Identification of Lorentz-boosted jets in the CMS experiment

Donato Troiano[1,2]

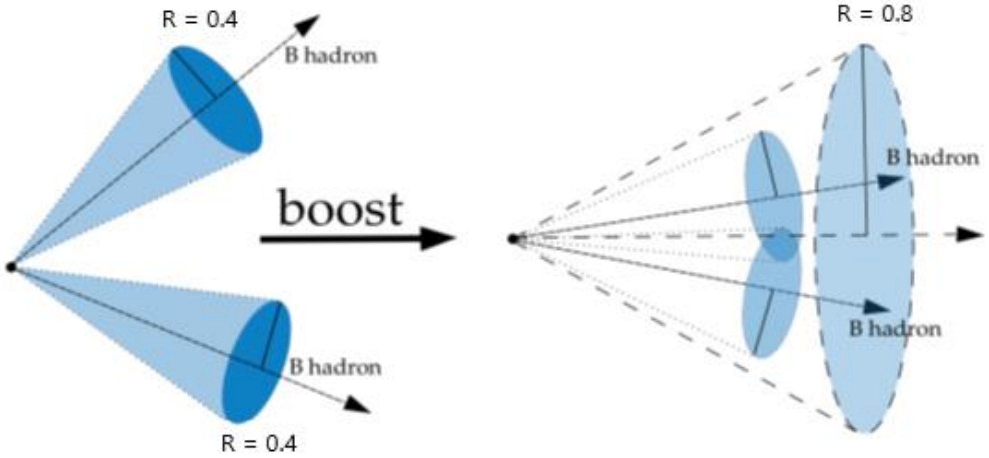
July 29, 2024

BOOST 2024

1 INFN of Bari

2 University of Bari

Boosted jets tagging



Hadronic decay products of high boosted particles are collimated.

- Particle reconstructed using one anti-kt (R=0.8) jet (AK8).

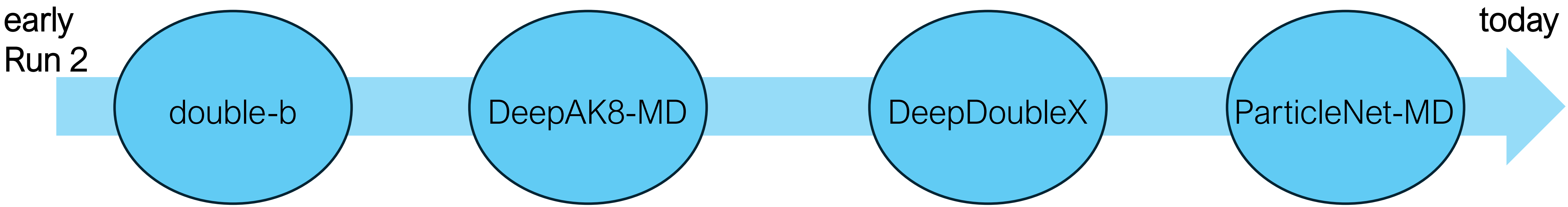
Techniques to reduce Pileup effect:

- Pileup mitigation algorithms: Pileup per particle identification;
- Jet grooming algorithms to remove soft and wide-angle radiation from the jet: Soft Drop.

ML tagging algorithm inputs: Particle Flow (PF) candidates (e. g. jet constituents) and Secondary vertices (SVs).

Before been used for measurements, tagging algorithms calibration from data needed .

Evolution of boosted jets tagging at CMS



double-b

- Boosted Decision Tree (BDT)
- jet inputs: tracks and SVs
- output classes: $H \rightarrow b\bar{b}$ vs QCD

DeepAK8-MD

- 1D Convolutional Neural Network (CNN)
- jet inputs: PF candidates and SVs
- output classes: $X \rightarrow b\bar{b}$ vs QCD (bbvsQCD), $X \rightarrow c\bar{c}$ vs QCD (ccvsQCD)

DeepDoubleX

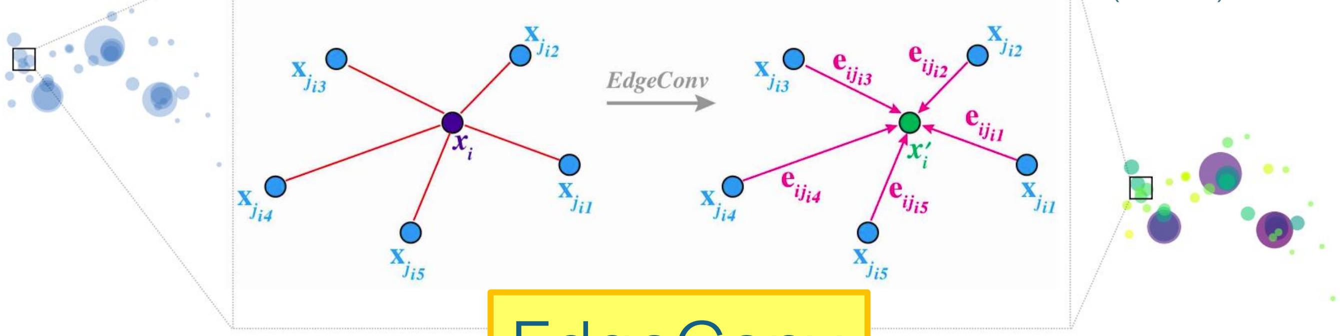
- 1D CNN + Recursive NN
- jet inputs: PF candidates and SVs
- output classes: $X \rightarrow b\bar{b}$ vs QCD (BvL), $X \rightarrow c\bar{c}$ vs QCD (CvL), $X \rightarrow b\bar{b}$ vs $X \rightarrow c\bar{c}$ (BvC)

ParticleNet-MD

- Dynamic Graph CNN (DGCNN)
- jet inputs: PF candidates and SVs
- state of art for Run 3 CMS

Jets as particle clouds: ParticleNet-MD

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



EdgeConv

ParticleNet-MD state-of-art for CMS boosted jet tagging.

- Graph based architecture describing the jet as a particle cloud (unordered sample).

EdgeConv block:

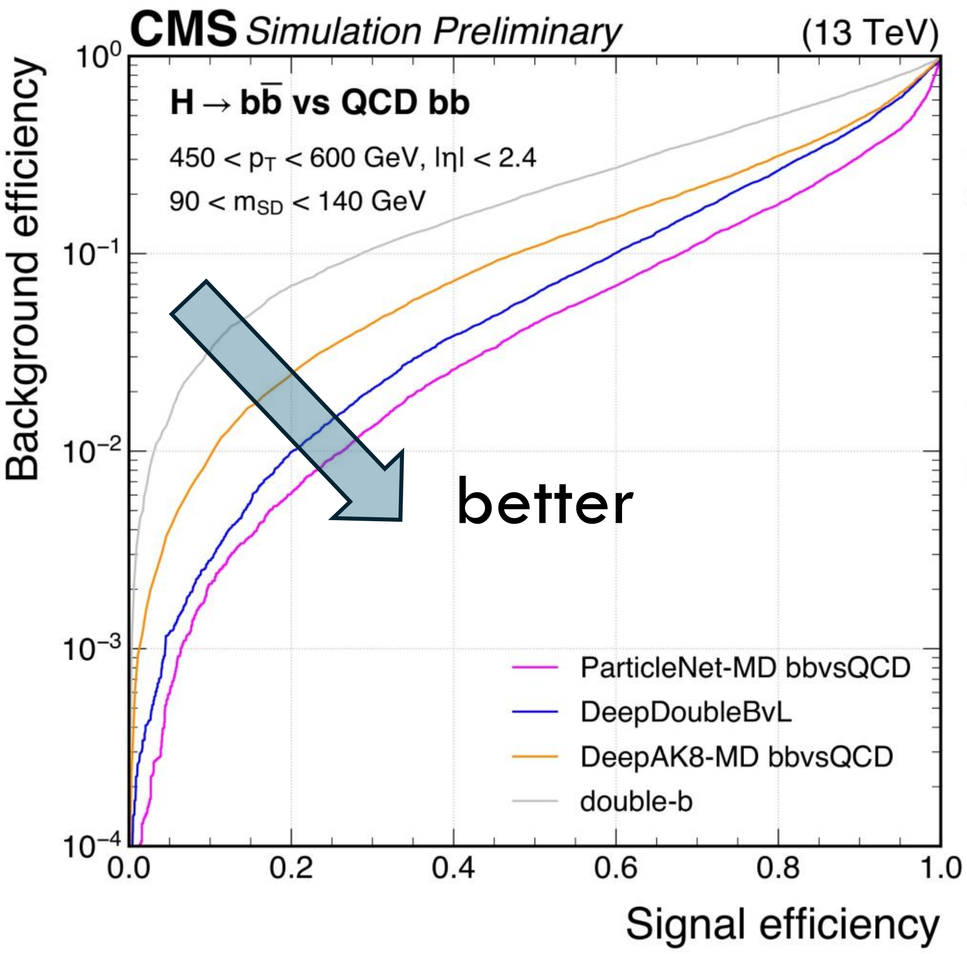
- NN module part of the ParticleNet architecture;
- New features vector associated to each jet constituent and based on the features of the *k*-nearest neighbors.

Mass decorrelation:

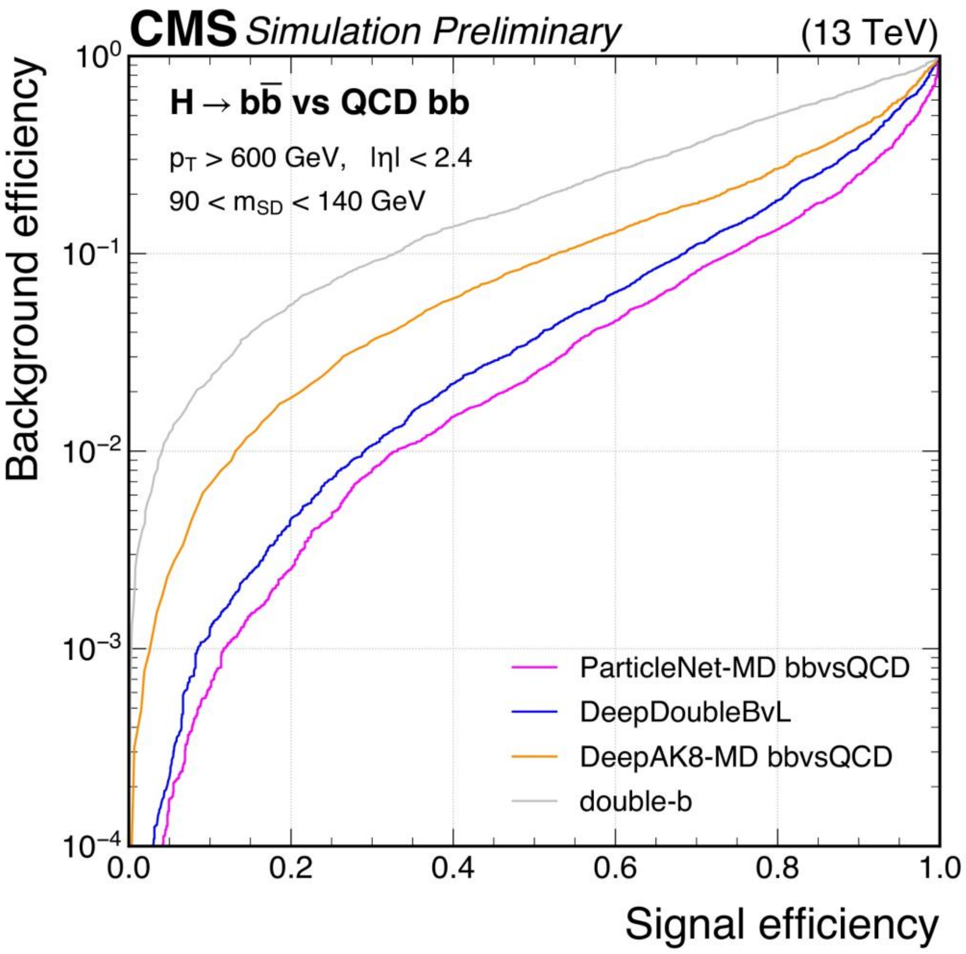
- Trained on Monte Carlo (MC) simulations containing boosted resonances (X) with a flat distributions in both of p_t and mass, as the signal sample, and the QCD multijet sample (reweighted to yield flat distributions) as the background sample.

