameter of R=0.8/1.5.
- **ParticleNet tagger**: A graph neural network based particle identification algorithm for identifying hadronic decays of highly Lorentz-boosted top quarks and W, Z, and Higgs bosons and classifying various decay modes. The network is trained using particle-flow candidates and secondary vertices associated with the AK8/AK15 jet. The "ParticleNet" neural network architecture [2-4] is used to process the input particle-flow candidates and secondary vertices in a permutation-invariant way.
- **ParticleNet-MD tagger**: A mass-decorrelated particle identification algorithm designed for identifying two-prong hadronic decays of highly Lorentz-boosted particles (e.g., $X \to b\bar{b}$, $X \to c\bar{c}$, $X \to q\bar{q}$). The tagger is trained on a set of signal jets including $X \to b\bar{b}$, $X \to c\bar{c}$, $X \to q\bar{q}$, and background QCD jets, where X is a variable-mass spin-0 particle. Jets from both signal and background samples are reweighted to yield flat distributions in both p_T and soft-drop mass (m_{SD}) so as to decorrelate the trained tagger variable with the jet soft-drop mass. The ParticleNet-MD algorithm outputs four probability-like scores: $p(X \to b\bar{b})$, $p(X \to c\bar{c})$, $p(X \to q\bar{q})$, and p(QCD). The $X \to b\bar{b}$ and $X \to c\bar{c}$ discriminant can be defined as $p(X \to b\bar{b})$ / $[p(X \to b\bar{b}) + p(QCD)]$ and $p(X \to c\bar{c})$ / $[p(X \to c\bar{c}) + p(QCD)]$, respectively.

Glossary

- **ParticleNet-MD X→$b\bar{b}$($c\bar{c}$) discriminant calibration**: A process to correct potential difference in X→$b\bar{b}$($c\bar{c}$) tagging efficiency between data and the simulation on a given working point of the ParticleNet-MD X→$b\bar{b}$($c\bar{c}$) tagger discriminant. The calibration strategy aims to correct the tagging efficiency in simulation to match that in data, by means of data-to-simulation scale factors, on a target phase-space (e.g., the phase-space for signal H→$b\bar{b}$($c\bar{c}$) jets). The resulting scale factors, usually measured in multiple jet p_T bins, are used to scale the simulated events to match the data.
- **Signal jets and proxy jets**: Signal jets correspond to the type of jets that defines the target phase-space for calibration, e.g., the H→$b\bar{b}$($c\bar{c}$) jets. When signal jets are technically intractable to be directly obtained from data, a set of proxy jets is utilised as a substitute for the signal jets. The proxy jets are obtained from data and are selected such to have similar characteristics to signal jets. The scale factors are then derived from the proxy jets by comparing the tagging efficiency in data and simulation. For the H→$b\bar{b}$($c\bar{c}$) jets, the gluon-splitting g→$b\bar{b}$($c\bar{c}$) jets from QCD multijet events with additional selections are used to build the proxy jet collection.

Glossary

- **BDT for scale factor derivation (sfBDT)**: A Boosted Decision Tree (BDT) classifier used for selecting a suitable set of proxy jets for scale factors derivation. It is the main tool of this new calibration method.

  The sfBDT is trained with jets from gluon-splitting g→b$\bar{b}$(c$\bar{c}$) in QCD multijet events, and is designed to separate jets with a clean composition of quarks, which more resembles the H→b$\bar{b}$(c$\bar{c}$) jets, against the ones with large contamination of extra gluons. Hence, a selection involving the sfBDT discriminant is capable to build a better proxy jet collection from g→b$\bar{b}$(c$\bar{c}$) jets, by vetoing jets with a high gluon contamination rate that exhibit different characteristics from the target signal H→b$\bar{b}$(c$\bar{c}$) jets.

  The gluon contamination rate is defined by a variable κ_g, which is the ratio of the scalar p_T sum of all final state gluons over the scalar p_T sum of all final-state gluons and quarks. The gluons and quarks are selected from the parton-level truth particles associated with a jet. The signal (background) jets are selected from the QCD g→b$\bar{b}$ or c$\bar{c}$ jets that satisfies $\kappa_g < 0.15$ ($\kappa_g > 0.85$). The input variables to the sfBDT involve the basic kinematics of the subjets and secondary vertices associated with the jet.

