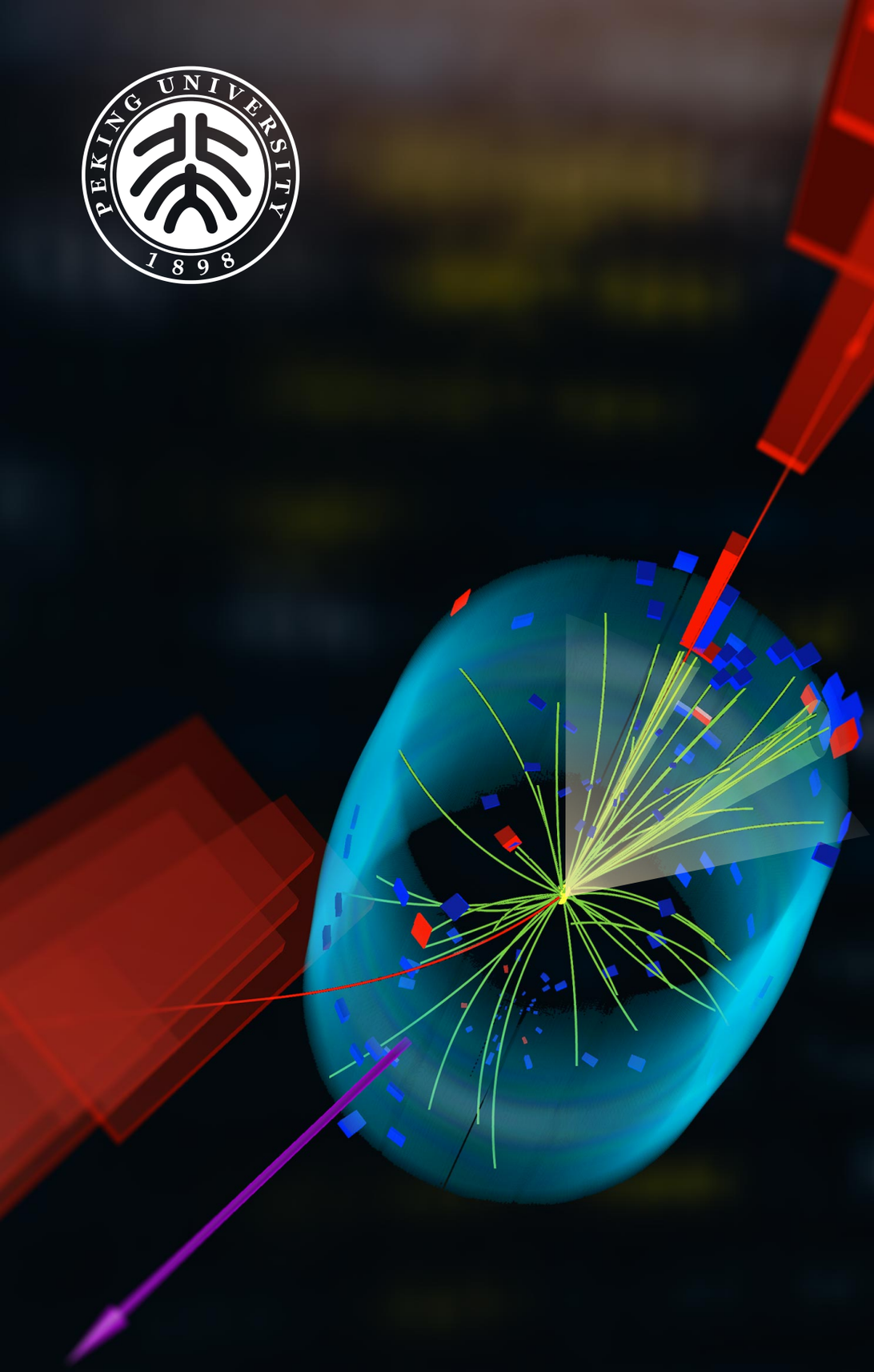


Performance of heavy-flavour jet identification in boosted topologies in CMS 13 TeV data

Congqiao Li (*Peking University*)

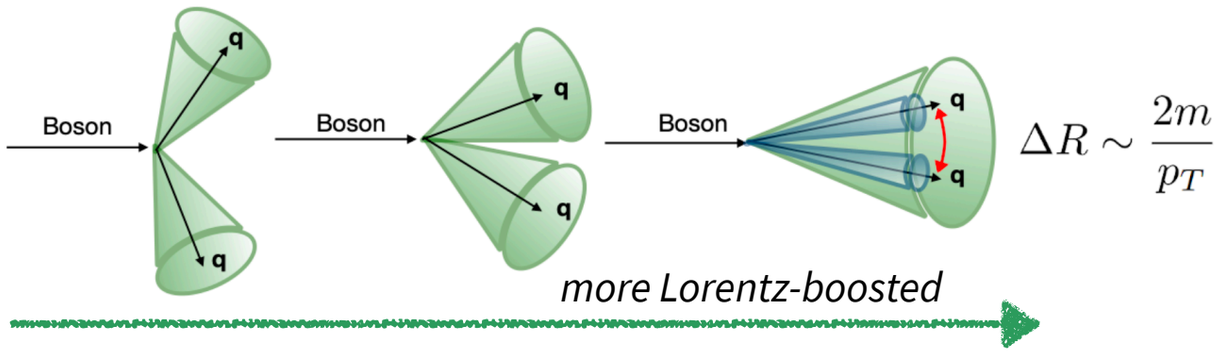
on behalf of the CMS Collaboration,
based on CMS-PAS-BTV-22-001

BOOST 2023 · Berkeley
31 July, 2023

Boosted heavy-flavour jet tagging

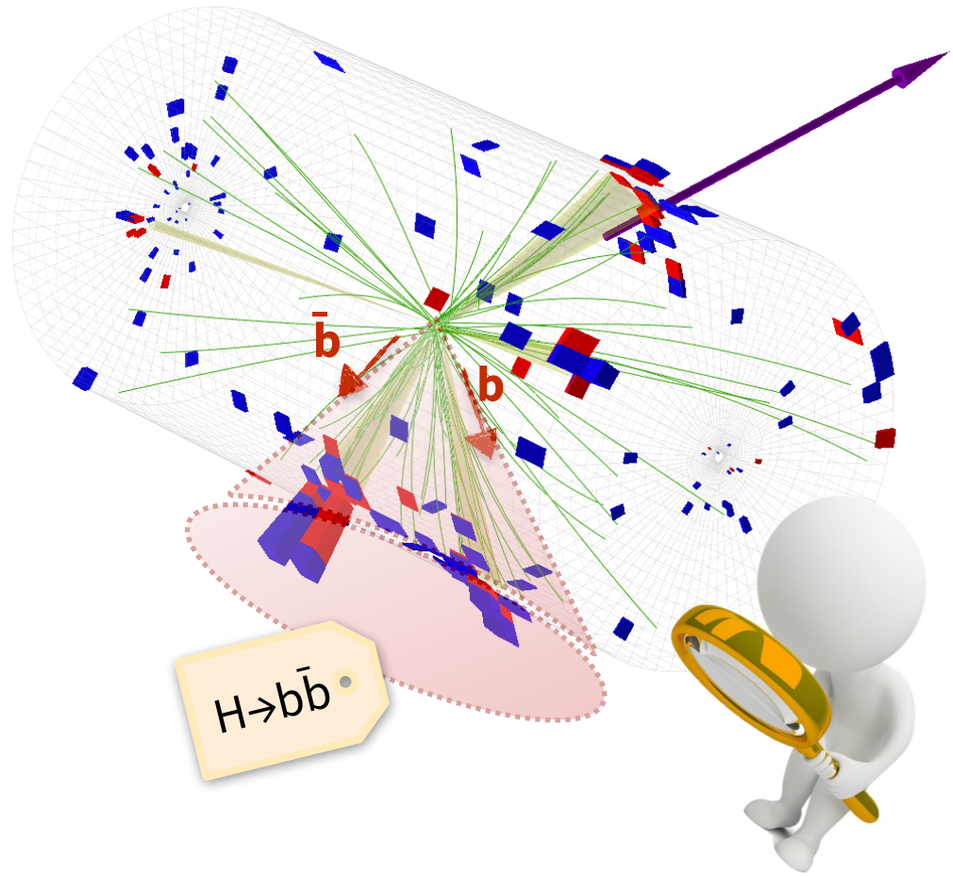
➔ Boosted topologies are crucial in LHC physics program

- highly Lorentz-boosted resonance
  → decay products are collimated



$$\Delta R \sim \frac{2m}{p_T}$$

A simulated Z(νν̄)H(bb̄) events in CMS



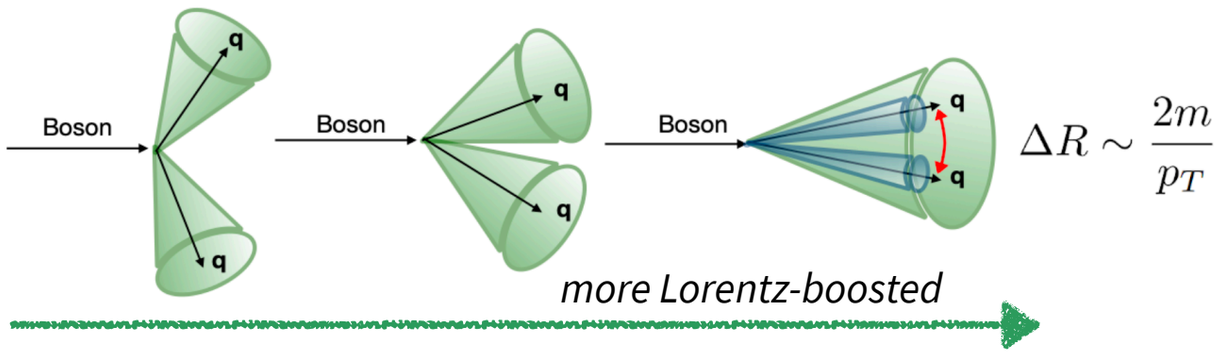
➔ Tagging the heavy resonance X to bb̄/cc̄ decay is an important technique for Higgs and BSM resonance search

- algorithms developed in CMS during Run 2:
  - ParticleNet-MD, DeepDoubleX, DeepAK8-MD, double-b
- calibrating the algorithm from data is a necessary step for physics measurements

Boosted heavy-flavour jet tagging

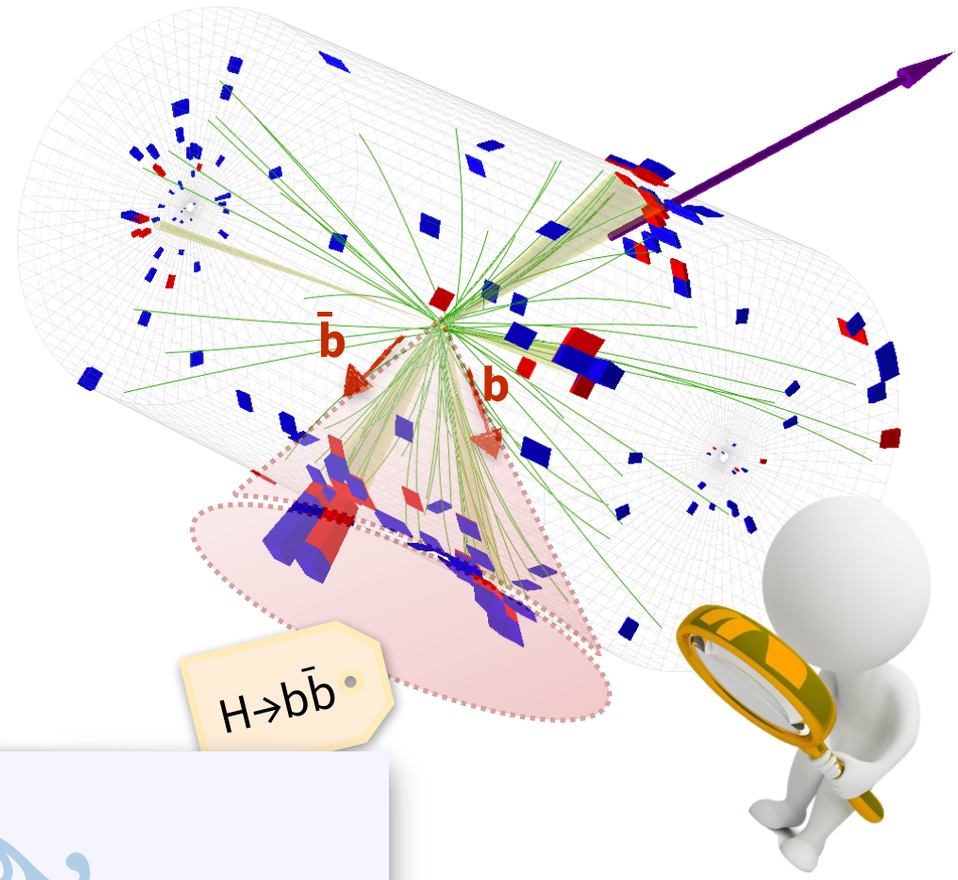
➔ Boosted topologies are crucial in LHC physics program

- ❖ highly Lorentz-boosted resonance
  → decay products are collimated



$$\Delta R \sim \frac{2m}{p_T}$$

A simulated Z(νν̄)H(bb̄) events in CMS



➔ Tagging the heavy resonance X to bb̄/cc̄ decay is an important technique for Higgs and BSM resonance

- ❖ algorith
  - Parti
    doub
- ❖ calibrati
  step for

Aim of this talk

- Introduce the X→bb̄/cc̄ taggers developed in CMS Run 2
- Highlight their performance on simulated jets
- Summarise and benchmark three calibration methods

NEW

Based on CMS-PAS-BTV-22-001

X→bb̄/cc̄ taggers: ParticleNet-MD

ParticleNet-MD | DeepDoubleX | DeepAK8-MD | double-b

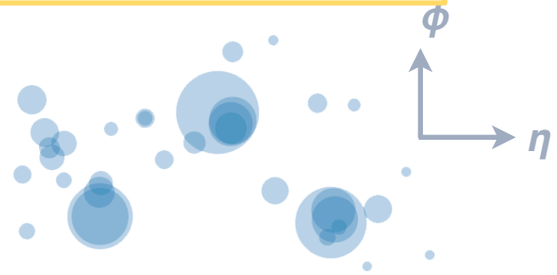
CMS-DP-2020-002

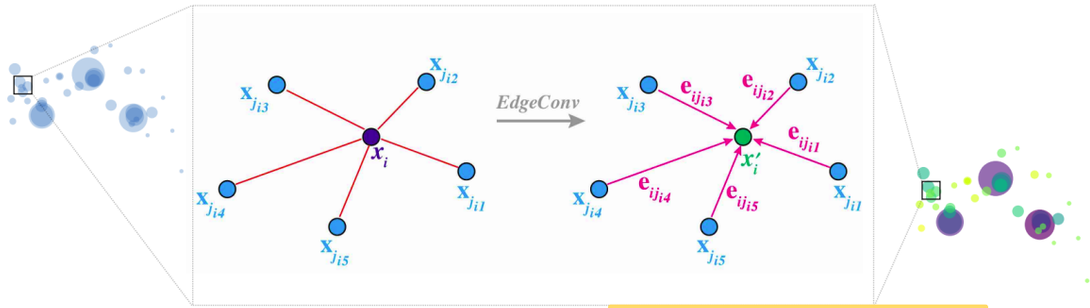
➔ Based on ParticleNet architecture

H. Qu, L. Gouskos, PRD 101, 056019 (2020)

- ❖ view jet as a point cloud
- ❖ local convolutional for message passing between neighbouring particles (EdgeConv)
- ❖ input: particle-flow (PF) candidate (i.e. jet constituents) + secondary vertex (SV) features

top jet image as a "point cloud"





EdgeConv operation

See full architecture in backup

X→bb̄/cc̄ taggers: ParticleNet-MD

ParticleNet-MD | DeepDoubleX | DeepAK8-MD | double-b

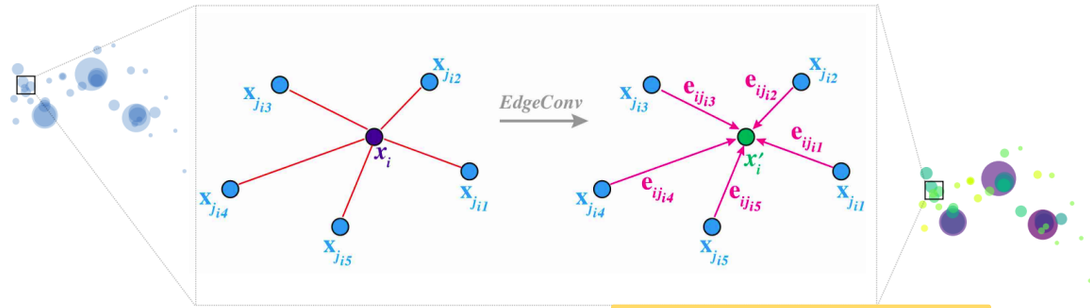
CMS-DP-2020-002

➔ Based on ParticleNet architecture

H. Qu, L. Gouskos, PRD 101, 056019 (2020)

- ❖ view jet as a point cloud
- ❖ local convolutional for message passing between neighbouring particles (EdgeConv)
- ❖ input: particle-flow (PF) candidate (i.e. jet constituents) + secondary vertex (SV) features

top jet image as a "point cloud"



EdgeConv operation

➔ The mass-decorrelated (MD) version

- ❖ trained on X→bb̄/cc̄/qq̄ signal and QCD background with flat and identical mass spectra

See full architecture in backup

Define discriminants:

$$\texttt{bbvsQCD} = \frac{p(\mathrm{X} \to \mathrm{b\bar{b}})}{p(\mathrm{X} \to \mathrm{b\bar{b}}) + p(\mathrm{QCD})}$$

$$\texttt{ccvsQCD} = \frac{p(\mathrm{X} \to \mathrm{c\bar{c}})}{p(\mathrm{X} \to \mathrm{c\bar{c}}) + p(\mathrm{QCD})}$$

X→bb̄/cc̄ taggers: ParticleNet-MD

ParticleNet-MD | DeepDoubleX | DeepAK8-MD | double-b

CMS-DP-2020-002

➔ Based on ParticleNet architecture

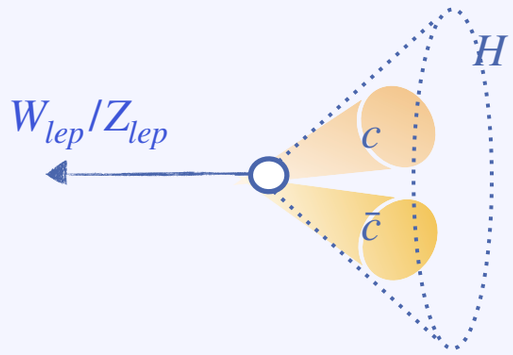
H. Qu, L. Gouskos, PRD 101, 056019 (2020)

- ❖ view jet as a point cloud

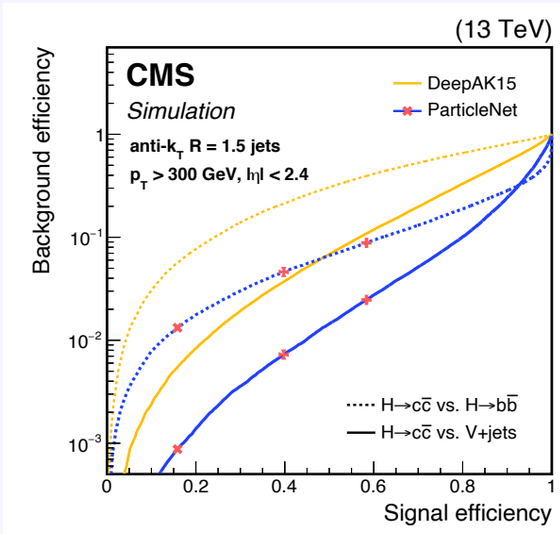
top jet image as a "point cloud"

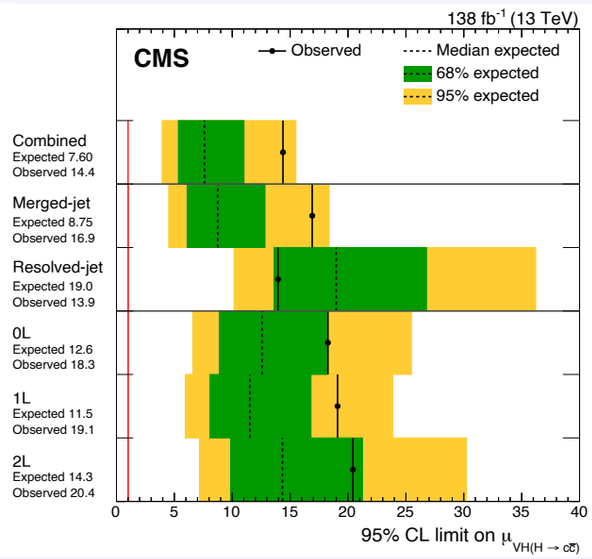
VH→cc search (merged region)

arXiv:2205.05550, accepted by PRL



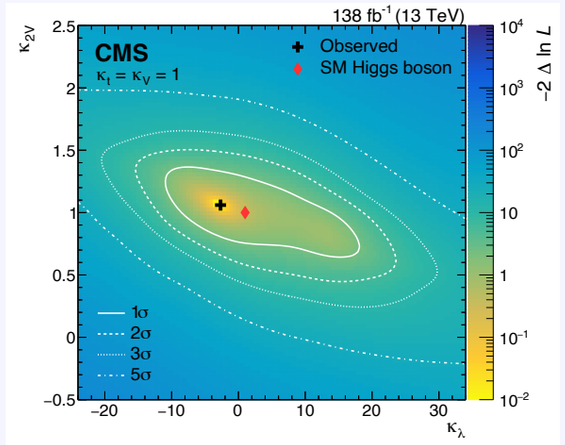
- With merged+resolved region combined, achieve most stringent direct limit on κ_c: $1.1 < |\kappa_c| < 5.5$

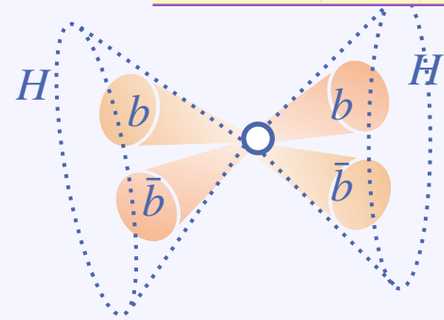




Boosted HH→4b search

PRL 131 (2023) 041803

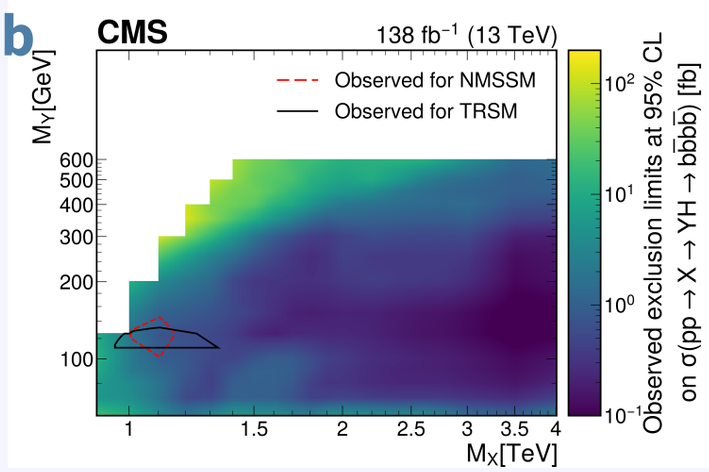




- First time excluding $\kappa_{2V} = 0$

Boosted X→YH→4b search

PLB 842 (2023) 137392



X→bb̄/cc̄ taggers: DeepDoubleX

ParticleNet-MD → **DeepDoubleX** → DeepAK8-MD → double-b

CMS-DP-2022-041

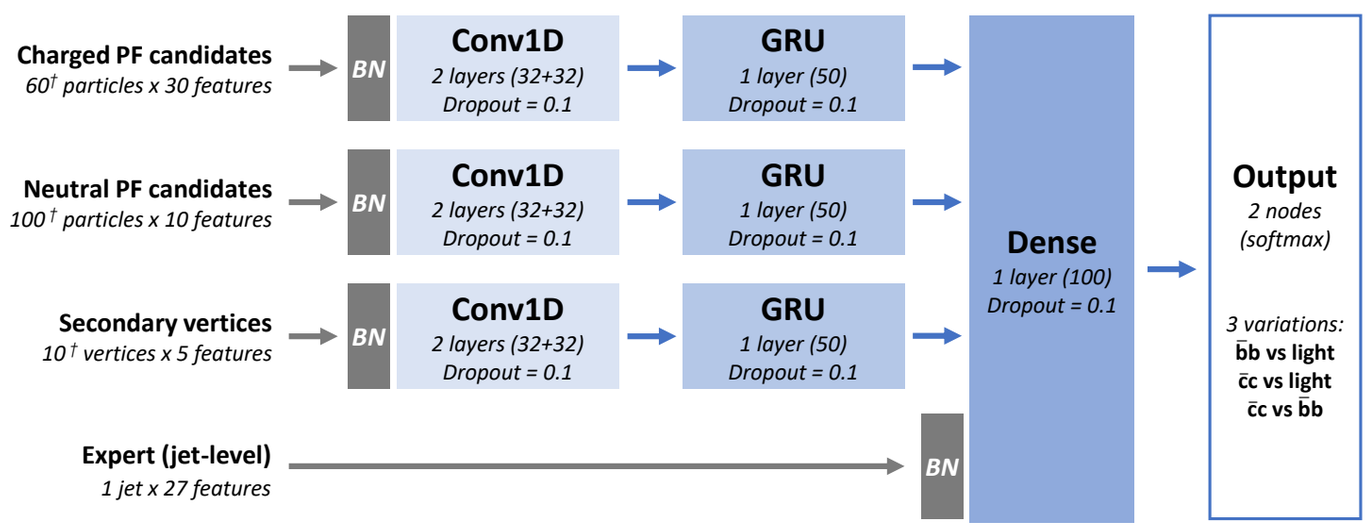
➔ DeepDoubleX:

- ❖ architecture: 1D CNN + RNN (with gated recurrent units, GRU)
- ❖ inputs: PF candidates + SV + jet global features
- ❖ irrelevant features pruned using layer-wise relevance propagation

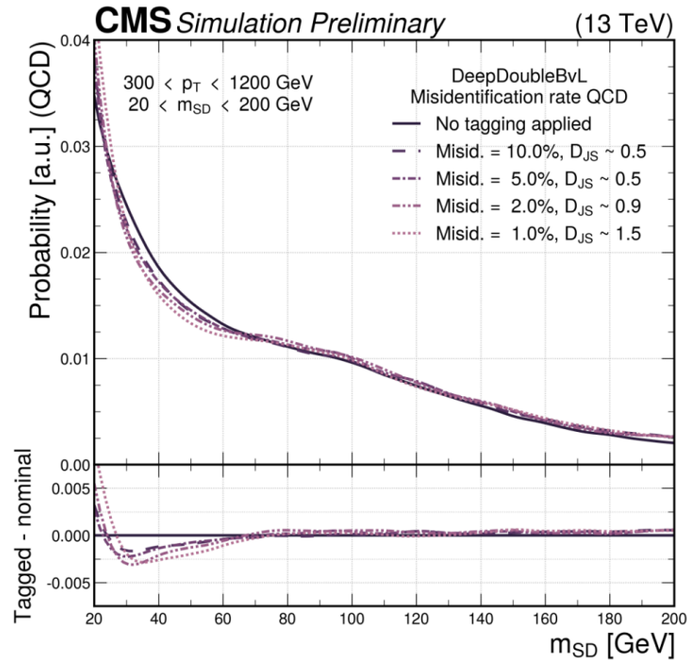
➔ Train three binary networks:

- ❖ X→bb̄ vs. QCD (BvL), X→cc̄ vs. QCD (CvL), X→bb̄ vs. X→cc̄ (CvB)
- ❖ studied tagger: BvL and CvL

NN architecture for DeepDoubleX



Mass decorrelation achieved by training on signal and background jets with uniform distribution



CMS-DP-2022-041

X→bb̄/cc̄ taggers: DeepDoubleX

ParticleNet-MD → **DeepDoubleX** → DeepAK8-MD → double-b

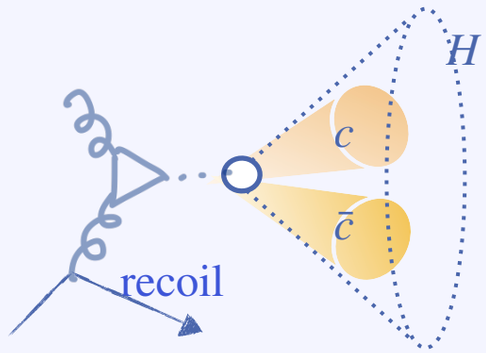
CMS-DP-2022-041

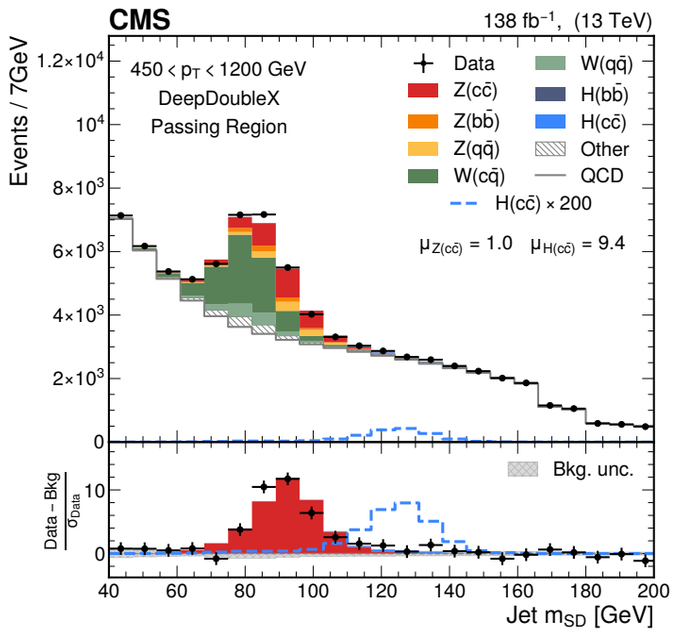
➔ DeepDoubleX:

- architecture: 1D CNN + RNN (with gated recurrent units, GRU)
- inputs: PF candidates + SV + jet global features

ggH→cc search (boosted region)

arXiv:2211.14181, accepted by PRL



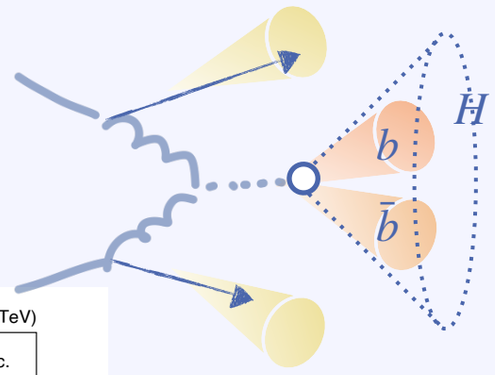


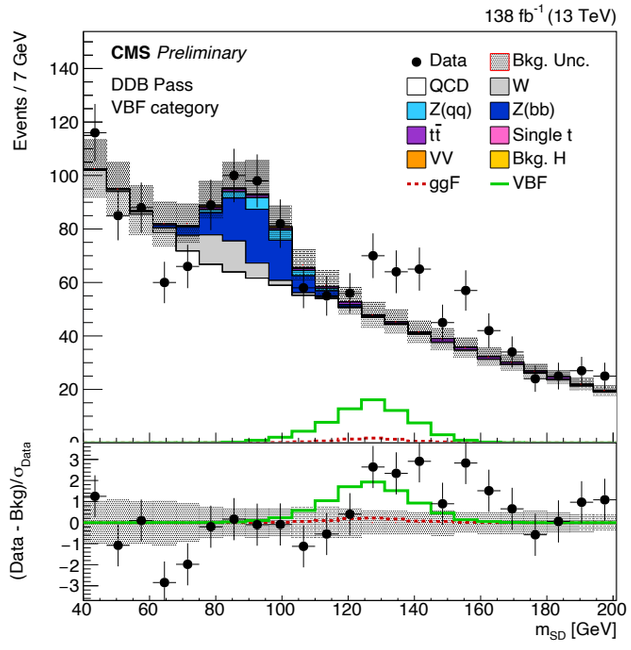
NEW

Boosted H→bb via VBF & ggF

CMS-PAS-HIG-21-020

See Jennet's talk tomorrow





CMS-DP-2022-041

X→bb̄/cc̄ taggers: DeepAK8-MD & double-b

ParticleNet-MD

DeepDoubleX

DeepAK8-MD

JINST 15 (2020) P06005

double-b

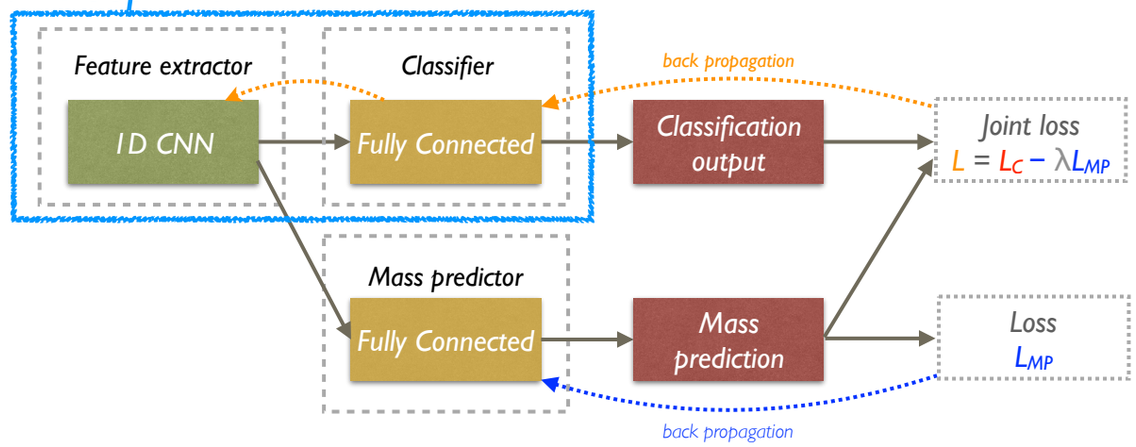
JINST 13 (2018) P05011

early Run-2 development

➔ DeepAK8-MD:

- architecture: 1D CNN
- inputs: same with ParticleNet-MD
- classes: t/H/Z/W with dedicated final states
- mass decorrelation: achieved by adversarial training

Nominal DeepAK8



Define discriminants:

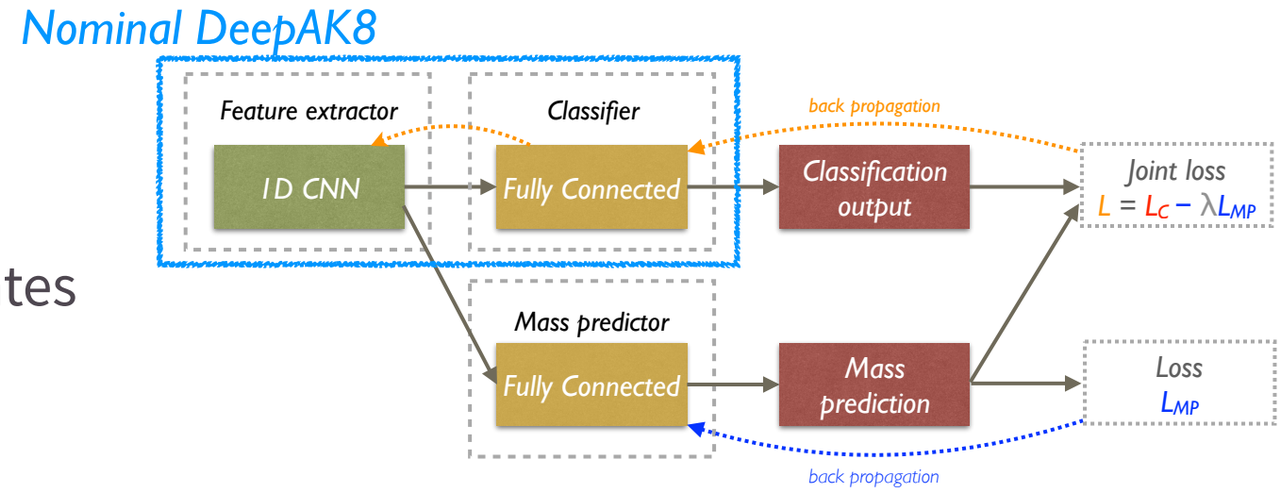
$$\texttt{bbvsQCD} = \frac{p(\mathrm{H} \to \mathrm{b\bar{b}}) + p(\mathrm{Z} \to \mathrm{b\bar{b}})}{p(\mathrm{H} \to \mathrm{b\bar{b}}) + p(\mathrm{Z} \to \mathrm{b\bar{b}}) + p(\mathrm{QCD})}$$

$$\texttt{ccvsQCD} = \frac{p(\mathrm{H} \to \mathrm{c\bar{c}}) + p(\mathrm{Z} \to \mathrm{c\bar{c}})}{p(\mathrm{H} \to \mathrm{c\bar{c}}) + p(\mathrm{Z} \to \mathrm{c\bar{c}}) + p(\mathrm{QCD})}$$

X→bb̄/cc̄ taggers: DeepAK8-MD & double-b

➔ DeepAK8-MD:

- ❖ architecture: 1D CNN
- ❖ inputs: same with ParticleNet-MD
- ❖ classes: t/H/Z/W with dedicated final states
- ❖ mass decorrelation: achieved by adversarial training



➔ double-b:

- ❖ a BDT for distinguishing H→bb̄ and QCD jets
- ❖ inputs: jet variables constructed from tracks and SVs
- ❖ mass decorrelation: by choosing input with weak correlation on p_T and mass

Define discriminants:

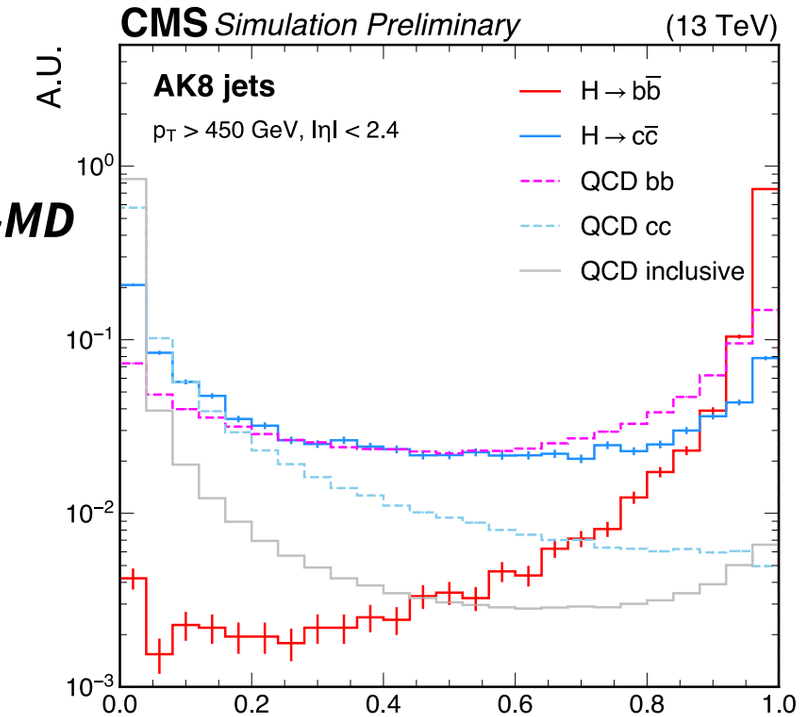
$$\texttt{bbvsQCD} = \frac{p(\mathrm{H}\to \mathrm{b\bar{b}}) + p(\mathrm{Z}\to \mathrm{b\bar{b}})}{p(\mathrm{H}\to \mathrm{b\bar{b}}) + p(\mathrm{Z}\to \mathrm{b\bar{b}}) + p(\mathrm{QCD})}$$

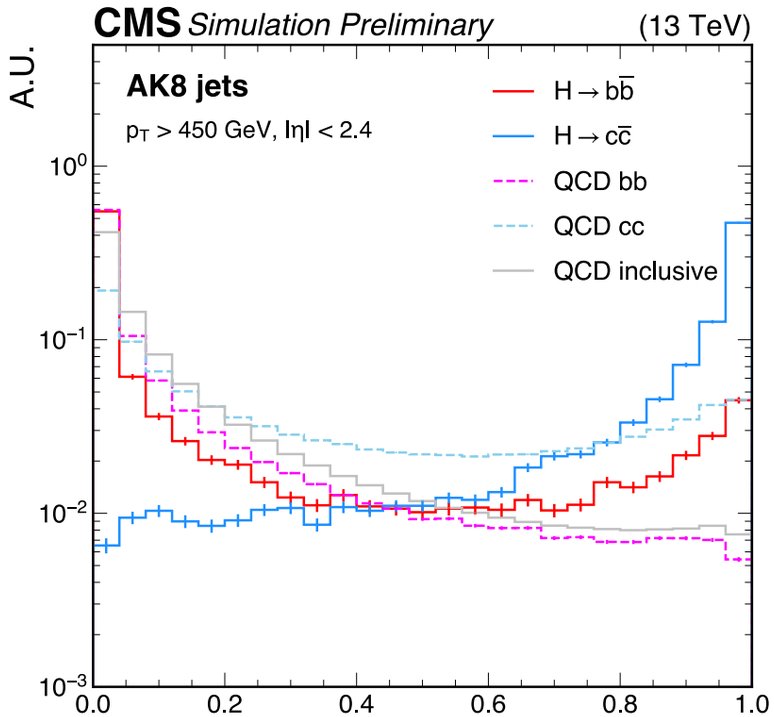
$$\texttt{ccvsQCD} = \frac{p(\mathrm{H}\to \mathrm{c\bar{c}}) + p(\mathrm{Z}\to \mathrm{c\bar{c}})}{p(\mathrm{H}\to \mathrm{c\bar{c}}) + p(\mathrm{Z}\to \mathrm{c\bar{c}}) + p(\mathrm{QCD})}$$

Tagger discriminants spectra

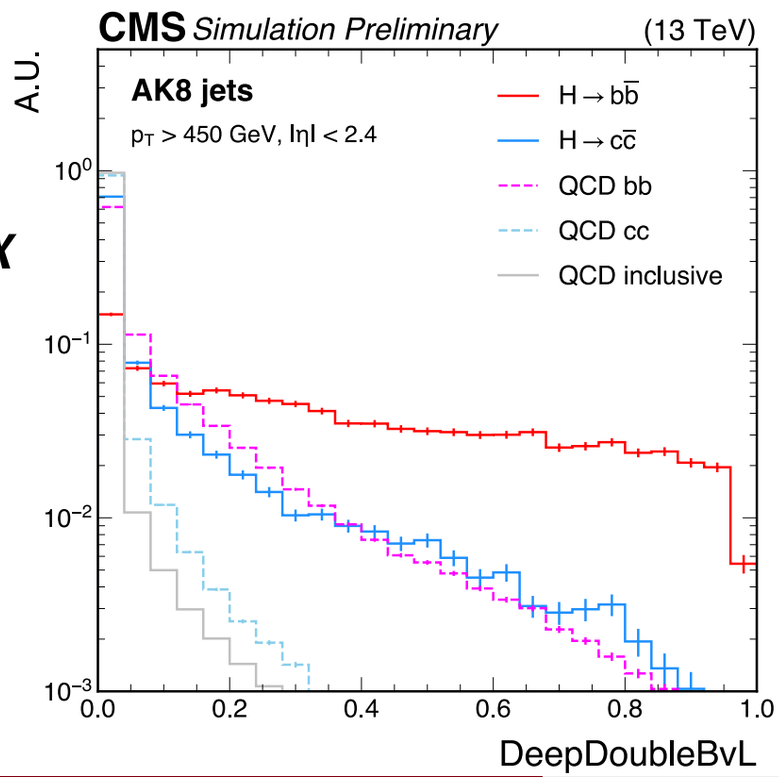
CMS-PAS-BTV-22-001

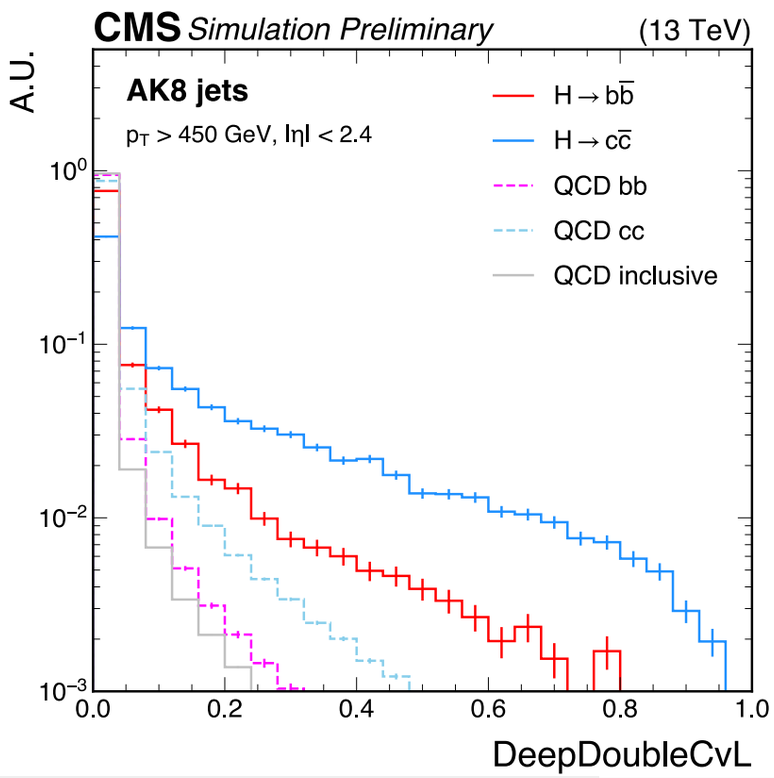
ParticleNet-MD





DeepDoubleX





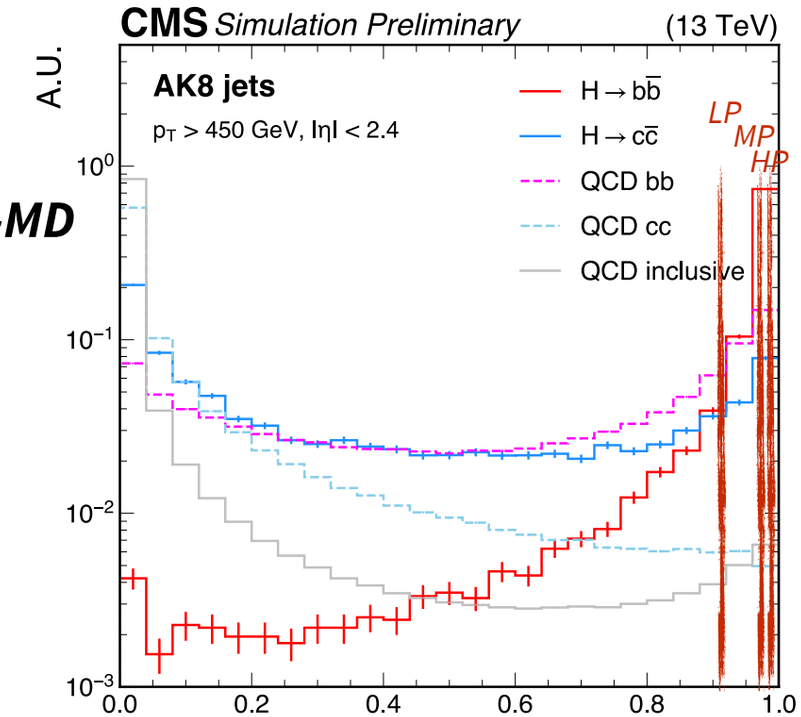
- **Signal: H→bb̄,** **H→cc̄** jets from ggH
- BKG: QCD jets
  - **QCD bb (cc)**: number of ghost-matched b and c hadrons: N_b=2 (N_b=0 & N_c=2)

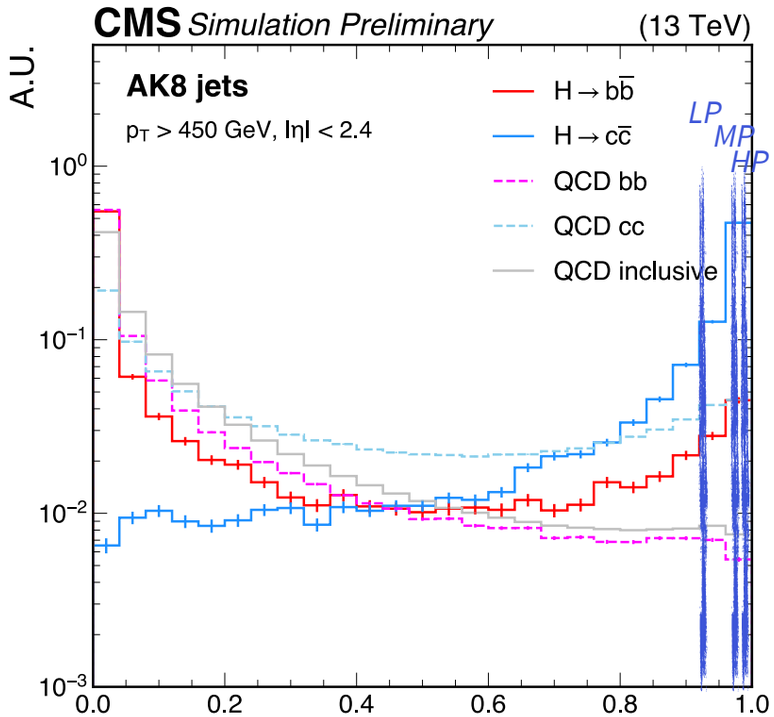
More taggers in backup

Tagger discriminants spectra

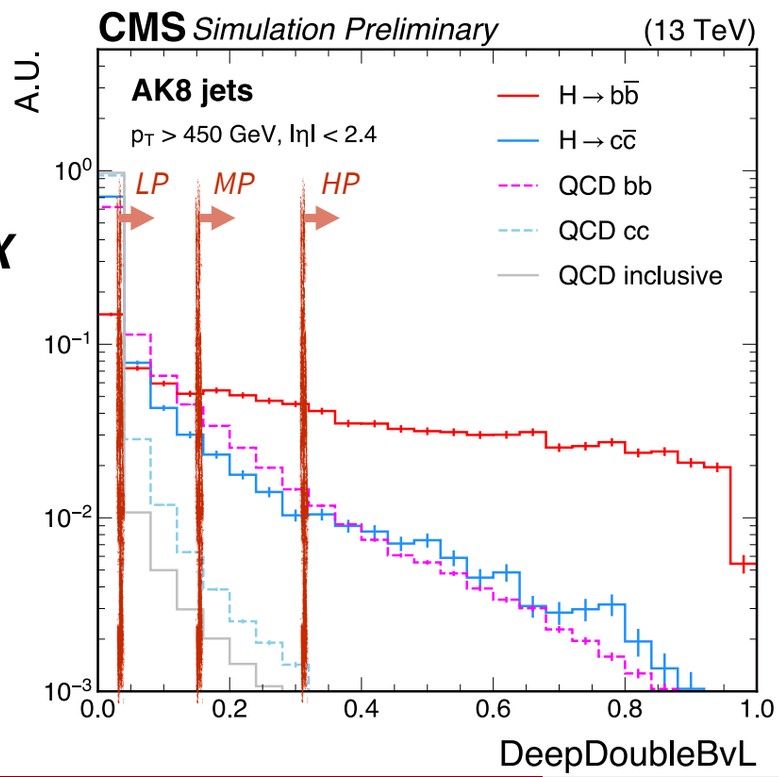
CMS-PAS-BTV-22-001

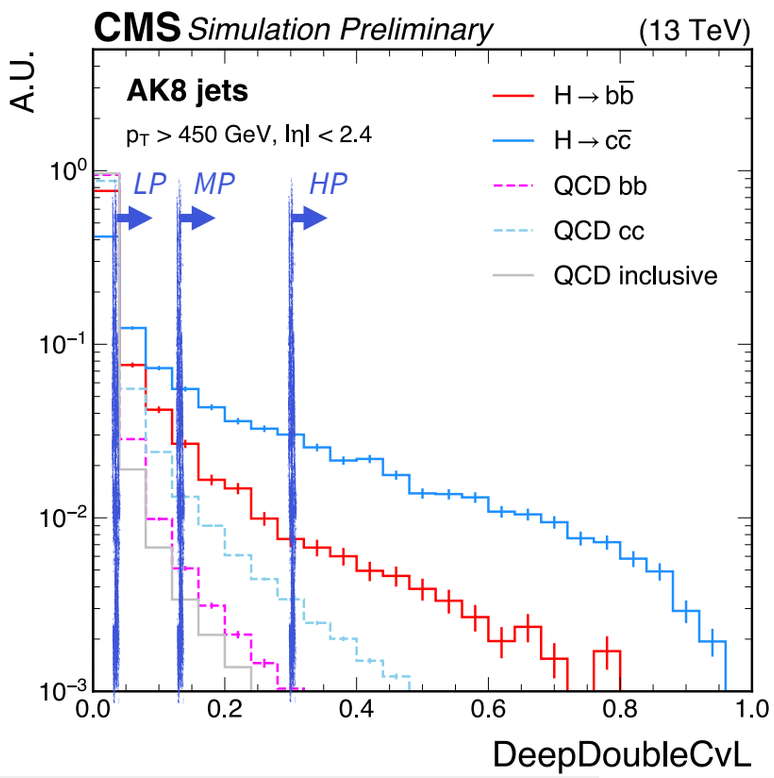
ParticleNet-MD





DeepDoubleX





- **Signal: H→bb̄,** **H→cc̄** jets from ggH
- BKG: QCD jets
  - **QCD bb (cc)**: number of ghost-matched b and c hadrons: N_b=2 (N_b=0 & N_c=2)

Determine working points (WPs) based on **H→bb̄,** **H→cc̄** jets efficiency

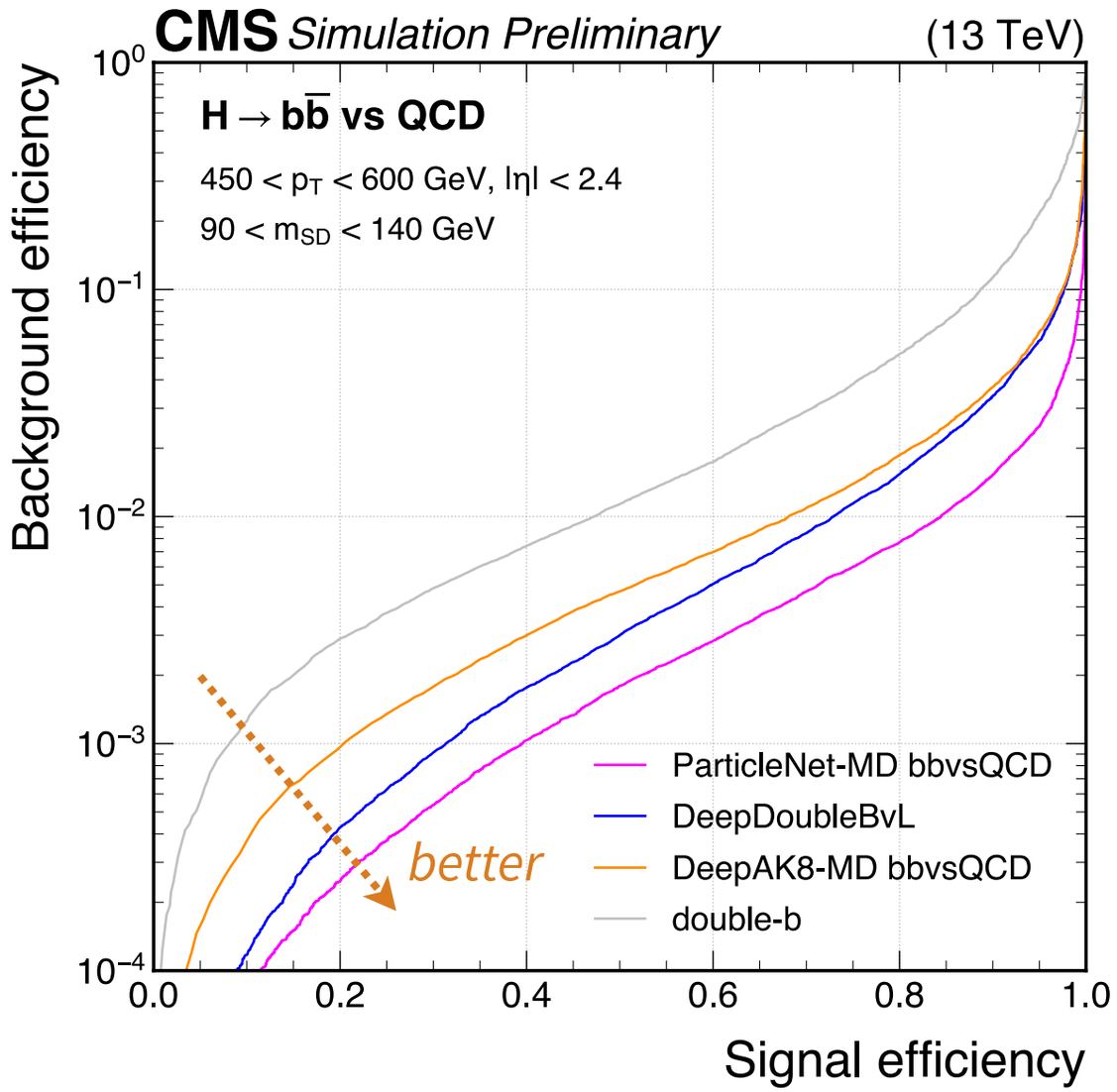
disc.	HP high purity	MP medium purity	LP low purity
X→bb̄	**40%**	**60%**	**80%**
X→cc̄	**15%**	**30%**	**50%**

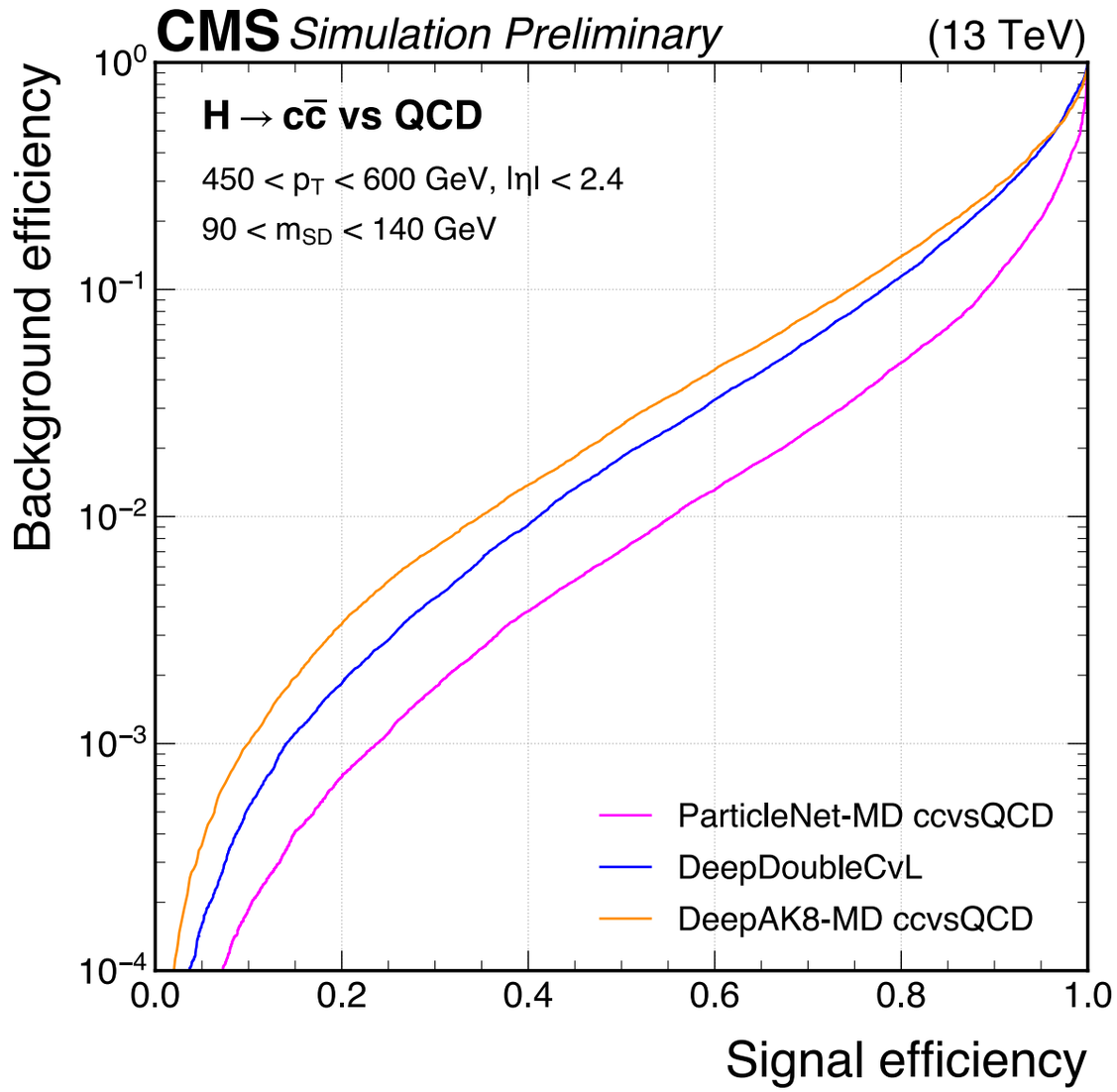
More taggers in backup

Performance on simulation

CMS-PAS-BTV-22-001

Low p_T: (450, 600) GeV, similar performance for high p_T (see backup)



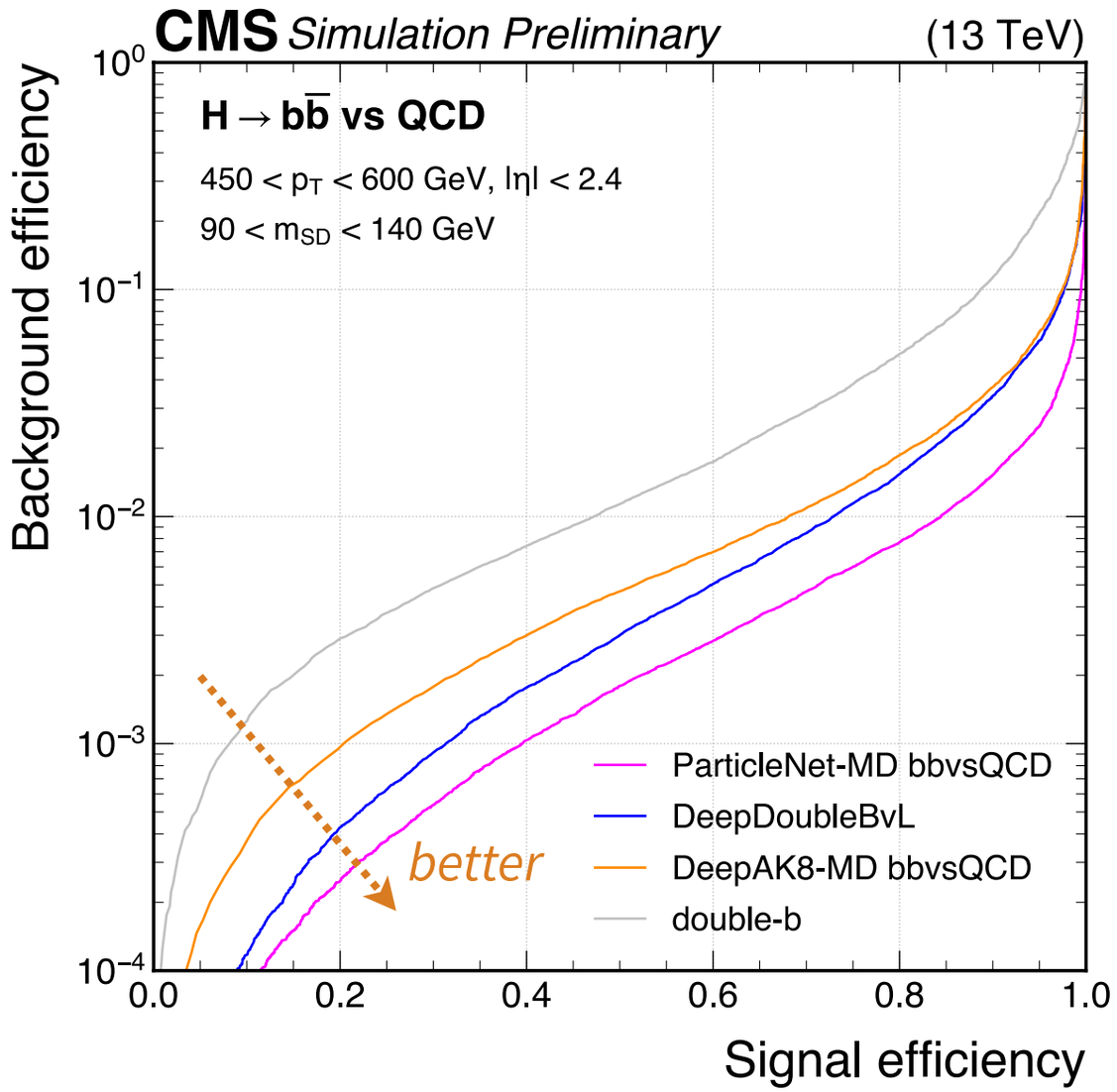


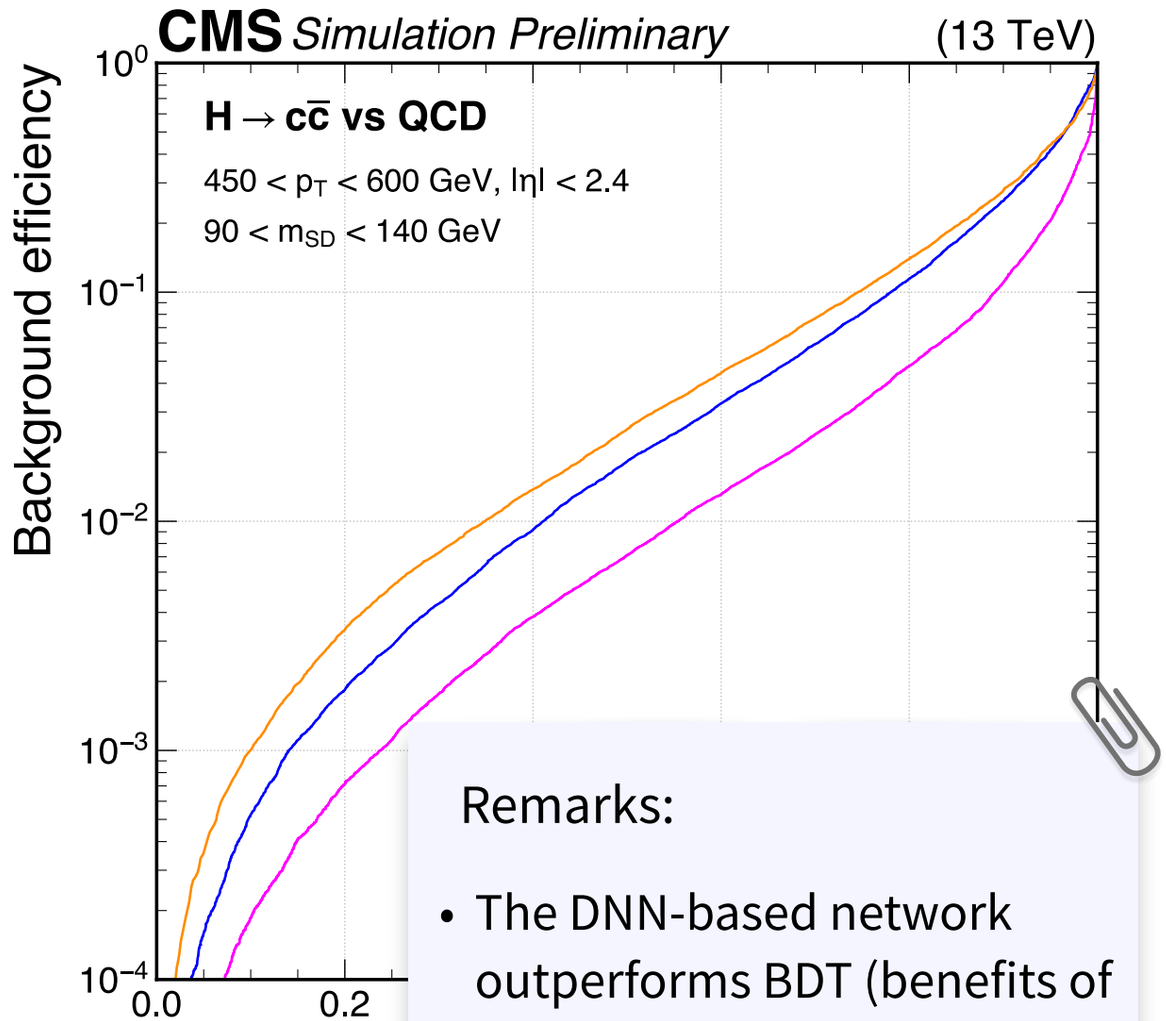
ROC comparison on H→bb̄ (cc̄) jets vs. inclusive QCD

Performance on simulation

CMS-PAS-BTV-22-001

Low p_T: (450, 600) GeV, similar performance for high p_T (see backup)





ROC comparison on H→bb̄ (cc̄) jets vs. inclusive QCD

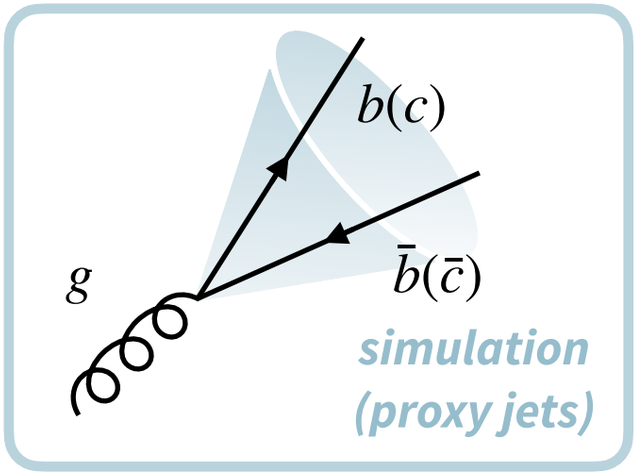
Remarks:

- The DNN-based network outperforms BDT (benefits of low-level features)
- DeepAK8-MD → DeepDoubleX → ParticleNet-MD: reflects advancement made in NN architecture

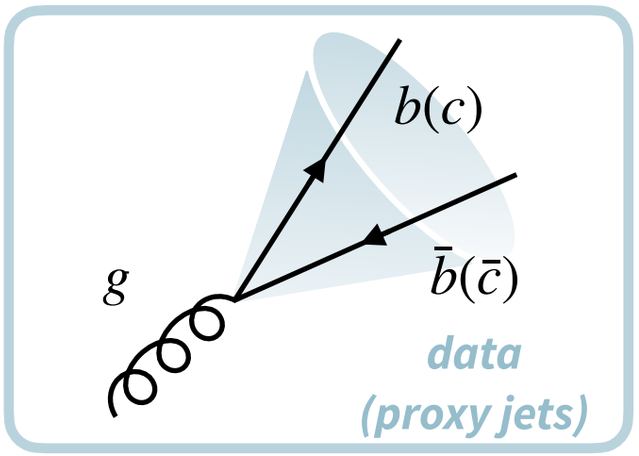
Three calibration methods

- sfBDT method
- μ-tagged method
- boosted Z method

sfBDT method: key concept

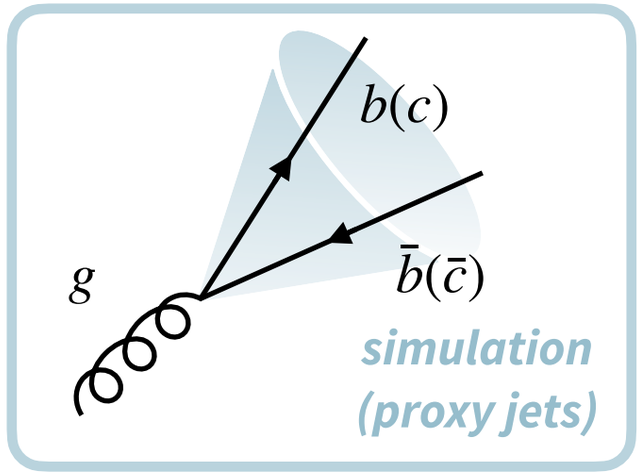


derive scale factor (SF)
then used as SF for H→b$\bar{b}$/c$\bar{c}$ jets



The general idea is to select **g→b$\bar{b}$/c$\bar{c}$** jets from QCD events as the proxy of **H→b$\bar{b}$/c$\bar{c}$**

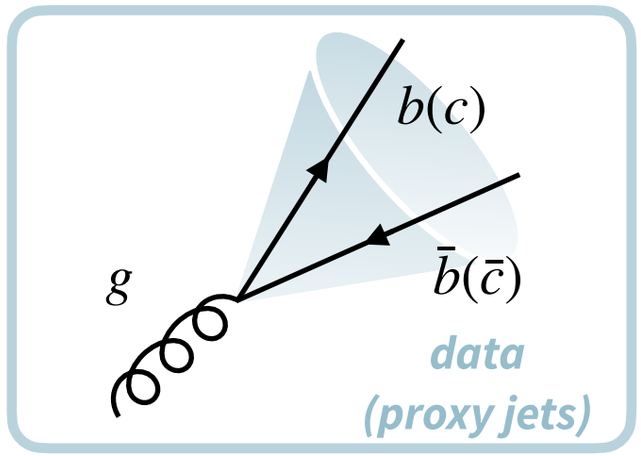
sfBDT method: key concept



selected jets from QCD multijet events

(requiring 2 SVs)

derive scale factor (SF)
then used as SF for H→bb̄/cc̄ jets

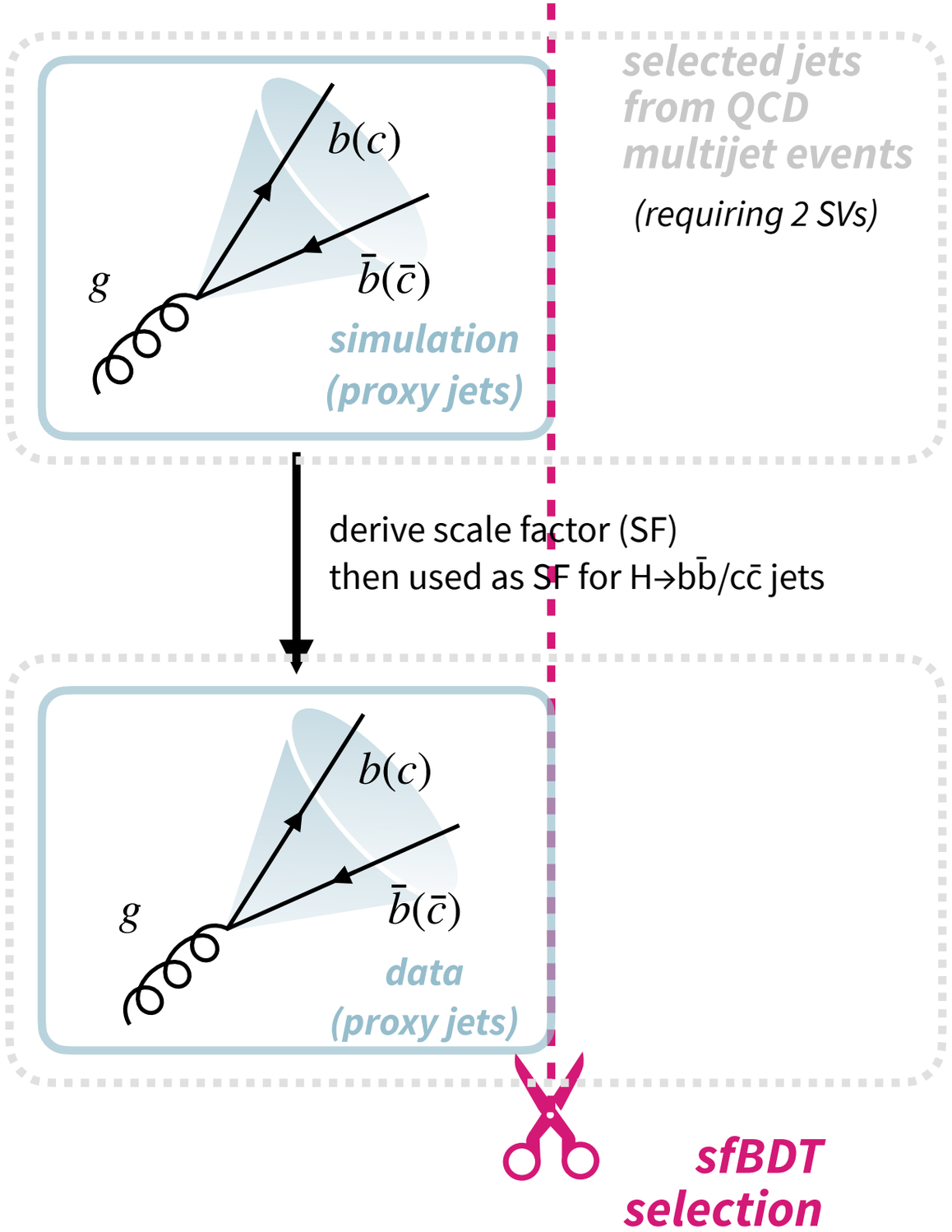


The general idea is to select **g→bb̄/cc̄** jets from QCD events as the proxy of **H→bb̄/cc̄**

Problem:
Preselected jets from QCD events, enriched in **g→bb̄/cc̄**, still do not closely resemble **H→bb̄/cc̄** jets.