ROC curve $b\bar{b}$ tagging performances (Run 2)

Low p_t: (450;600) GeV



High p_t: (600;∞) GeV



CMS-PAS-BTV-22-001

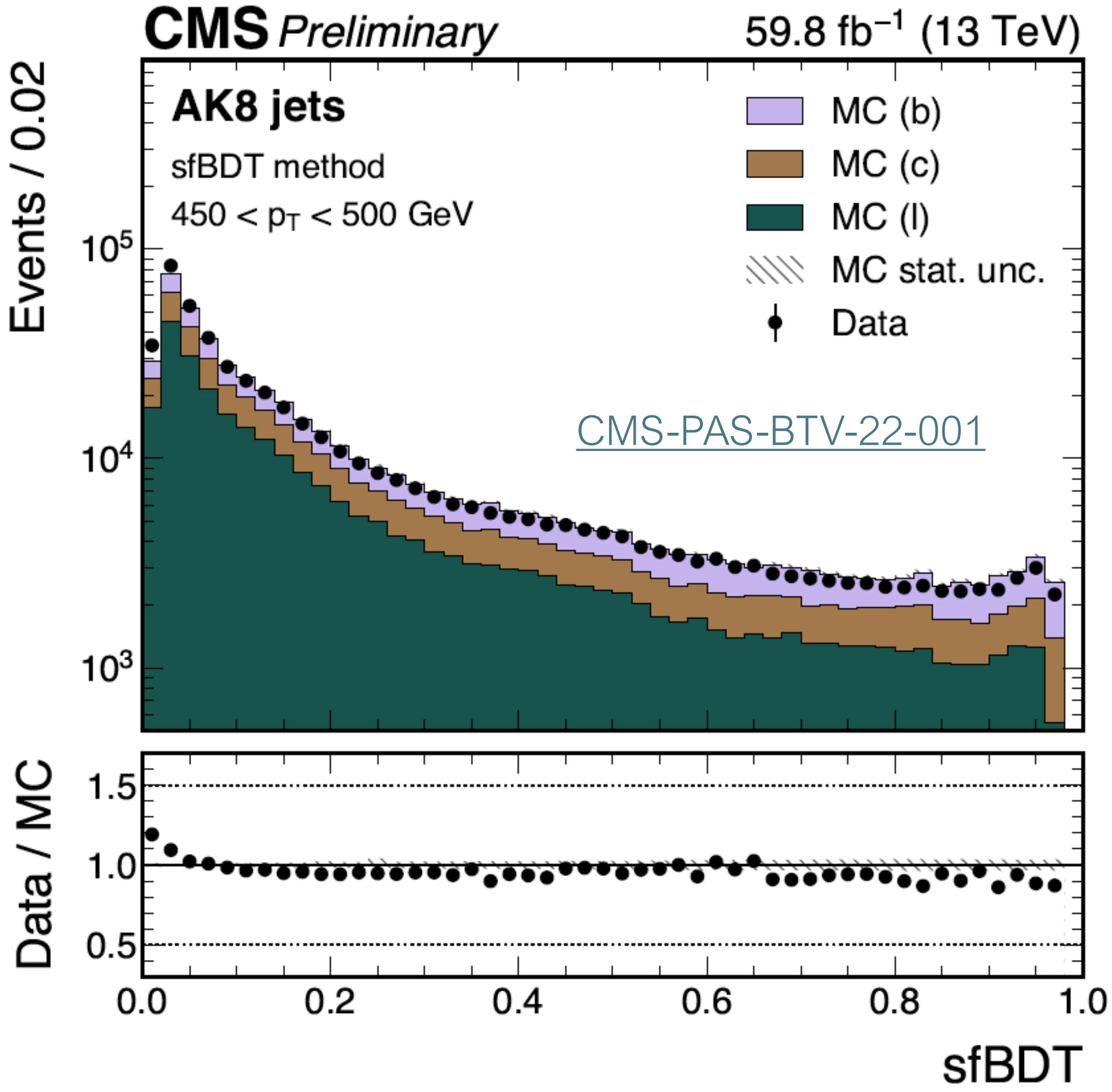
$$\text{PNet-MD}_{\text{bbvsQCD}} = \frac{p(X \rightarrow b\bar{b})}{p(X \rightarrow b\bar{b}) + p(QCD)}$$

Calibration methods

- Three methods used to measure the efficiency of tagging-algorithms in data:
  - sfBDT method;
  - μ-tagged method;
  - boosted Z method.
- Result: scale factors (SFs) obtained as the ratio between the jet tagging efficiency in data and in simulated samples.
  - $SF = \varepsilon_{data}(p_t) / \varepsilon_{sim.}(p_t)$
- SFs obtained for each method in three p_t bins:
  - (450, 500), (500, 600), and (600, +∞) GeV.

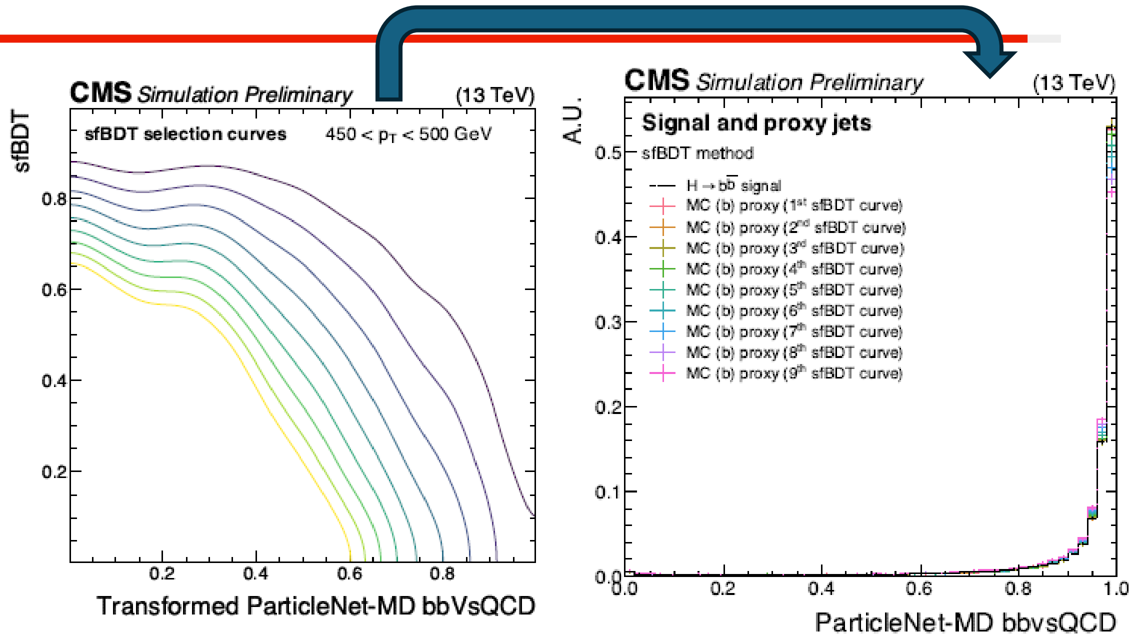
sfBDT calibration method: design of sfBDT

- Signal: $g \rightarrow b\bar{b}/c\bar{c}$ jets from QCD events like $H \rightarrow b\bar{b}/c\bar{c}$ events.
- Background: remaining QCD events.
- Trained a BDT (named sfBDT)
  - Inputs: costituents + SVs + jet N-subjettiness τ_{31} (three-prong jets like event → low τ_{31})
- Jets with at least one matched b hadron labelled as "b".
- Jets not labelled "b" with at least one matched c hadron labelled as "c".
- Remaining jets labelled as "l" type.

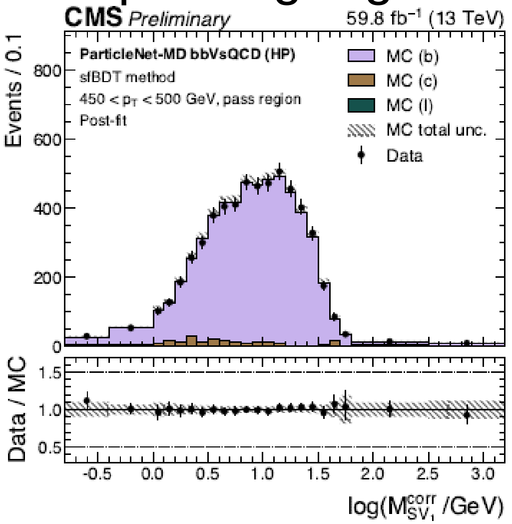


sfBDT calibration method: SF evaluation

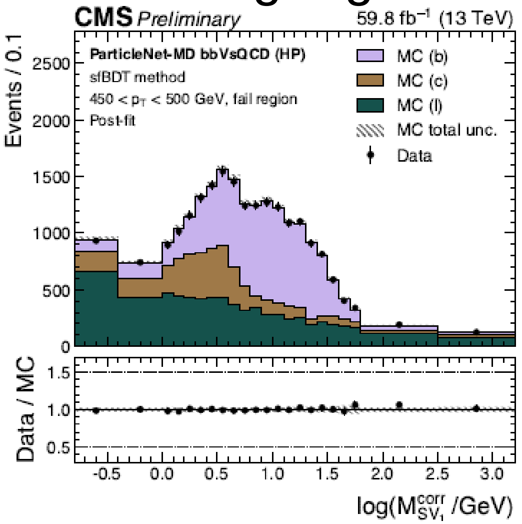
- sfBDT thresholds as function of the tagger score such that the distribution of the tagger discriminant of the proxy jets matches that of the signal jets.
- Tagger transformed (selection $X > X_0$ corresponding to jet selection efficiency of $1 - X_0$ of the tagger).
- Select proxy jet : $g \rightarrow b\bar{b}/c\bar{c}$ passing the sfBDT thresholds for a tagger transformed value.
- Scale factors by a simultaneous fit on the mass in the “pass” and “fail” region of a tagger working point (WP).
  - Each flavour template assigned a free-floating SF in fit.
- Multiple sfBDT curves are selected to achieve various selection efficiencies.