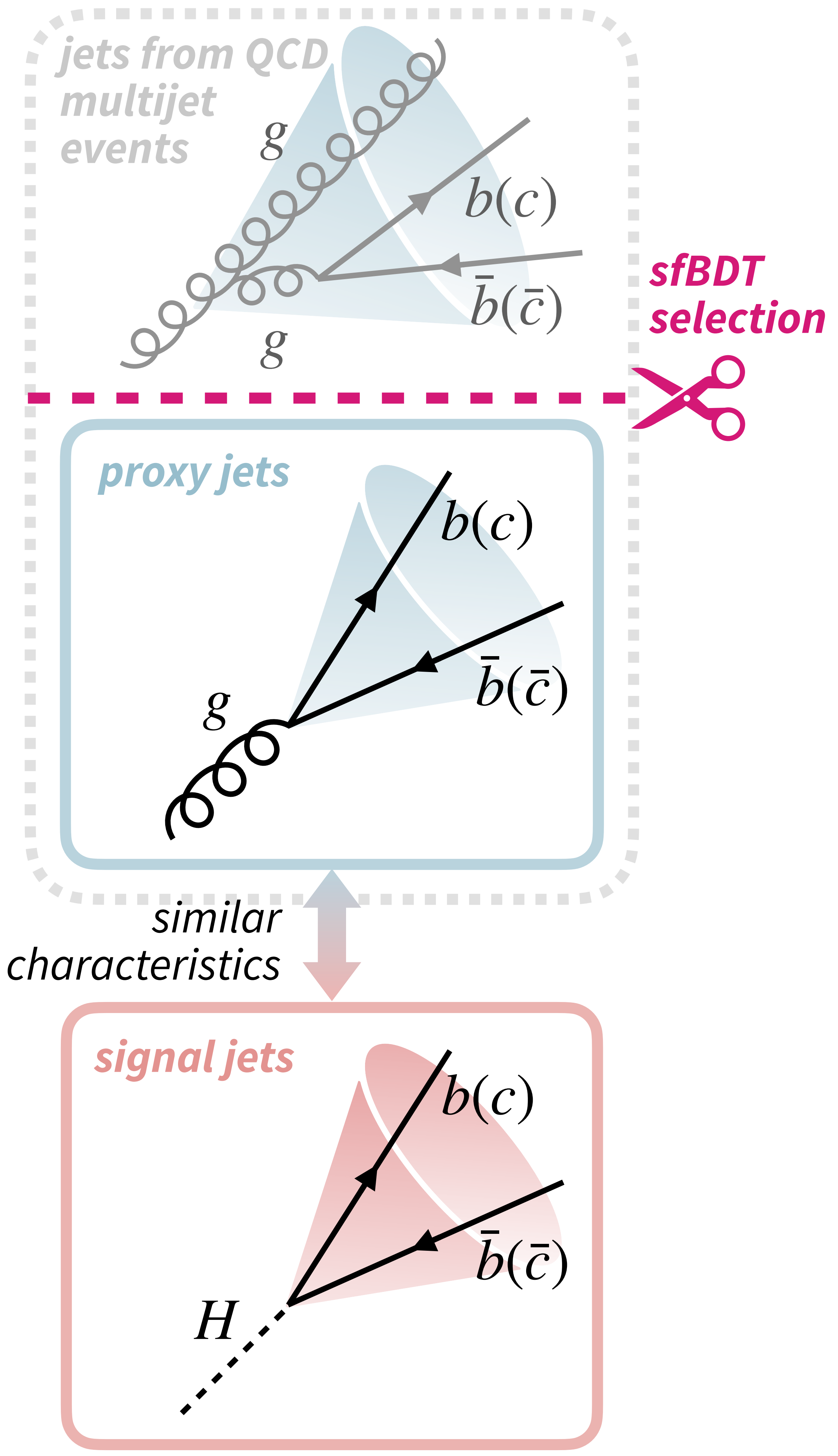
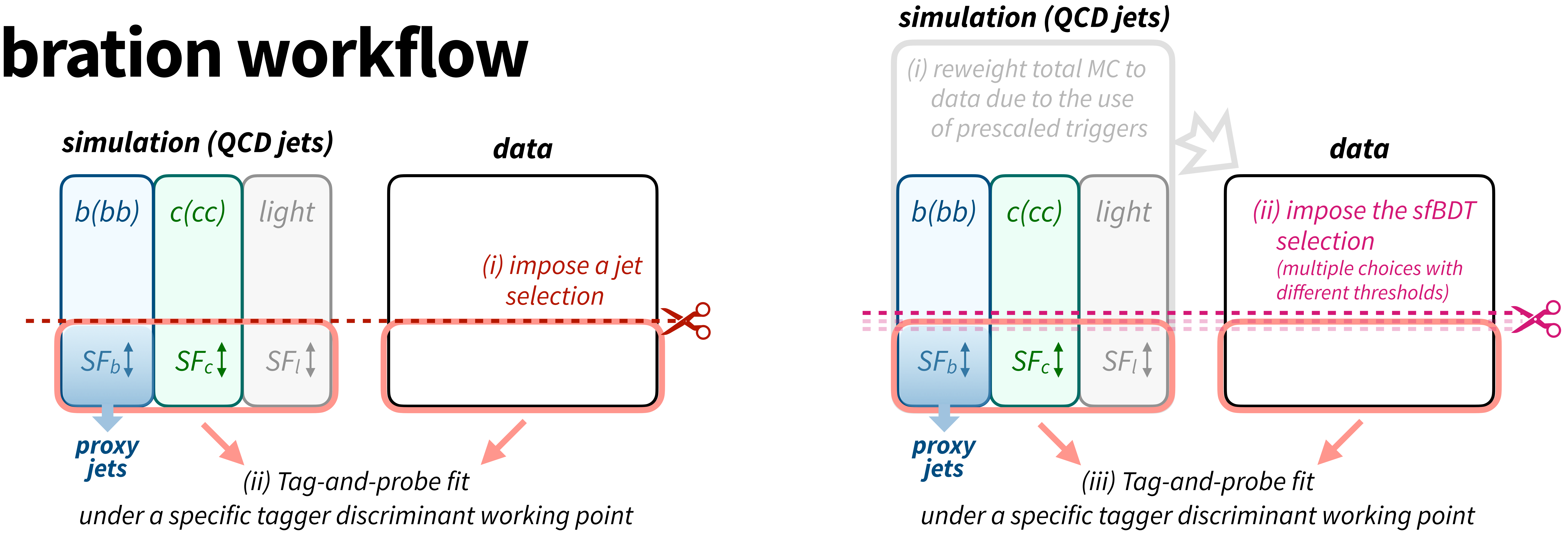


Fig. 1. Illustration of the calibration method and the effect of the sfBDT variable.