→ Further selection needed

sfBDT method: key concept

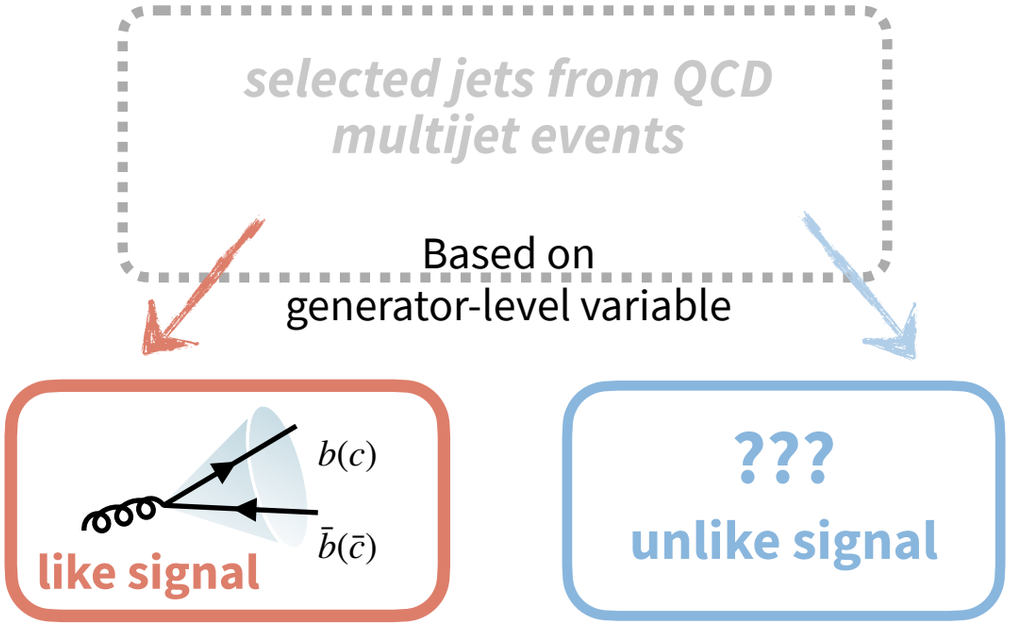


The general idea is to select **g→bb̄/cc̄** jets from QCD events as the proxy of **H→bb̄/cc̄**

Problem:
Preselected jets from QCD events, enriched in **g→bb̄/cc̄**, still do not closely resemble **H→bb̄/cc̄** jets.
→ Further selection needed

Solution:
Train a BDT variable which selects more signal-like jets from QCD

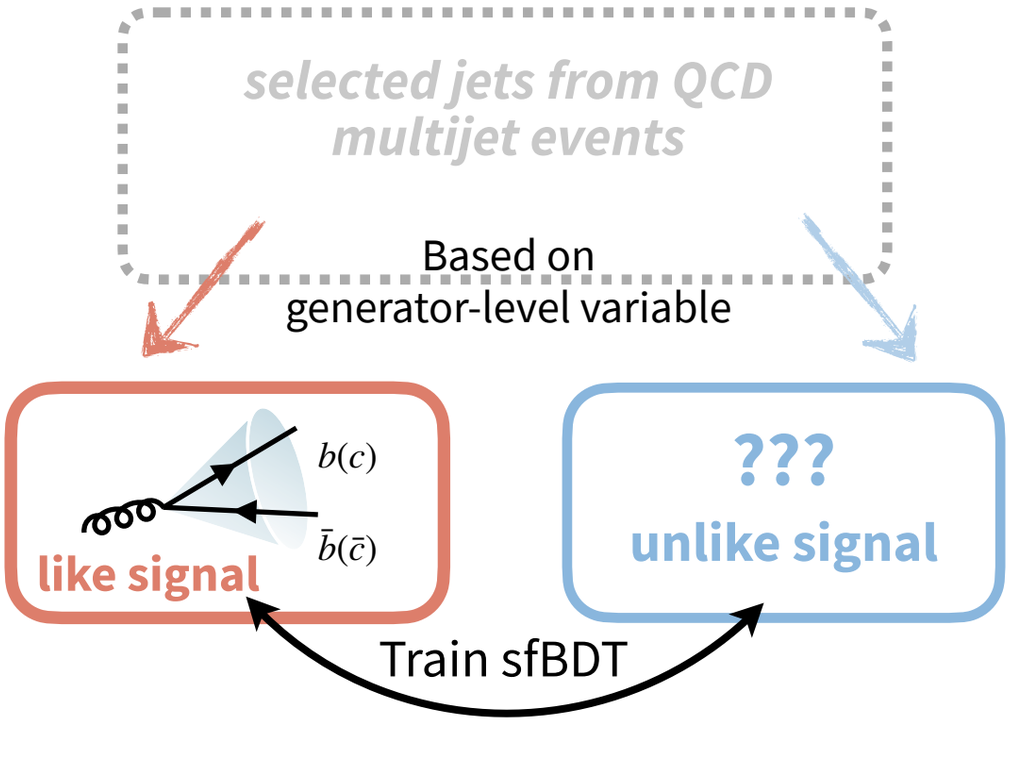
sfBDT method: design of sfBDT



Select two regions using a generator-level variable

- one resembles more signal **H→b$\bar{b}$/c$\bar{c}$** jets, one does not

sfBDT method: design of sfBDT

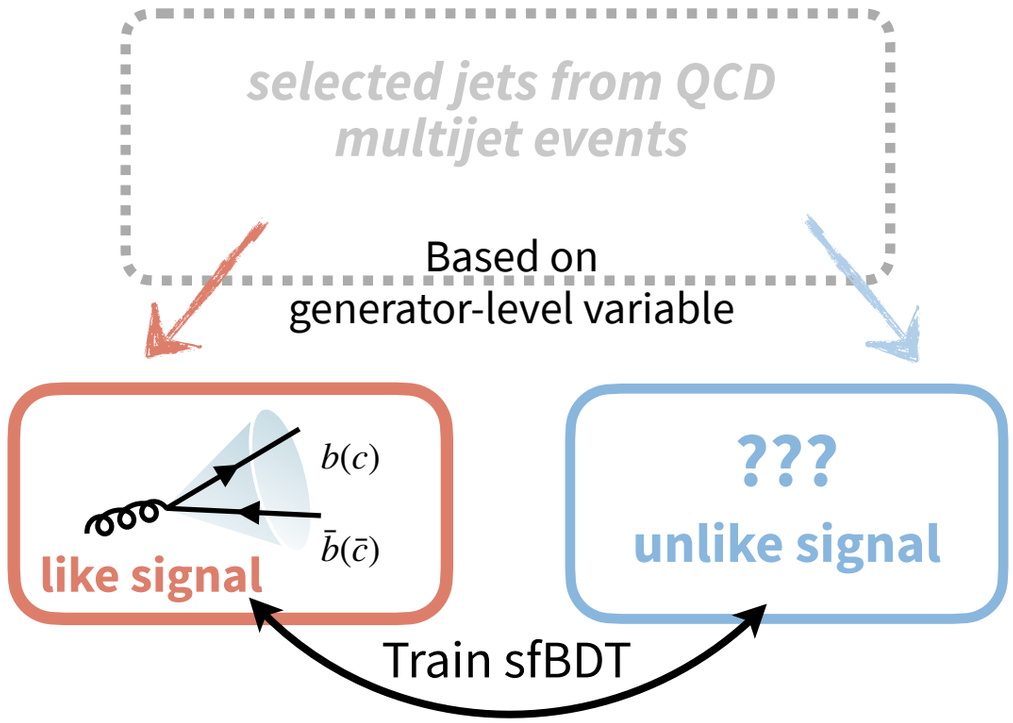


Select two regions using a generator-level variable

- one resembles more signal **H→b$\bar{b}$/c$\bar{c}$** jets, one does not

Train a BDT (named sfBDT) against each other, using reconstructed-level variables as input
(jet N-subjettiness, track/SV features)

sfBDT method: design of sfBDT



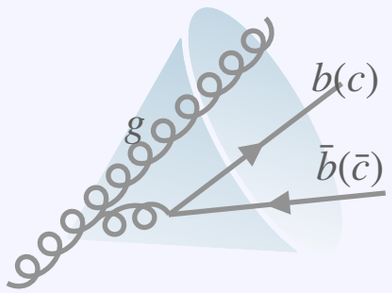
Select two regions using a generator-level variable

- one resembles more signal **H→b$\bar{b}$/c$\bar{c}$** jets, one does not

Train a BDT (named sfBDT) against each other, using reconstructed-level variables as input
(jet N-subjettiness, track/SV features)

Designing the generator-level variable:

Design 1: Published in CMS-DP-2022-005



- QCD jets are likely to be contaminated with extra gluons
- Define variables from **partons**

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

Design 2: this talk CMS-PAS-BTV-22-001

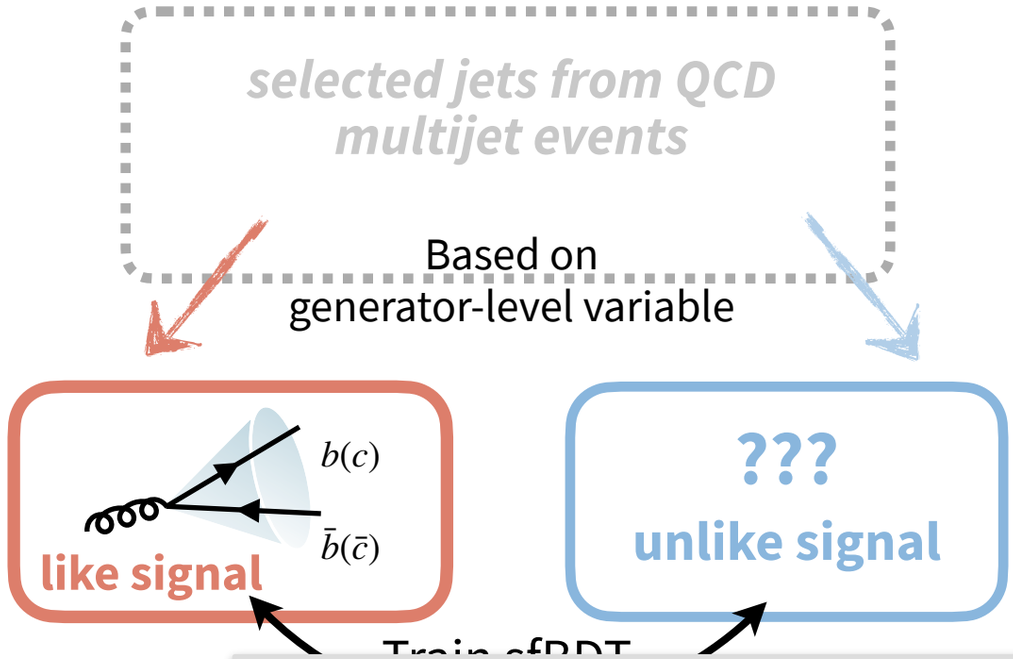
NEW

- Consider a way to take into account all q/g emissions
- Define variables from the list of the **first generation of hadrons** (before hadron decays)

Find the N-subjettiness τ_{31} (on hadrons) helpful to evaluate the multi-prongness of a QCD jet

smaller τ_{31} → less complexity in prongness → signal-like

sfBDT method: design of sfBDT



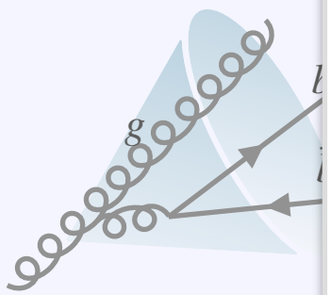
Select two regions using a generator-level variable

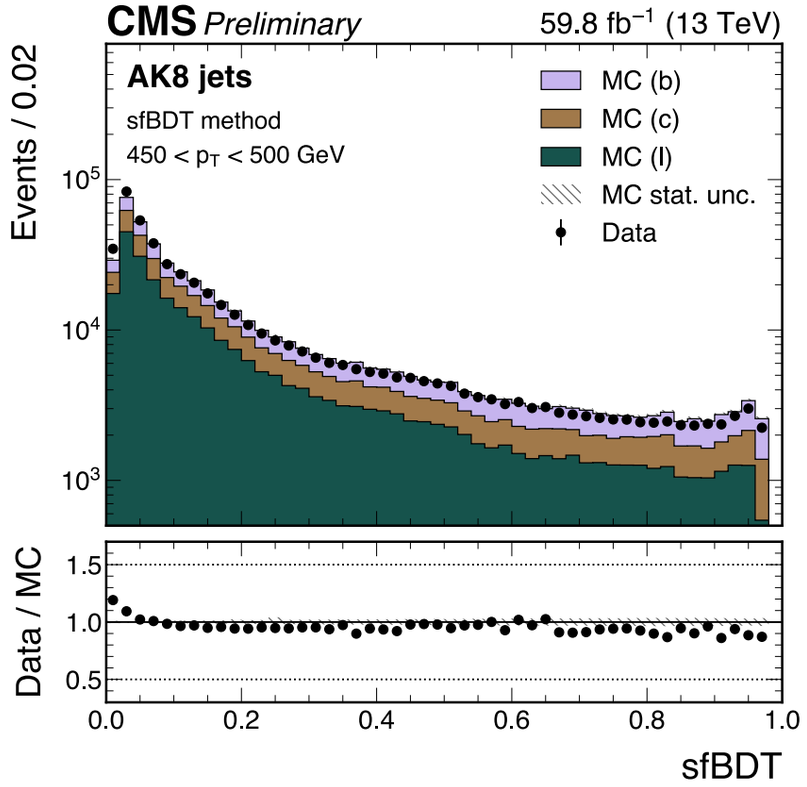
- one resembles more signal **H→b$\bar{b}$/c$\bar{c}$** jets, one does not

Train a BDT (named sfBDT) against each other, using reconstructed-level variables as input
(jet N-subjettiness, track/SV features)

Designing th

Design 1:



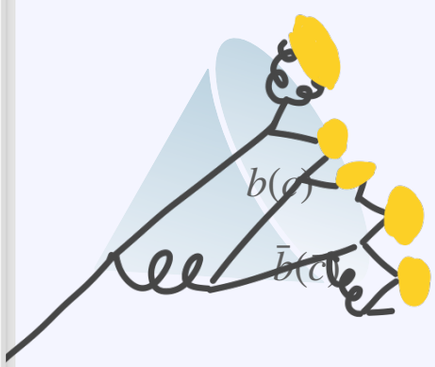


MC categorised to b/c/l based on the number of ghost-matched b and c hadrons

Design 2:

this talk CMS-PAS-BTV-22-001

NEW



- Consider a way to take into account all q/g emissions
- Define variables from the list of the **first generation of hadrons** (before hadron decays)

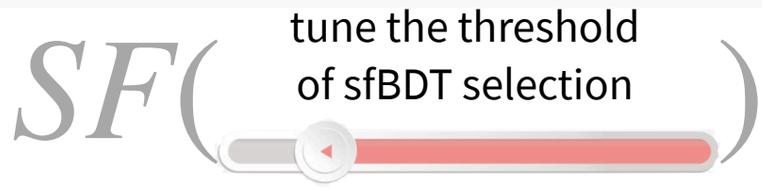
Find the N-subjettiness τ_{31} (on hadrons) helpful to evaluate the multi-prongness of a QCD jet

smaller τ_{31} → less complexity in prongness → signal-like

sfBDT method: selection via sfBDT

Key logic:

- sfBDT as a handle to “tune” proxy—signal similarity
- evaluate the SF’s dependence on sfBDT selection - extract an extra uncertainty from it



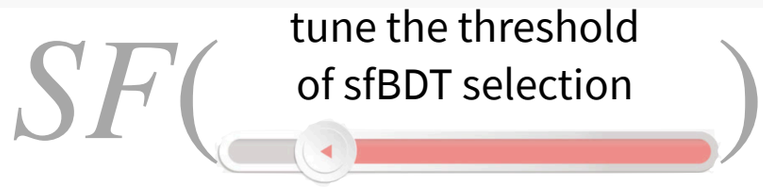
sfBDT method: selection via sfBDT

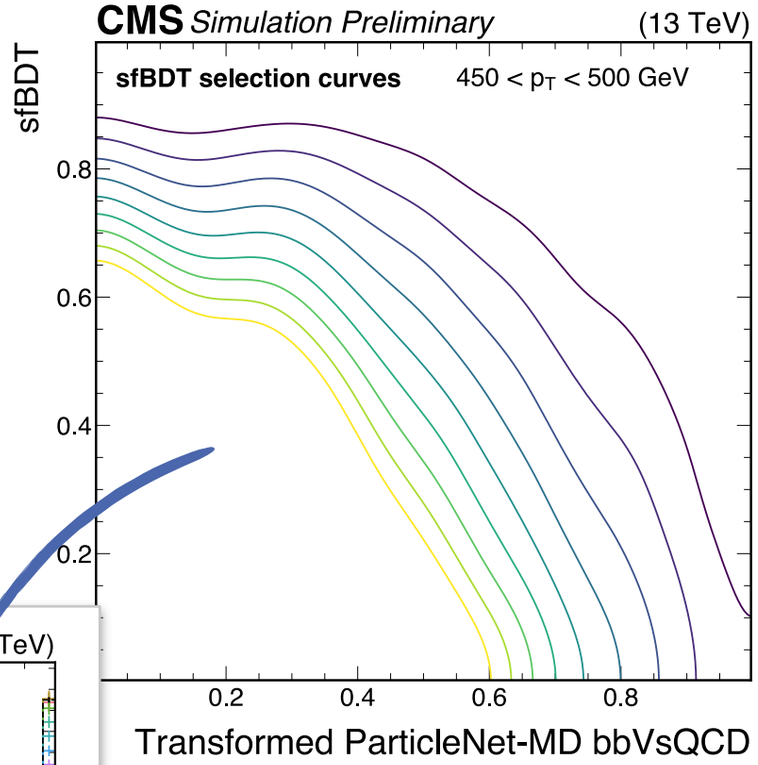
Key logic:

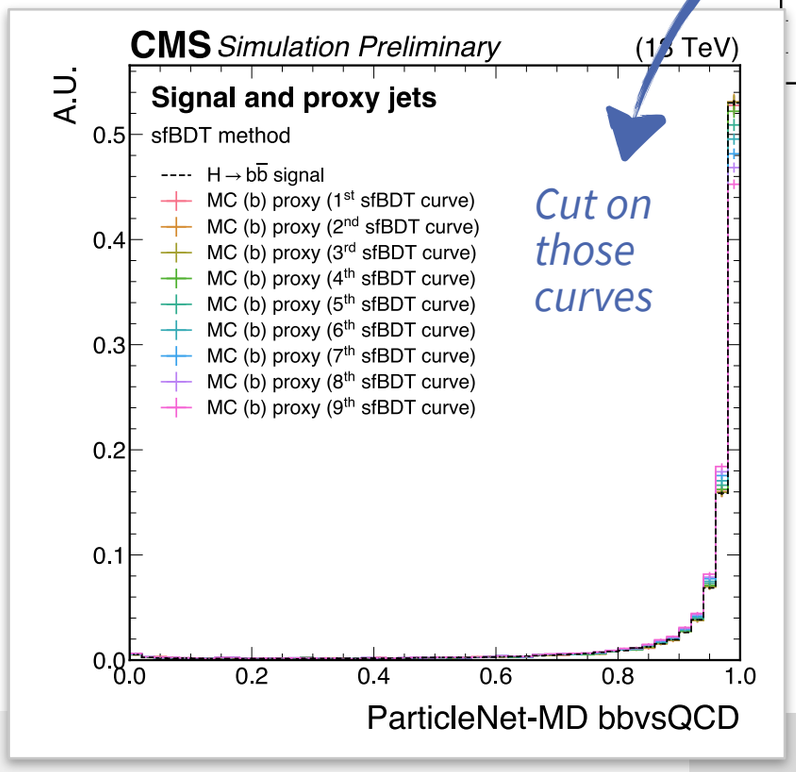
- sfBDT as a handle to "tune" proxy—signal similarity
- evaluate the SF's dependence on sfBDT selection - extract an extra uncertainty from it

Design ***a systematic*** workflow that applies to all taggers

- step1: transform the tagger to a uniform shape
- step2: define **a set of sfBDT curves** on the sfBDT-tagger plane, that match the proxy tagger shape to the signal (but with different selection efficiency)







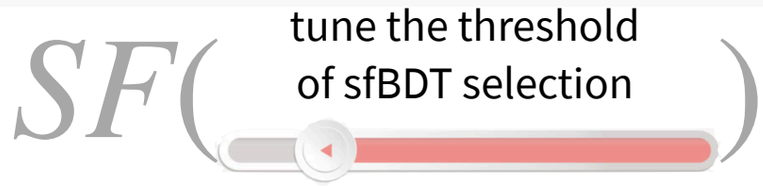
sfBDT method: selection via sfBDT

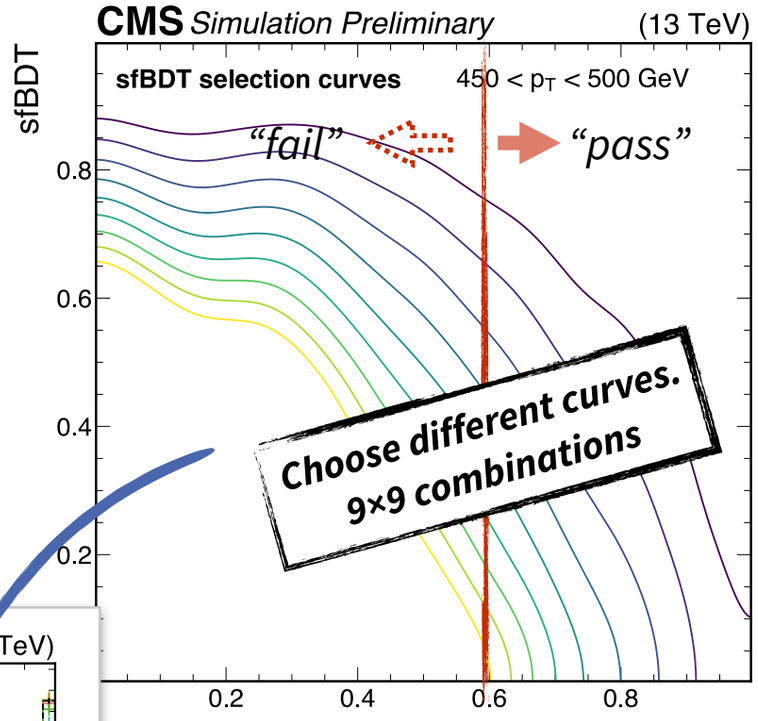
Key logic:

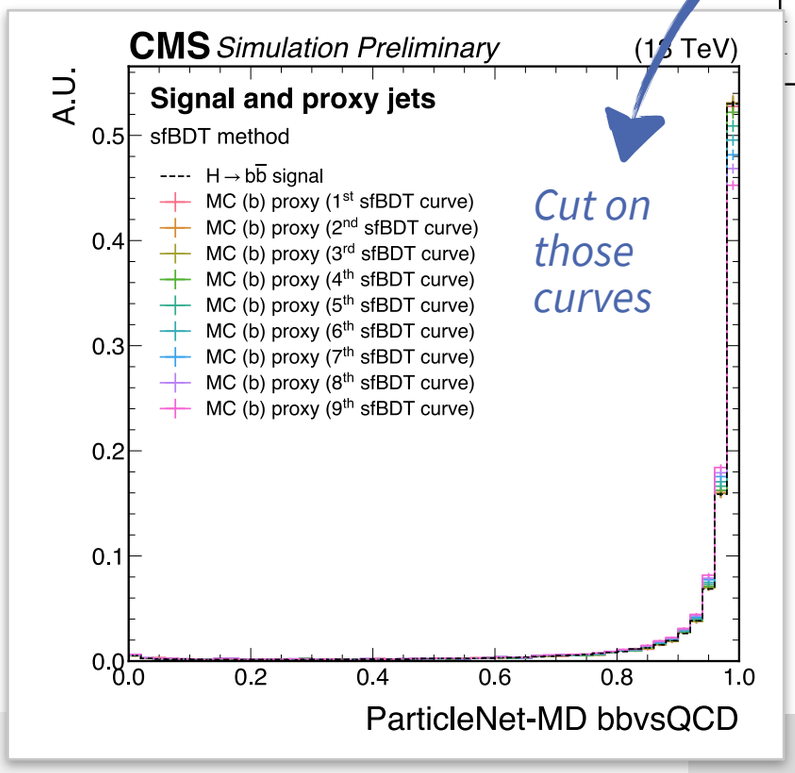
- sfBDT as a handle to "tune" proxy—signal similarity
- evaluate the SF's dependence on sfBDT selection - extract an extra uncertainty from it

Design ***a systematic*** workflow that applies to all taggers

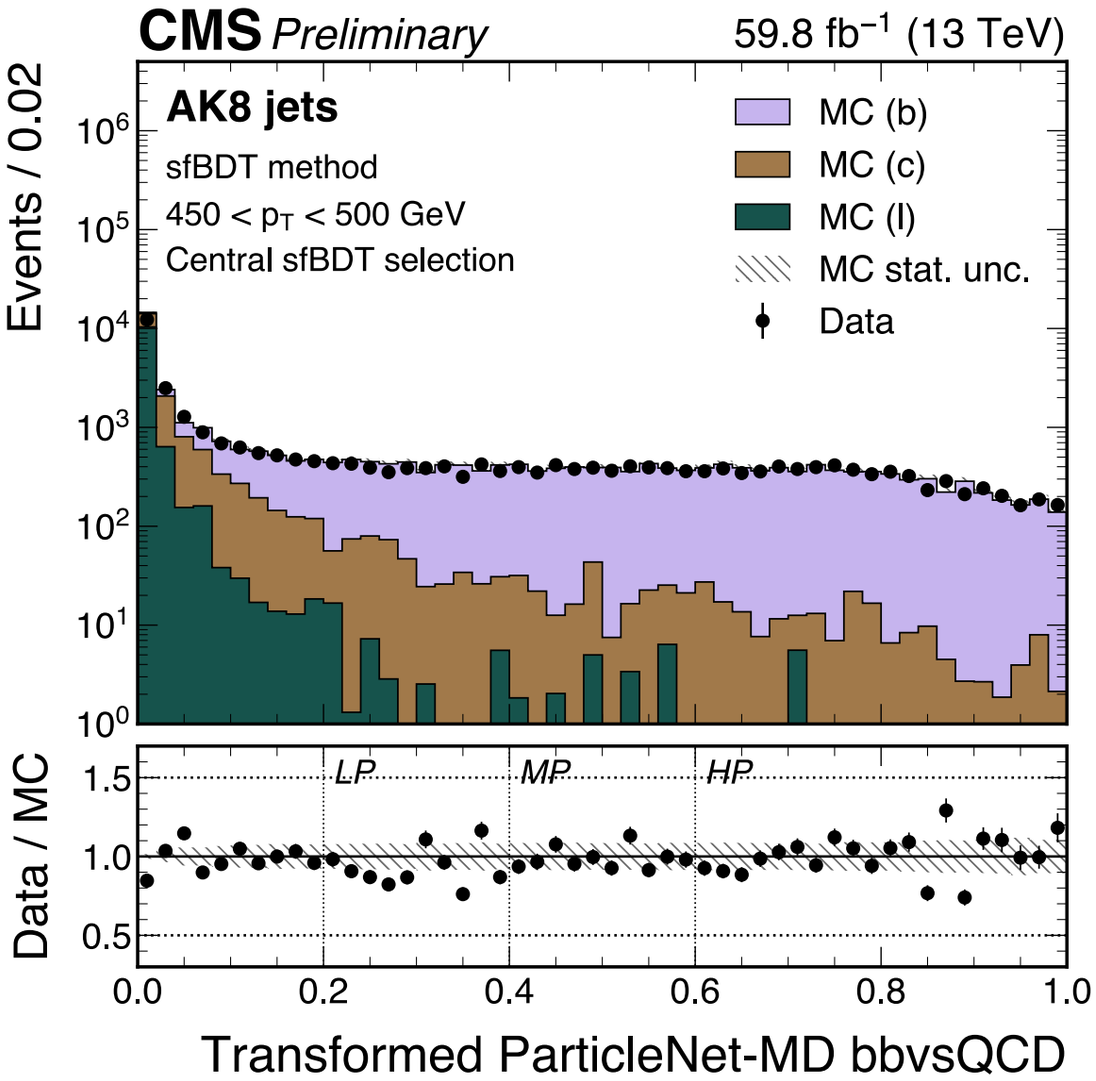
- step1: transform the tagger to a uniform shape
- step2: define **a set of sfBDT curves** on the sfBDT-tagger plane, that match the proxy tagger shape to the signal (but with different selection efficiency)
- step3: select on those curves to derive nominal SFs; **select different curves for the "pass" and "fail" tagger** region to modify the proxy—signal tagger shape similarity
- step4: from the deviation of SFs, we obtain an external uncertainty assigned to the final SF

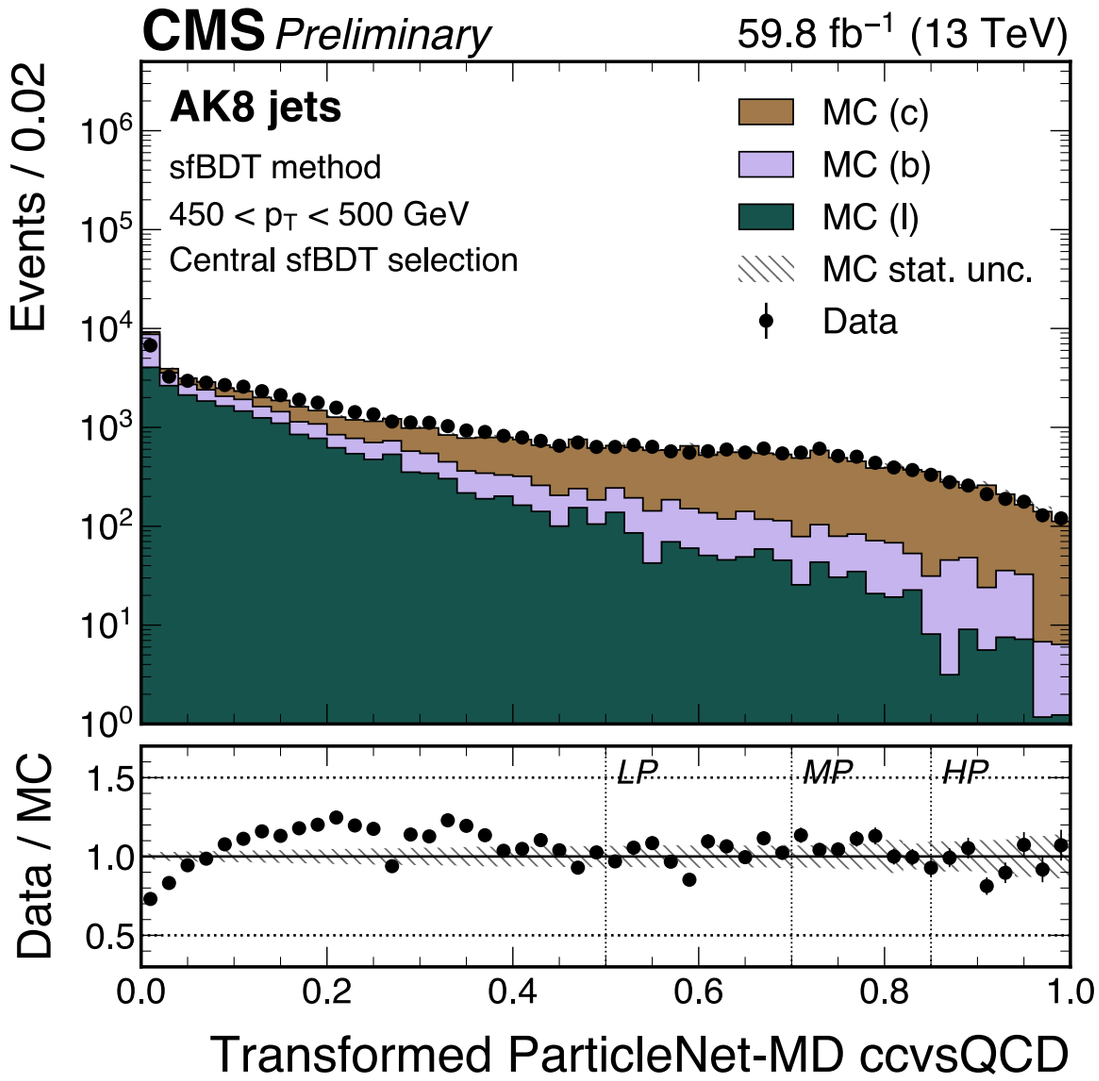






sfBDT method: selection via sfBDT

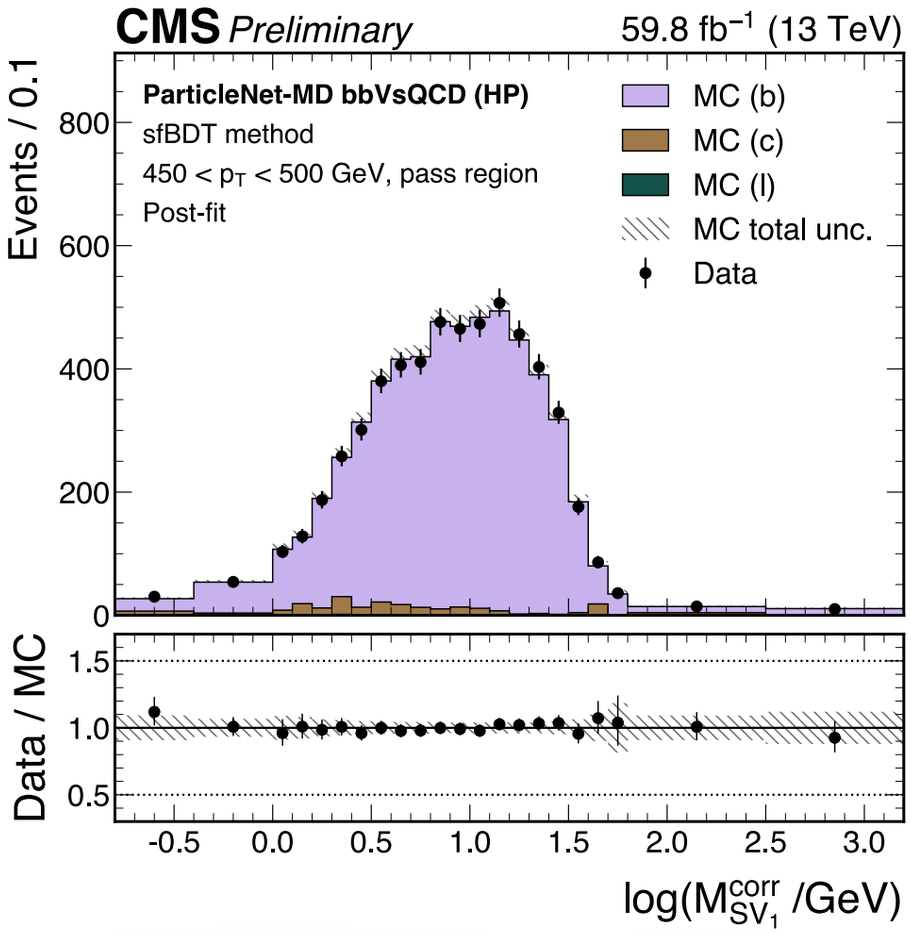


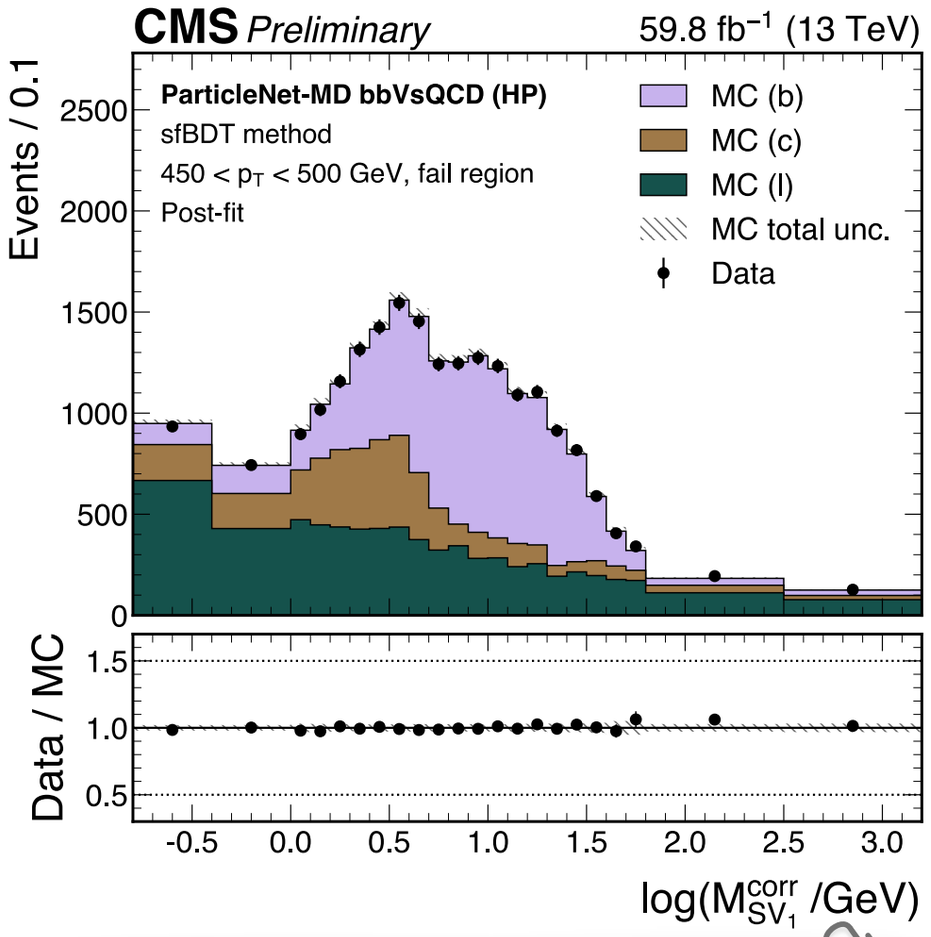


Data and MC distribution on transformed tagger discriminant,
applying the central sfBDT selection