passing region



failing region

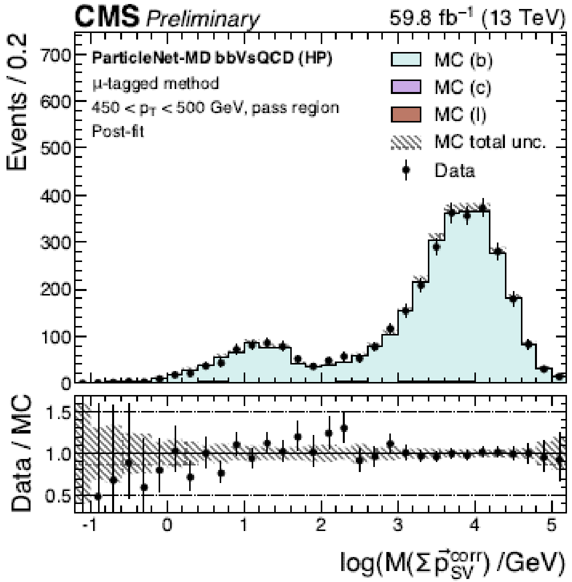


μ-tagged calibration method

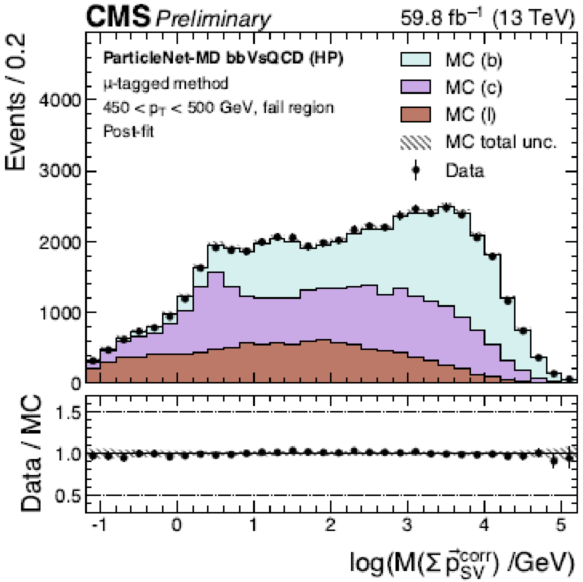
CMS-PAS-BTV-22-001

- Select proxy jet: $g \rightarrow b\bar{b}/c\bar{c}$ jets from QCD matching a soft muon.
  - Hadronised final state initiated from a heavy-flavour quark has ~20% probability of generating an electron or muon.
  - jet N-subjettiness $\tau_{21} < 0.3$ (two-prong jets like event → low τ_{21}).
- Simultaneous fit on the mass distribution in the “pass” and “fail” region of a tagger WP.
  - Each flavour template assigned a free-floating SF in fit.

passing region



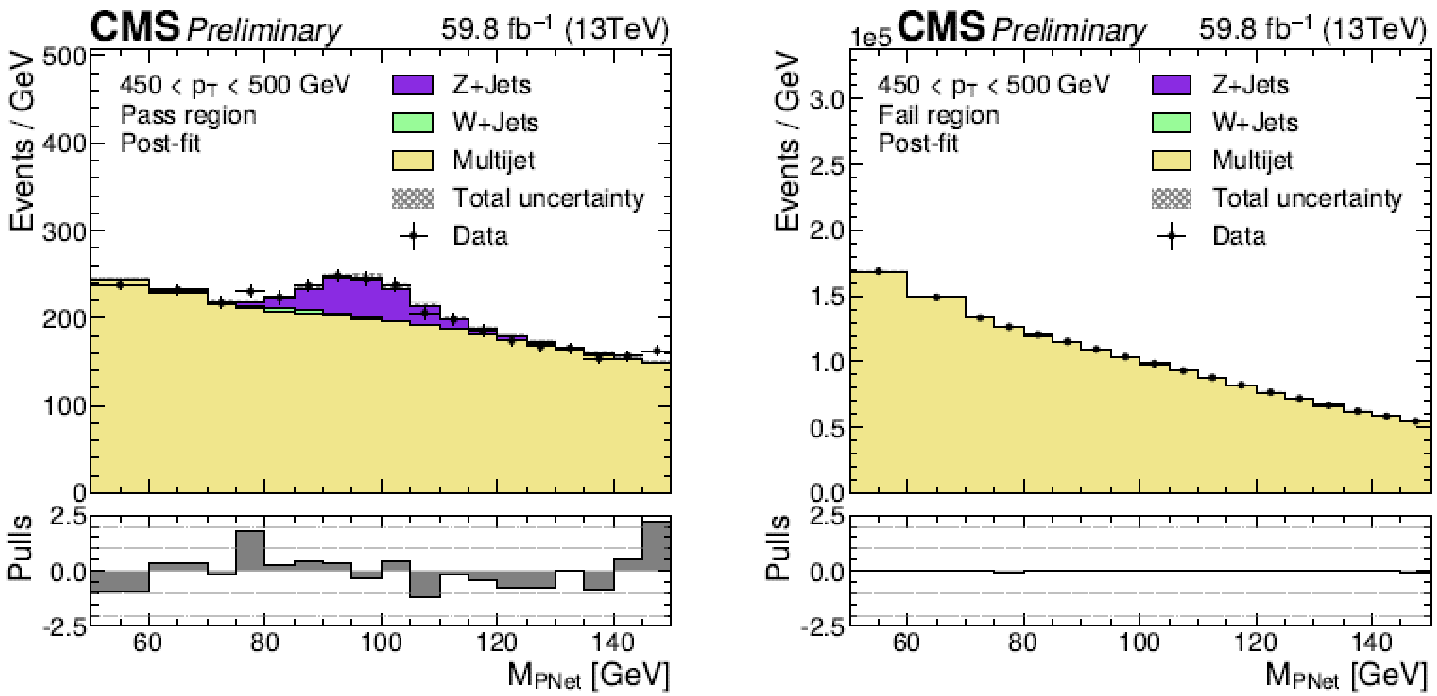
failing region



boosted Z calibration method

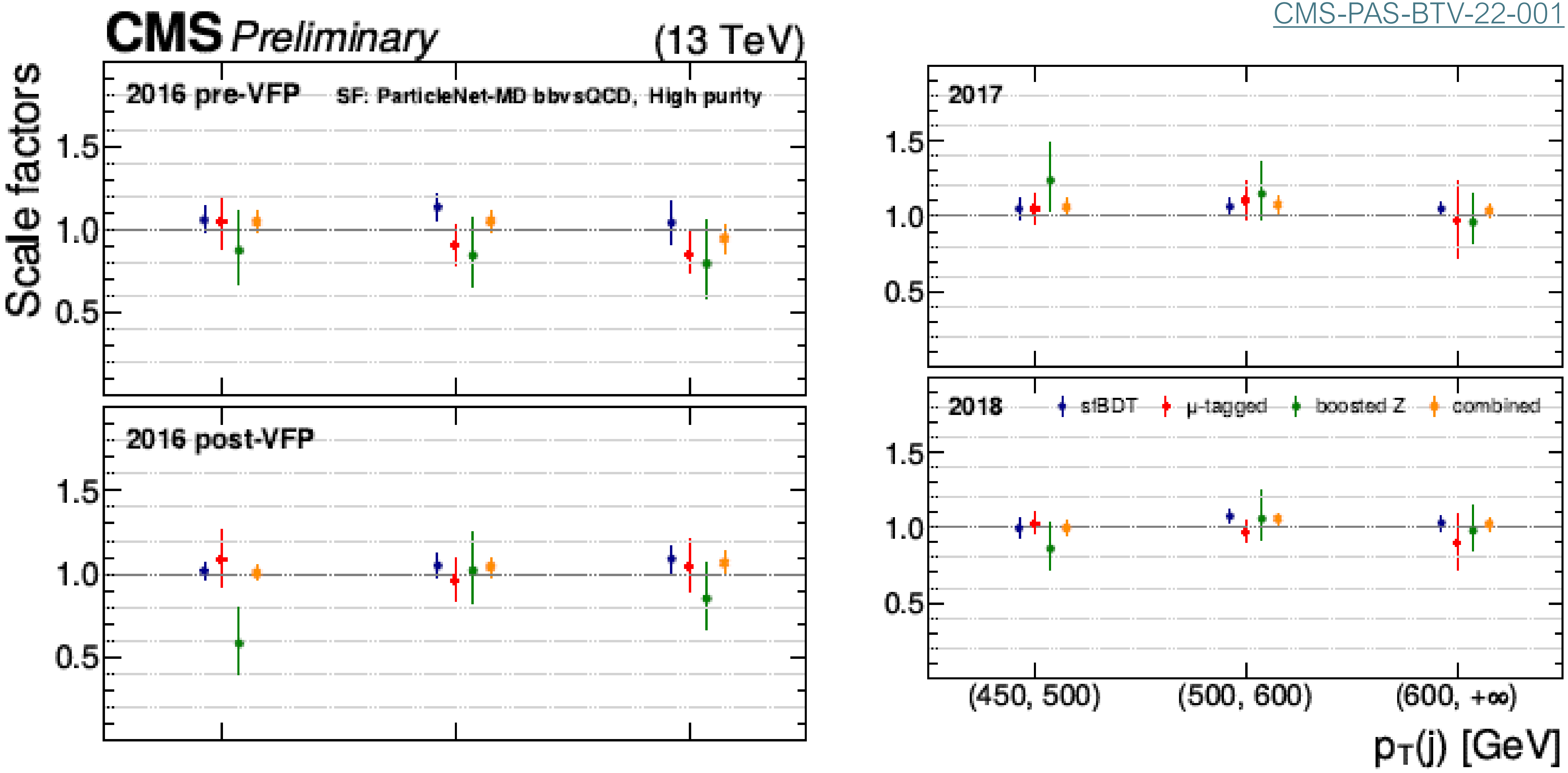
- Select proxy jet: $Z \rightarrow b\bar{b}$.
  - measure the SF at Z peak on top of the QCD multijet background.
- QCD estimated data-driven.
- Simultaneous fit on the mass in the “pass” and “fail” region of a tagger WP.