Calibration workflow


simulation (QCD jets)
data
b(bb)
c(cc)
light
(i) impose a jet selection
SF_b
SF_c
SF_l
proxy jets
(ii) Tag-and-probe fit
under a specific tagger discriminant working point
simulation (QCD jets)
(i) reweight total MC to data due to the use of prescaled triggers
data
b(bb)
c(cc)
light
(ii) impose the sfBDT selection
(multiple choices with different thresholds)
SF_b
SF_c
SF_l
proxy jets
(iii) Tag-and-probe fit
under a specific tagger discriminant working point

(a) Conventional workflow

(b) New workflow

Fig. 2. The schematic workflow of the new $b\bar{b}/c\bar{c}$ calibration approach (right) compared with the conventional approach (left). A general description of both methods goes as follows. All QCD jets from the Monte Carlo (MC) simulation are categorised into b(bb), c(cc), and the light classes based on the truth-level matching with the b- and c-hadrons. The proxy of $H\to b\bar{b}$ or $H\to c\bar{c}$ jets is built from the b(bb) or c(cc) class with a specific selection. The remaining MC jets are fitted to the corresponding data with the tag-and-probe method, under the specific tagger discriminant working point. Three free-floating rate parameters SF_b, SF_c, SF_l are assigned to the three classes respectively. The fit is performed individually on multiple jet p_T bins. The post-fit parameter SF_b (SF_c) is then regarded as the scale factor for the $H\to b\bar{b}(c\bar{c})$ signal jets in the context of $b\bar{b}(c\bar{c})$ calibration.

Calibration workflow

The new method in Fig. 2 (b) has improvements with respect to the conventional approach Fig. 2 (a) (adopted in e.g., Ref. [5]) in three aspects:

(1) The simple selection on jet-level variables for building the proxy, which is used in the previous approaches, was not adequate for the new generation of algorithms. To this end, the sfBDT is developed to improve the selection of proxy jets. The effect of the selection on sfBDT is illustrated in Figs. 3 and 4.

(2) To increase the statistical precision, up to two leading fatjets from the multijet events are selected as jet candidates. Besides, prescaled high-level triggers with smaller H_T thresholds are adopted to data and MC, and hence a reweighting from MC to data is applied before all selections.

(3) For each calibration point, example pre-fit and post-fit plots are presented in Fig. 5, and a summary of the systematic uncertainties is shown in Table 1. Besides these uncertainty terms considered in each individual fit, the dependence of the resulting scale factor with different choices of the sfBDT selection threshold is also studied. A dedicated uncertainty is developed to cover this effect, as detailed in Fig. 6.

A summary of the scale factors in the context of ParticleNet-MD $b\bar{b}/c\bar{c}$ calibration, both for the AK8 and AK15 jets, is provided in this note.