More taggers in backup

sfBDT method: fit





Post-fit templates for calibrating ParticleNet-MD bbvsQCD (HP)

see backup for ccvsQCD

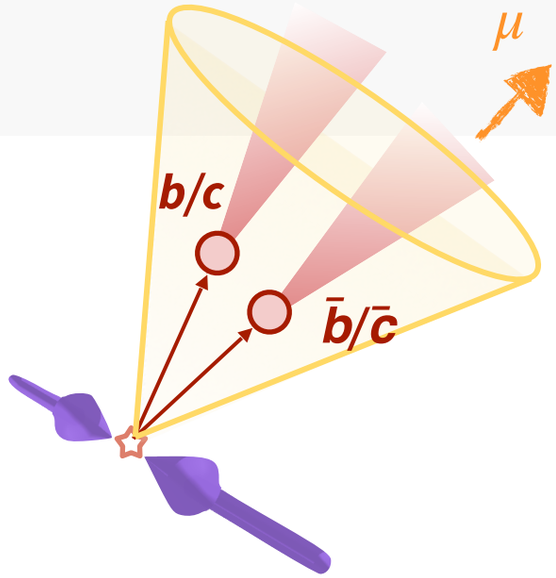
- Fit on the SV mass (log scale) to distinguish the b/c/light templates
- each flavour template assigned a free-floating SF in fit
- simultaneous fit in the “pass” and “fail” tagger region

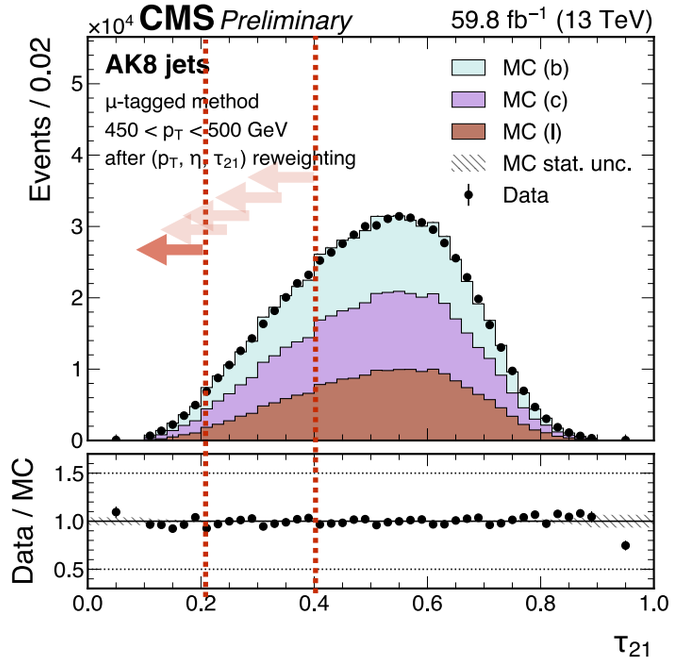
For clarity, read the full workflow in backup

μ-tagged method: concept

➔ Select proxy jet:

- ❖ **g→bb̄/cc̄** jets from QCD, but matched with a soft muon
  - b/c flavour content enriched
- ❖ incorporate online trigger selection: requiring a soft muon in AK4/AK8 jets
  - largely orthogonal to the phase space explored in the sfBDT method
- ❖ **τ_{21} variable as a tune** to modify signal—proxy similarity
  - for main fit: apply $\tau_{21} < 0.3$

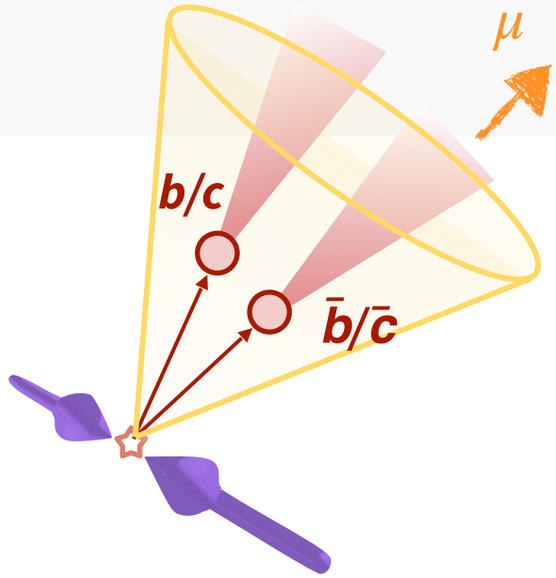


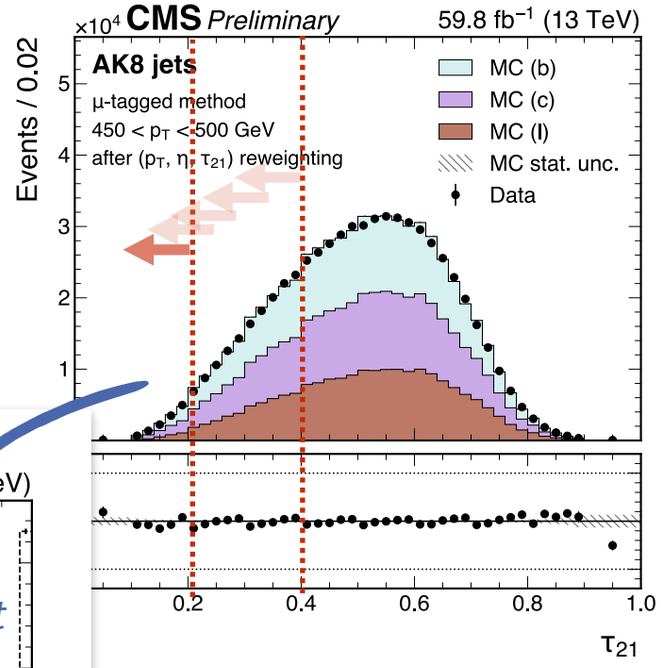


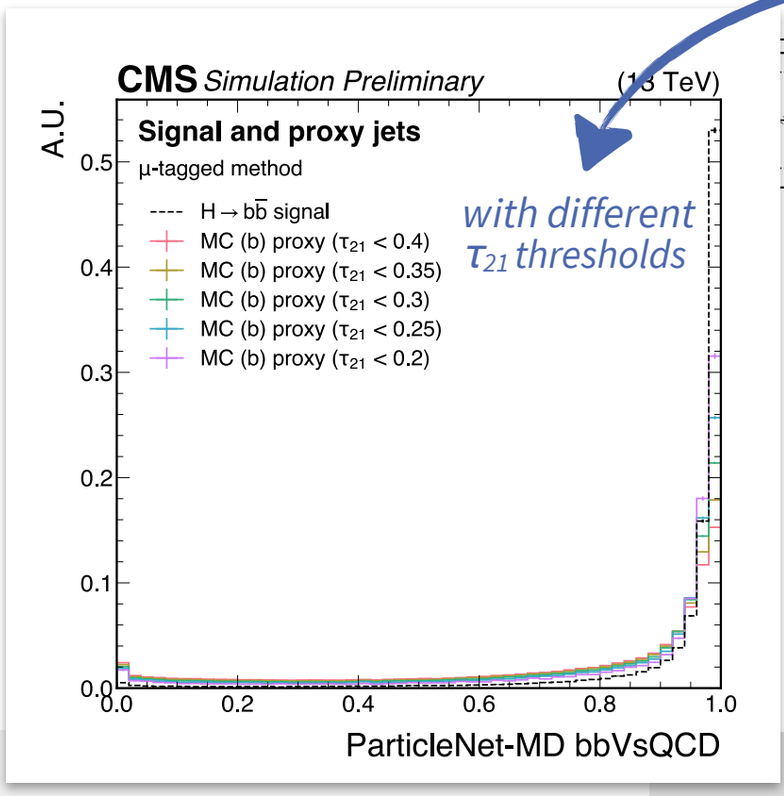
μ-tagged method: concept

➔ Select proxy jet:

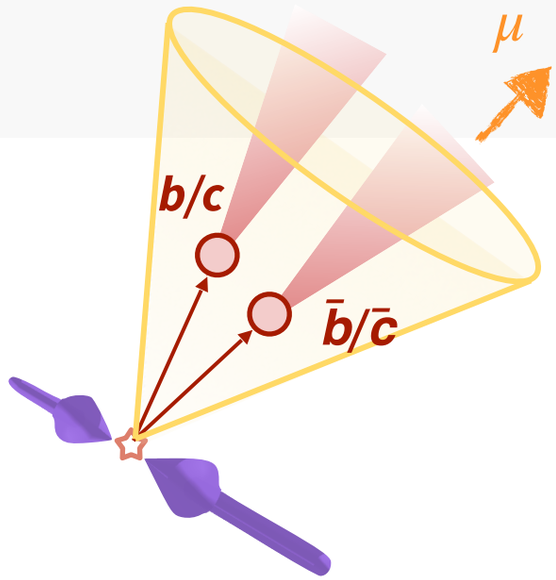
- ❖ **g→bb̄/cc̄** jets from QCD, but matched with a soft muon
  - ▸ b/c flavour content enriched
- ❖ incorporate online trigger selection: requiring a soft muon in AK4/AK8 jets
  - ▸ largely orthogonal to the phase space explored in the sfBDT method
- ❖ **τ_{21} variable as a tune** to modify signal—proxy similarity
  - ▸ for main fit: apply $\tau_{21} < 0.3$





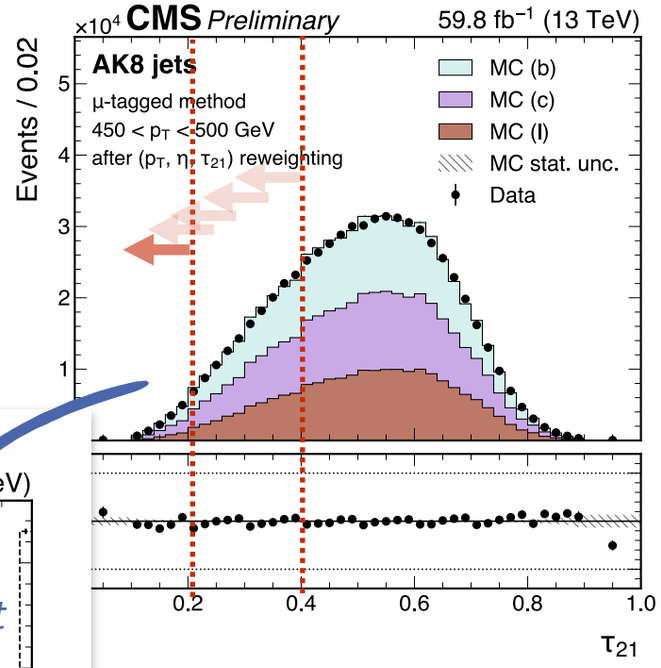


μ-tagged method: concept



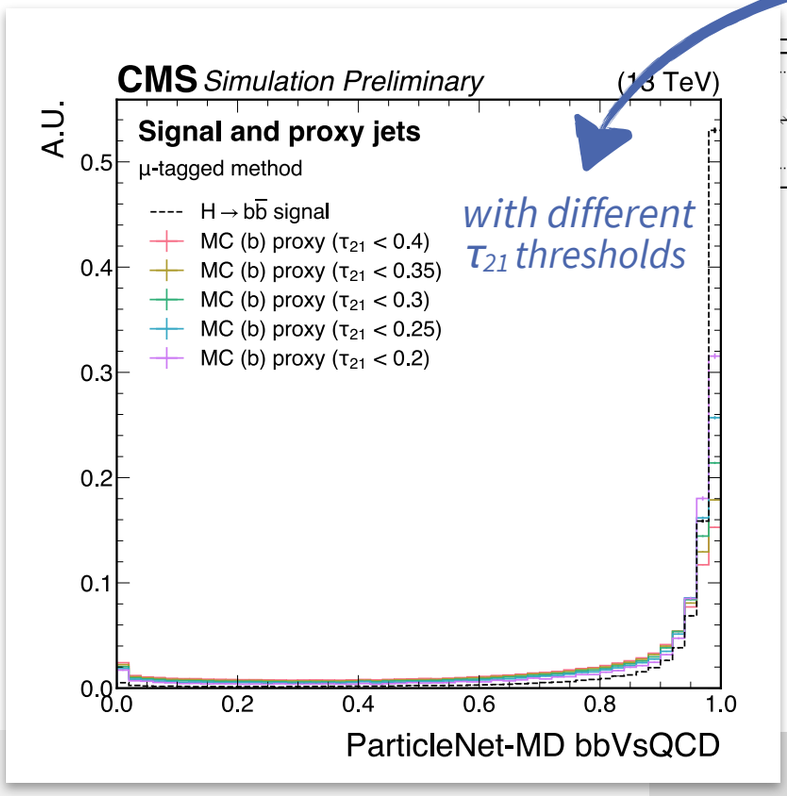
➔ Select proxy jet:

- ❖ **g→bb̄/cc̄** jets from QCD, but matched with a soft muon
  - ▸ b/c flavour content enriched
- ❖ incorporate online trigger selection: requiring a soft muon in AK4/AK8 jets
  - ▸ largely orthogonal to the phase space explored in the sfBDT method
- ❖ **τ_{21} variable as a tune** to modify signal—proxy similarity
  - ▸ for main fit: apply $\tau_{21} < 0.3$

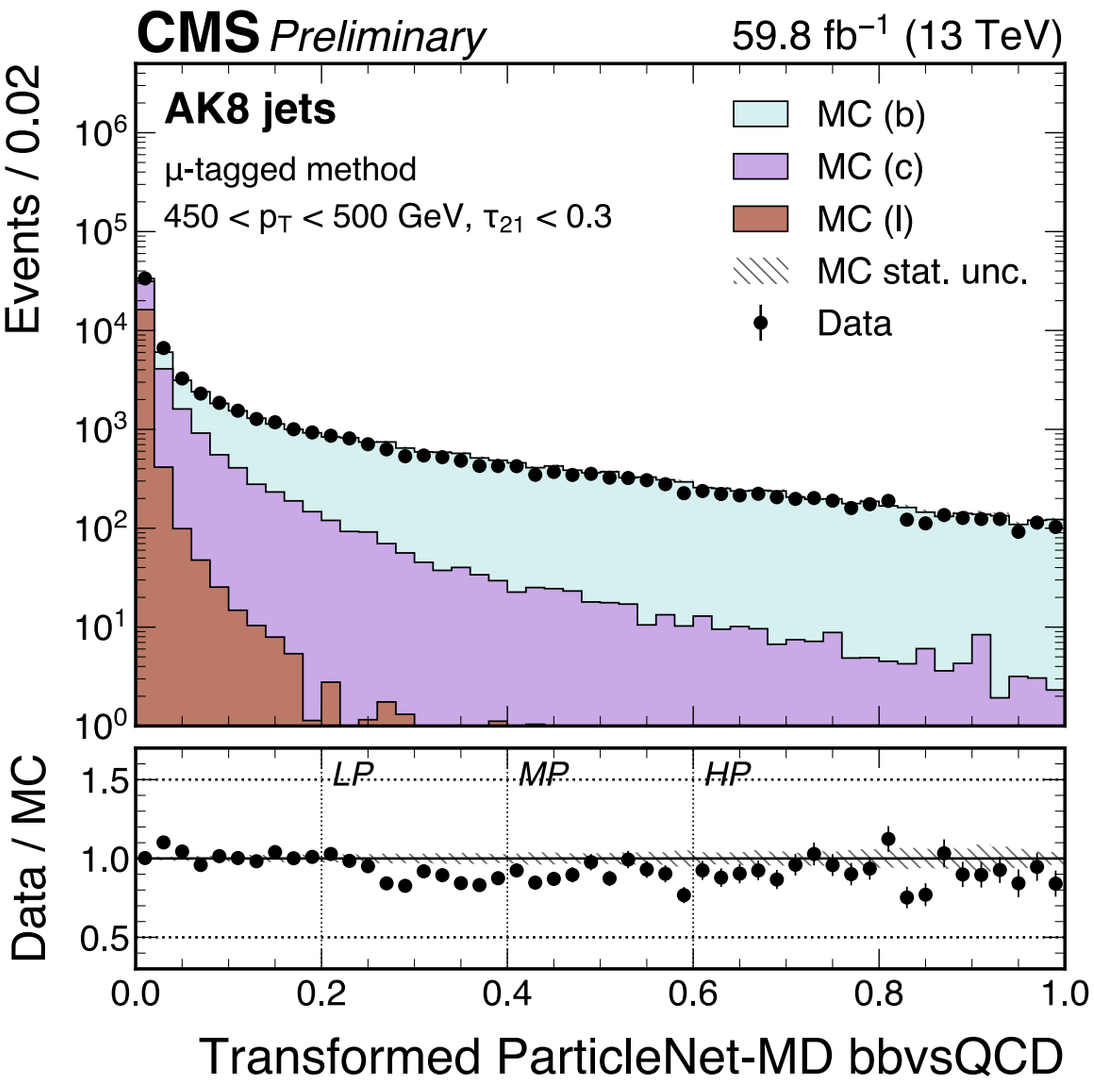


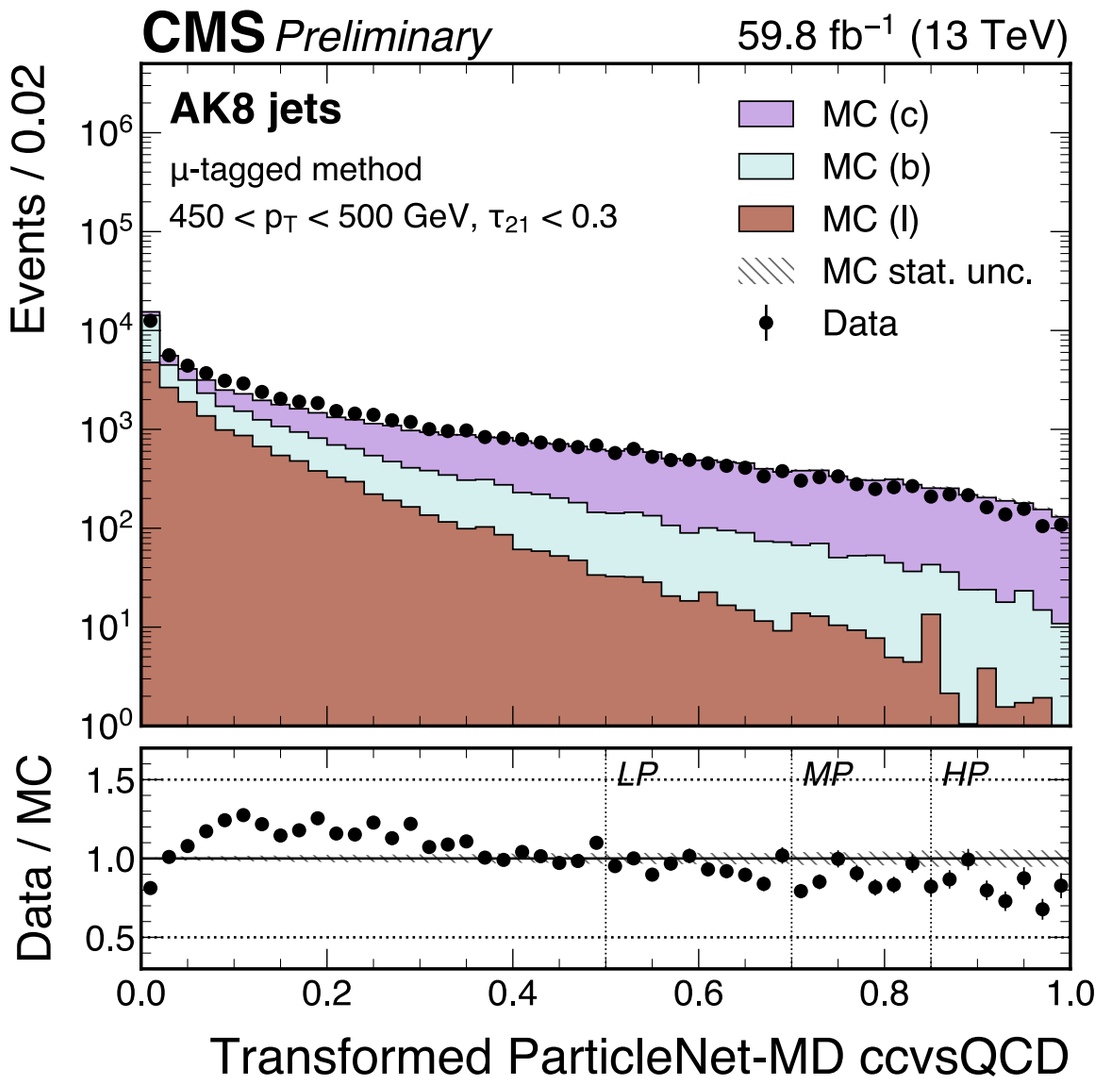
➔ Feasibility of the proxy

- ❖ taggers not trained with particle ID input
  → suitable to calibrate using the μ-tagged proxy



μ-tagged method: selected jets

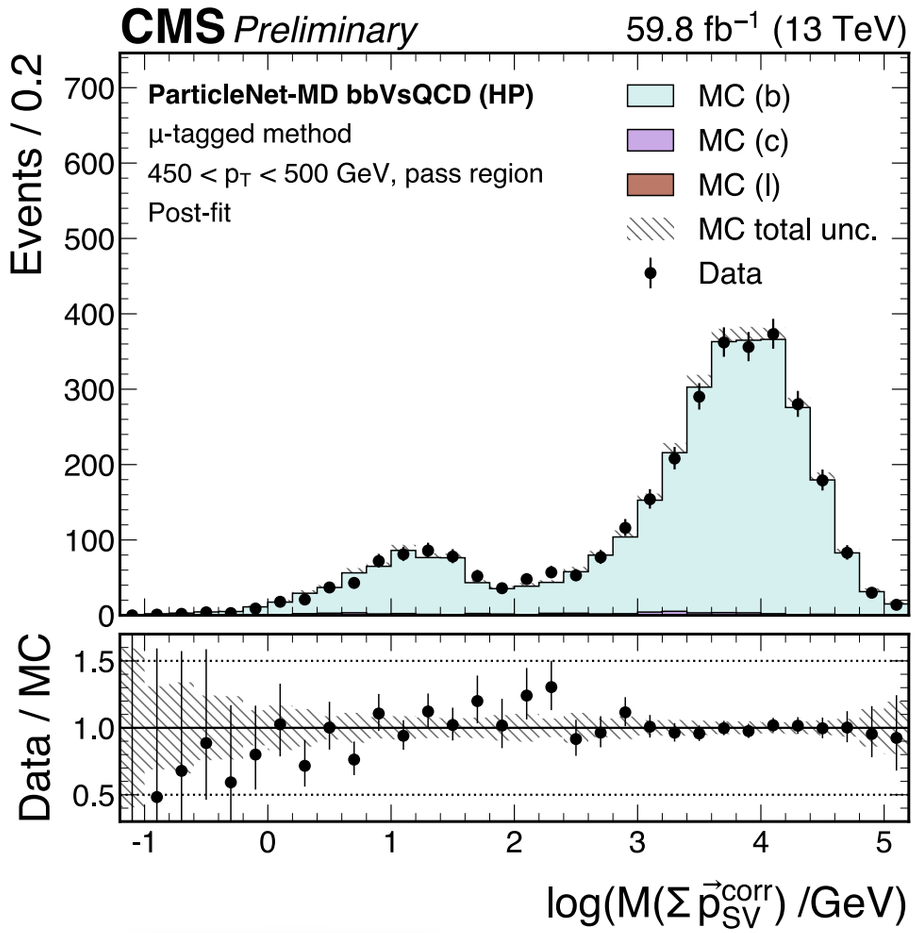


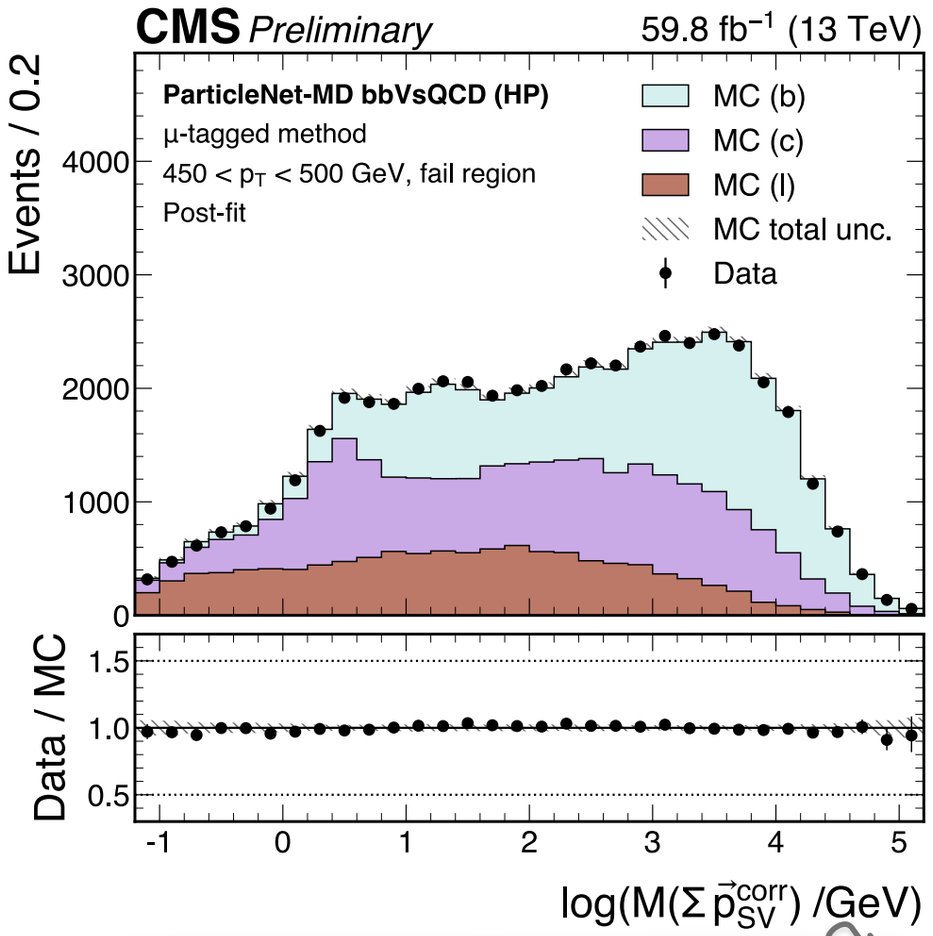


Data and MC distribution applying the μ-tagged requirement and the central τ_{21} < 0.3 cut

More taggers in backup

μ-tagged method: fit





Post-fit templates for calibrating ParticleNet-MD bbvsQCD (HP)

see backup for ccvsQCD

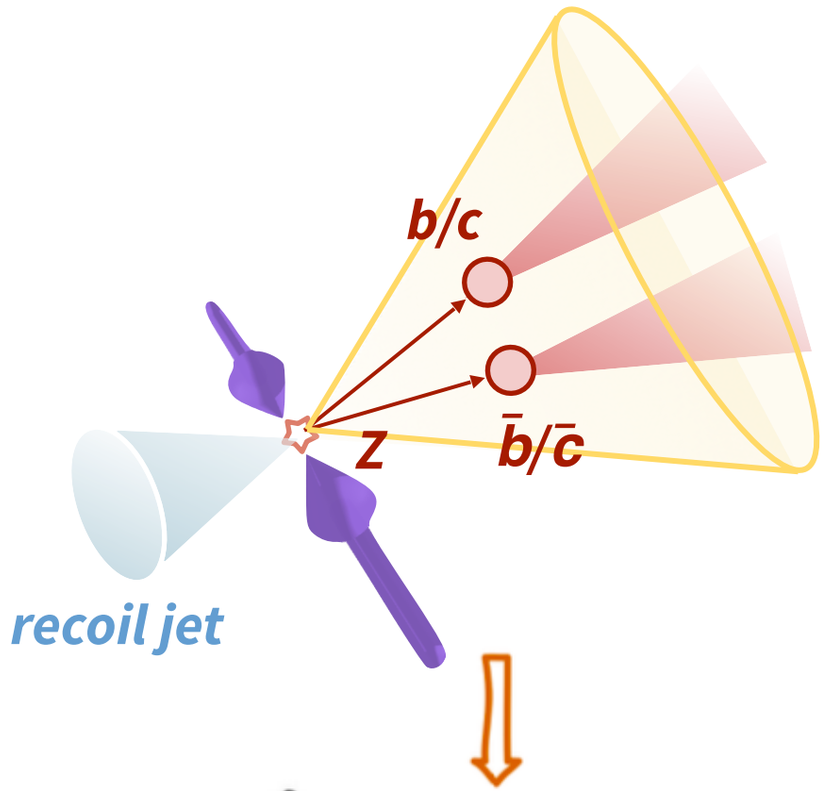
- Fit on the invariant mass of SV 4-vectors (log scale)
- each flavour template assigned a free-floating SF in fit
- simultaneous fit in the “pass” and “fail” region

For clarity, read the full workflow in backup

Boosted Z method: concept

➔ Proxy jets: Z→bb jets

- ❖ Z→bb jet is a better proxy to H→bb jet
- ❖ measure the SF at Z peak on top of the overwhelming QCD BKG



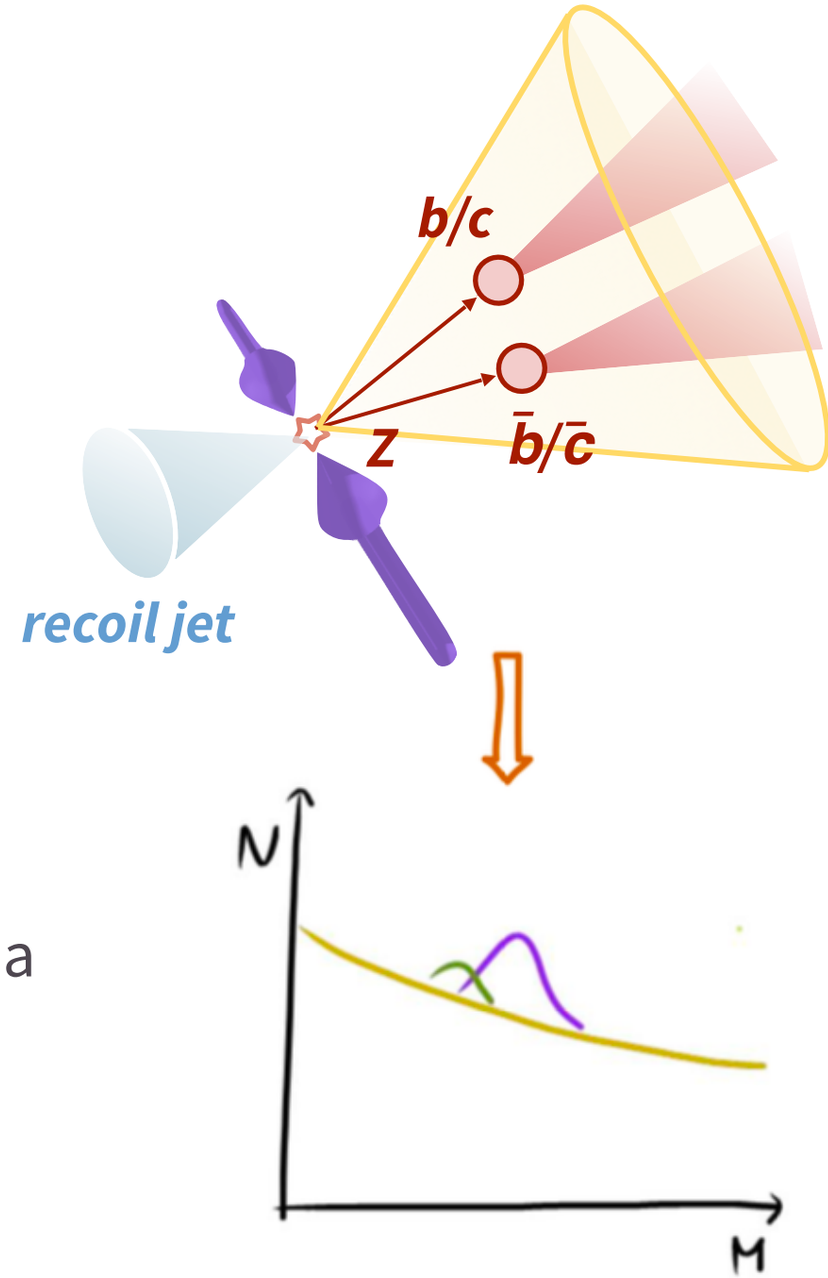
Boosted Z method: concept

➔ Proxy jets: Z→bb jets

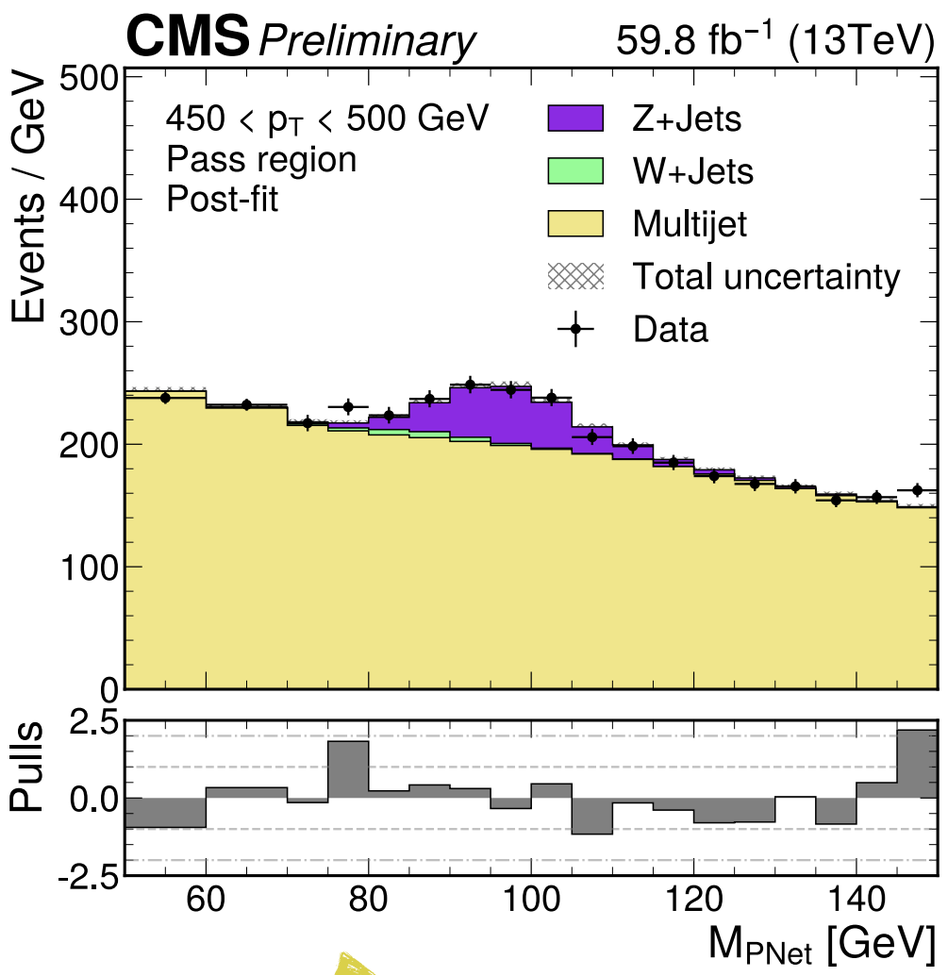
- ❖ Z→bb jet is a better proxy to H→bb jet
- ❖ measure the SF at Z peak on top of the overwhelming QCD BKG