CMS-PAS-BTV-22-001



Calibrations combination

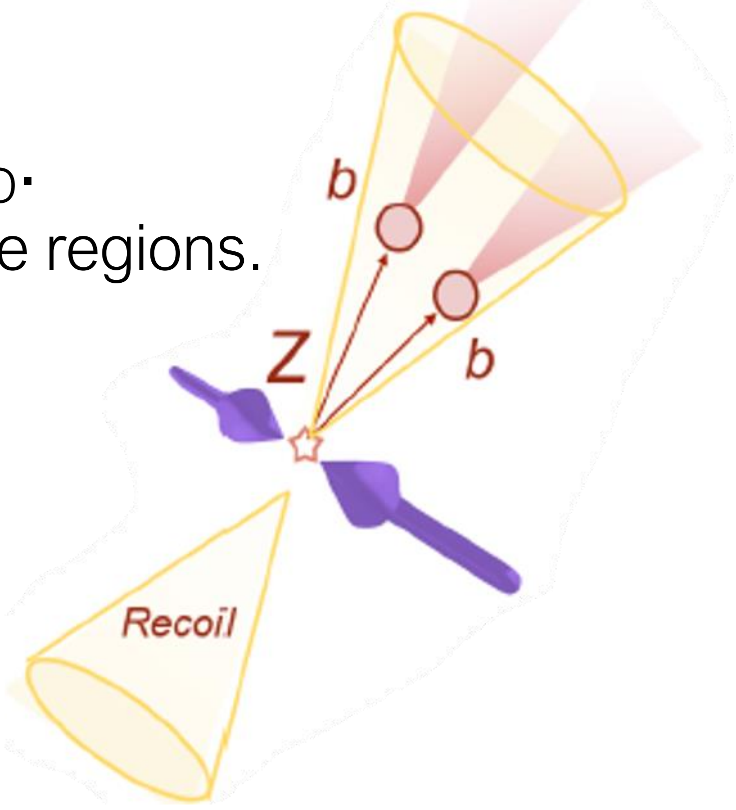
CMS-PAS-BTV-22-001



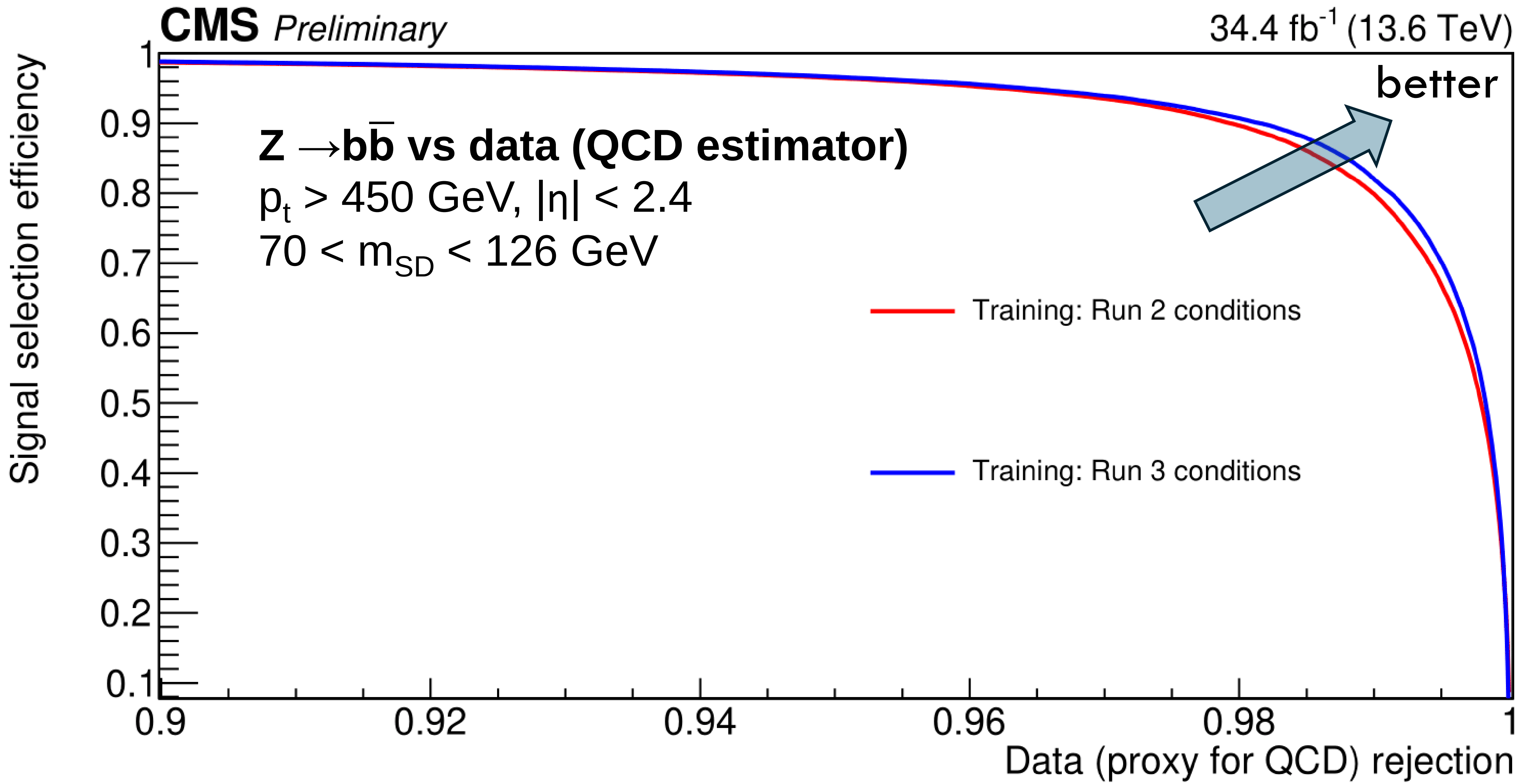
ParticleNet-MD validation for 2022 data

- ParticleNet-MD optimized for Run 3. CMS-DP-2024-055
- Validation done on event containing high boosted $Z \to b\bar{b}$ + jets.
- Likelihood fit to check the agreement between 2022 data and Standard Model (SM) prediction ($Z\to q\bar{q}$, $W \to q\overline{q'}$ and QCD).
- **Event selection:** boosted $Z \to q\bar{q}$ + jets
- **Events categorized in five region of** PNet-$MD_{bbvsQCD}$.
  - Likelihood fit independent in each of the four highest score regions.
  - Lowest score region (<0.641) not included in the fit.

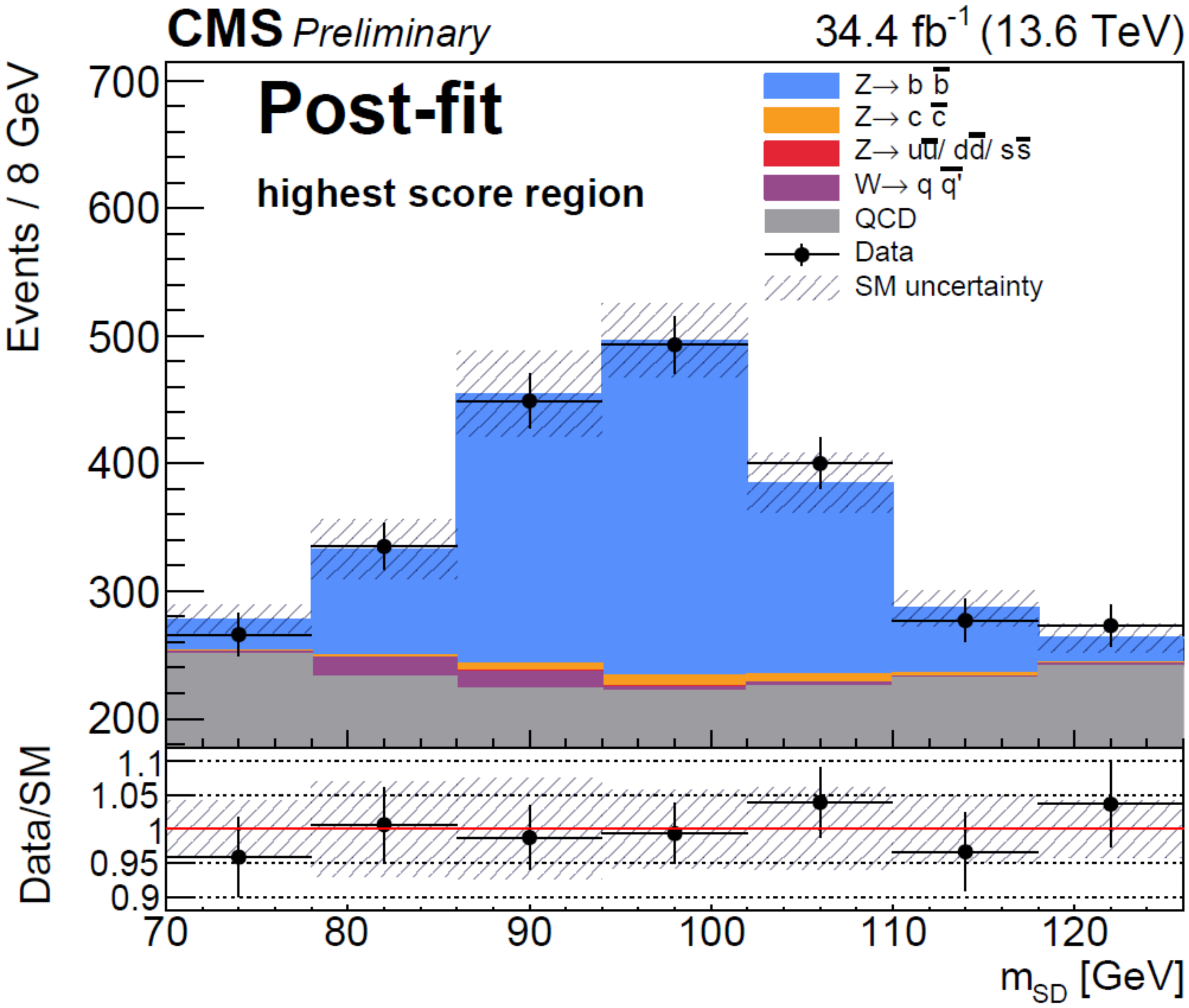
	PNet-$MD_{bbvsQCD}$ range
4th highest score region	$0.641 < \text{PNet-MD}_{bbvsQCD} \leq 0.875$
3rd highest score region	$0.875 < \text{PNet-MD}_{bbvsQCD} \leq 0.957$
2nd highest score region	$0.957 < \text{PNet-MD}_{bbvsQCD} \leq 0.988$
highest score region	$0.988 < \text{PNet-MD}_{bbvsQCD} \leq 1$