sfBDT selection

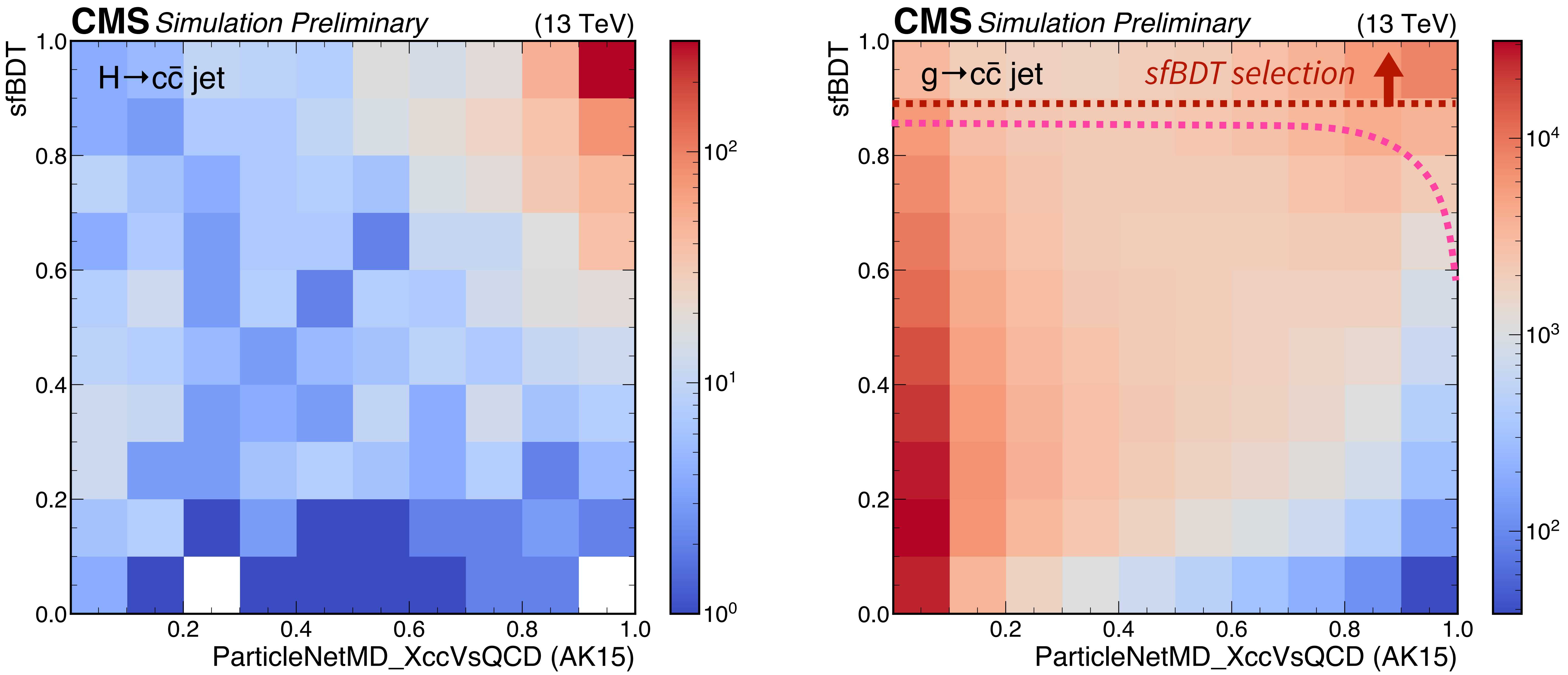


Fig. 3. The example 2D histogram on the tagger discriminant (ParticleNet-MD X→cc̄ on AK15 jets) versus the sfBDT score, for the H→cc̄ (left) and the g→cc̄ (right) jets. The g→cc̄ histogram has two enriched regions, one in the top-right region that resembles H→cc̄ signal jets, and one in the bottom-left corner representing jets with more gluon contamination. A selection involving the sfBDT helps to select a dedicated phase-space that improve the proxy and signal jet similarity.

Two possible types of selection are considered depending on the specific signal and proxy condition: a straight cut on the sfBDT variable (red dashed line), and an sfBDT cut with thresholds depending on the discriminant value (pink dashed curve) designed in the spirit to preserve more signal-like, high-discriminant-score jets after the selection.