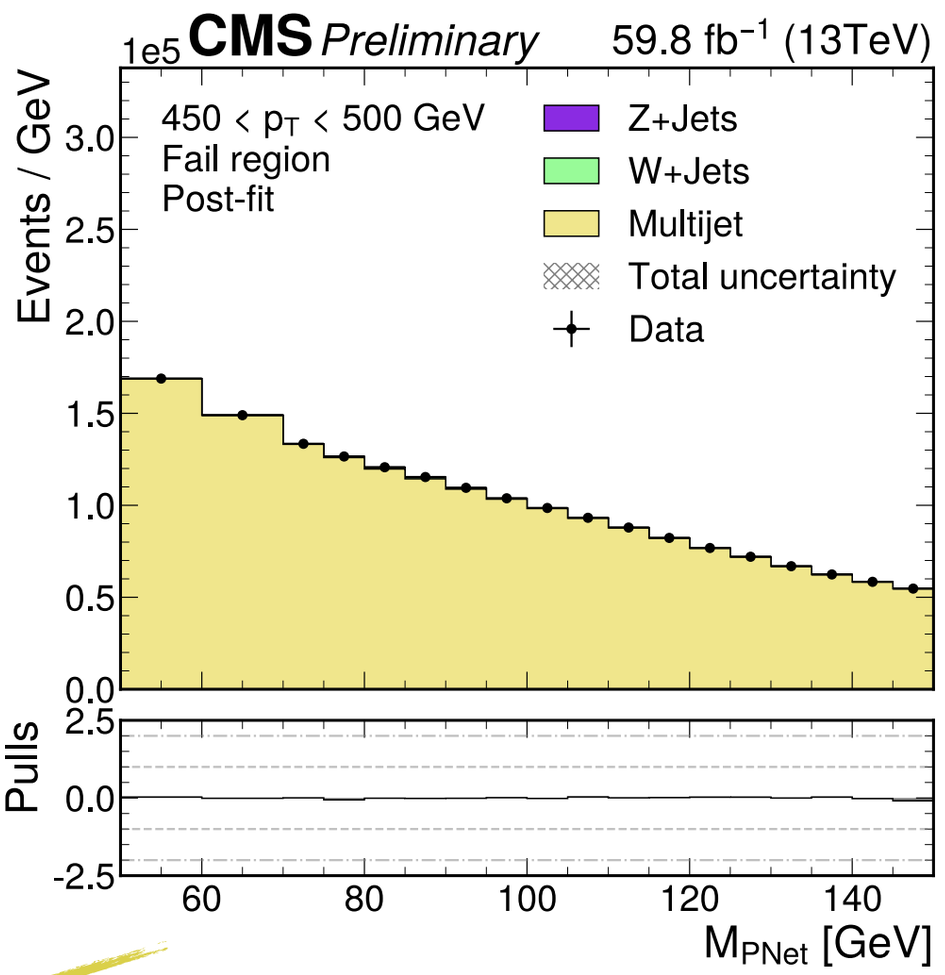
➔ Signal and background estimation

- ❖ QCD yield in “fail” tagger region estimated from data
- ❖ transferred to the QCD yield in “pass” region by a fitted polynomial on jet p_T and “mass”
  - ▸ “mass”: a regressed mass via ParticleNet DNN architecture [See CMS-DP-2021-017]
- ❖ signal: rely on accurate modelling of the Z+jet process (NLO correction applied)



Boosted Z method: fit





Post-fit templates for calibrating ParticleNet-MD bbvsQCD (HP)

For clarity, read the full workflow in backup

QCD "fail" → "pass" transfer ratio

$$R_{\mathrm{P/F}} = \sum_{m,n=0}^{m+n\leq o \,\wedge\, n<3} k_{m,n} (M_{\mathrm{PNet}})^m (p_{\mathrm{T}})^n,$$

- Fit on the ParticleNet-based regressed mass
- three SFs for Z peak (in 3 p_T bins), one for W peak
- simultaneous fit in the "pass" and "fail" region, and 3 p_T bins

Systematic uncertainties (sfBDT & μ-tagged)

sfBDT method

relative contribution to overall uncertainty

Uncertainty source	$\Delta SF / (\Delta SF)_{tot}$
Statistical	49%
Theory	
Fraction of jet flavours	30%
ISR and FSR in parton shower	10%
Experimental	
Effect of varying sfBDT thresholds	38%
Effect of applying "reweighting schemes"	64%
Jet energy scale and resolution	6.4%
Integrated luminosity	1.0%
Pileup reweighting	7.4%

μ-tagged method

Uncertainty source	$\Delta SF / (\Delta SF)_{tot}$
Statistical	34%
Theory	
Fraction of jet flavours	42%
ISR and FSR in parton shower	14%
QCD jet modelling	6.5%
Experimental	
Effect of varying τ_{21} thresholds	69%
Effect of "simulation-to-data reweighting"	28%
Jet energy scale and resolution	5.0%
Integrated luminosity	3.5%
Pileup reweighting	3.9%

Varying the yields of b/c/light flavour jets by 20%, separately

→ compensate for the mismodelling of flavour proportion in QCD simulation

Systematic uncertainties (sfBDT & μ-tagged)

sfBDT method

relative contribution to overall uncertainty

Uncertainty source	$\Delta SF / (\Delta SF)_{tot}$
Statistical	49%
Theory	
Fraction of jet flavours	30%
ISR and FSR in parton shower	10%
Experimental	
Effect of varying sfBDT thresholds	38%
Effect of applying "reweighting schemes"	64%
Jet energy scale and resolution	6.4%
Integrated luminosity	1.0%
Pileup reweighting	7.4%

μ-tagged method

Uncertainty source	$\Delta SF / (\Delta SF)_{tot}$
Statistical	34%
Theory	
Fraction of jet flavours	42%
ISR and FSR in parton shower	14%
QCD jet modelling	6.5%
Experimental	
Effect of varying τ_{21} thresholds	69%
Effect of "simulation-to-data reweighting"	28%
Jet energy scale and resolution	5.0%
Integrated luminosity	3.5%
Pileup reweighting	3.9%

Varying the yields of b/c/light flavour jets by 20%, separately

→ compensate for the mismodelling of flavour proportion in QCD simulation

An external uncertainty is assigned to "cover the span" of derived SFs, when the selection (sfBDT or τ_{21}) varies

→ a substantial uncertainty is obtained from this source

Systematic uncertainties (sfBDT & μ-tagged)

sfBDT method

relative contribution to overall uncertainty

Uncertainty source	$\Delta SF / (\Delta SF)_{tot}$
Statistical	49%
Theory	
Fraction of jet flavours	30%
ISR and FSR in parton shower	10%
Experimental	
Effect of varying sfBDT thresholds	38%
Effect of applying "reweighting schemes"	64%
Jet energy scale and resolution	6.4%
Integrated luminosity	1.0%
Pileup reweighting	7.4%

μ-tagged method

Uncertainty source	$\Delta SF / (\Delta SF)_{tot}$
Statistical	34%
Theory	
Fraction of jet flavours	42%
ISR and FSR in parton shower	14%
QCD jet modelling	6.5%
Experimental	
Effect of varying τ_{21} thresholds	69%
Effect of "simulation-to-data reweighting"	28%
Jet energy scale and resolution	5.0%
Integrated luminosity	3.5%
Pileup reweighting	3.9%

Varying the yields of b/c/light flavour jets by 20%, separately

→ compensate for the mismodelling of flavour proportion in QCD simulation

An external uncertainty is assigned to "cover the span" of derived SFs, when the selection (sfBDT or τ_{21}) varies

→ a substantial uncertainty is obtained from this source

A residue mismodelling on the SV mass is observed, especially in the sfBDT method, which is difficult to fix by using alternative simulation

→ adopt a conservative approach: **reweight directly on the fitted mass variable in "pass+fail" region** to check the change of SF

Systematic uncertainties (boosted Z)

boosted Z method

relative contribution to overall uncertainty

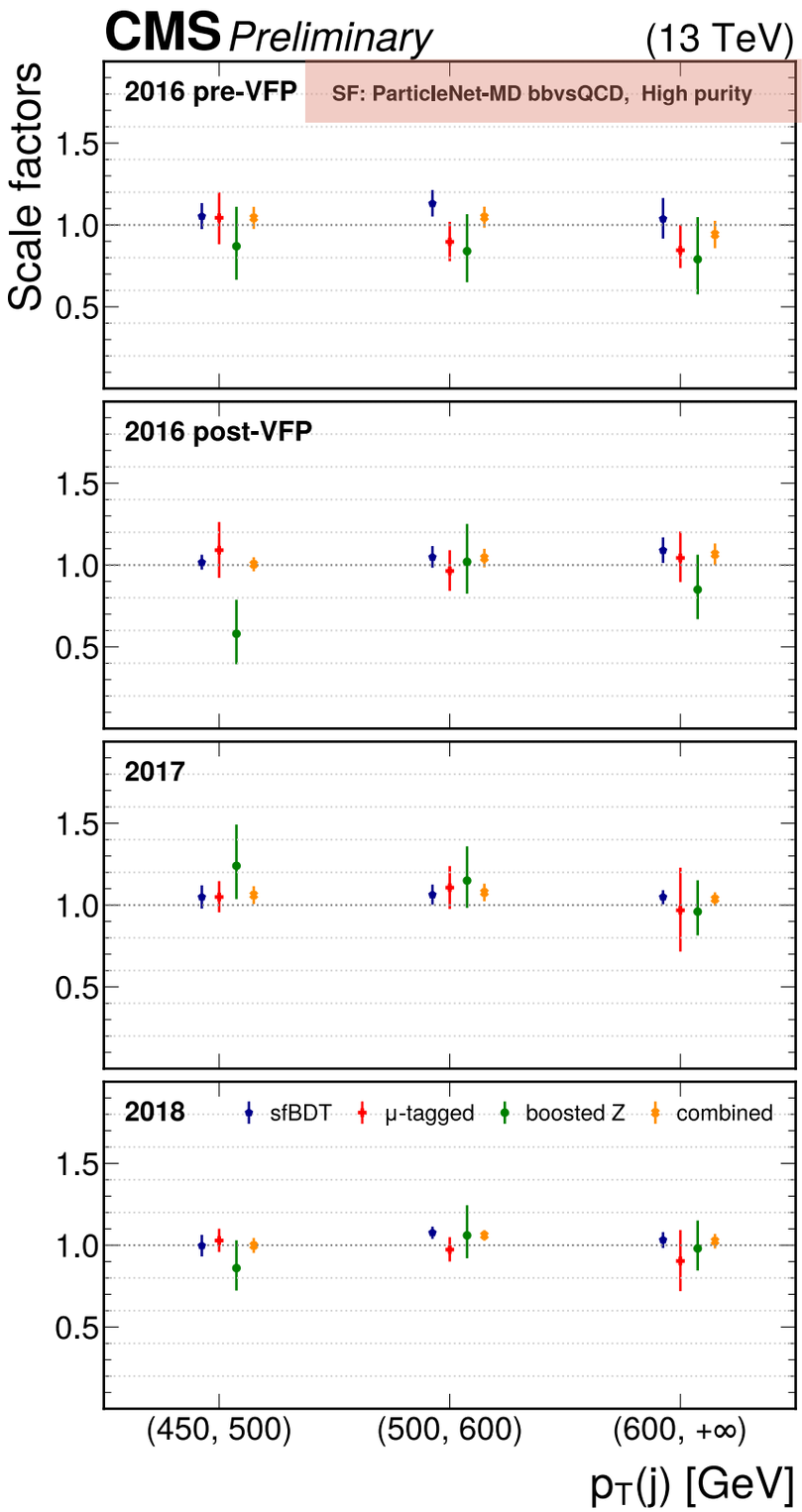
Uncertainty source	$\Delta SF / (\Delta SF)_{tot}$
Statistical	67%
Theory	
ISR and FSR in parton shower	45%
NLO corrections	43%
PDF uncertainties	5.5%
Experimental	
Jet mass scale and resolution	6.9%
Jet energy scale and resolution	25%
Trigger effiency	5.4%
Integrated luminosity	11%
Pileup reweighting	2.5%

Large statistical uncertainty for determining $R_{P/F}$ is a major limiting power

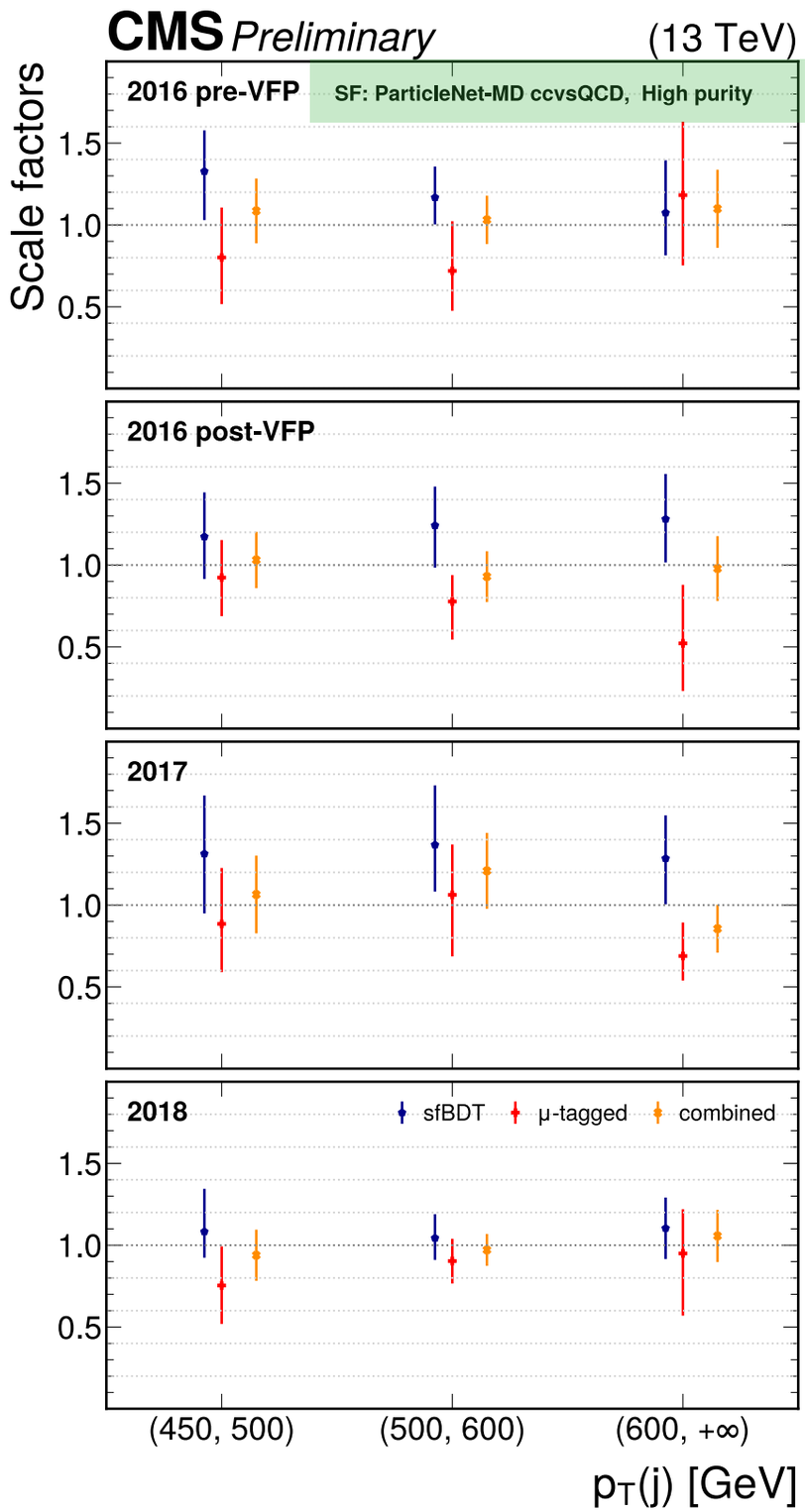
Theoretical uncertainties on Z+jets affect the Z peak yield, directly resulting to a larger SF uncertainty

Results and combination

ParticleNet-MD bbvsQCD (HP)



ParticleNet-MD ccvsQCD (HP)

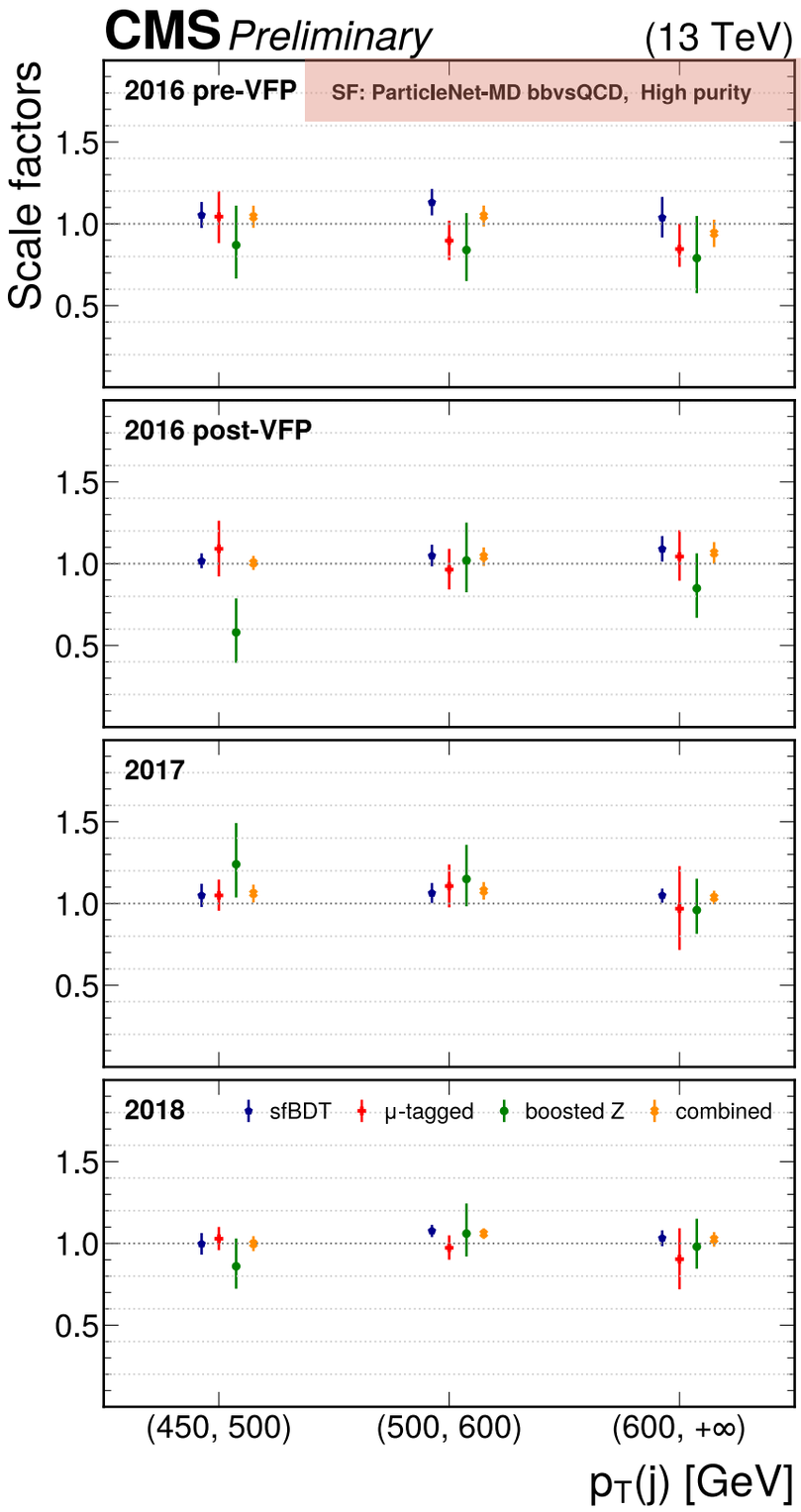


- Combination among methods via **BLUE**
  (best linear unbiased estimate)
  - treating common uncertainty sources as correlated, and others as uncorrelated

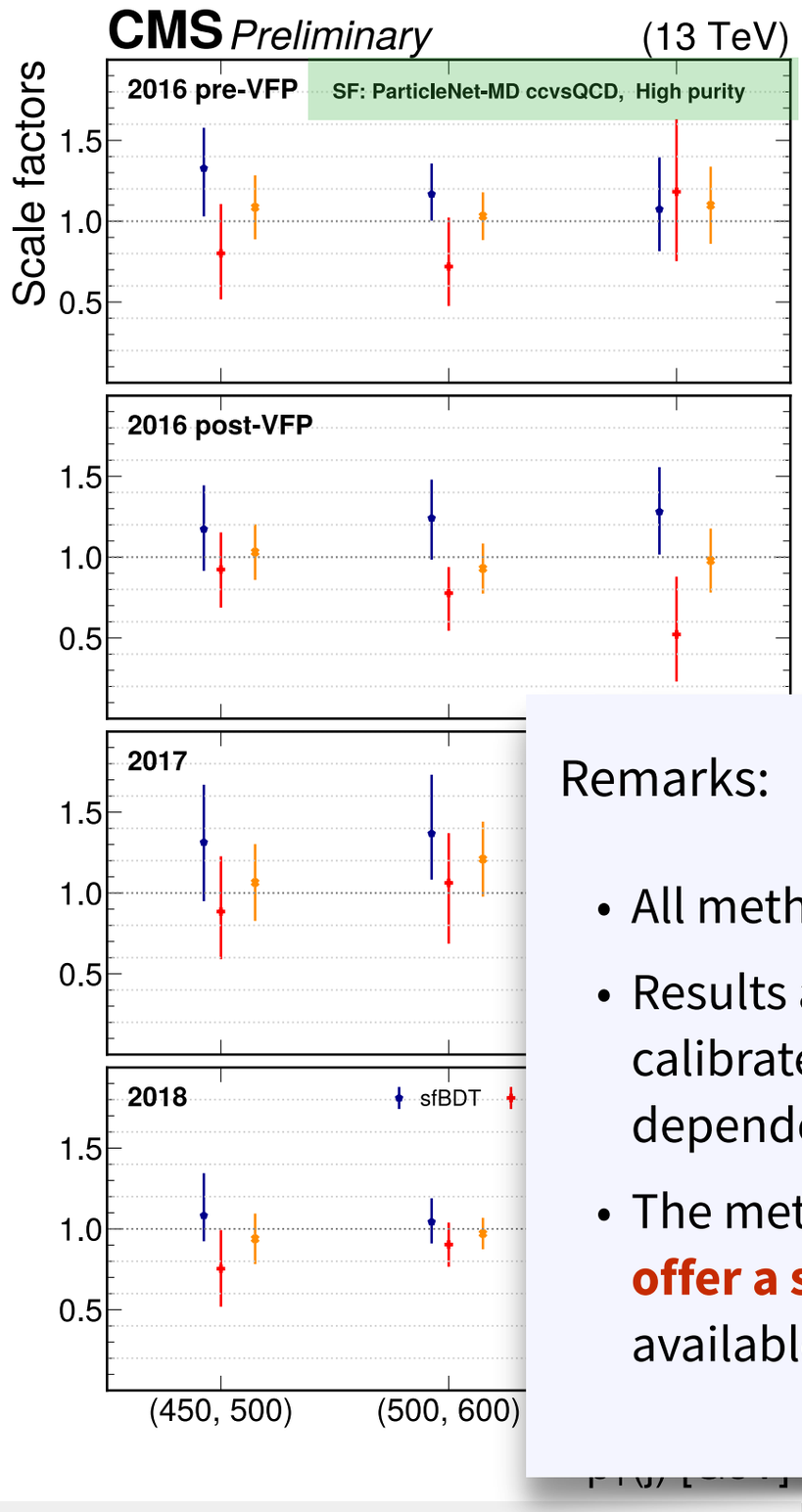
Full SF results in backup

Results and combination

ParticleNet-MD bbvsQCD (HP)



ParticleNet-MD ccvsQCD (HP)



- Combination among methods via **BLUE**
  (best linear unbiased estimate)
  - treating common uncertainty sources as correlated, and others as uncorrelated

Full SF results in backup

Remarks:

- All methods yield consistent results
- Results also consistent with previous calibrated SFs, on the central SF, year-dependence, and level of uncertainties
- The methods summarised in our works **offer a systematic approach** to handle all available X→b$\bar{b}$/c$\bar{c}$ taggers in CMS

Summary

CMS-PAS-BTV-22-001

➔ Present a comprehensive summary of X→b$\bar{b}$/c$\bar{c}$ taggers:

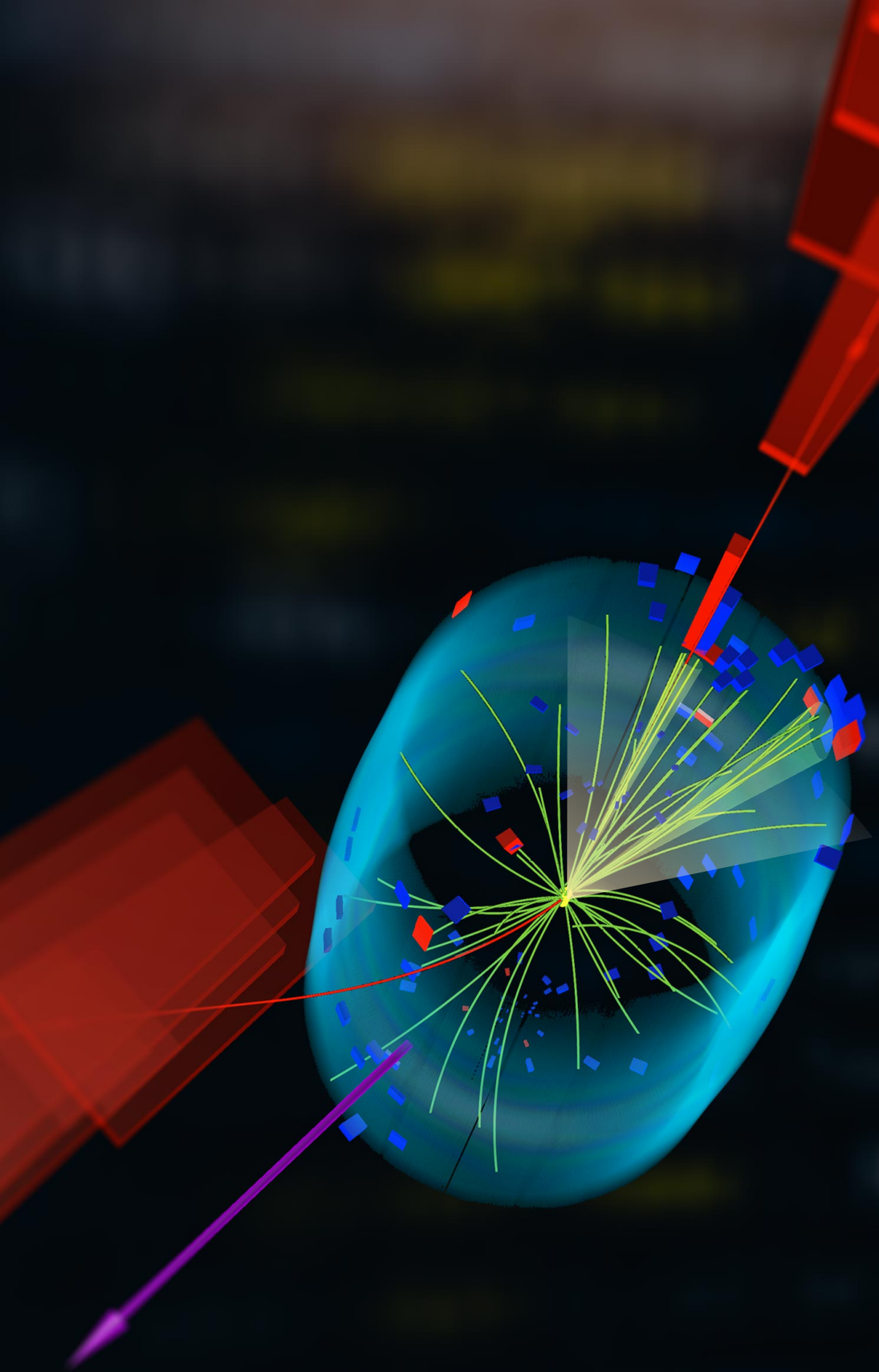
- ❖ a review of taggers developed in CMS Run 2
- ❖ performance comparison in ROC curves
- ❖ calibration methods
  - ▸ sfBDT methods / μ-tagged methods / boosted Z method

➔ Showcase individual and combined SF measurements

- ❖ consistency found among methods, and with previous results

➔ Benchmark the calibration methods for CMS Run 2 and Run 3

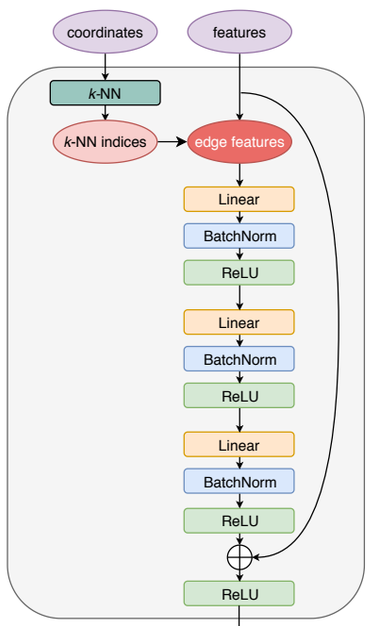
- ❖ developed to handle a variety of tagger discriminants
- ❖ marks the maturity of boosted-jet tagging and calibration technique at this stage



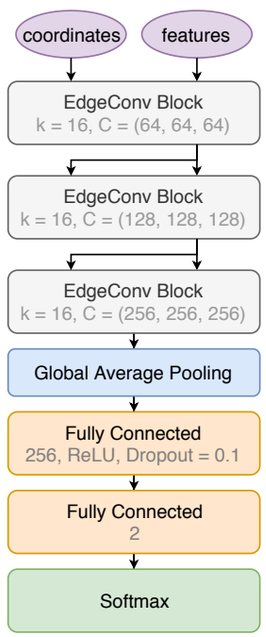
Backup

ParticleNet architecture

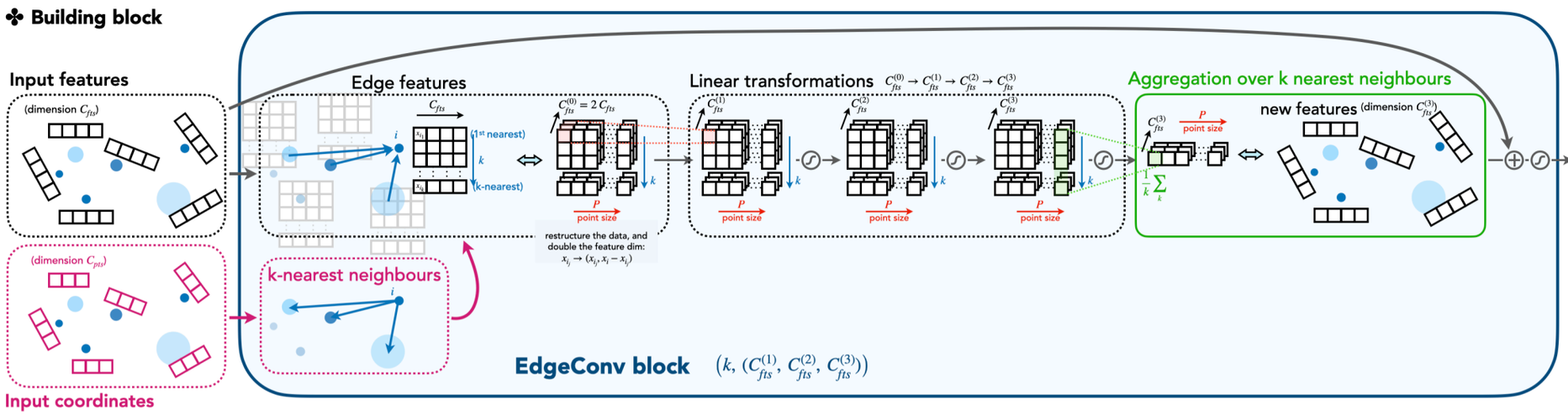
EdgeConv block

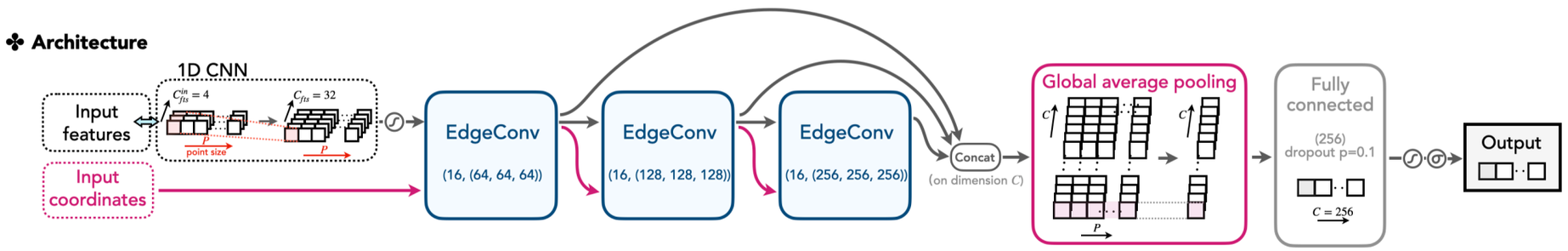


ParticleNet full architecture



H. Qu, L. Gouskos, PRD 101, 056019 (2020)



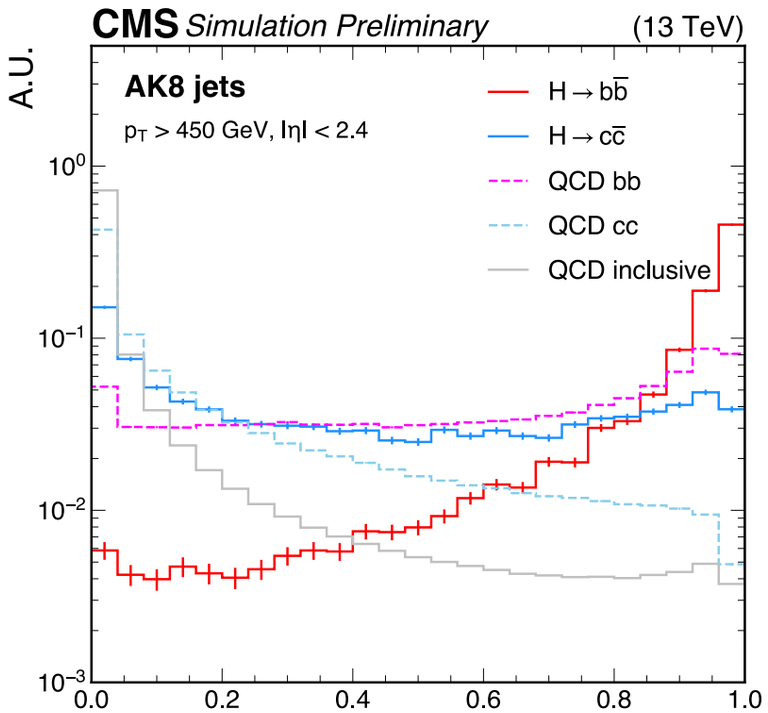


ParticleNet's full architecture

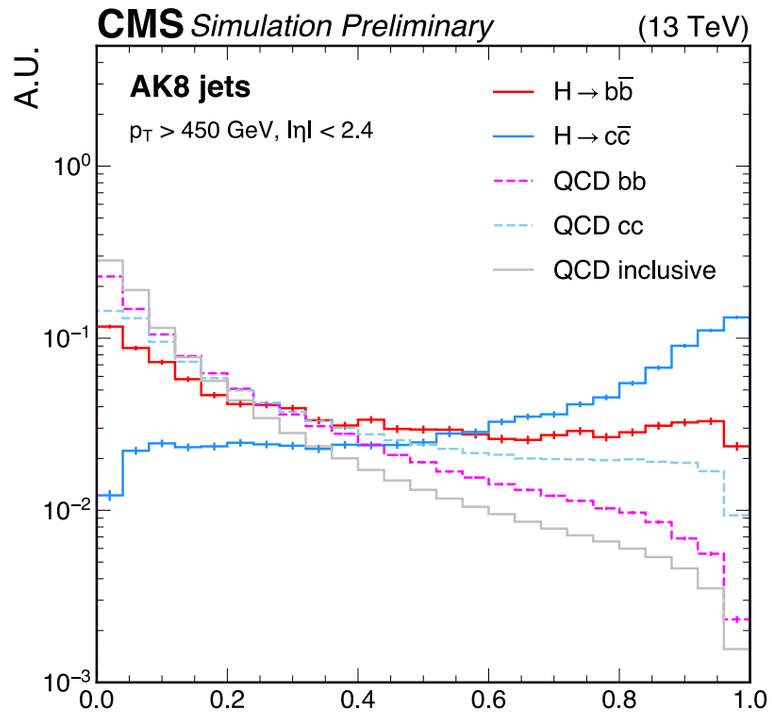
https://cms-ml.github.io/documentation/inference/particlenet.html