ROC curve $b\bar{b}$ tagging performances

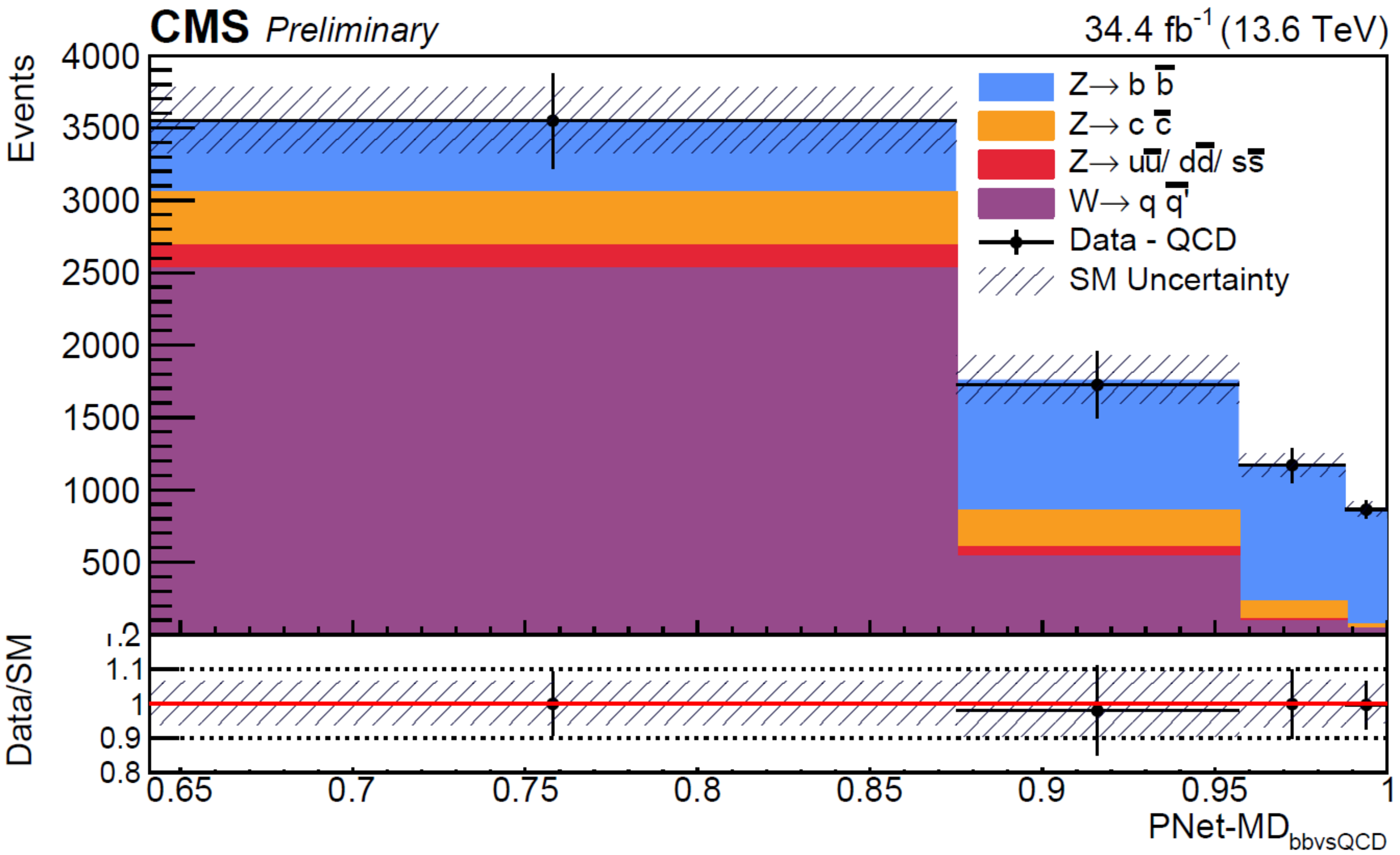


$0.988 < \text{PNet-MD}_{bbvsQCD} \leq 1$



- Post-fit data to prediction comparison in the highest score region.
- Upper plot: soft-drop mass (m_{SD}) distributions of data (black markers) and prediction (stack of QCD, W $\rightarrow q\bar{q}'$ and $Z \rightarrow q\bar{q}$, with the latter split on the basis of its decay products) of the leading-p_t AK8.
- Lower plot: ratio between data and prediction.
- Good data-prediction agreement.
- Z-peak visible in the data distribution.

PNet-MD$_{bbvsQCD}$ score post-fit distribution



- PNet-MD$_{bbvsQCD}$ distribution of the stack of $W\to q\bar{q}'$ and $Z\to q\bar{q}$ (split on the basis of its decay products) yield, and the data, once the QCD contribution is subtracted (Data - QCD), of the leading-p_t AK8.
- Bottom pad: ratio between Data-QCD and the stack.
- Good data-prediction agreement.
- The amount of $W\to q\bar{q}'$ and $Z\to q\bar{q}$ events, with q (light) or c quark, decreases increasing the score value.

ParticleNet-MD validation results

	$Z \to bb$	$Z \to cc$	$Z \to$ uu/ dd/ ss
$0.641 < \text{PNet-MD}_{bbvsQCD} \leq 0.875$	48.8%	35.5%	15.7%
$0.875 < \text{PNet-MD}_{bbvsQCD} \leq 0.957$	73.9%	20.5%	5.6%
$0.957 < \text{PNet-MD}_{bbvsQCD} \leq 0.988$	87.7%	10.9%	1.4%
$0.988 < \text{PNet-MD}_{bbvsQCD} \leq 1$	96.6%	3.18%	0.26%

- Percentage on the total number of $Z \to q\bar{q}$ (q = u, d, s, c, b) events depending on the quark flavours.
- Mis-identified $Z \to b\bar{b}$ percentage decreases as $\text{PNet-MD}_{bbvsQCD}$ increases, reaching less than 4%.

New Jet tagging developments

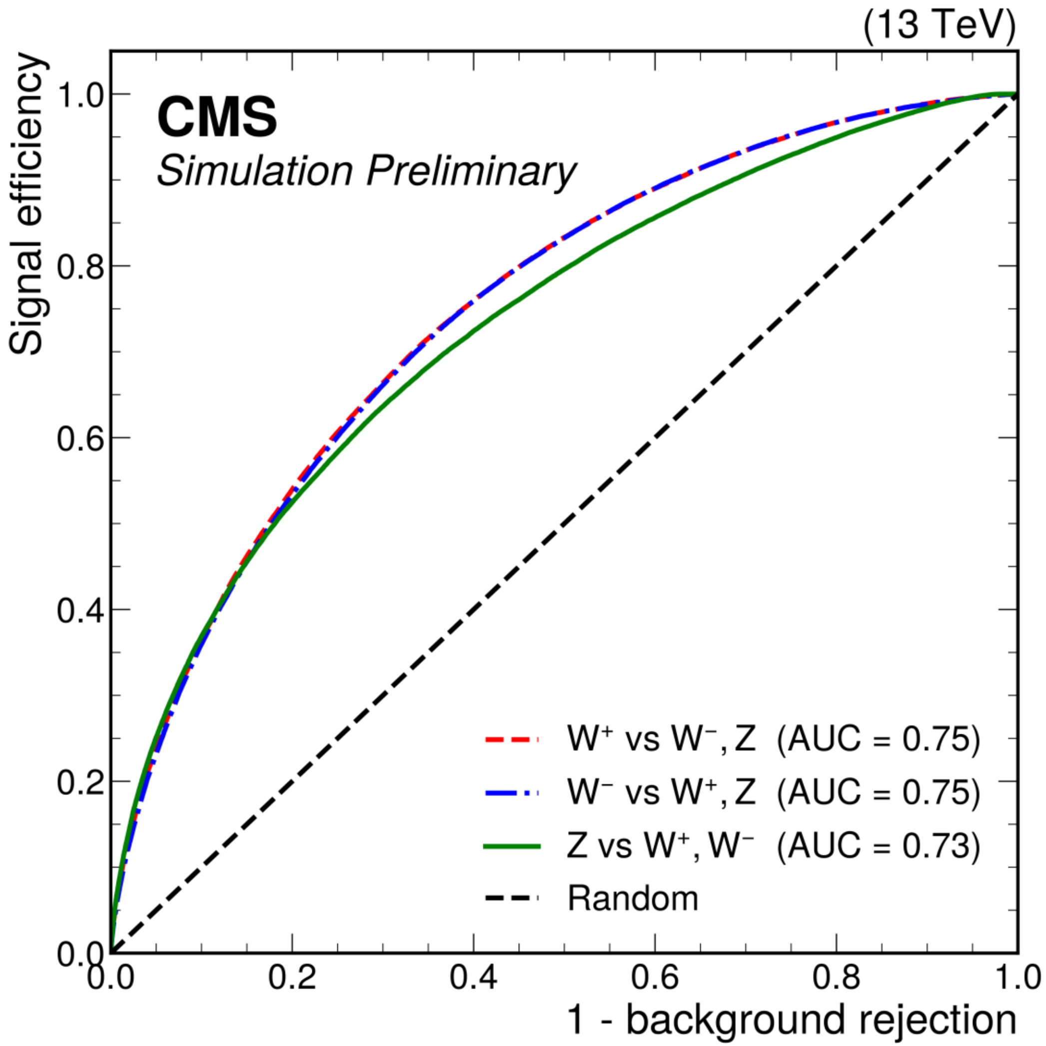
- Charge identification
  - More information in K. Tauqeer poster

- Variable-R jet clustering
  - More information in G. Milella poster

Jet charge tagger

- DGCNN based on the ParticleNet architecture predicting the charge of the AK8 jet.
    - Discriminate W^+ and W^- bosons from Z boson.
- Training samples:
    - semileptonic $t\bar{t}$ MC simulation is used to get a sample with W^+ and W^- jets;
    - Z+jets MC simulation.
- Validation done on a region enriched of semileptonic $t\bar{t}$ events.
    - Good data-MC agreement.