Effect of sfBDT selection

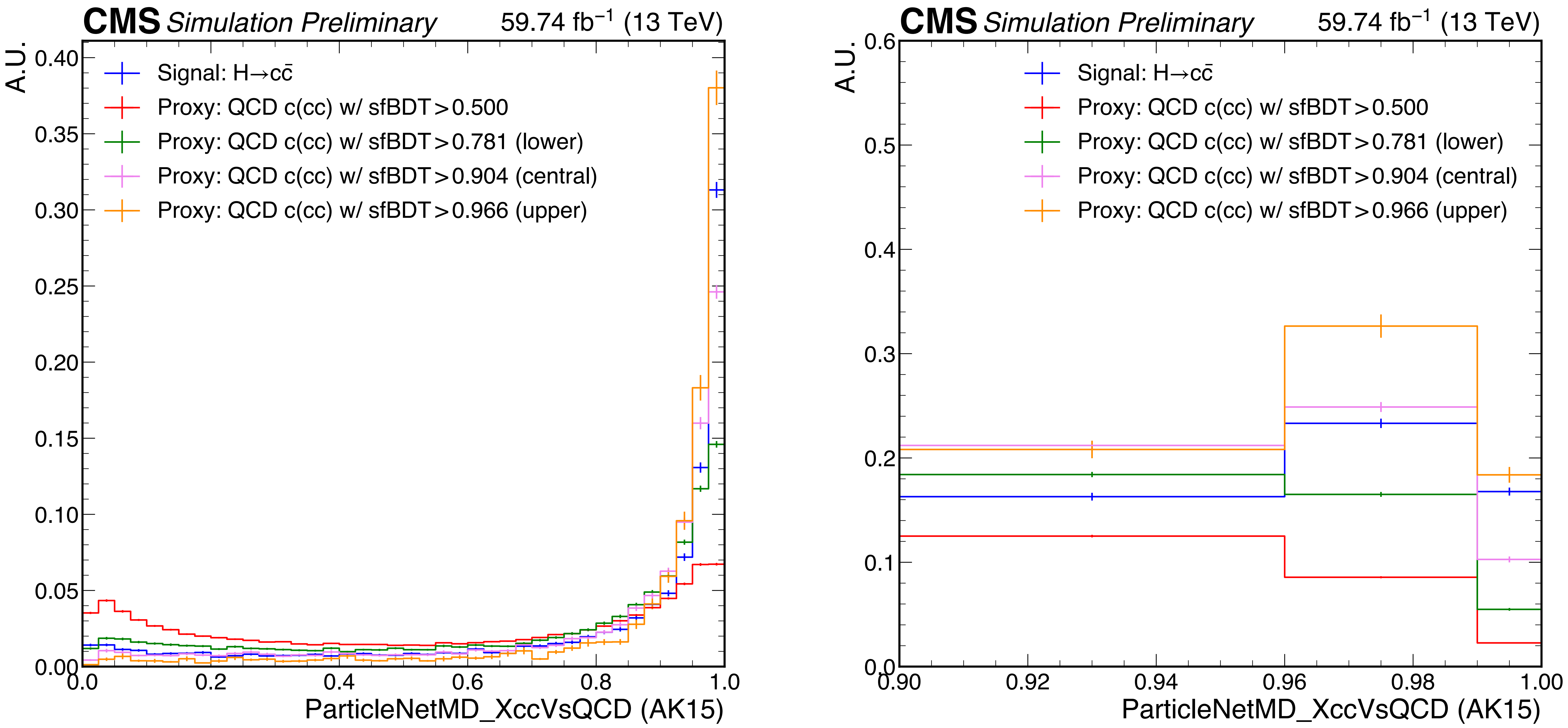


Fig. 4. The example discriminant shape (ParticleNet-MD X→cc̄ on AK15 jets) on equal-width binning (left) and analysis-defined working points (right) of the signal and proxy jets, where the signal jets are H→cc̄ jets under this case, and proxy jets are selected from the c(cc) class from QCD jets passing an sfBDT selection with different thresholds. The histograms demonstrate that the sfBDT is capable of tuning the tagger shape and improving the similarity between the signal and proxy shape. According to the variation of the proxy shape, a list containing 11 sfBDT selection choices is determined by an algorithm such that the dynamic range of the proxy shape can cover the signal shape.

Pre-fit and post-fit result

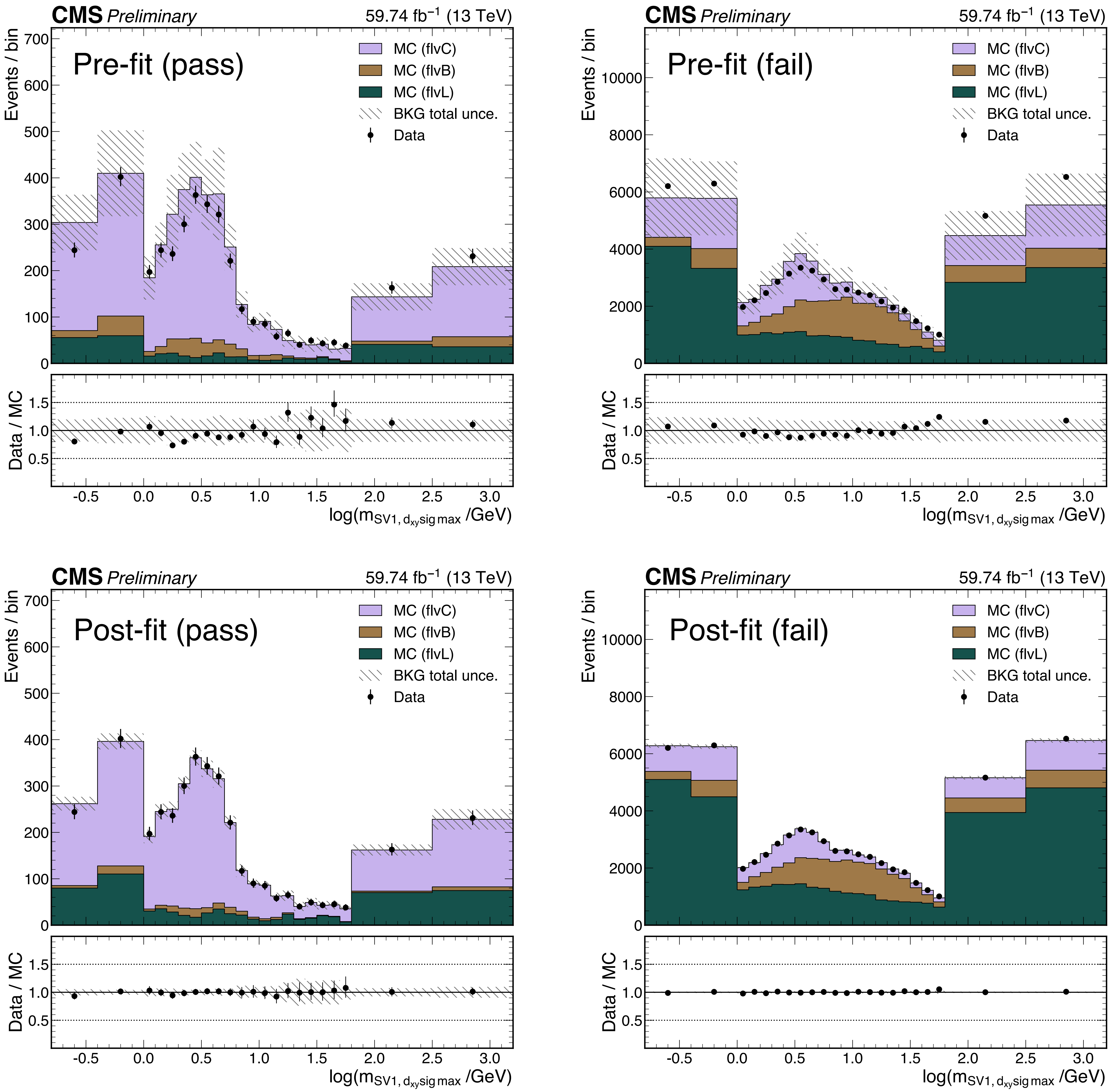


Fig. 5. The example pre-fit (top) and post-fit (bottom) plots for a single calibration point with one specific sfBDT selection adopted. The distributions in the pass (left) and fail (right) regions of a specified tagger discriminant working point are shown.