More spectra

DeepAK8-MD

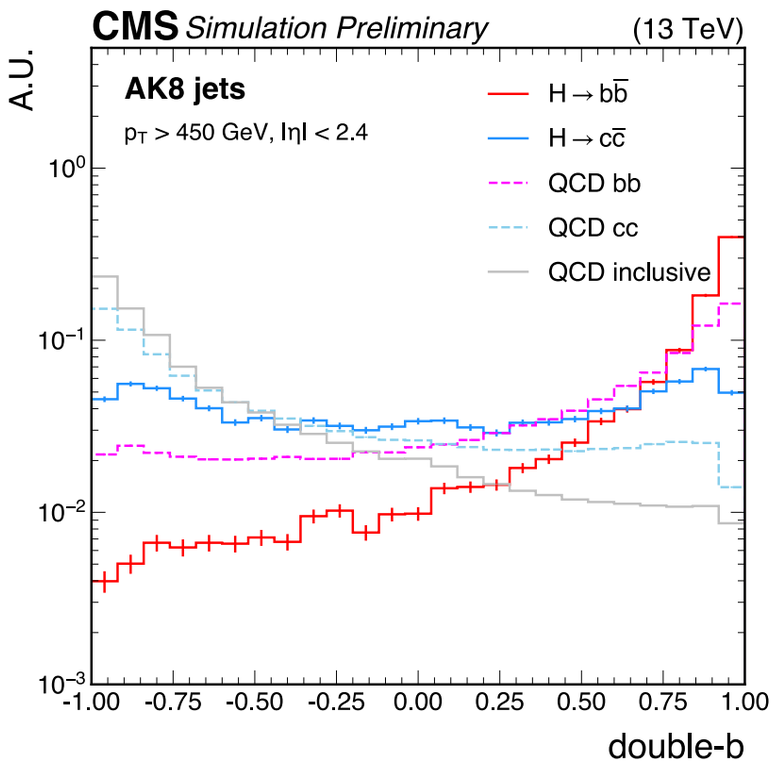


DeepAK8-MD bbvsQCD



DeepAK8-MD ccvsQCD

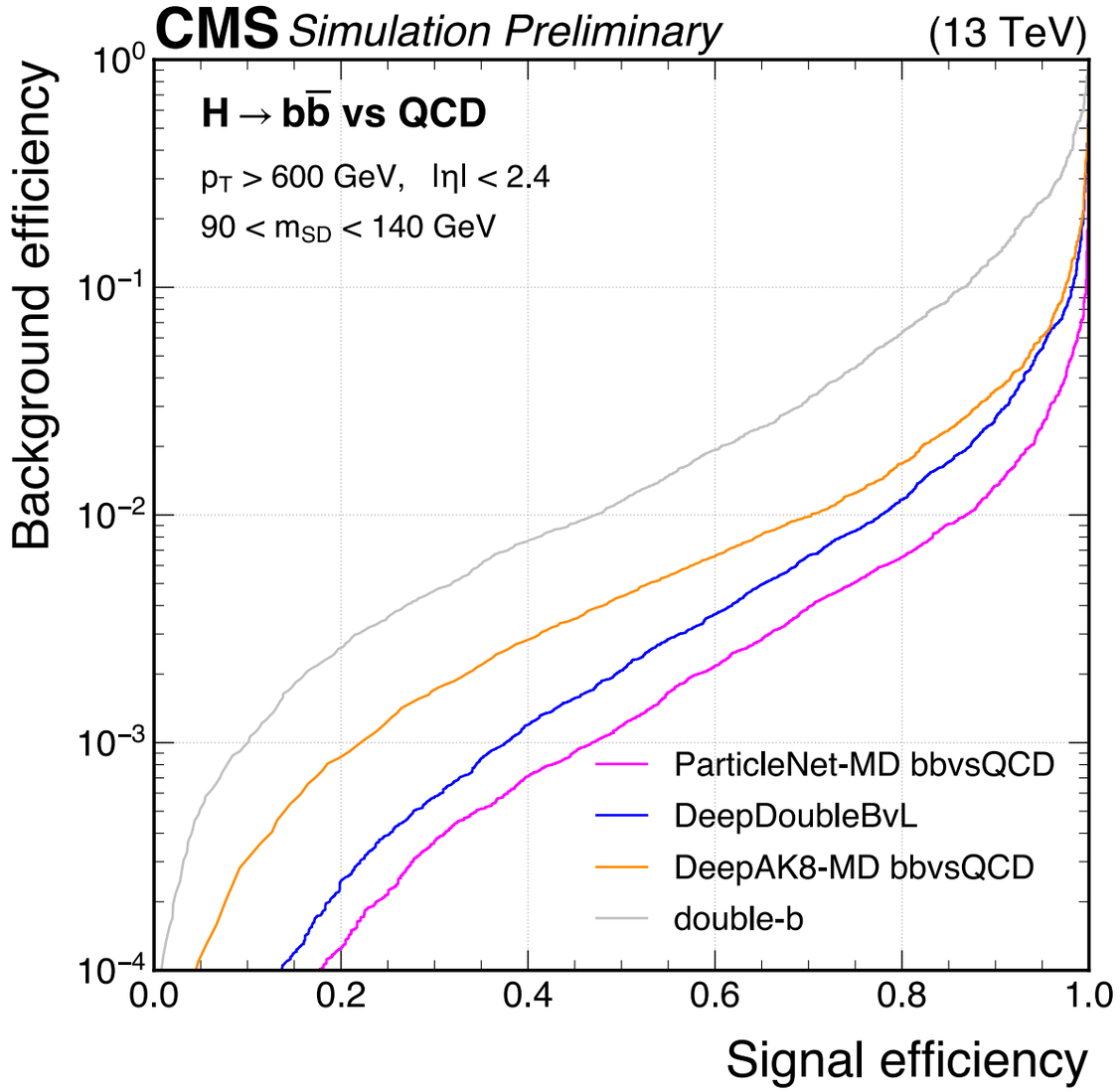
double-b

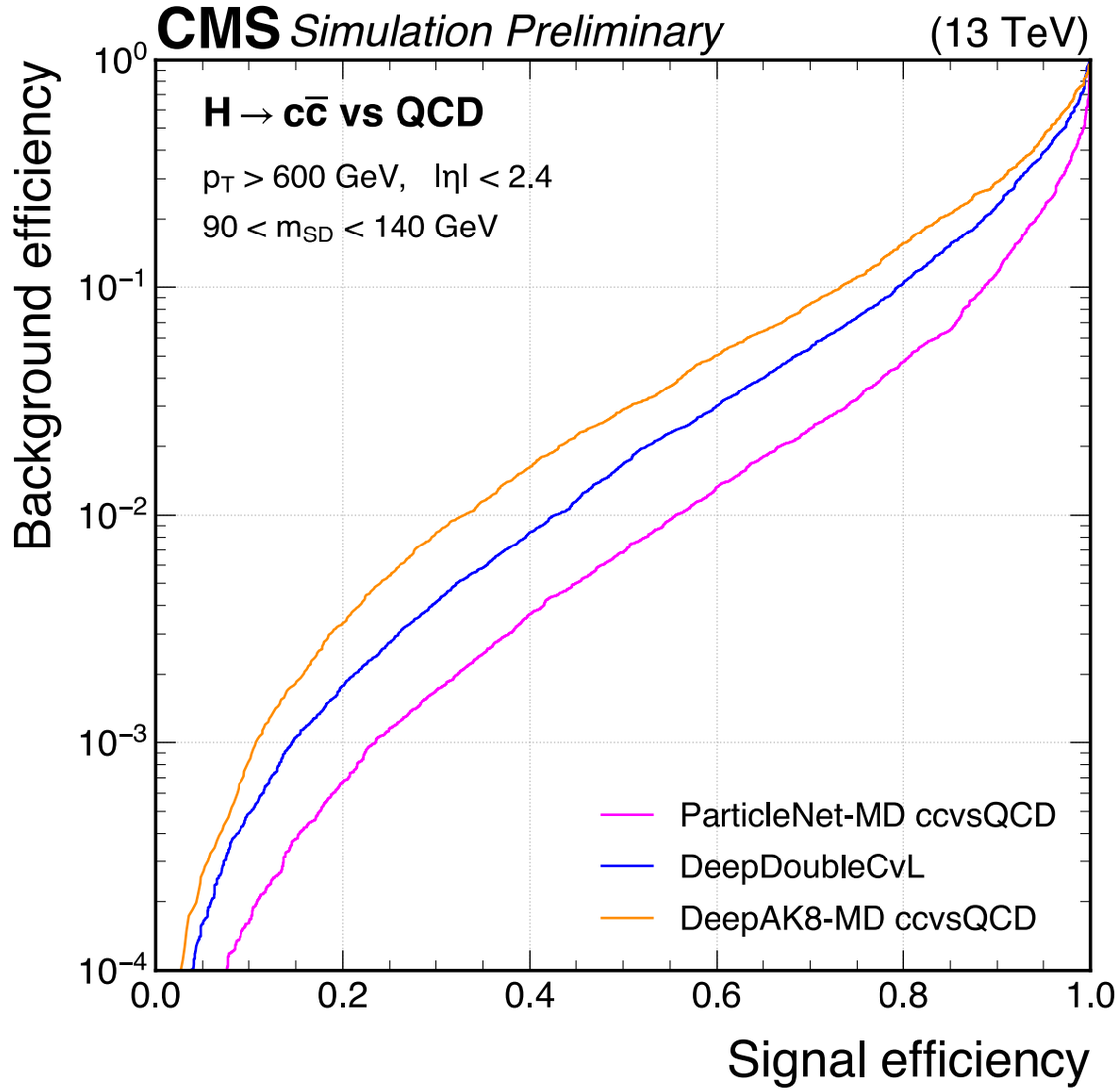


Performance on simulation

CMS-PAS-BTV-22-001

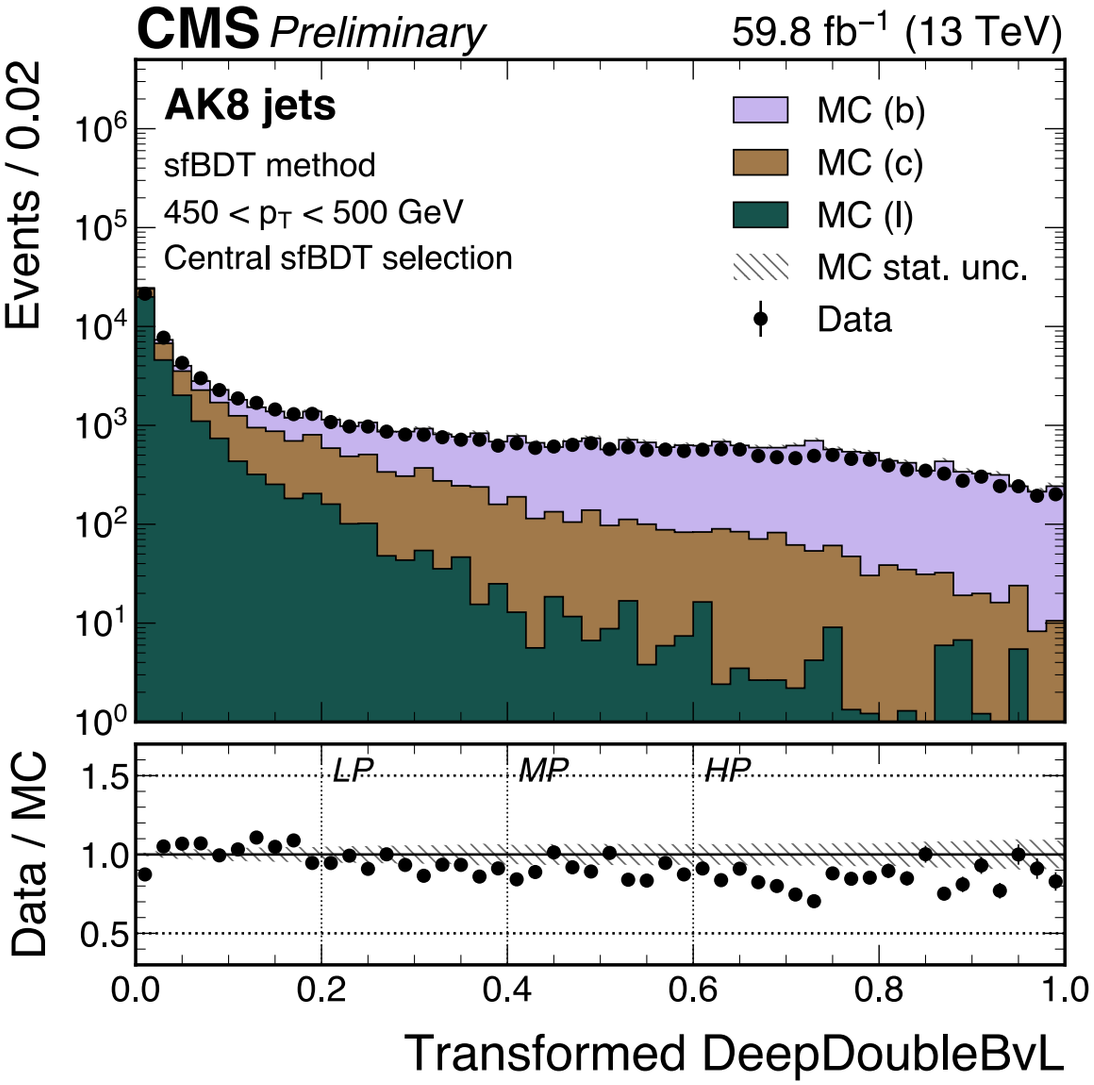
high p_T: (600, +∞) GeV

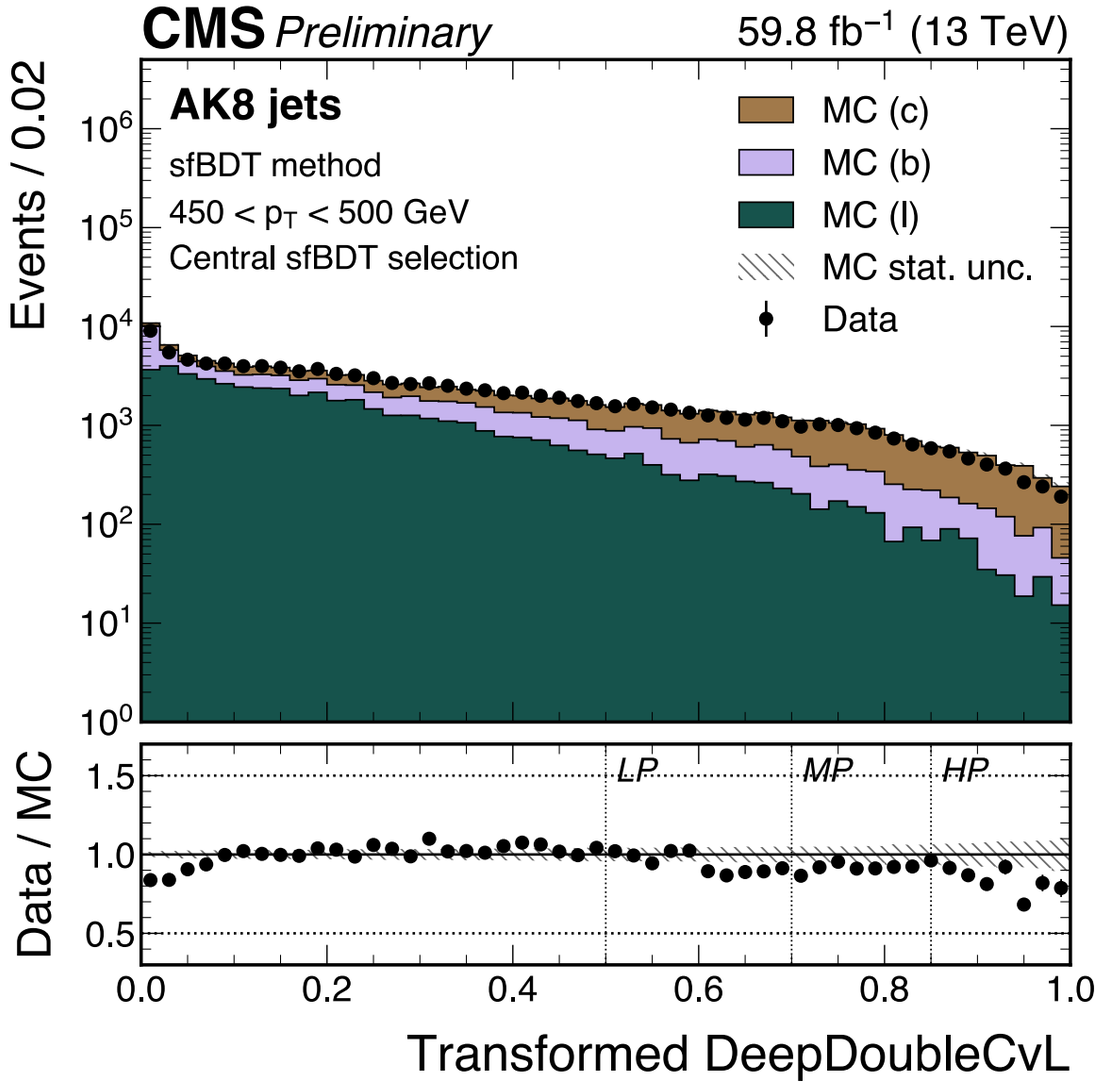




ROC comparison on H→bb̄ (cc̄) jets vs. inclusive QCD

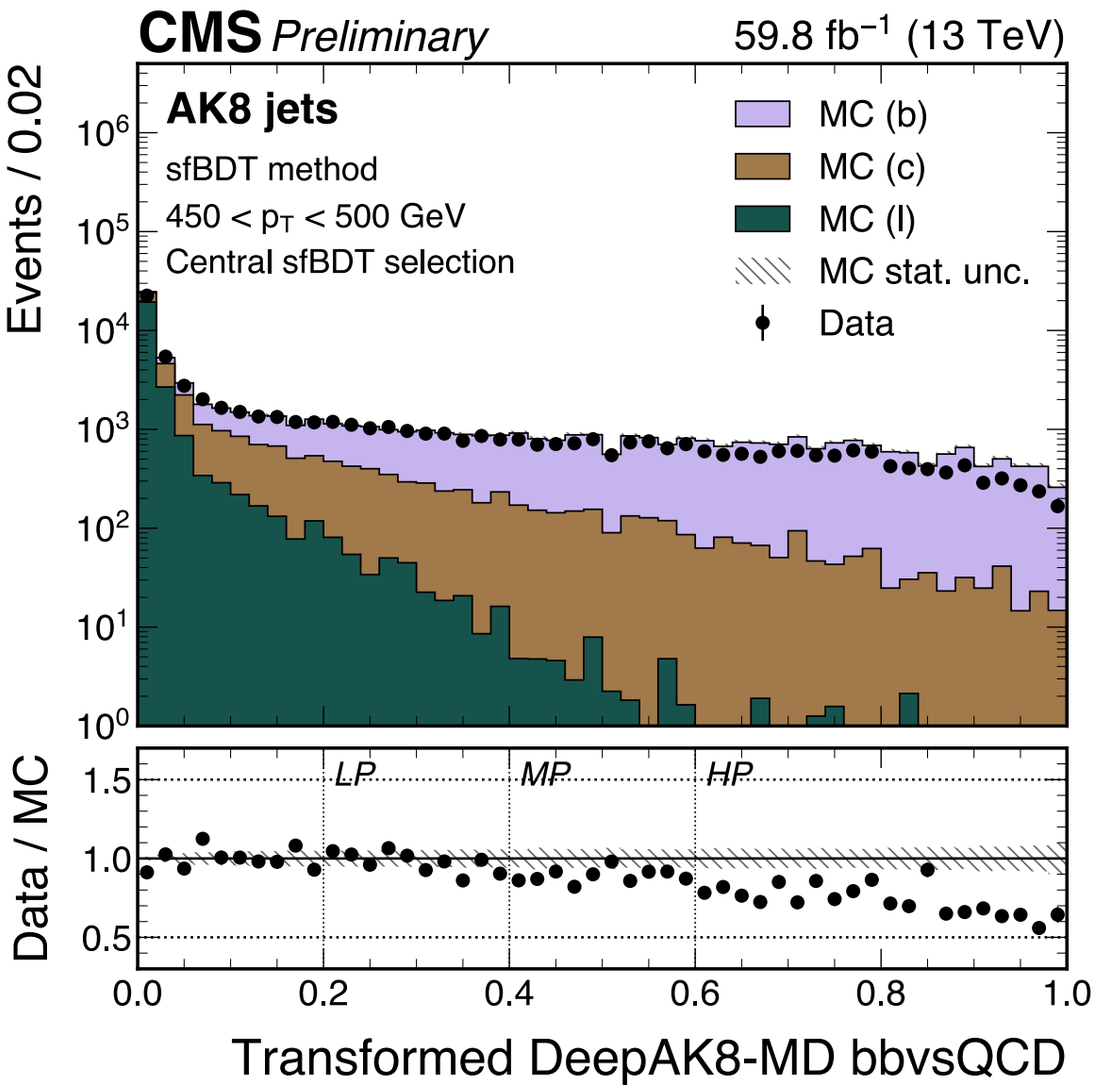
sfBDT method: selection via sfBDT

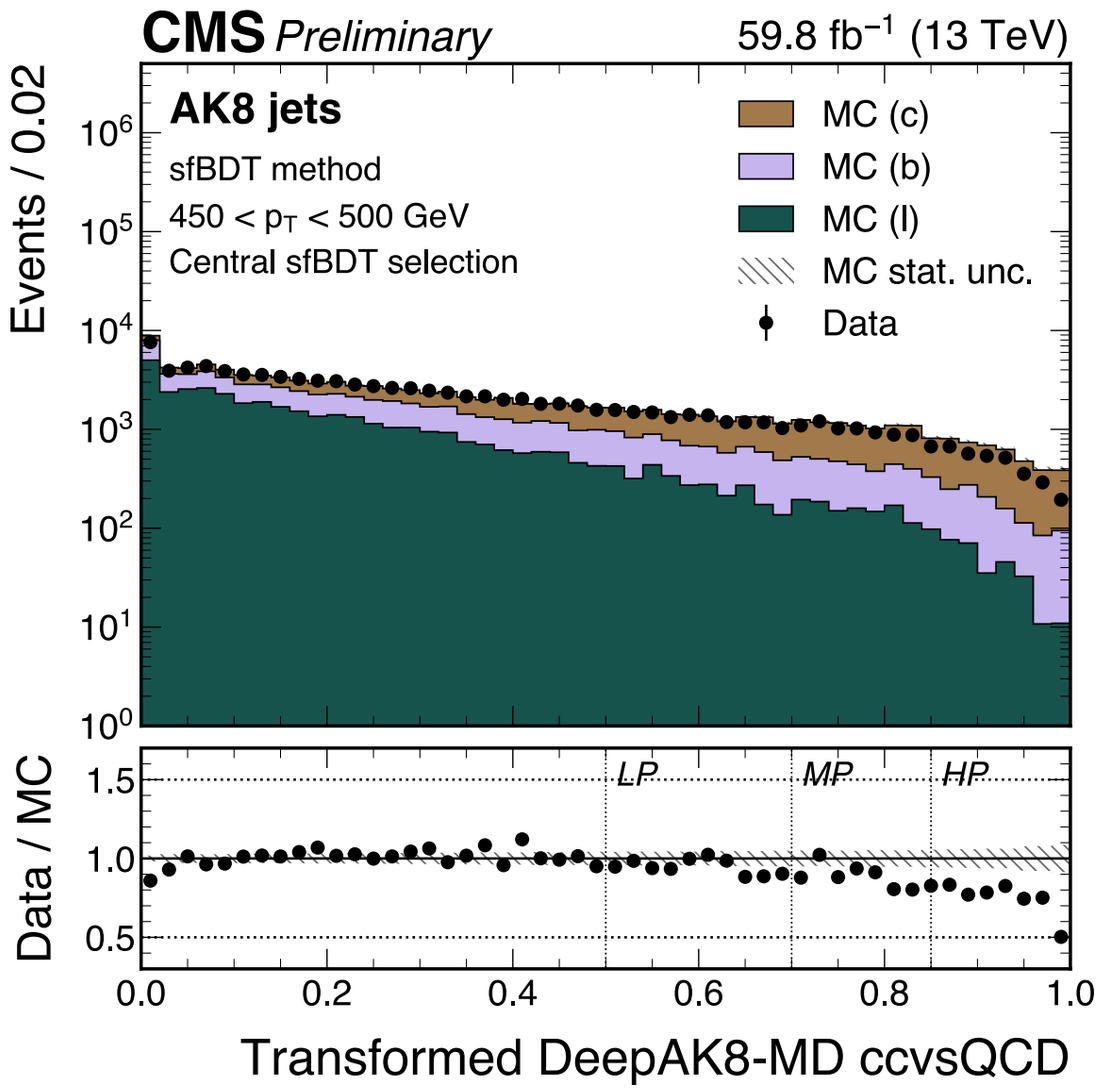




Data and MC distribution on transformed tagger discriminant,
applying the central sfBDT selection

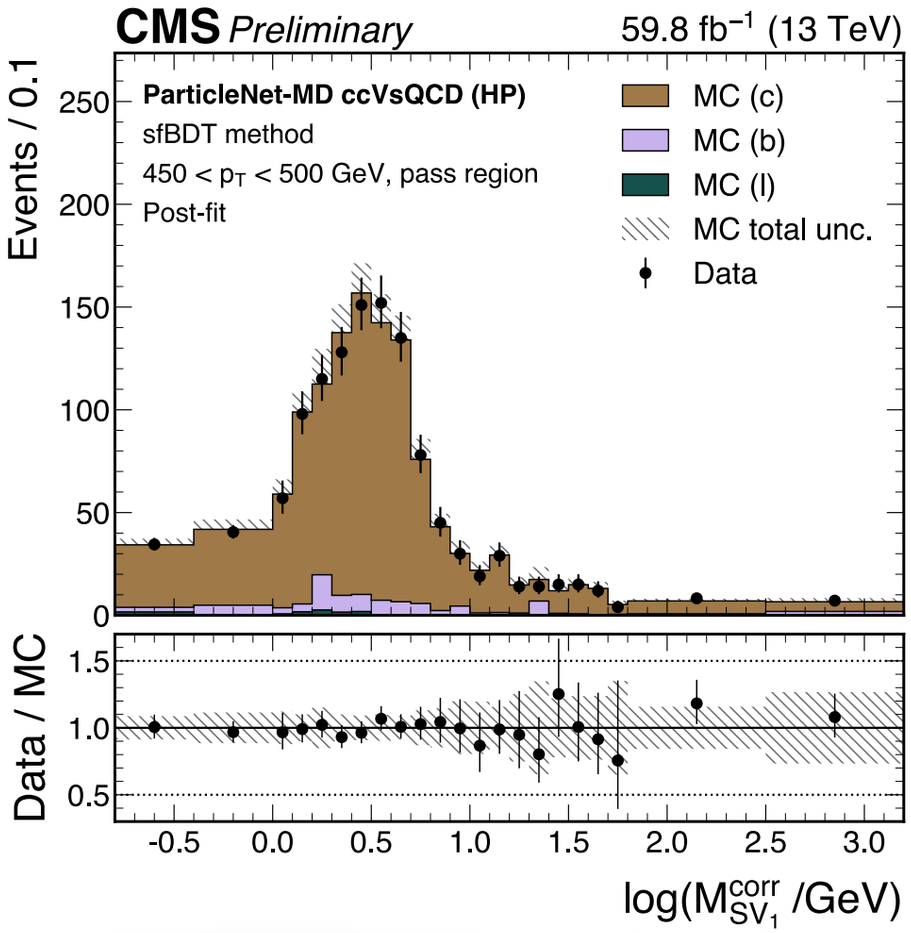
sfBDT method: selection via sfBDT

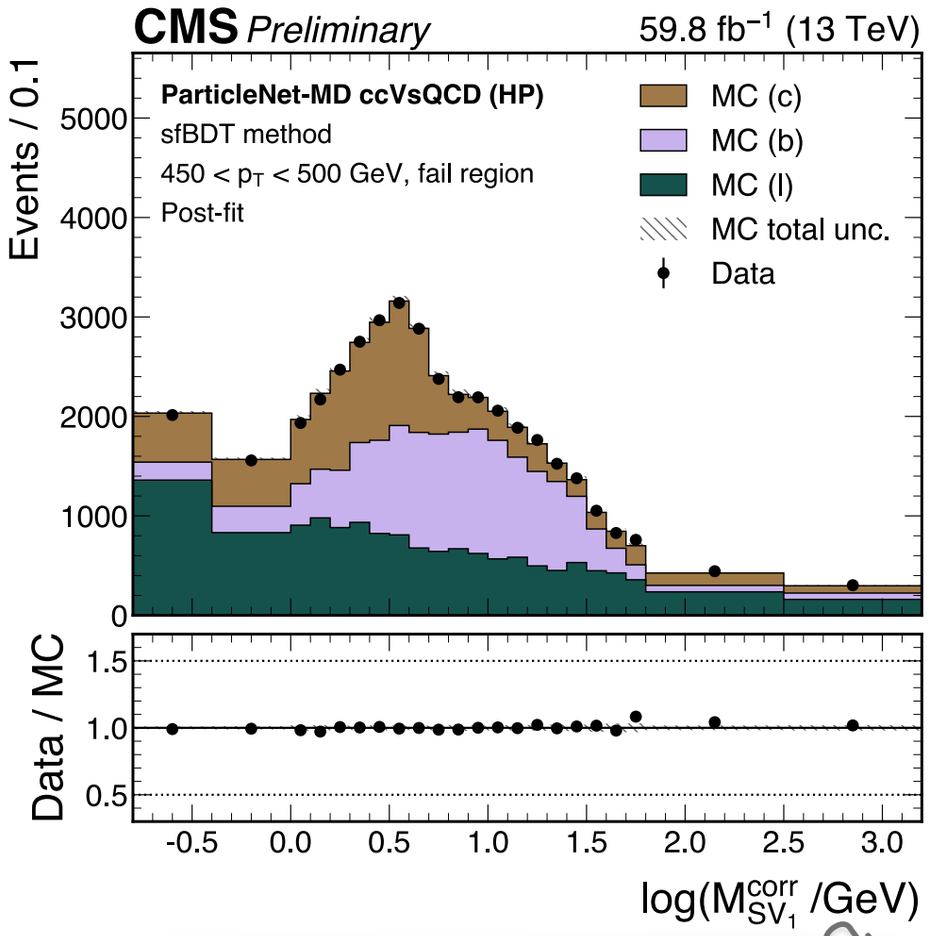




Data and MC distribution on transformed tagger discriminant,
applying the central sfBDT selection

sfBDT method: fit





Post-fit templates for calibrating ParticleNet-MD ccvsQCD (HP)

- Fit on the SV mass (log scale) to distinguish the b/c/light templates
- each flavour template assigned a free-floating SF in fit
- simultaneous fit in the “pass” and “fail” region

sfBDT method: full workflow

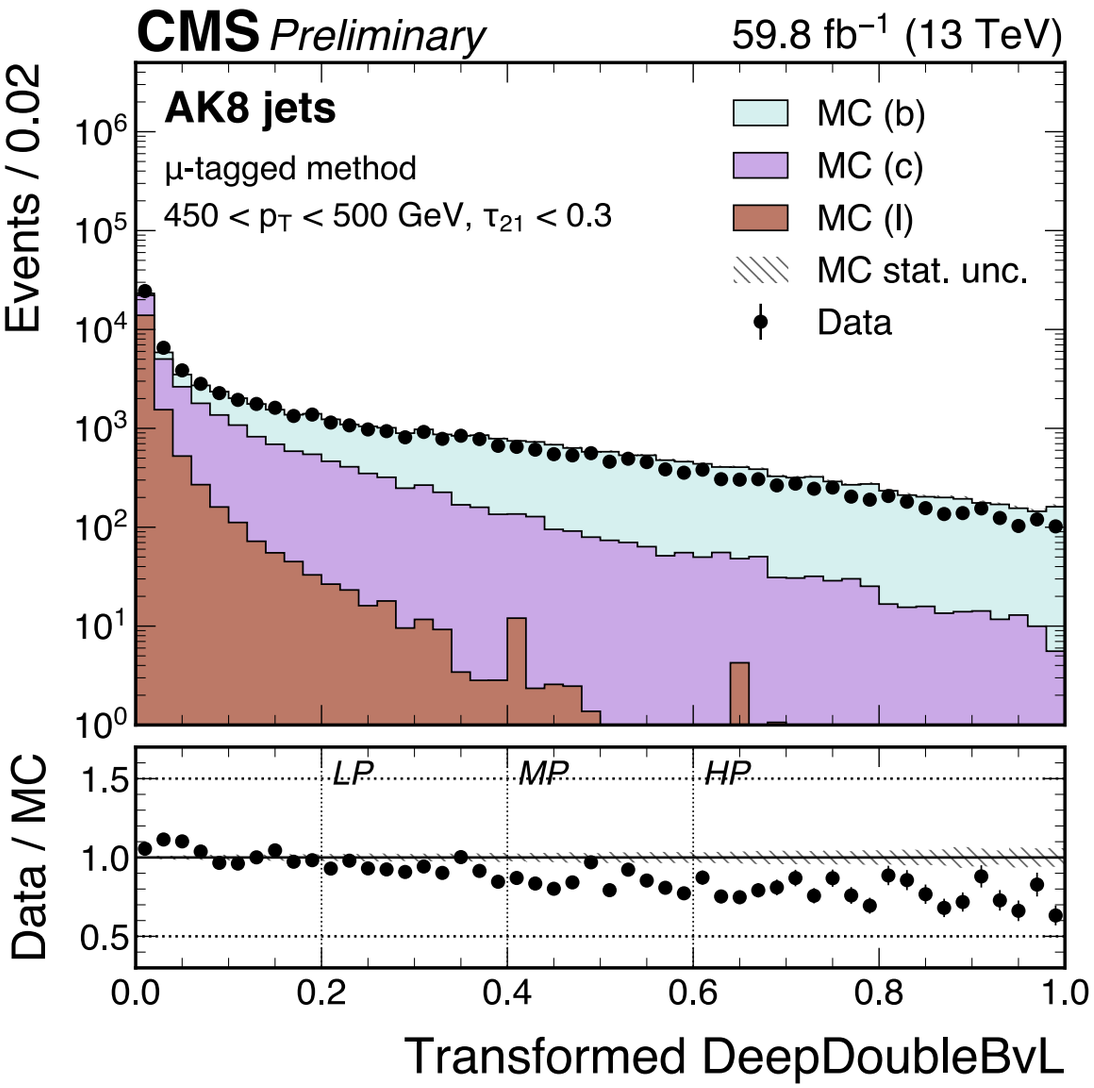
See also in CMS-PAS-BTV-22-001

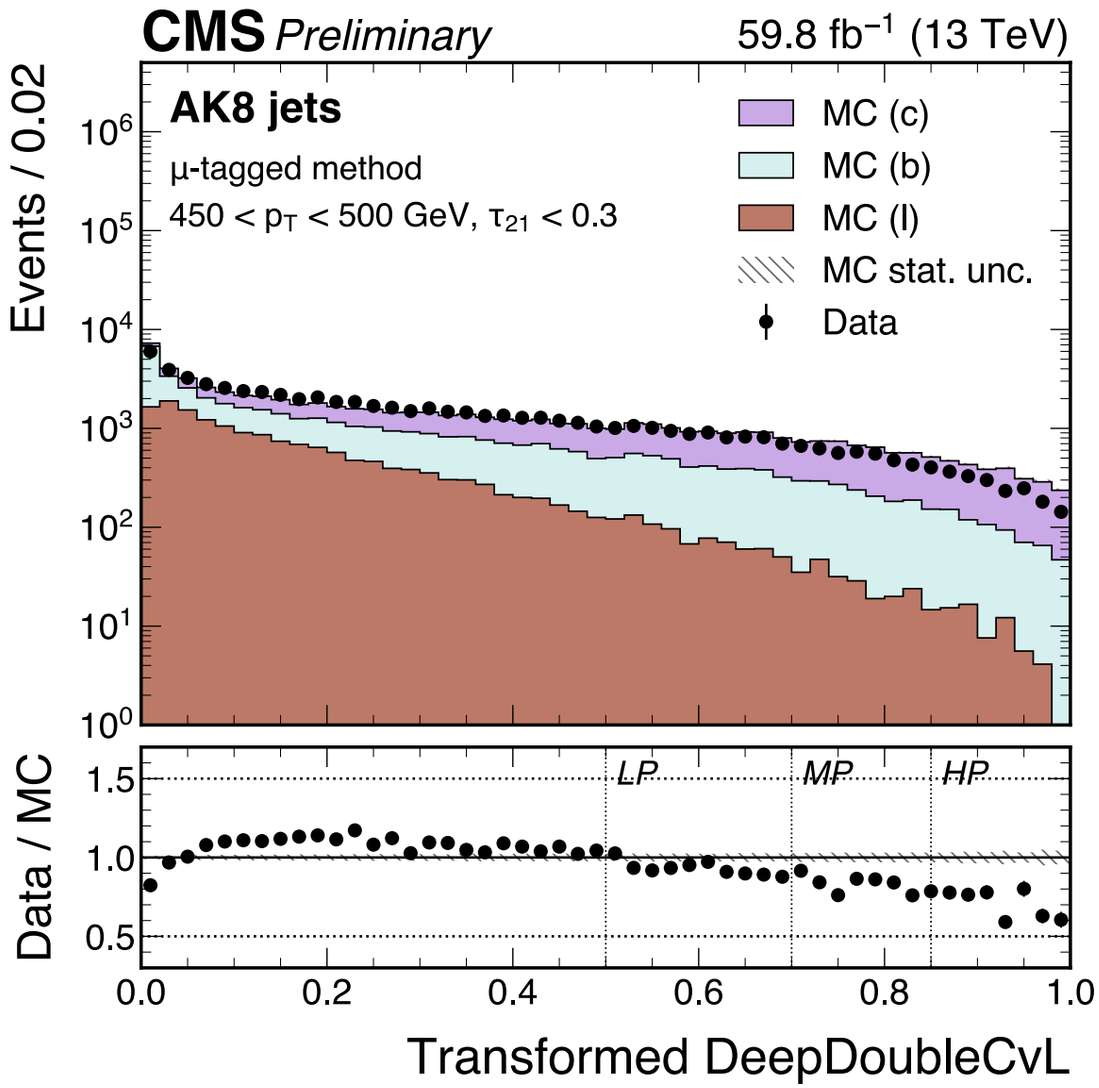
➔ **Target proxy jets**: g→b$\bar{b}$/c$\bar{c}$ jets from QCD multijet events, selected by a BDT trained on QCD jets

➔ **Workflow**:

- ❖ select events with logical OR of H_T triggers with multiple thresholds
- ❖ select both leading and subleading jet (if it exists), passing $p_T > 200$ GeV, $|\eta| < 2.4$, and $500 < m_{SD} < 200$ GeV, and requires N(matched-SV) ≥ 2
- ❖ reweight total simulated events to data on (jetIndex, H_T, p_T) bins
- ❖ all simulated jets are classified to b, c, and light flavour (based on the number of ghost-associated b and c hadrons, N_b and N_c)
- ❖ *[individual step]* train the sfBDT using a special QCD sample (enriched in bbb/ccc by production), with jets selected in the same criteria and requiring $N_b \geq 2$ or ($N_b = 0$ & $N_c \geq 2$), and the gen-level τ_{31} on hadrons applied to define signal and BKG jets for BDT training. With the sfBDT, for each target discriminant, determine the sfBDT curves on the 2D sfBDT—transformed tagger plane
- ❖ apply jet p_T selection for a dedicated fit point
- ❖ apply selection via the sfBDT curve, with 9×9 combinations in total
- ❖ perform fit on $\log(M_{\mathrm{SV}_1}^{\mathrm{corr}}/\mathrm{GeV})$: fit simultaneously in the “pass” and “fail” tagger region, with three SFs assigned to the b/c/light categories in the “pass” region. Target SF is SF_b (SF_c) for calibrating X→b$\bar{b}$ (c$\bar{c}$) tagger
- ❖ assign two sources of external uncertainties:
  - merge the target SFs (with all varied sfBDT selections) into one: the uncertainty in the final SF is larger than that of the individual fit result, and its central value is positioned in the middle of the entire SF series.
  - consider the deviation of SF as an external uncertainty, from the nominal SF to the SF derived in the schemes to reweight on the fit variable, or the sfBDT discriminant

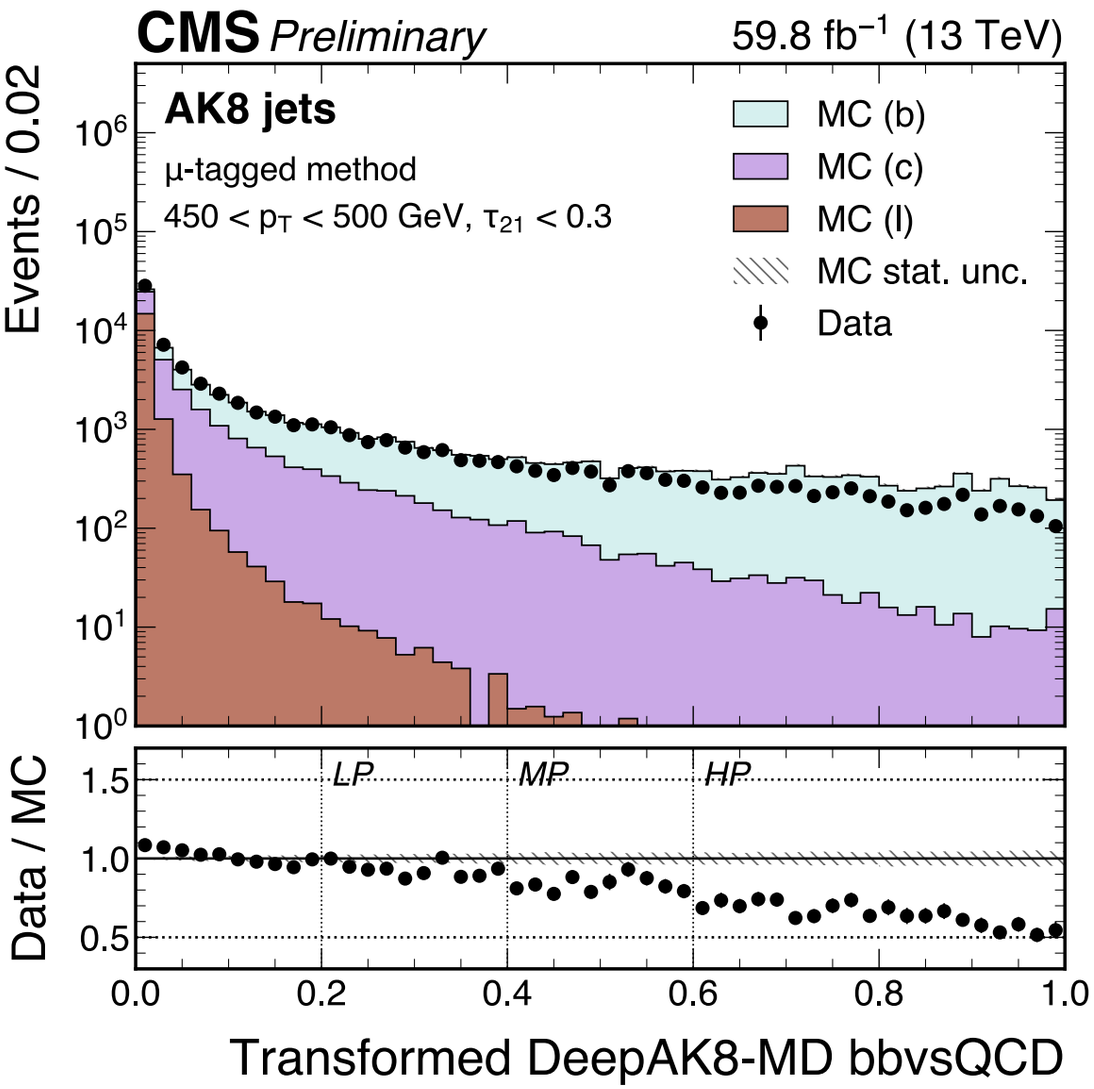
μ-tagged method: selected jets

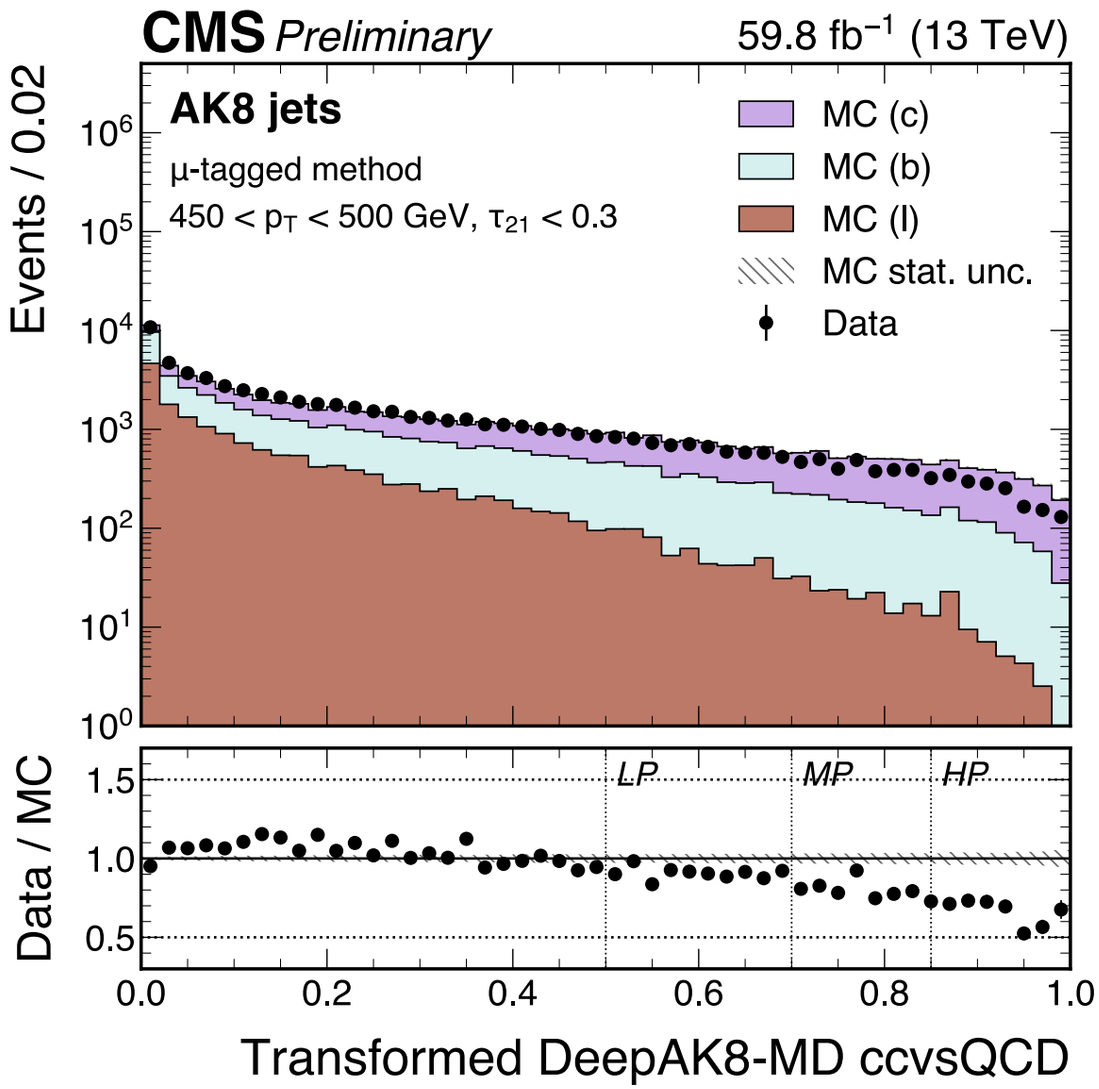




Data and MC distribution applying the μ-tagged requirement and the central τ_{21} < 0.3 cut

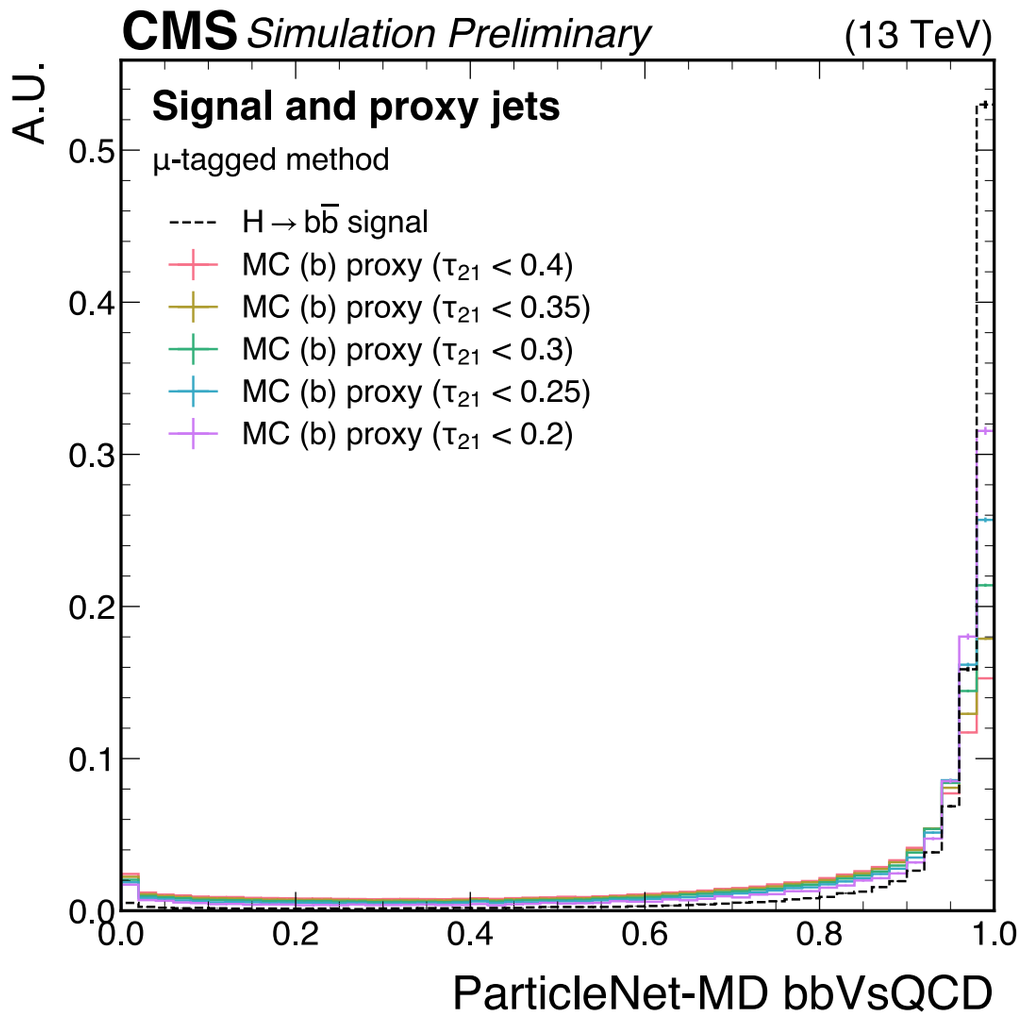
μ-tagged method: selected jets

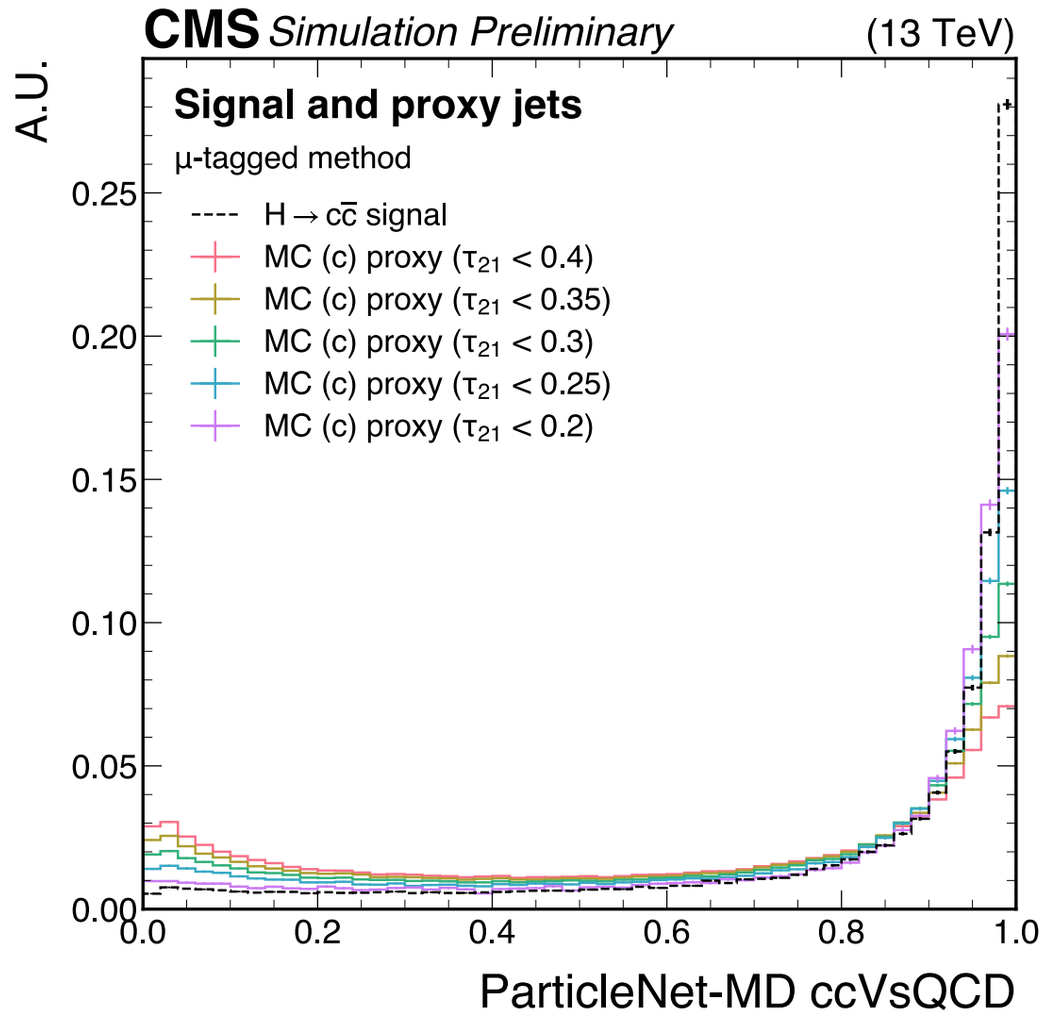




Data and MC distribution applying the central τ_{21} < 0.3 cut

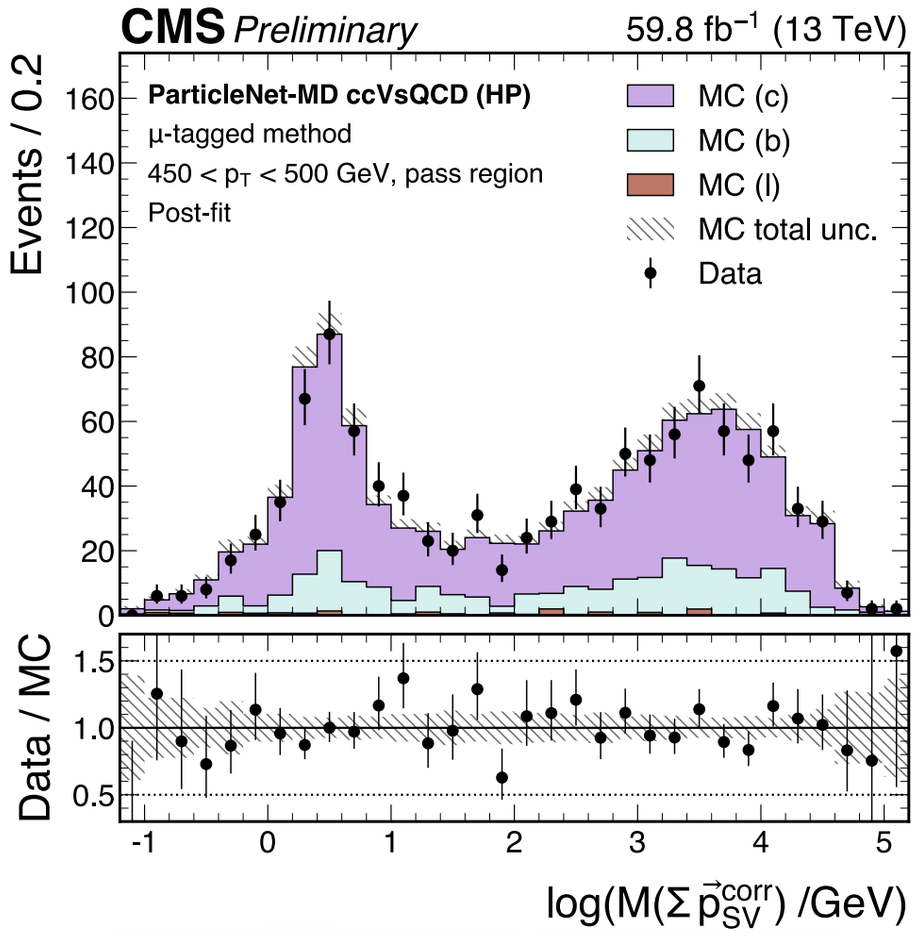
μ-tagged method: proxy—signal similarity

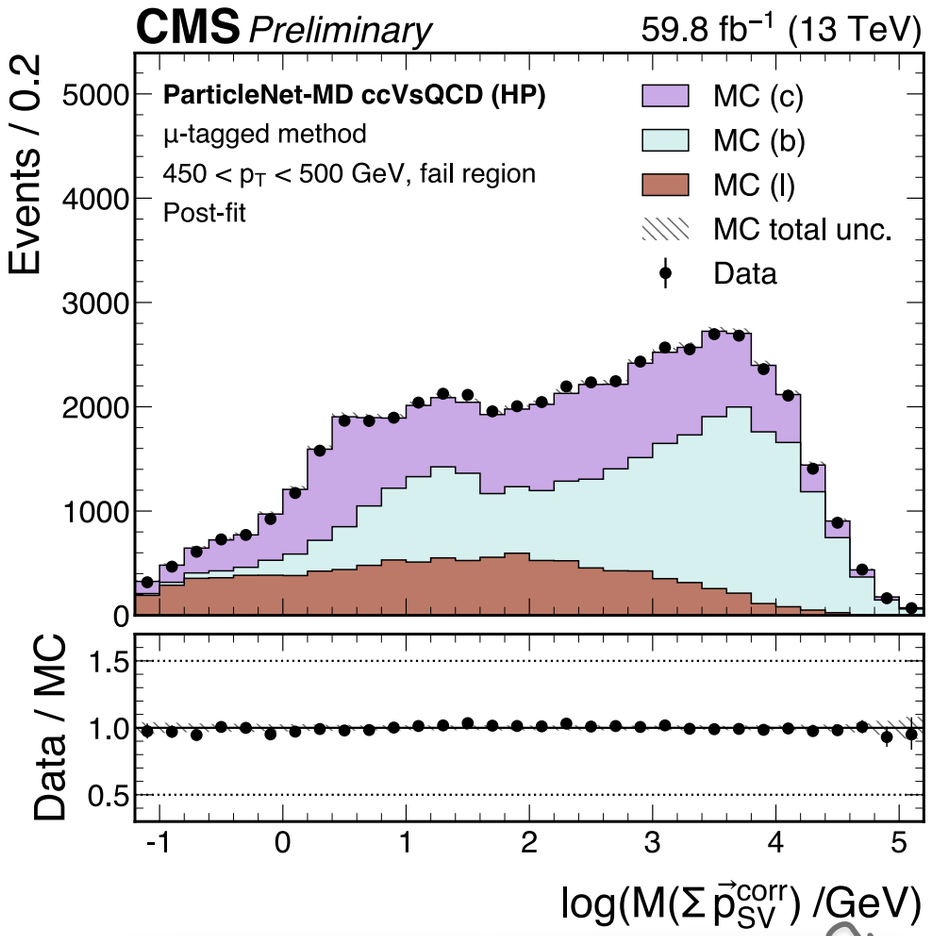




Signal and proxy similarity when different τ_{21} selections are applied

μ-tagged method: fit





Post-fit templates for calibrating ParticleNet-MD ccvsQCD (HP)

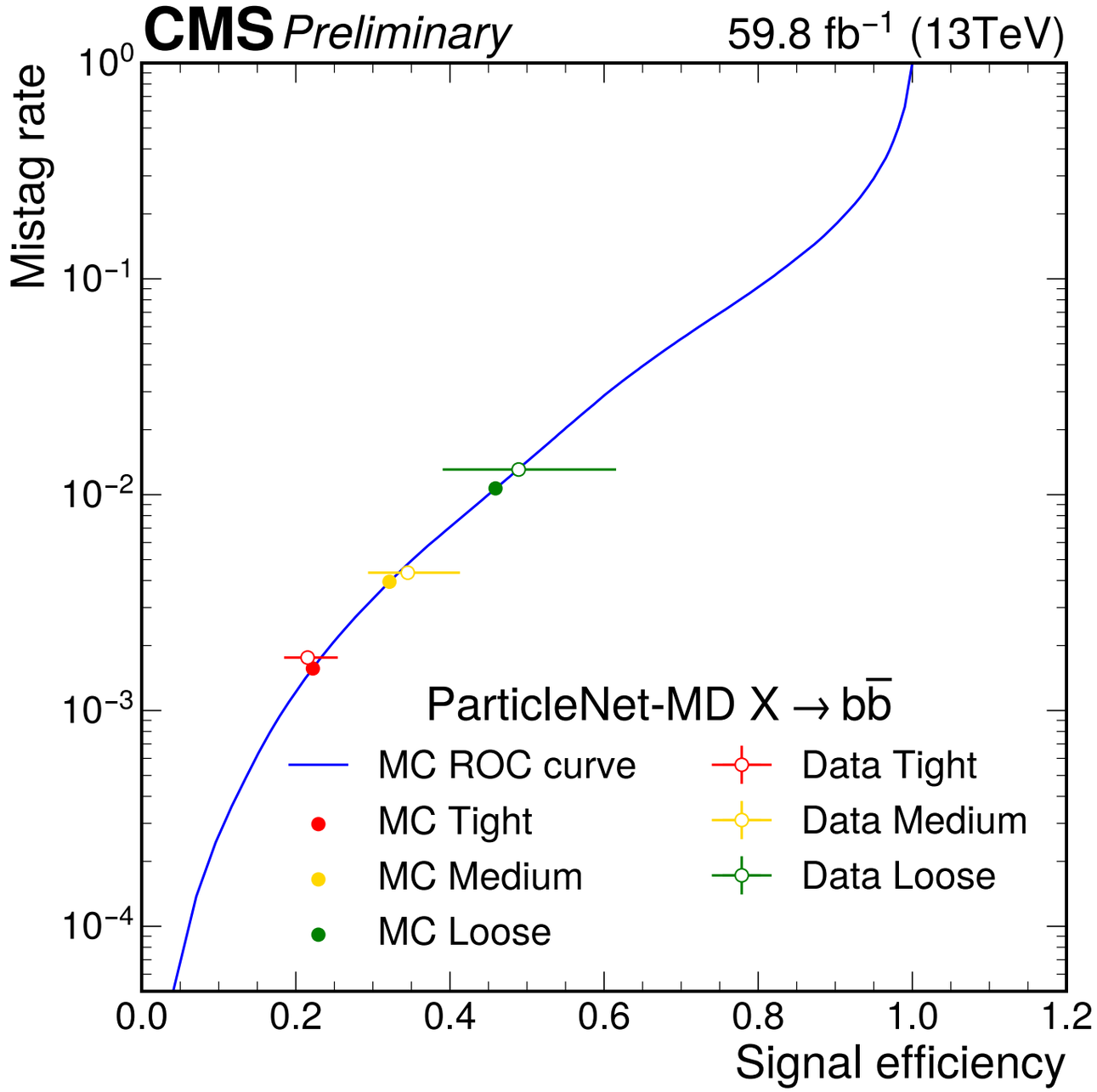
- Fit on the invariant mass of SV 4-vectors (log scale)
- each flavour template assigned a free-floating SF in fit
- simultaneous fit in the “pass” and “fail” region

μ-tagged method: full workflow

See also in CMS-PAS-BTV-22-001

- **Target proxy jets:** g→b$\bar{b}$/c$\bar{c}$ jets from QCD multijet events (triggered by the existence of a soft muon in jets), with the additional offline requirements on the soft muon and N-subjettiness ratio τ_{21}
- **Workflow**:
  - select events with online triggers: requiring AK4 or AK8 jet with $p_T > 300$, and including a muon with $p_T > 5$ GeV
  - select both leading and subleading jets (if exists) with offline $p_T > 350$ GeV, $|\eta| < 2.4$, $m_{SD} > 40$ GeV
  - apply a jet-based reweighting from QCD MC to data subtracting the t$\bar{t}$ and single top and V+jets MC, on the (p_T, η, τ_{21})
    (note: reweighting on τ_{21} mitigate the mismodelling of gluon splitting jets by PYTHIA-based QCD MC, and reduce its discrepancy with the MG-based QCD MC on tau21 and all tagger discriminants)
  - all simulated jets are classified to b, c, and light flavour (based on the number of ghost-associated b and c hadrons, N_b and N_c)
  - apply jet p_T selection for a dedicated fit point
  - apply the selection via τ_{21}: the main scheme has $\tau_{21} < 0.3$, and it is varied from 0.2 to 0.4 in the auxiliary fit
  - perform fit on $\log(M(\sum \vec{p}_{\mathrm{SV}}^{\,\mathrm{corr}})\ /\mathrm{GeV})$: fit simultaneously in the “pass” and “fail” tagger region, with three SFs assigned to the b/c/light categories in the “pass” region. Target SF is SF_b (SF_c) for calibrating X→b$\bar{b}$ (c$\bar{c}$) tagger
  - assign two sources of external uncertainties:
    - merge the target SFs (with different τ_{21} selections) into one, by expanding the uncertainty of the nominal fitted SF to cover the maximum deviation of the SFs.
    - consider the deviation of SF as an external uncertainty, from the nominal SF to the SF derived in the scheme to reweight on the fit variable

Boosted Z method: ROC from data

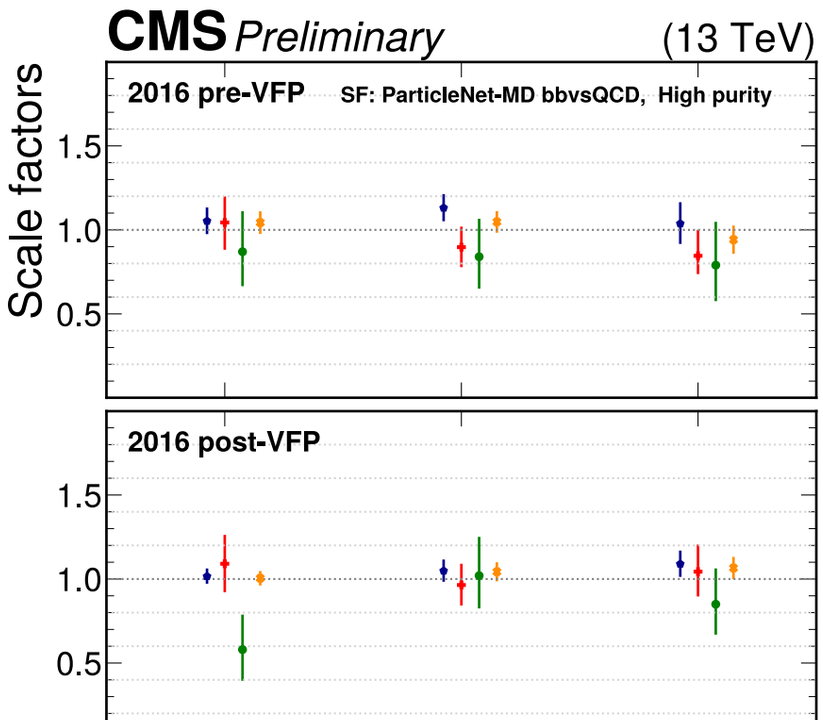


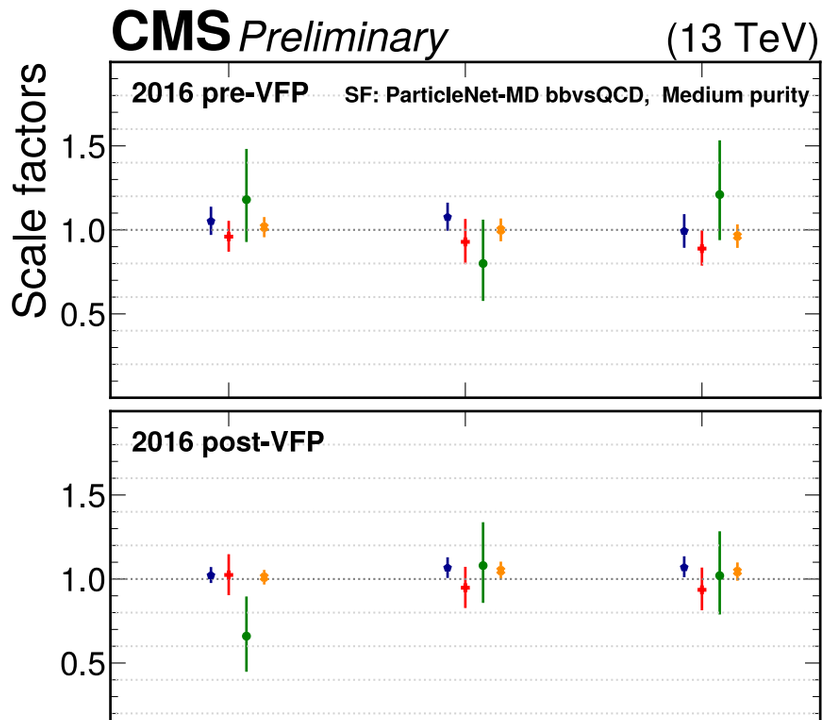
Boosted Z method: full workflow

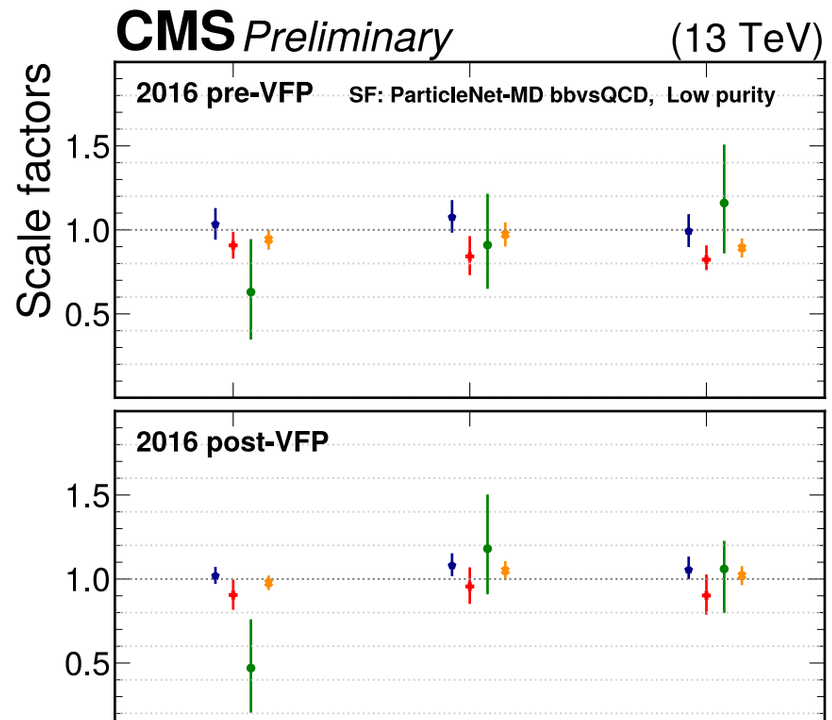
See also in CMS-PAS-BTV-22-001

- **Target proxy jets:** $Z\to b\bar{b}$ jets from Z+jets events
- **Workflow**:
  - select events with online triggers: a combination of requirements on the jet p_T, trimmed mass, and event H_T
  - select the leading AK8 jet with $p_T > 450$ GeV, $|\eta| < 2.4$, $M_{PNet} > 40$ GeV as the target jet; presence of a subleading AK8 jet with $p_T > 200$ GeV and $|\eta| < 2.4$ as the recoil jet is required
  - veto events with loose electrons/muons, and events with b-tagged AK4 jets (to suppress $t\bar{t}$ events)
  - Perform the fit on the 2D binned histogram of (M_{PNet}, p_T)
    - Z+jets and W+jets modelled by MC
    - QCD in the "fail" region estimated directly from data, after subtracting other MC processes
    - QCD in the "pass" region estimated from the "fail" region through a fitted polynomial
    - three SFs assigned to Z+jets (3 p_T bins) and one SF to W+jets, in the "pass" region
    - the target SF is that for the Z+jets since $Z\to b\bar{b}$ is the predominated contribution in the "pass region"
- Note: the boosted Z method is only used to derive SFs for:
  - ParticleNet-MD bbvsQCD in HP, MP, LP; DeepDoubleBvL in HP, MP; double-b in HP
  - DeepAK8-MD bbvsQCD has a residue mass correlation which sculpts the QCD background in the "pass" region, making the method invalid