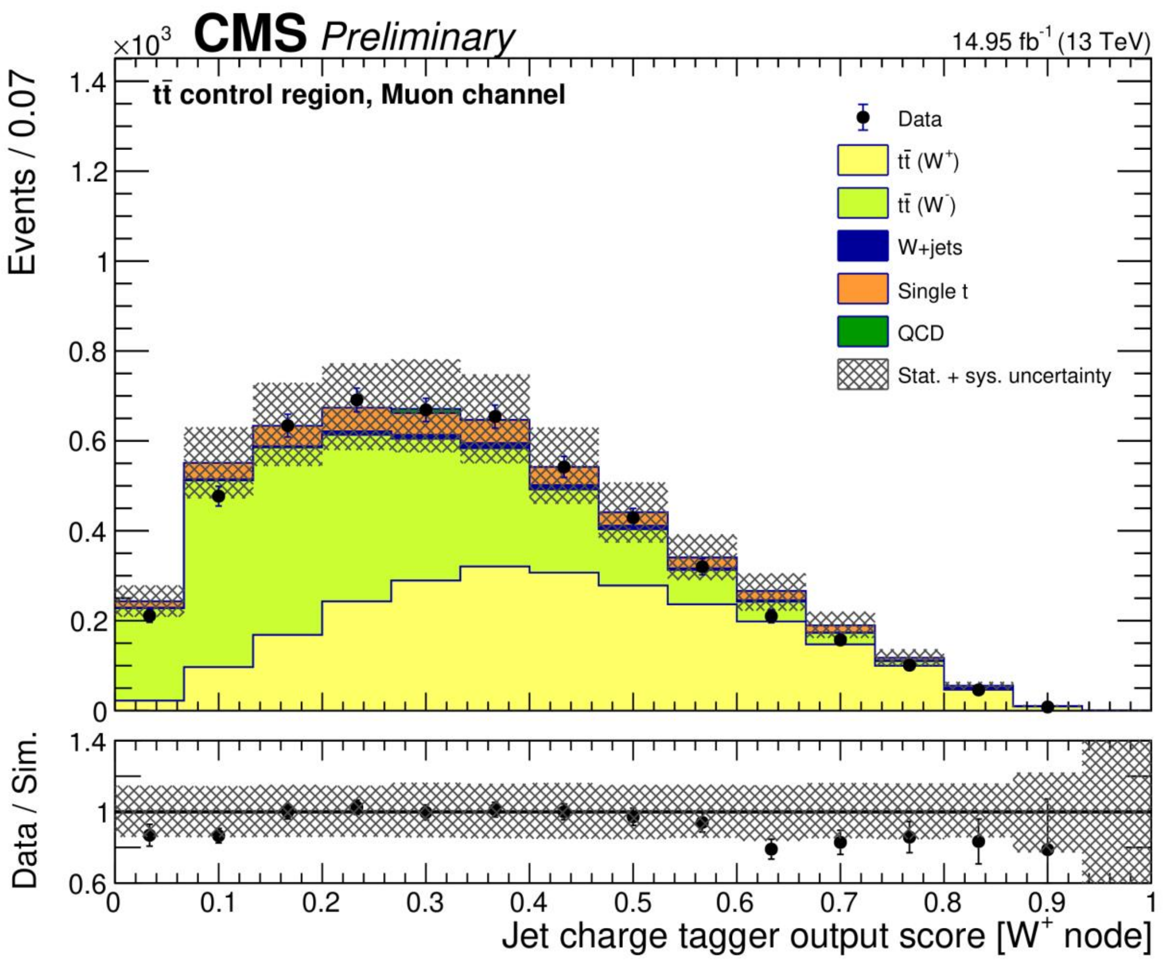
CMS-DP-2024-044



Jet charge tagger

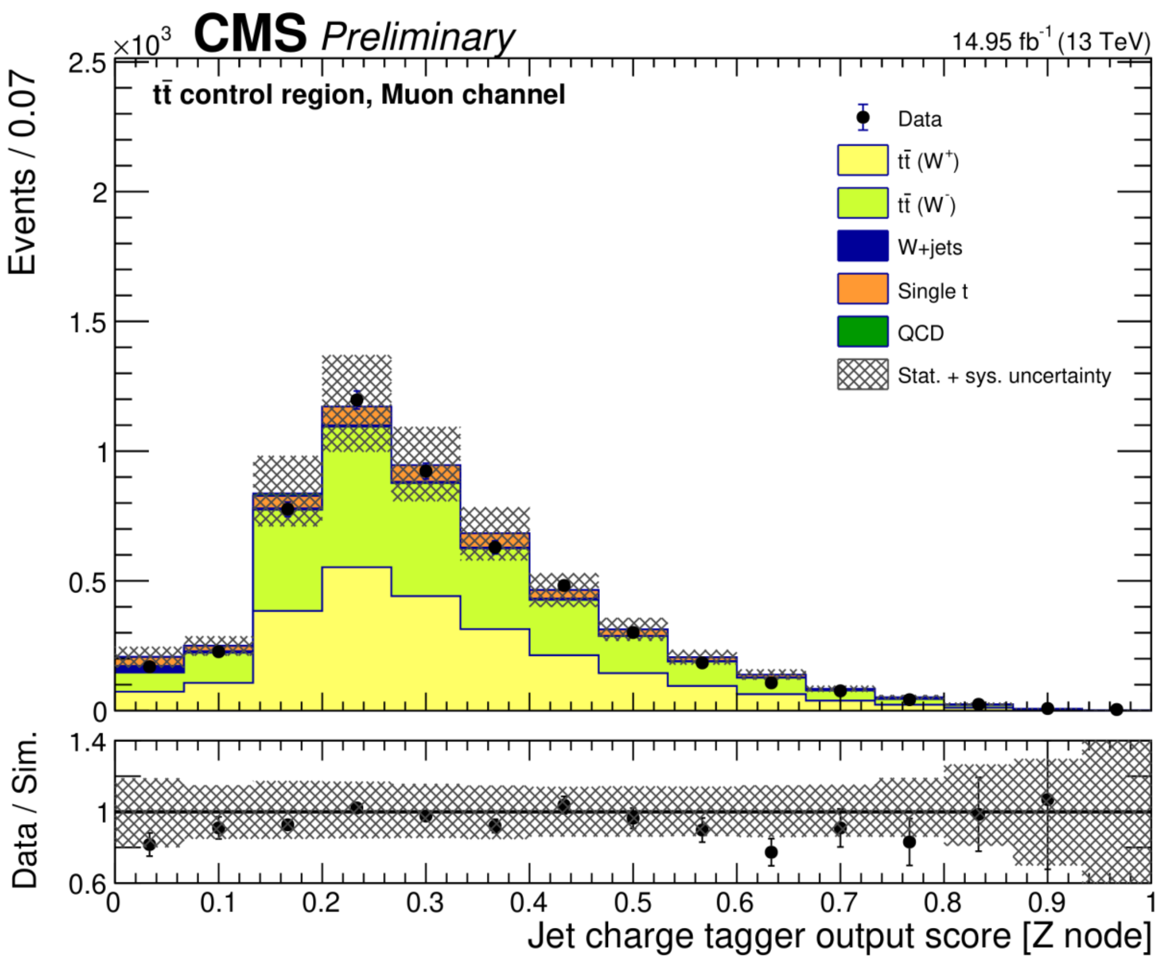
semileptonic $t\bar{t}$ enriched region

output score: W^+



semileptonic $t\bar{t}$ enriched region

output score: Z (not expected Z bosons)



Top tagging with variable sized jets

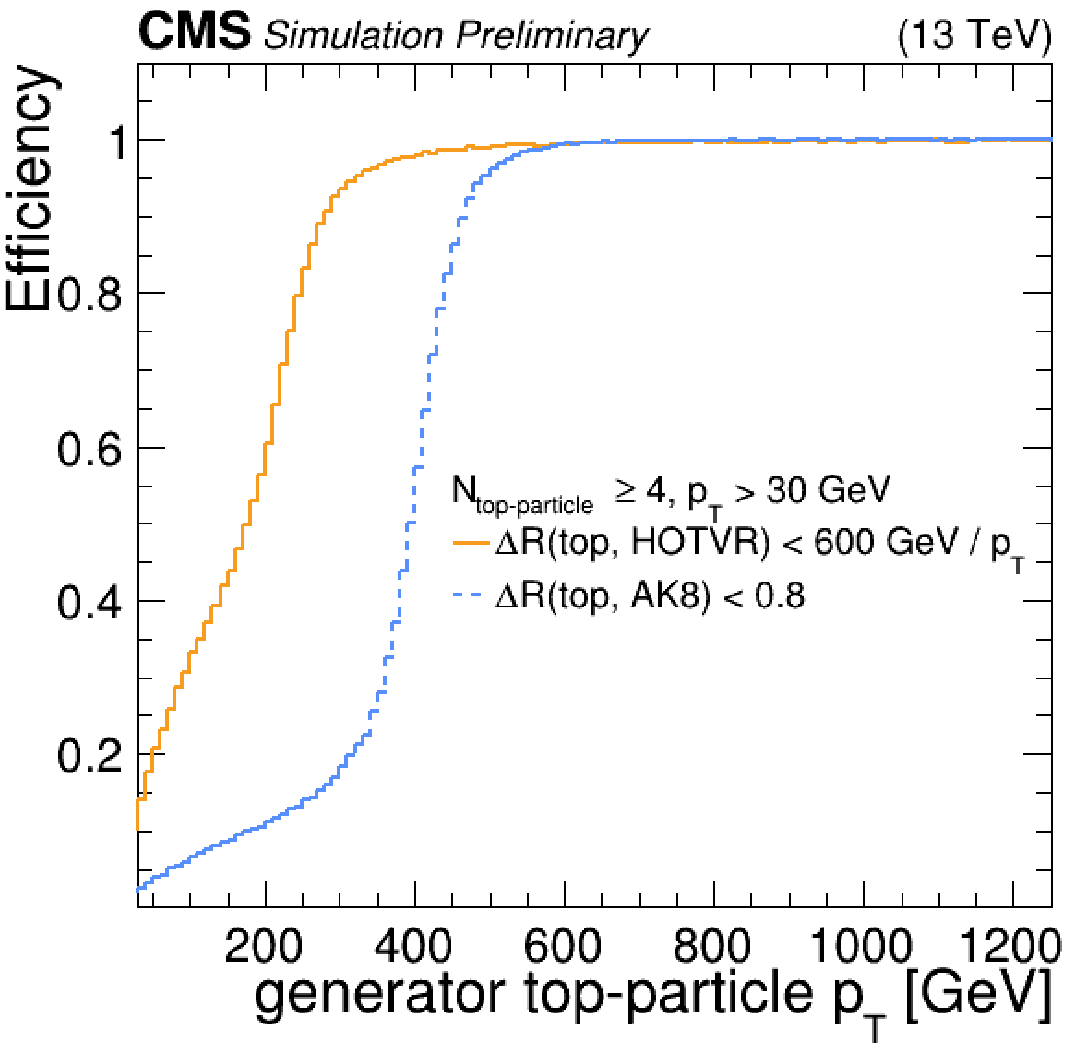
CMS-DP-2024-038

Heavy Object Tagger with Variable Radius

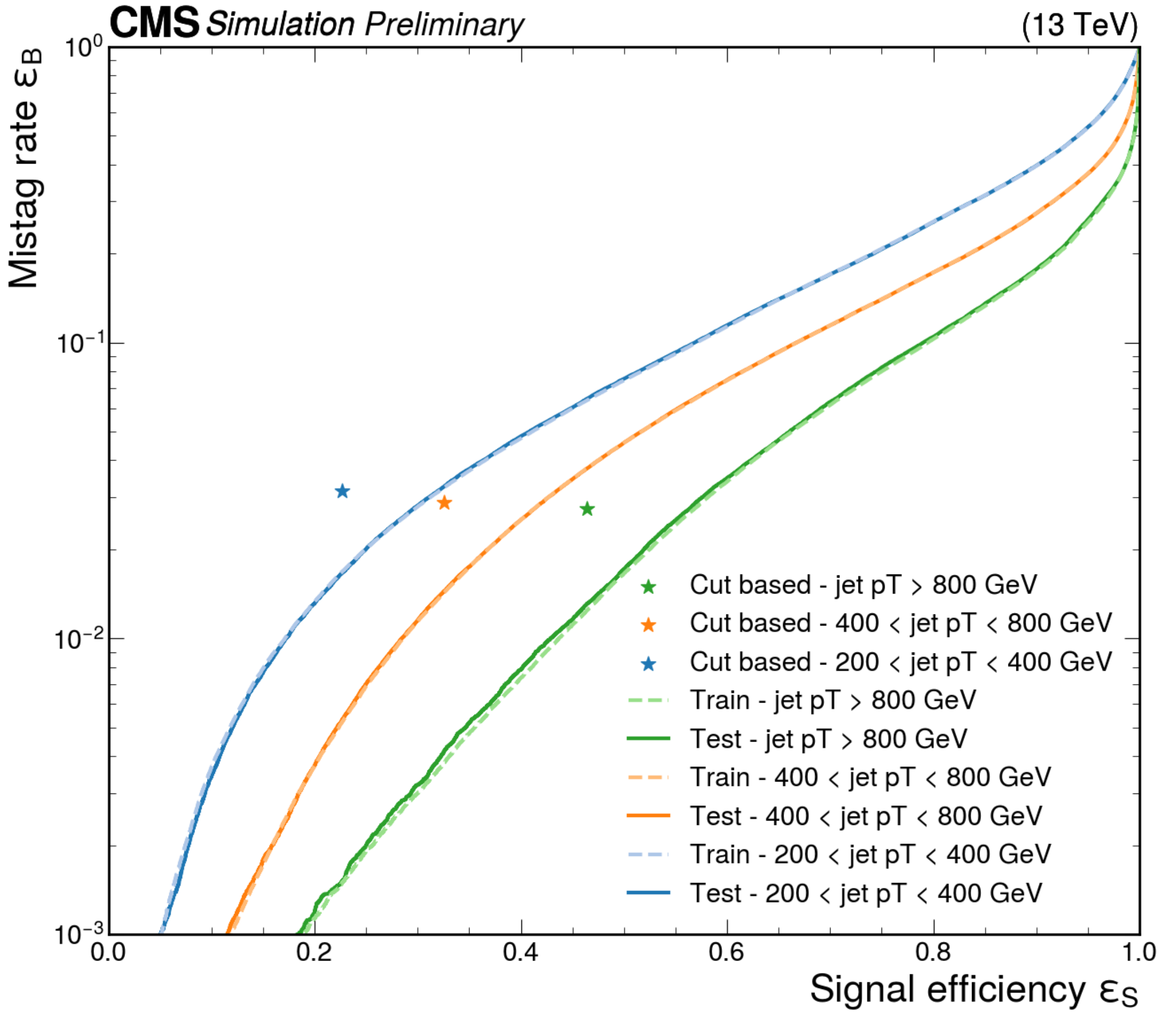
- $R = 600 / p_t$ (R min 0.1, R max 1.5).
- Useful for 4 top final states where the top quark is not completely boosted ($200 < p_t < 800$ GeV).
- Efficiency as the ratio between the generated top quarks matching a reconstructed jet within ΔR and all the generated top quarks.