The fit variable is log(m_{SV1}), where SV_1 stands for the leading secondary vertex associated with the jet that has the highest impact parameter d_{xy} significance. The design of the fit variable and the binning ensures that the three MC flavour templates (i.e., b(bb), c(cc), and light) are as distinct as possible to obtain a stable fit result.

Systematic uncertainties

Source	Uncertainties on three flavour templates		
	b(bb)	c(cc)	light
Luminosity	1.2–2.5%	1.2–2.5%	1.2–2.5%
Pileup reweighting	<0.5%	<0.6%	<1.9%
sfBDT variable data-to-MC reweighting	<0.2%	<0.2%	<0.2%
ISR parton shower uncertainty	1–3%	4–6%	3–5%
FSR parton shower uncertainty	2–6%	8–12%	17–20%
Fragmentation uncertainty on bottom quarks	14–16%	—	—
Fragmentation uncertainty on charm quarks	—	13–16%	—
Fragmentation uncertainty on light quarks	—	—	20%

Table 1. Summary of the systematic uncertainties included in an individual fit.

Dependence of scale factors with sfBDT

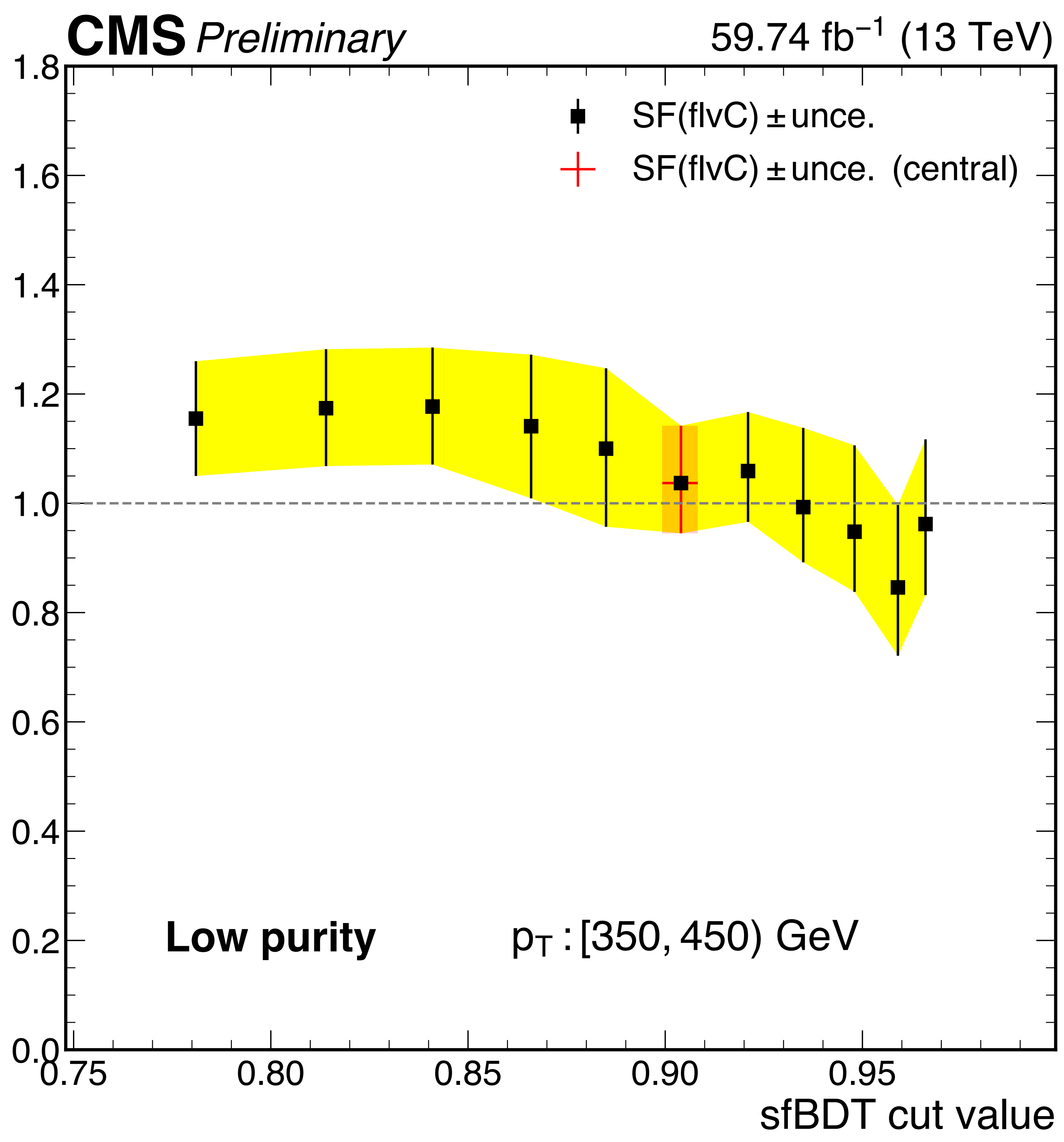


Fig. 6. The example plot to show the target scale factor as a function of the sfBDT selection thresholds. In a specific calibration point, each of the 11 sfBDT choices as introduced in Fig. 4 is used to define the proxy jet collection and then perform a fit to produce a target scale