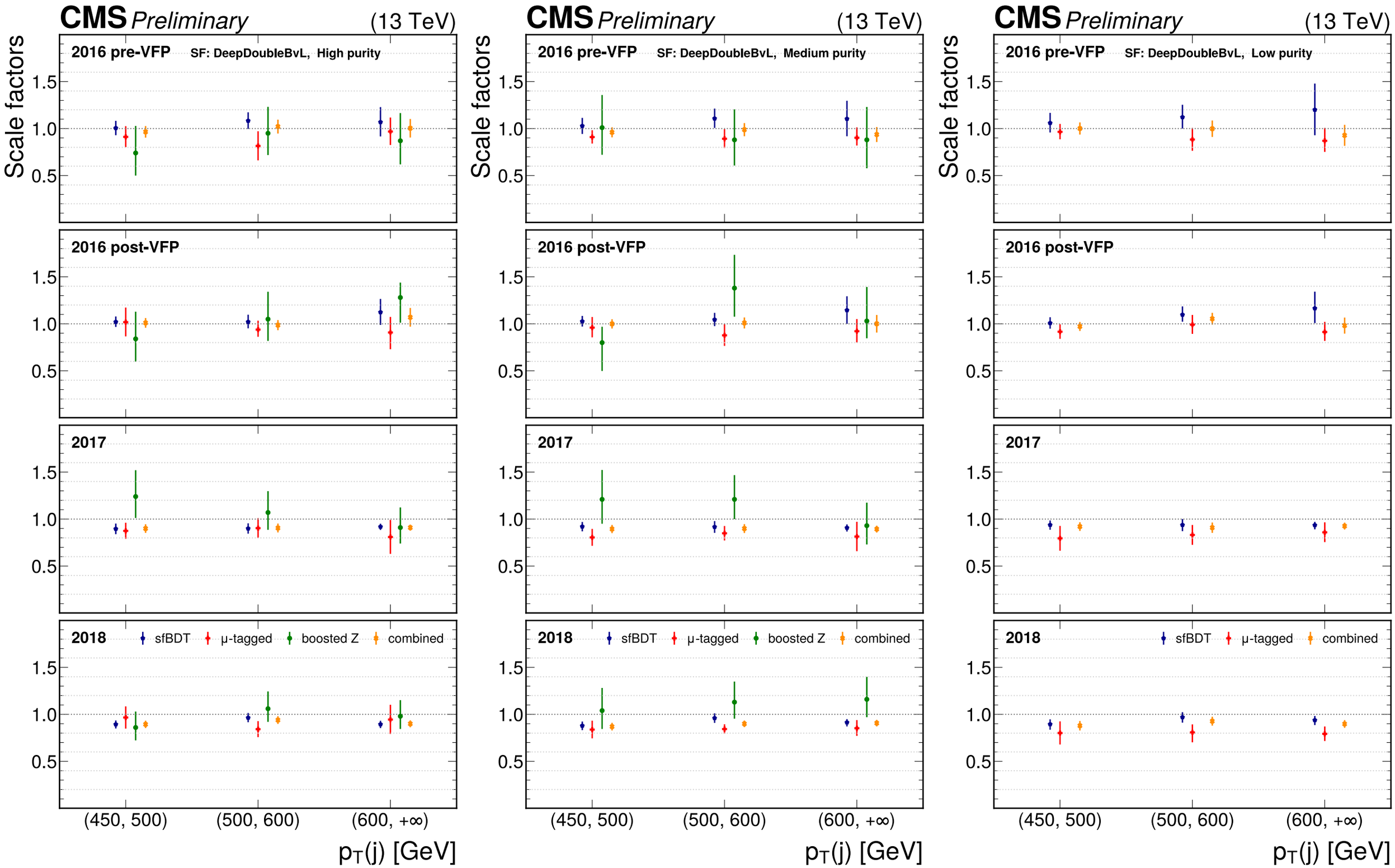
Full results: ParticleNet-MD bbvsQCD



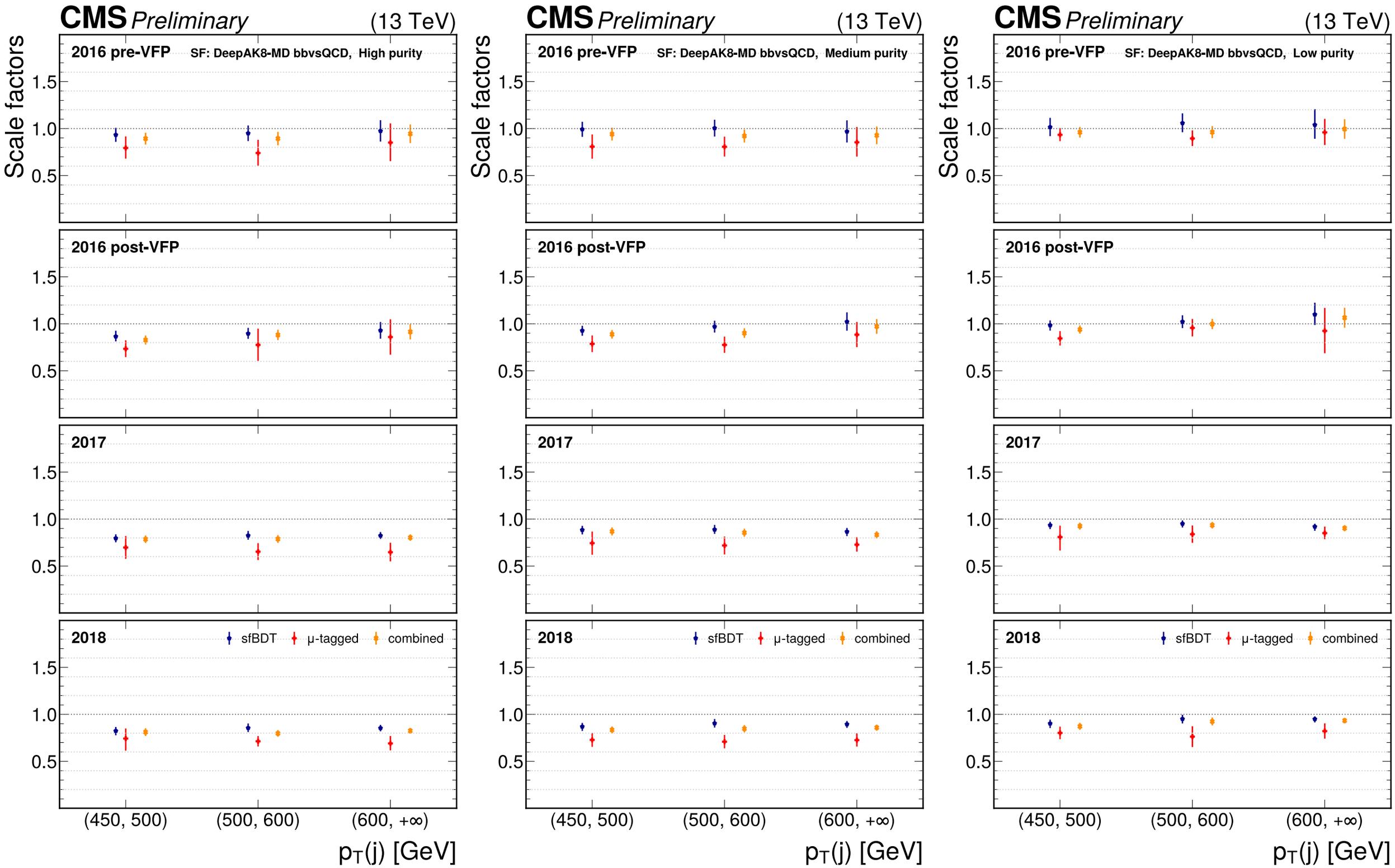




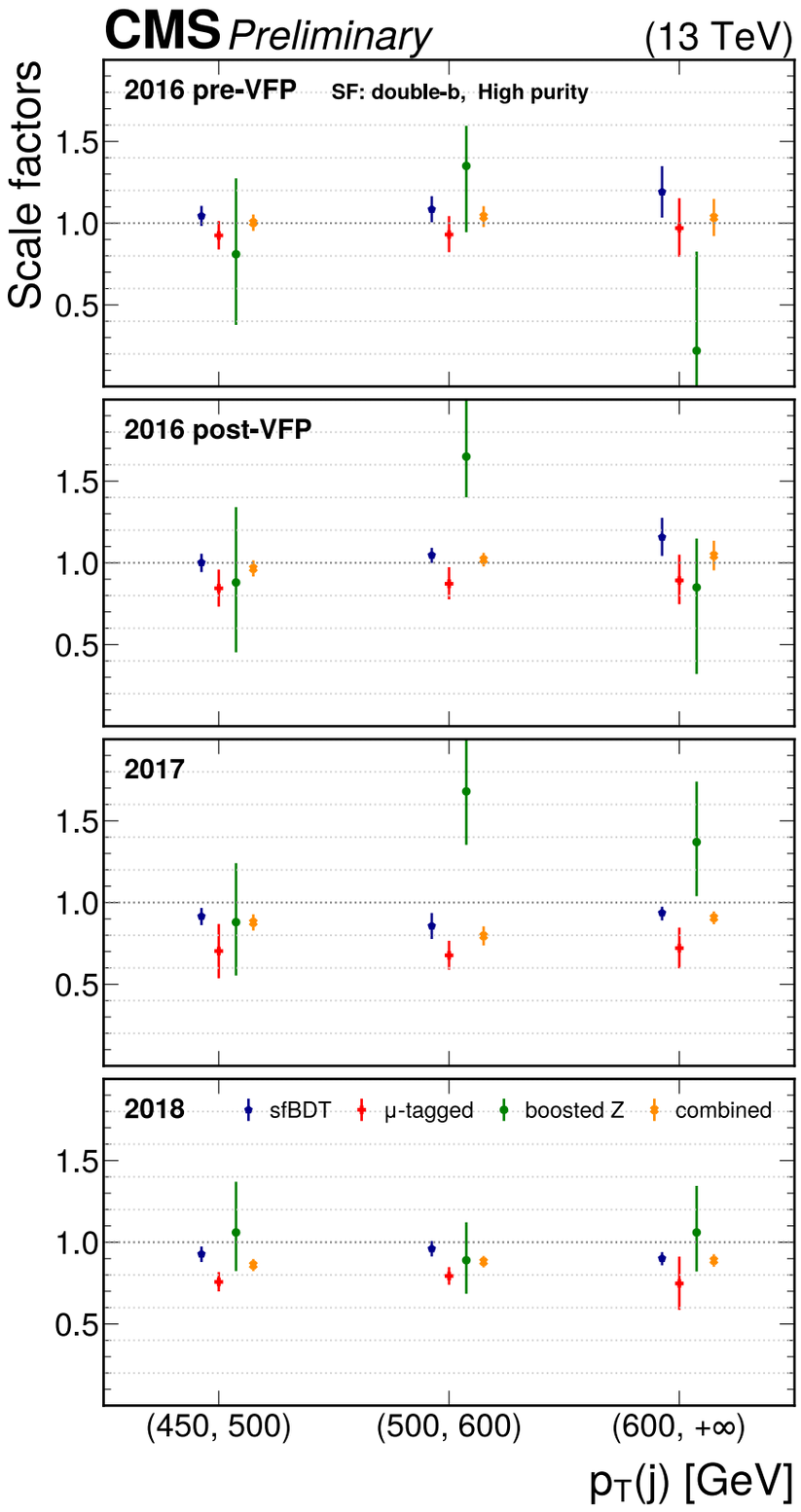
Full results: DeepDoubleBvL

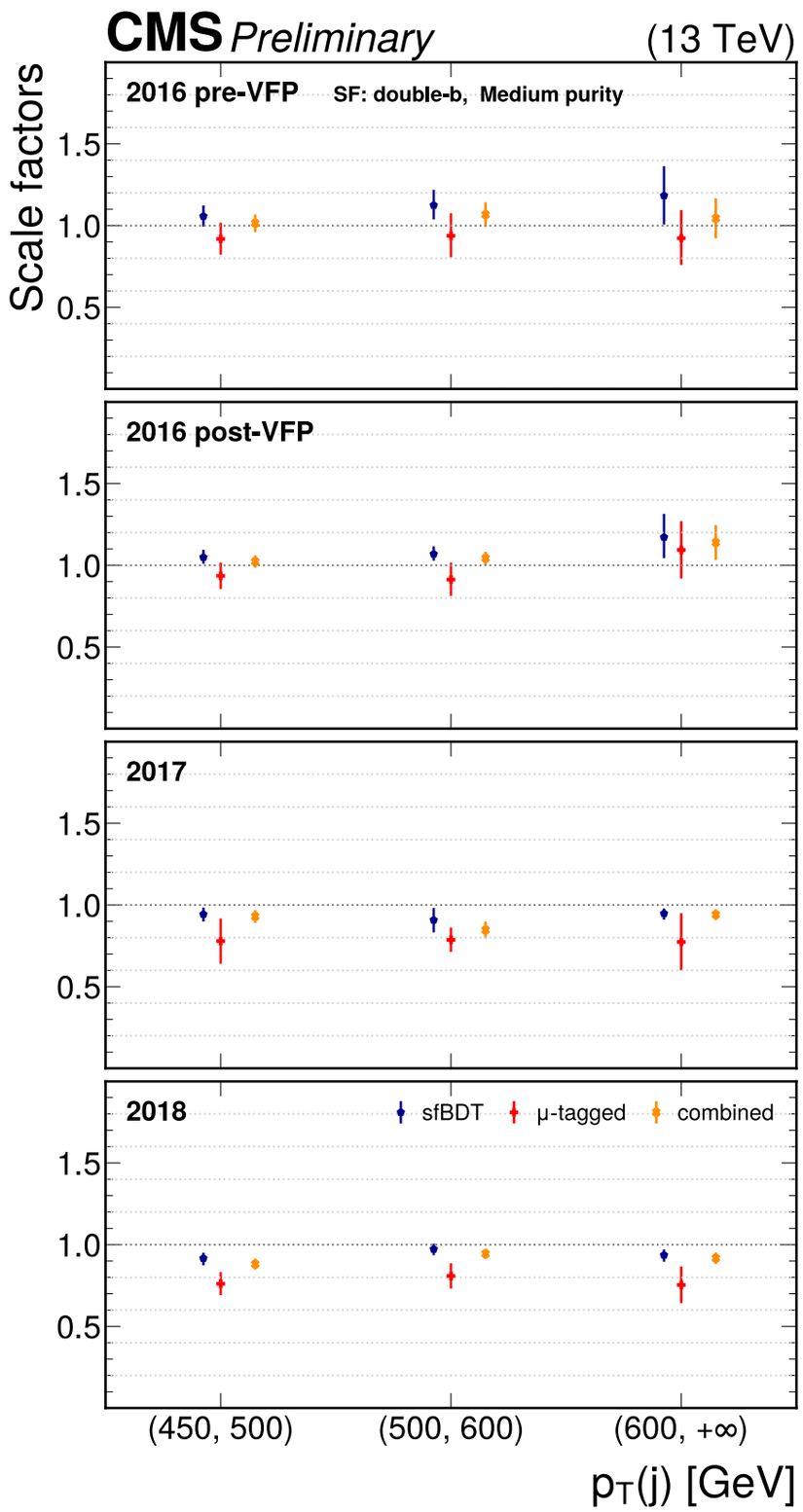


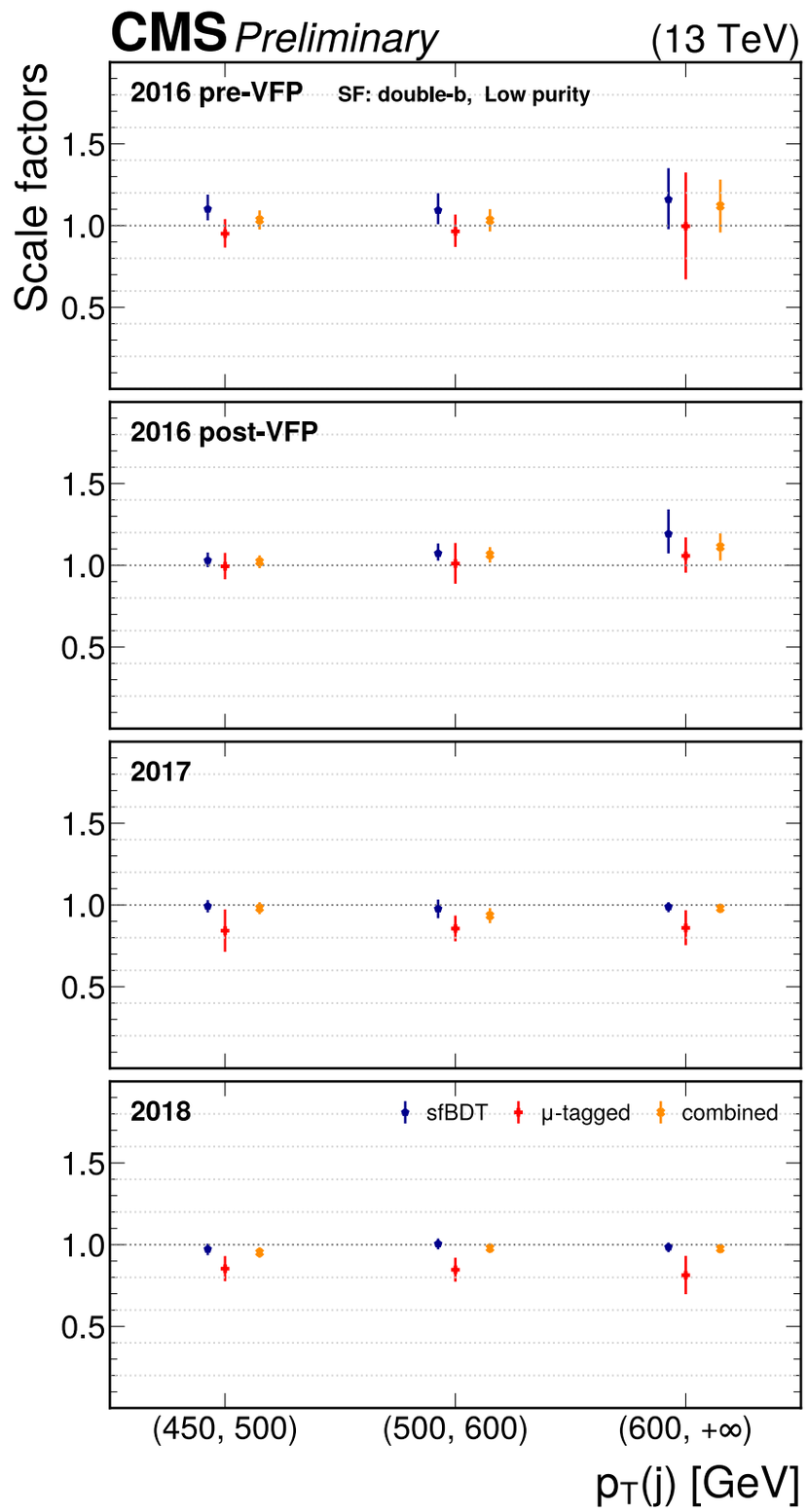
Full results: DeepAK8-MD bbvsQCD



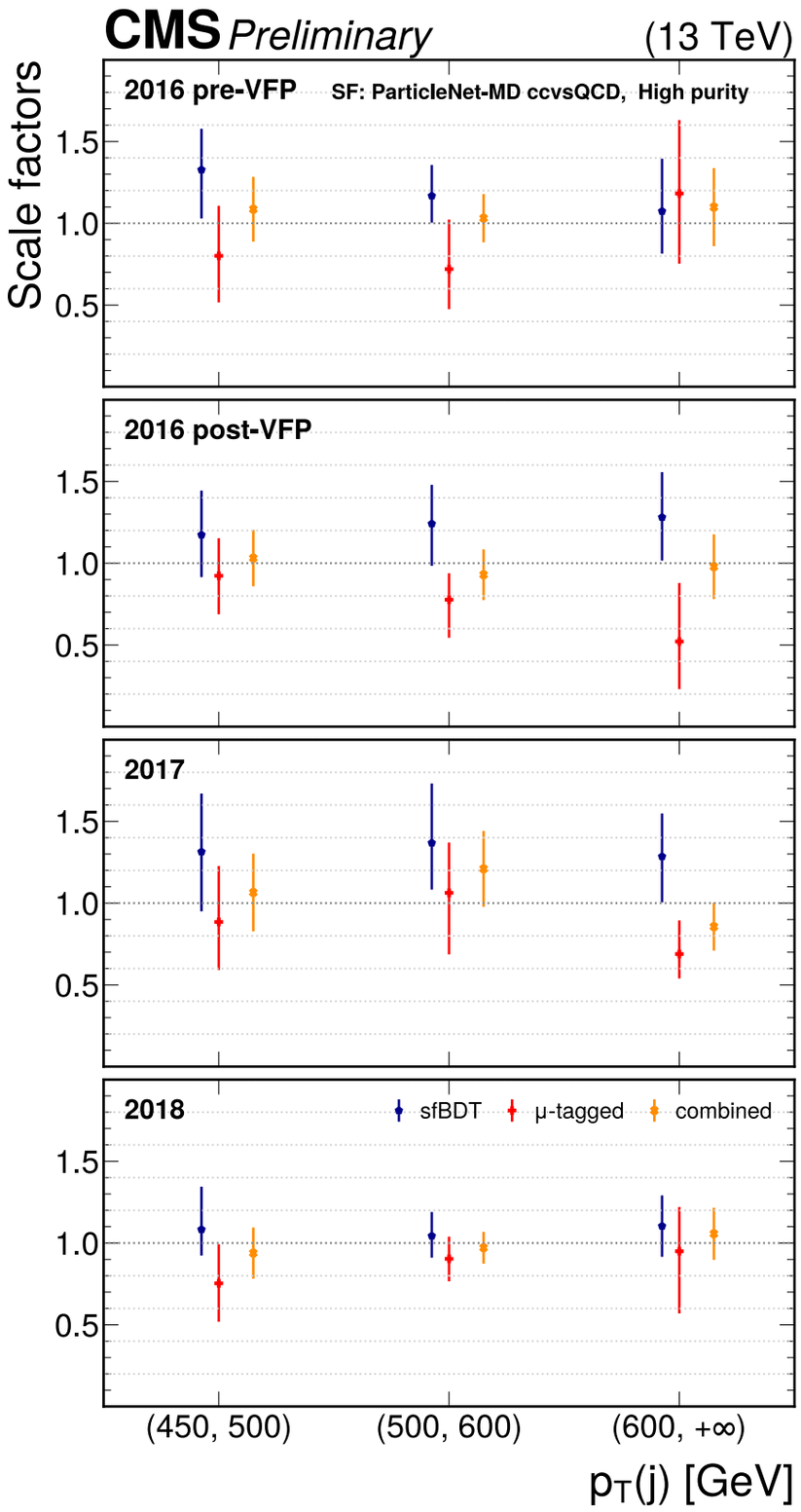
Full results: double-b

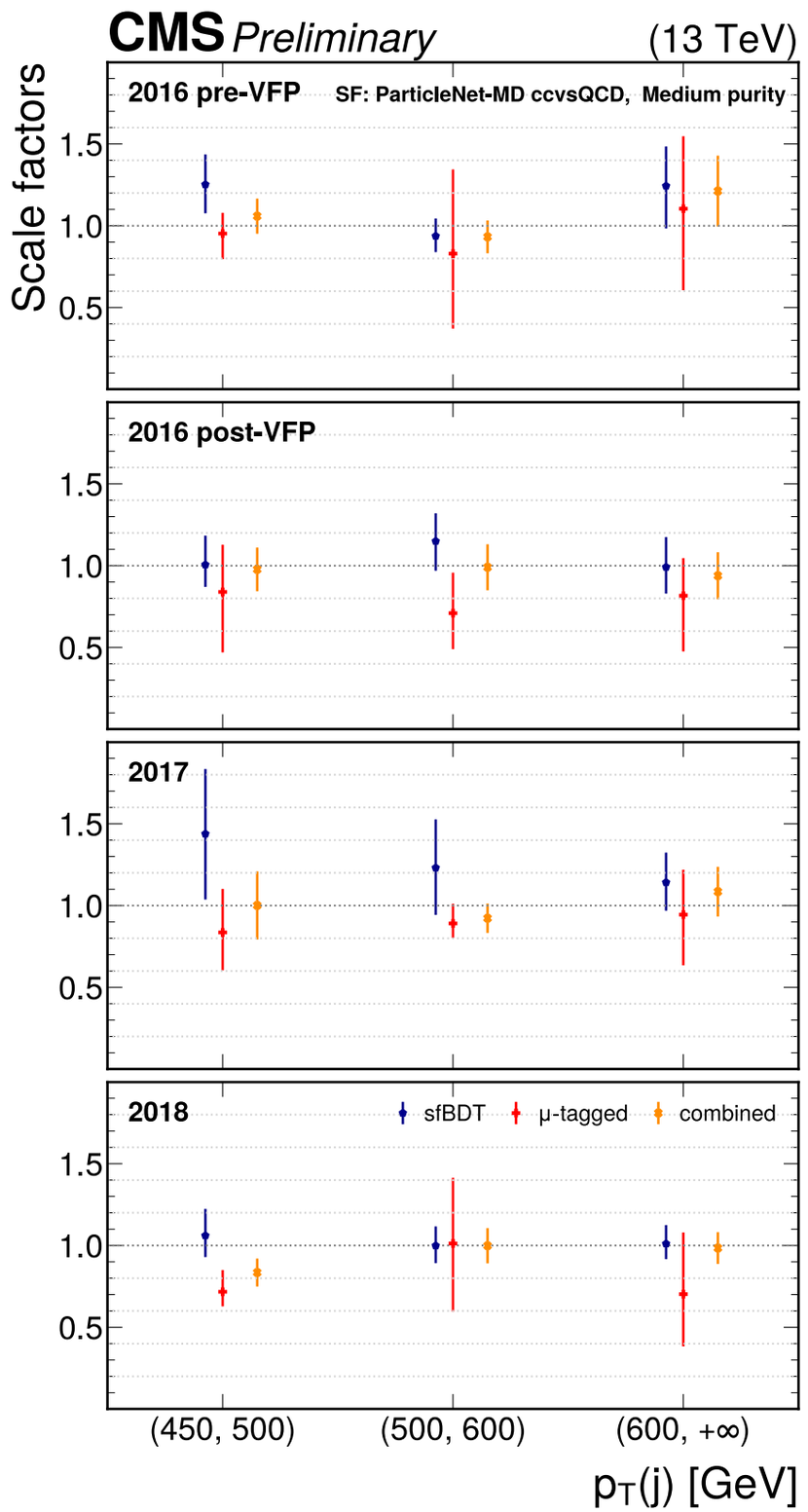


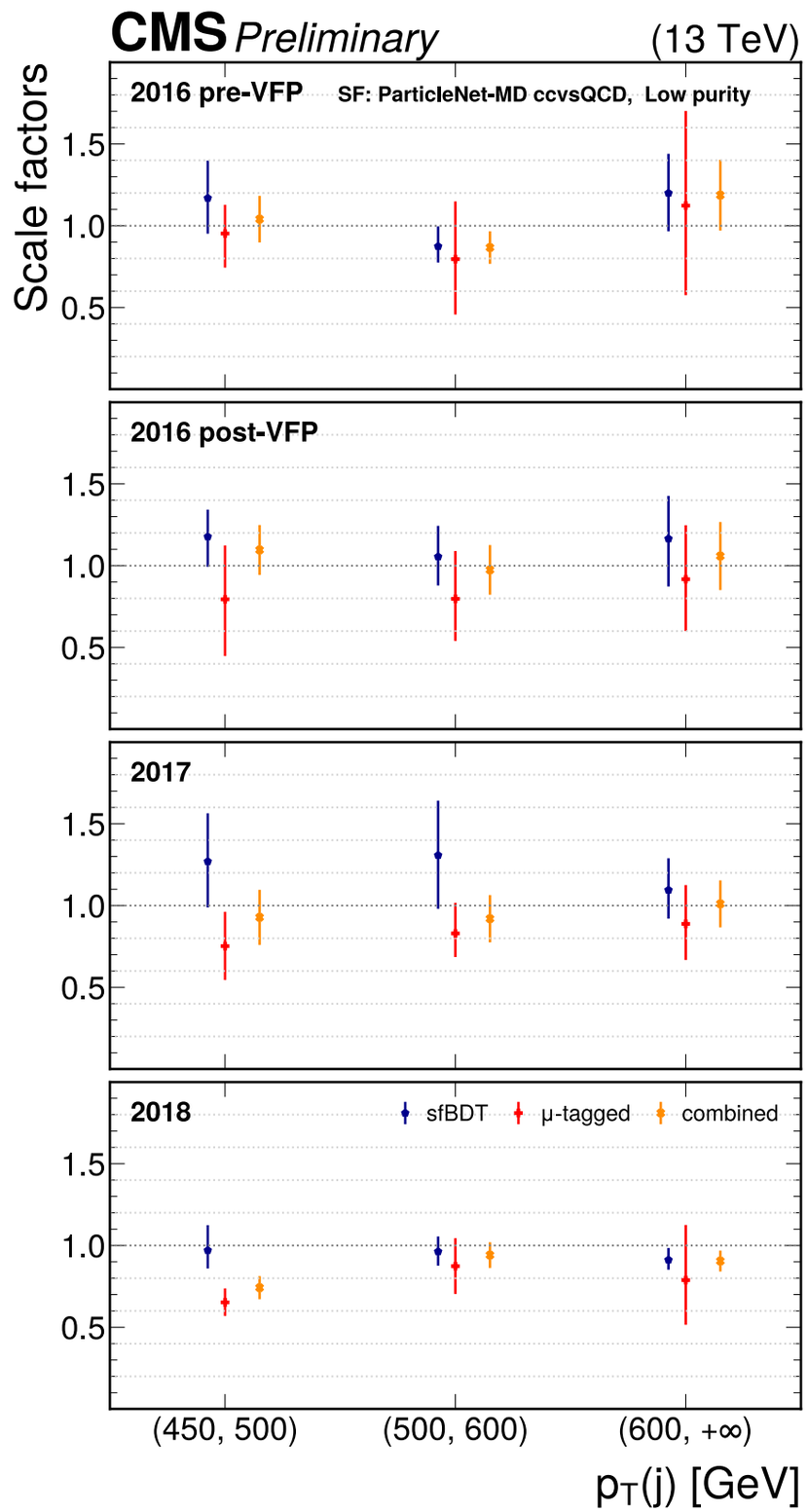




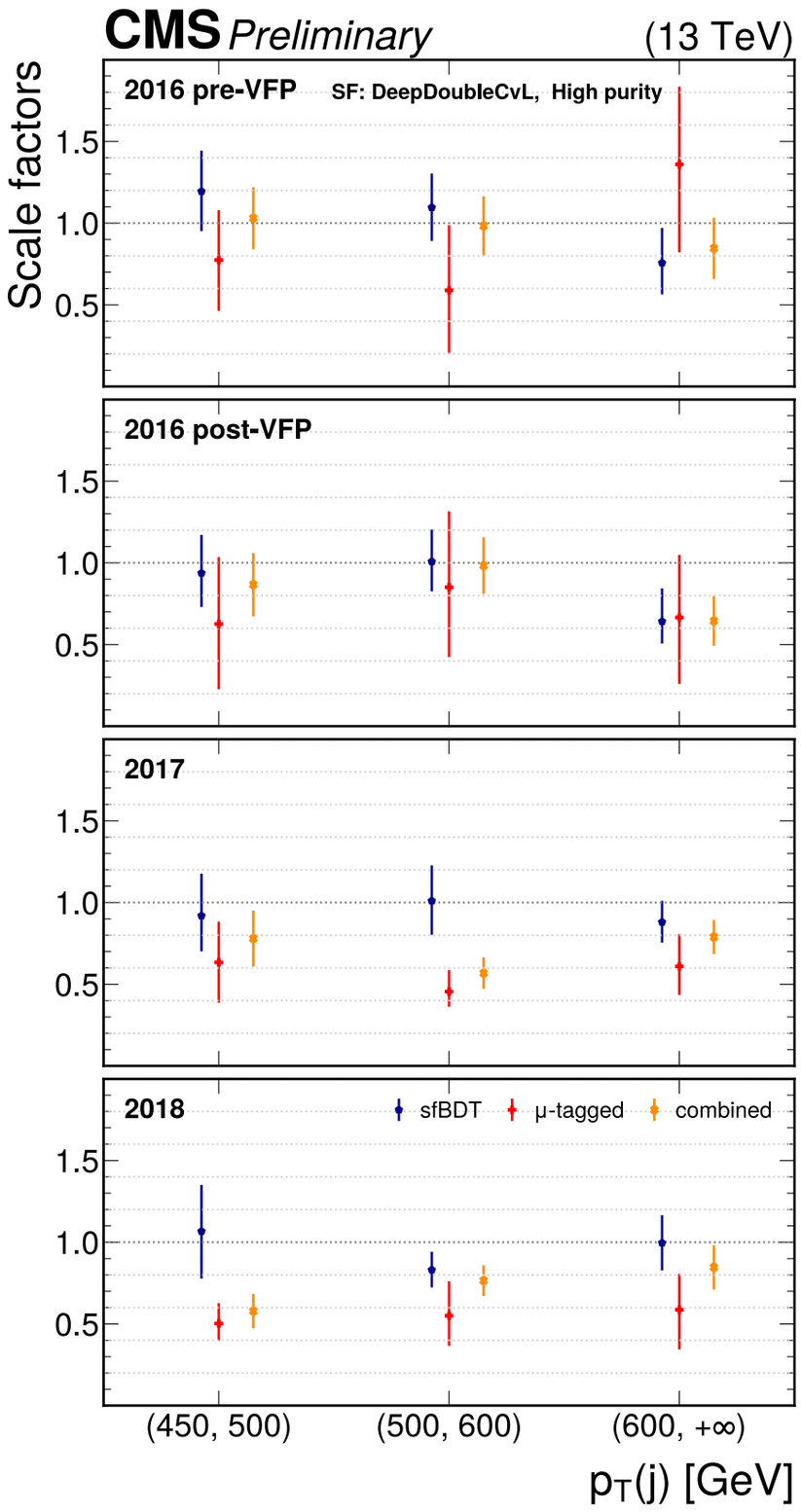
Full results: ParticleNet-MD ccvsQCD

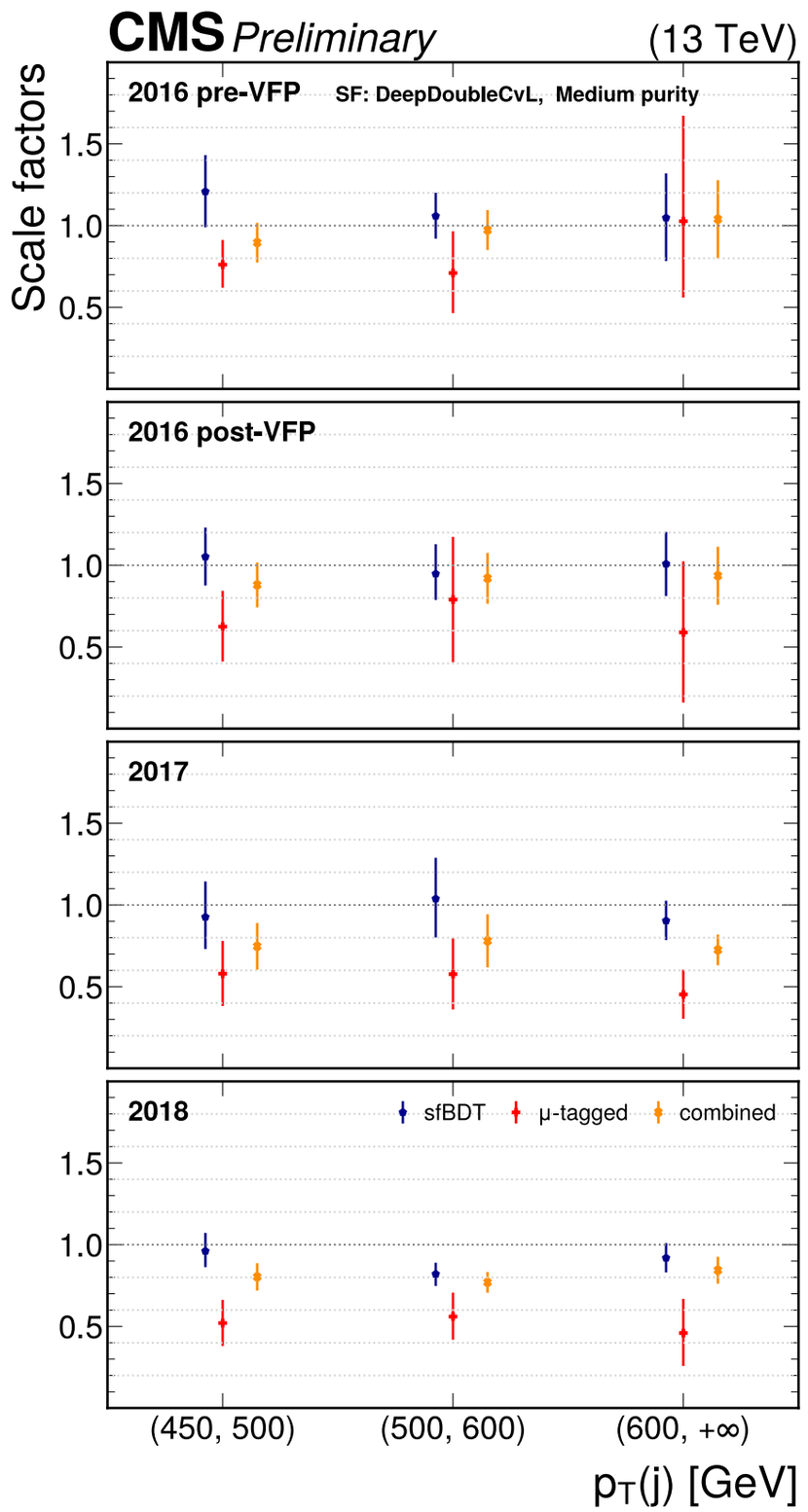


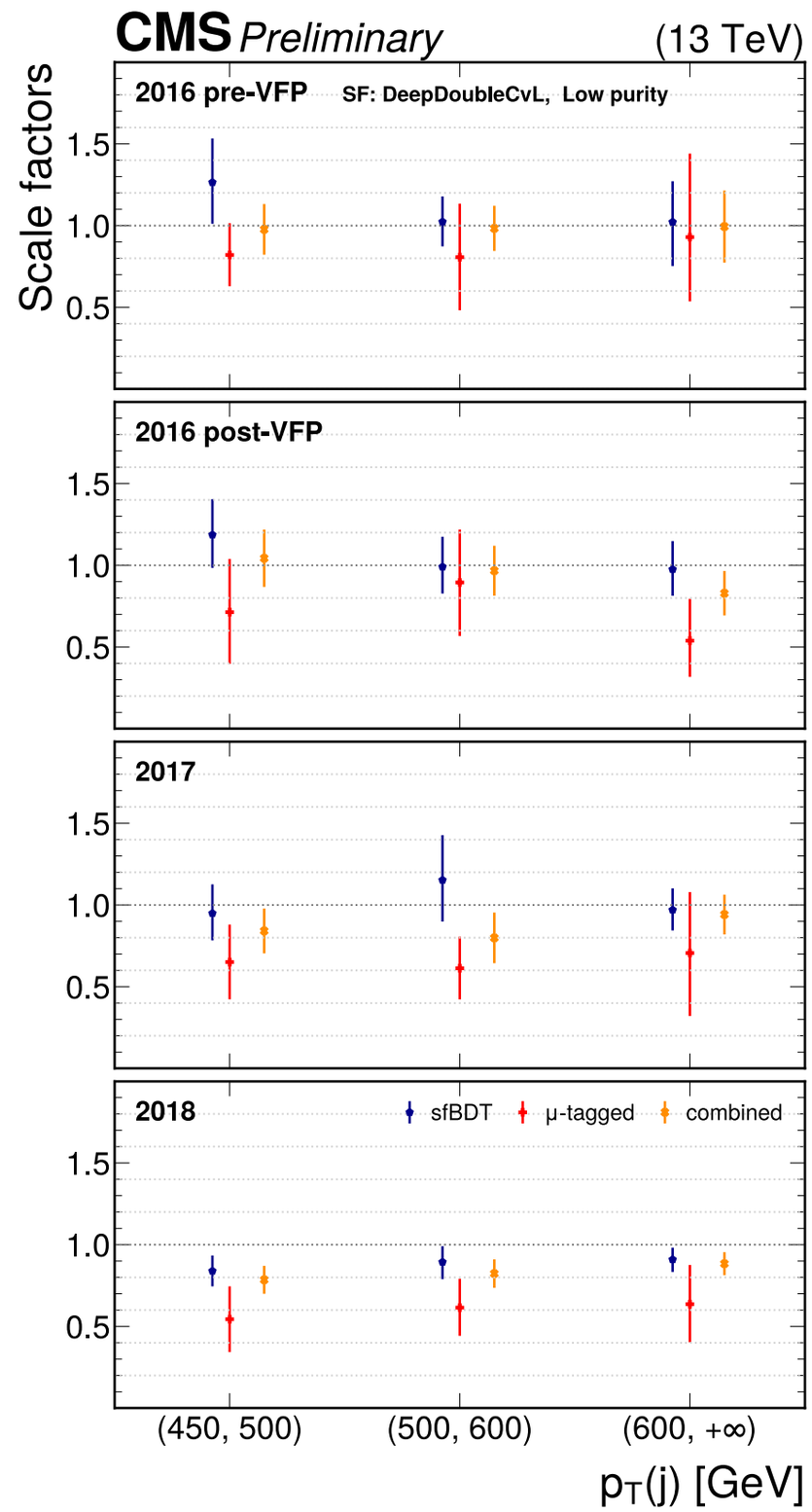




Full results: DeepDoubleCvL







Full results: DeepAK8-MD ccvsQCD