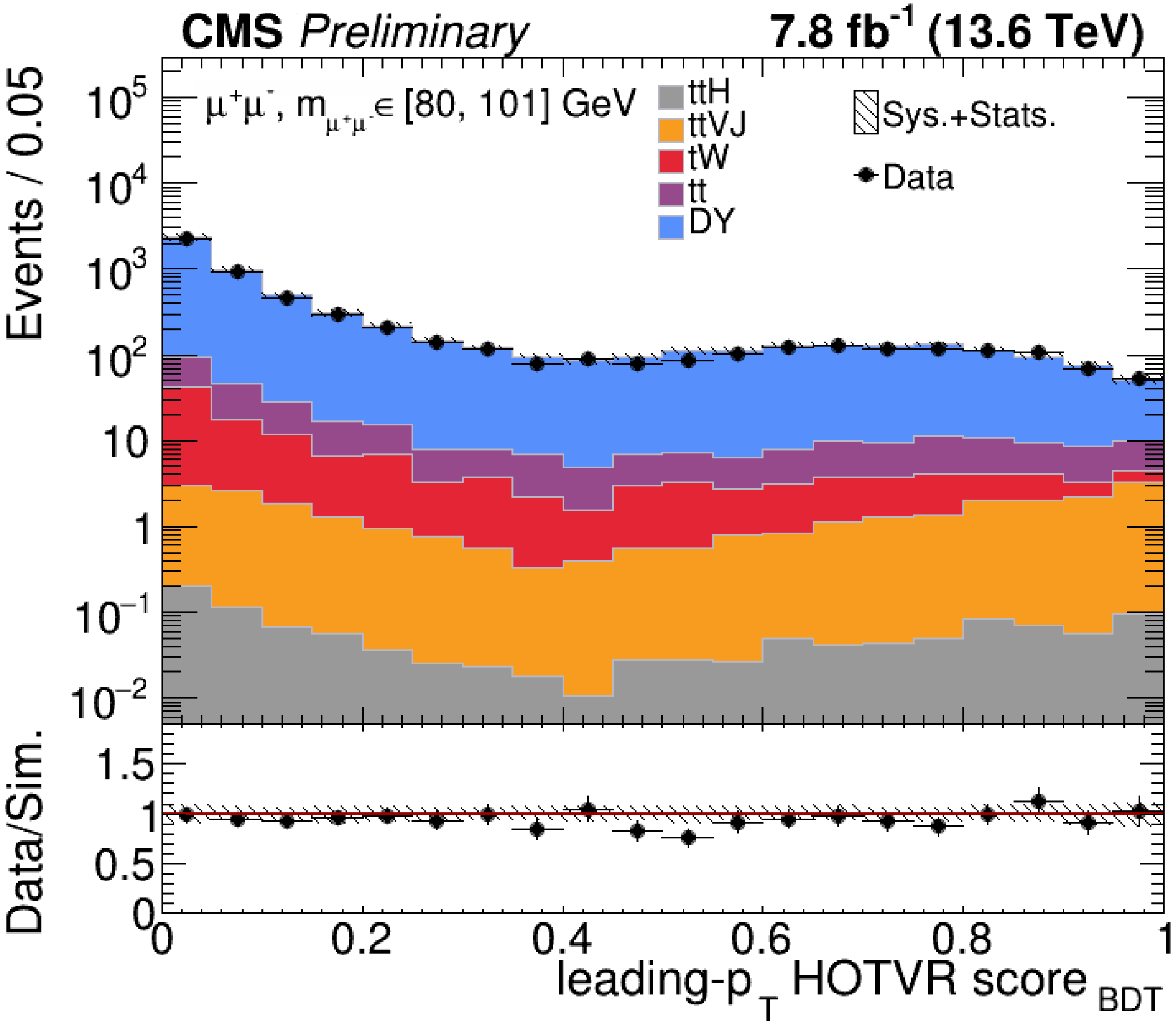
Developed a BDT to distinguish top quarks from QCD:

- Training on QCD multijet and the ttZ to simulate the background and the signal, respectively;
- Tested on a Z+jets enriched selection
  - Two opposite sign leptons ($80 < m_{\ell\ell} < 101$ GeV) + ≥1 HOTVR



Top tagging with variable sized jets





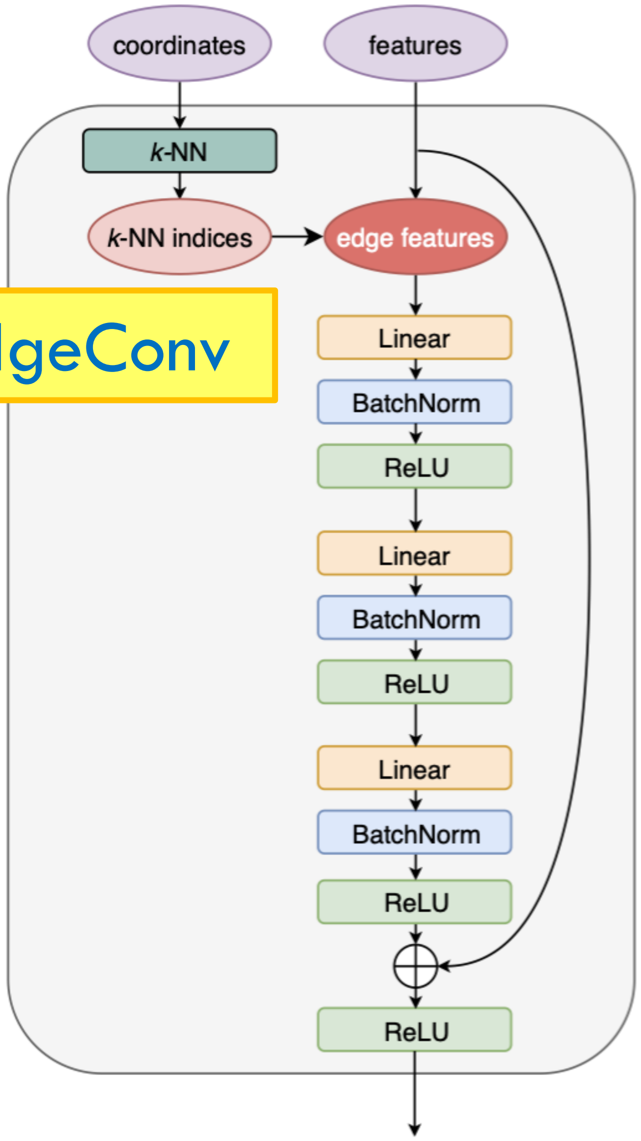
Conclusions

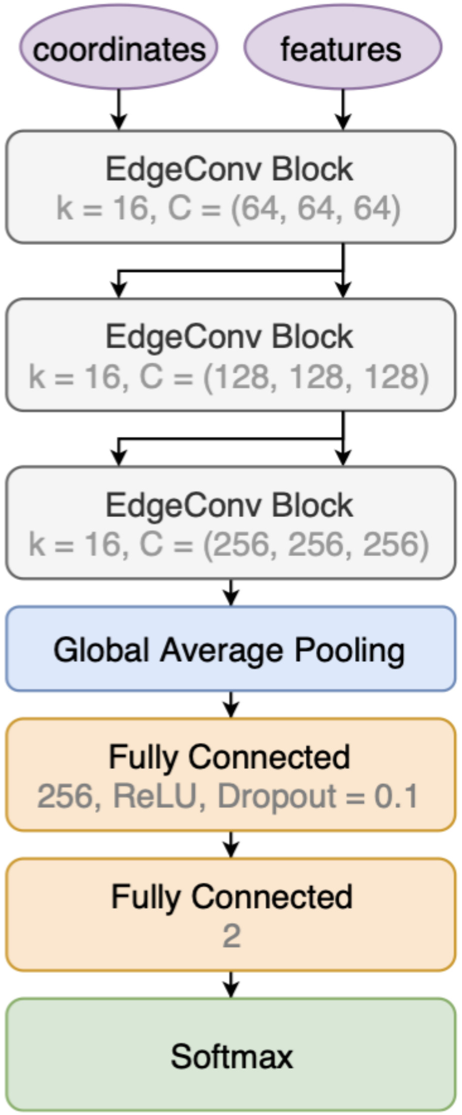
- Summary the CMS boosted jet tagging state of art.
  - Brief description of ParticleNet-MD.
  - CMS Run 2 ROC curve.
  - Tagger calibration methods.
  - ParticleNet-MD best boosted jet tagging for Run 2.
  - ParticleNet-MD optimized for Run 3.
- Shown ParticleNet-MD validation for $Z \rightarrow b\bar{b}$-like events for 2022 data.
  - Improvement in the ROC curve between the Run 3 – Run 2 training.
  - Good data-prediction agreement.
  - Z-peak visible in data distribution at high score.
  - $Z \rightarrow b\bar{b}$ purity increases as the score increases (96.6% at the highest score region).
- Summary of the new developed jet tagging algorithms.
  - Jet charge tagger.
  - Heavy Object Tagger with Variable Radius.