factor. The maximum distance between all 11 scale factors with the central value (in red colour) is taken as an additional uncertainty term, which will contribute to the final uncertainty of the scale factor. This additional term aims at covering the variation of the scale factor when the sfBDT threshold varies.

Besides, in the case when the tagger-discriminant-dependent sfBDT threshold (introduced in Fig. 3, the pink curve) is used to define the sfBDT selection, an extra set of 11 sfBDT selections corresponding to the straight version of cuts are also used to define the proxy and proceed with the scale factor derivation. Thus, a total of 22 scale factors is obtained and used for deriving the maximum distance uncertainty.

Summary of scale factors (AK8 jets)

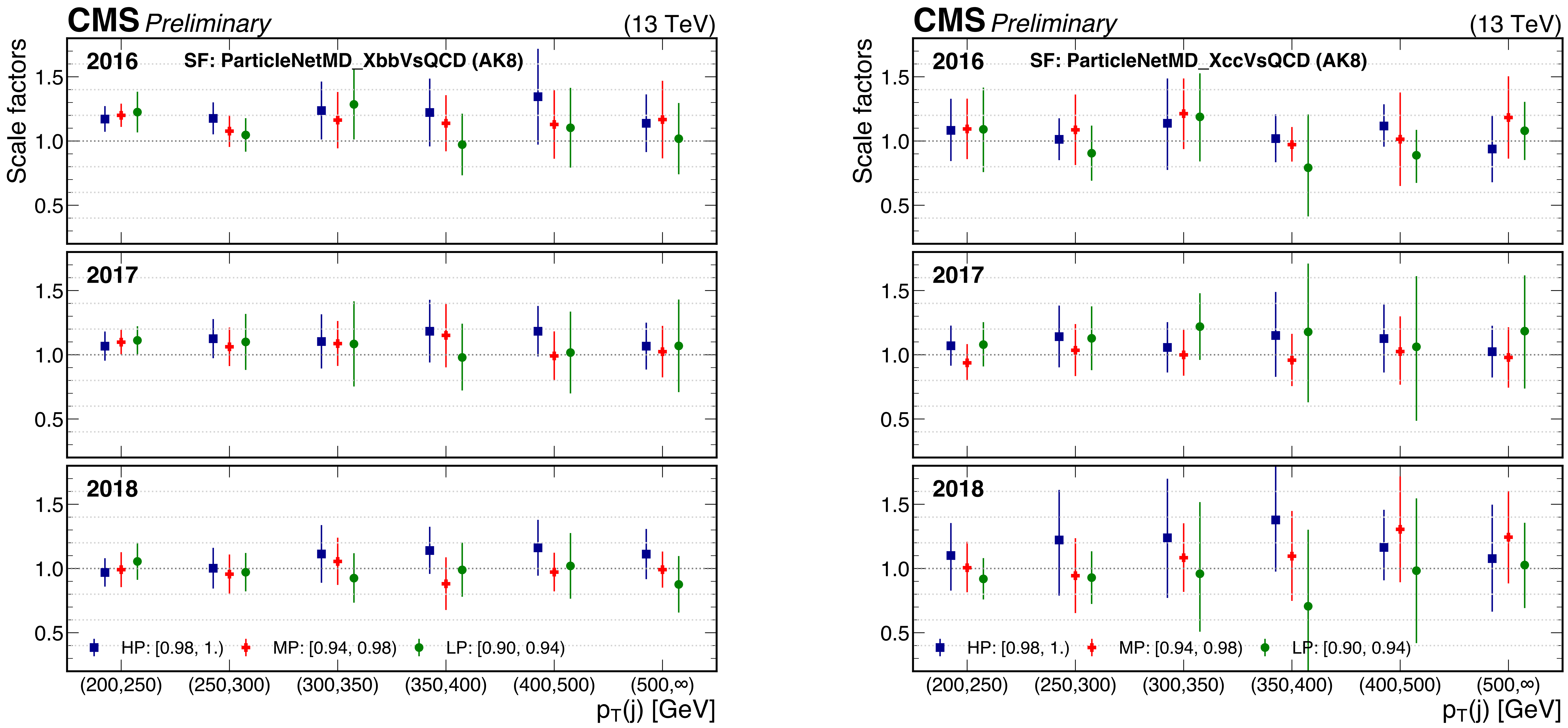


Fig. 7. Summary of the scale factors in the context of ParticleNet-MD AK8 X→$b\bar{b}$ (left) and X→$c\bar{c}$ (right) discriminant calibration, for H→$b\bar{b}$ and H→$c\bar{c}$ jets respectively. The results are derived in the three data taking years of Run 2, in six exclusive jet p_T ranges, and in three exclusive tagger discriminant working points, denoted as High Purity (HP), Medium Purity (MP), and Low Purity (LP) with a detailed definition in the plot.

Summary of scale factors (AK15 jets)

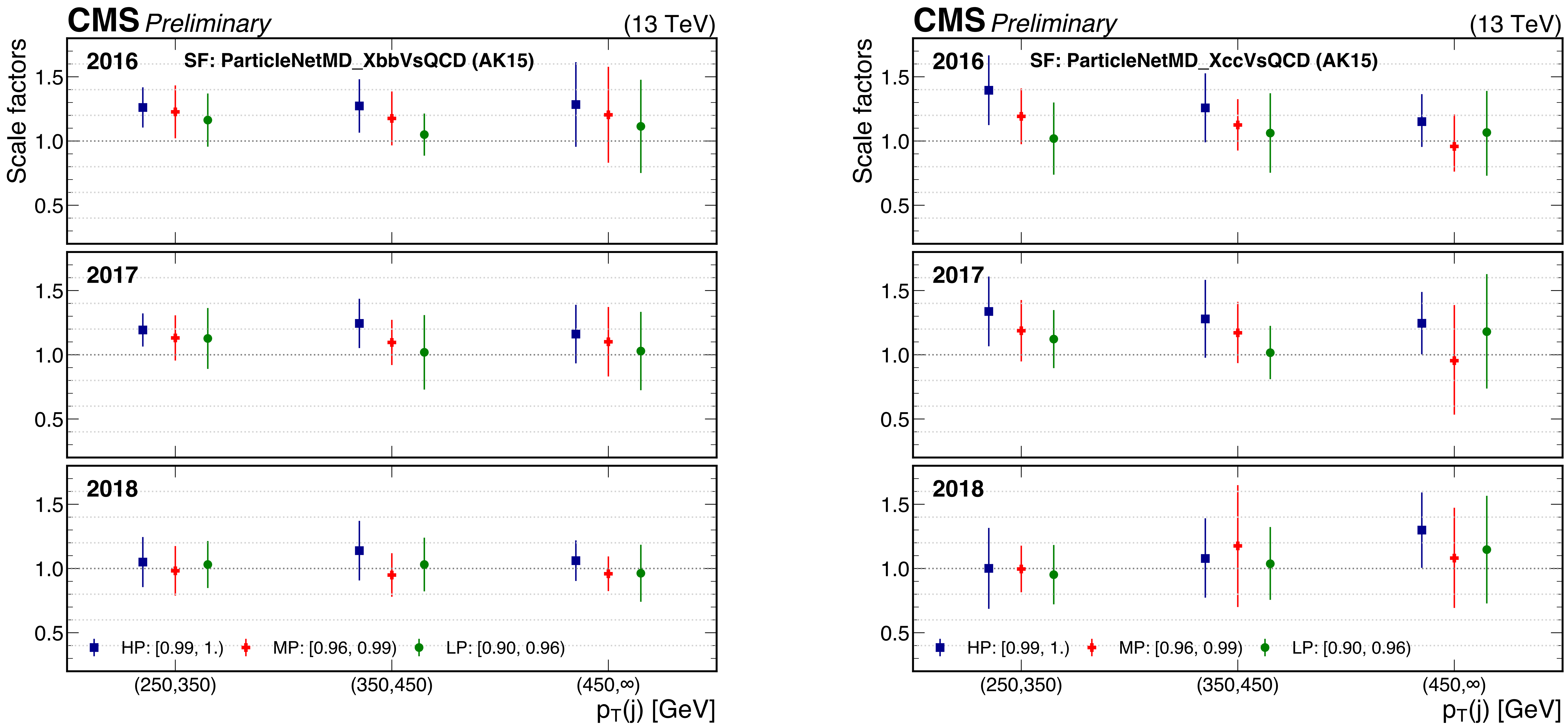


Fig. 8. Summary of the scale factors in the context of ParticleNet-MD AK15 X→$b\bar{b}$ (left) and X→$c\bar{c}$ (right) discriminant calibration, for H→$b\bar{b}$ and H→$c\bar{c}$ jets respectively. The results are derived in the three data taking years of Run 2, in three exclusive jet p_T ranges, and in three exclusive tagger discriminant working points, denoted as High Purity (HP), Medium Purity (MP), and Low Purity (LP) with a detailed definition in the plot.

Reference

[1] M. Cacciari, G. P. Salam and G. Soyez, “The anti-kt jet clustering algorithm,” JHEP **0804** (2008) 063.

[2] H. Qu and L. Gouskos, “Jet Tagging via Particle Clouds,” Phys. Rev. D **101** (2020) 056019.

[3] CMS Collaboration, “Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques,” CMS DP 2020/002.

[4] CMS Collaboration, “Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques,” JINST **15** (2020) P06005.

[5] CMS collaboration, “A search for the standard model Higgs boson decaying to charm quarks,” JHEP **2003** (2020) 131.