Backup

ParticleNet architecture

EdgeConv





ParticleNet

Binary outputs:

$$\text{PNet-MD}_{\text{bbvsQCD}} = \frac{p(X \to b\bar{b})}{p(X \to b\bar{b}) + p(QCD)}$$

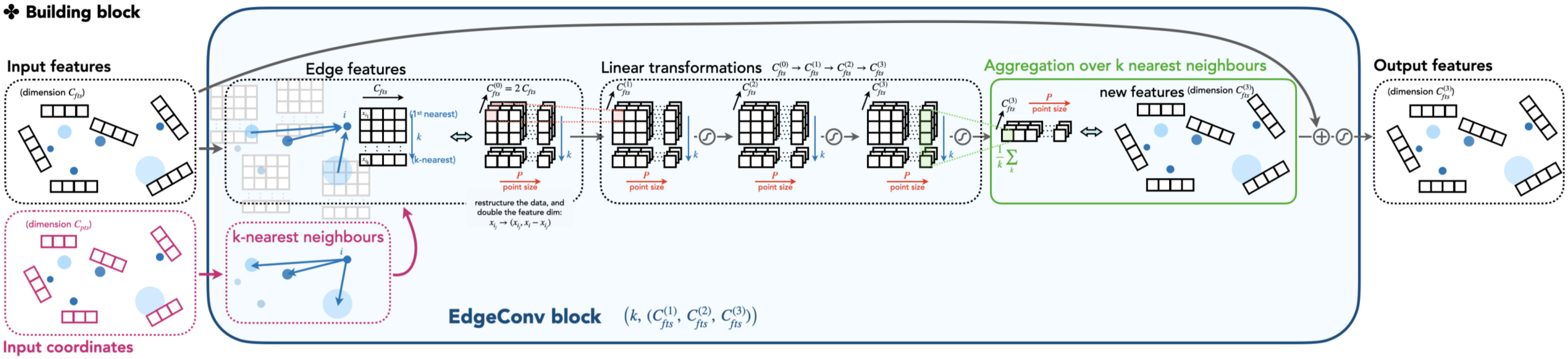
$$\text{PNet-MD}_{\text{ccvsQCD}} = \frac{p(X \to c\bar{c})}{p(X \to c\bar{c}) + p(QCD)}$$

$$\text{PNet-MD}_{\text{qqvsQCD}} = \frac{p(X \to q\bar{q})}{p(X \to q\bar{q}) + p(QCD)}$$

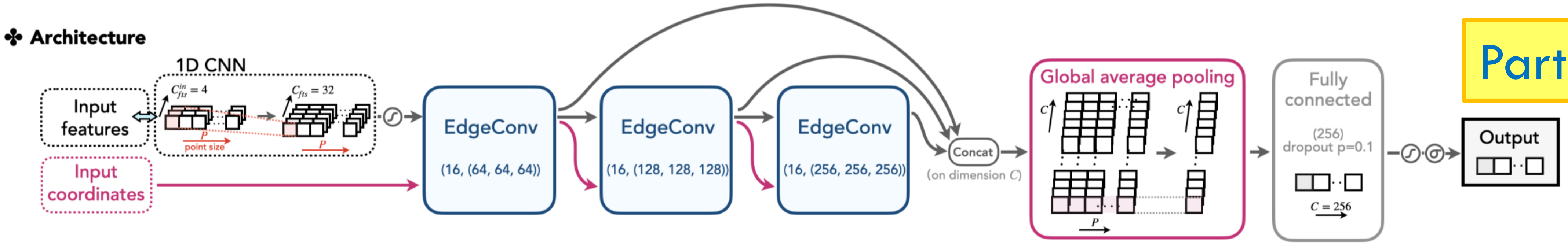
$$\text{PNet-MD}_{\tau\tau\text{vsQCD}} = \frac{p(X \to \tau\bar{\tau})}{p(X \to \tau\bar{\tau}) + p(QCD)}$$

ParticleNet architecture

EdgeConv

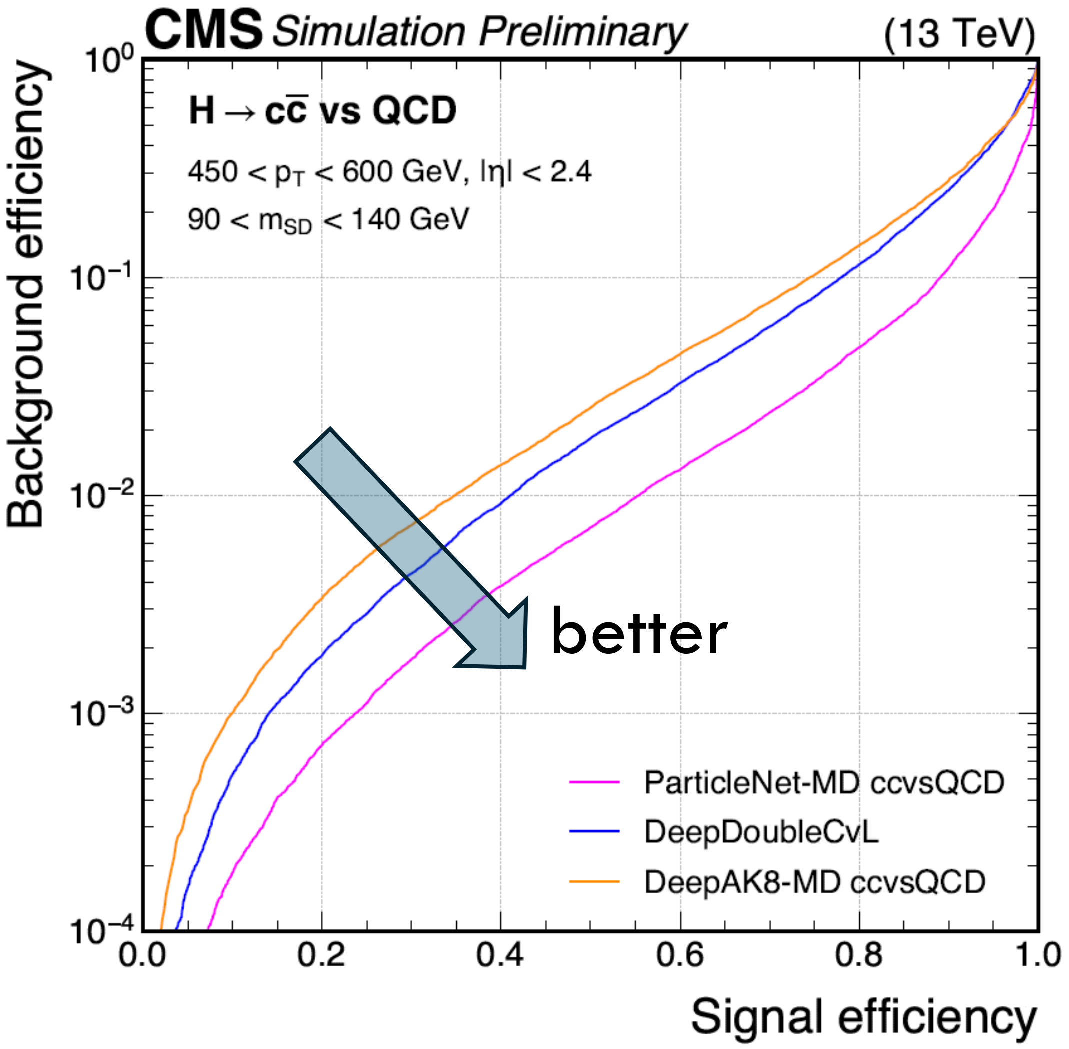


ParticleNet

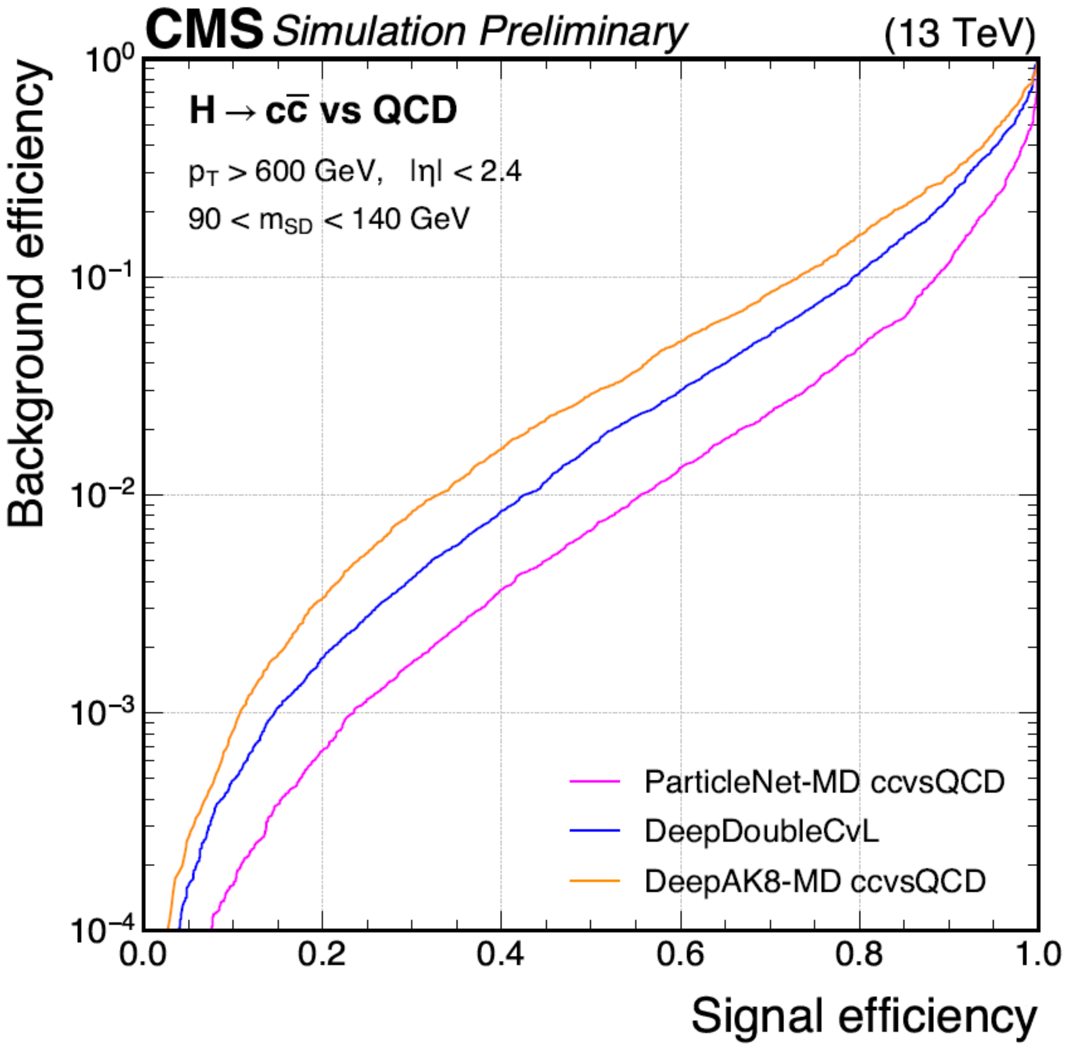


ROC curve $c\bar{c}$ tagging performances (Run 2)

Low p_t: (450;600) GeV



High p_t: (600;∞) GeV

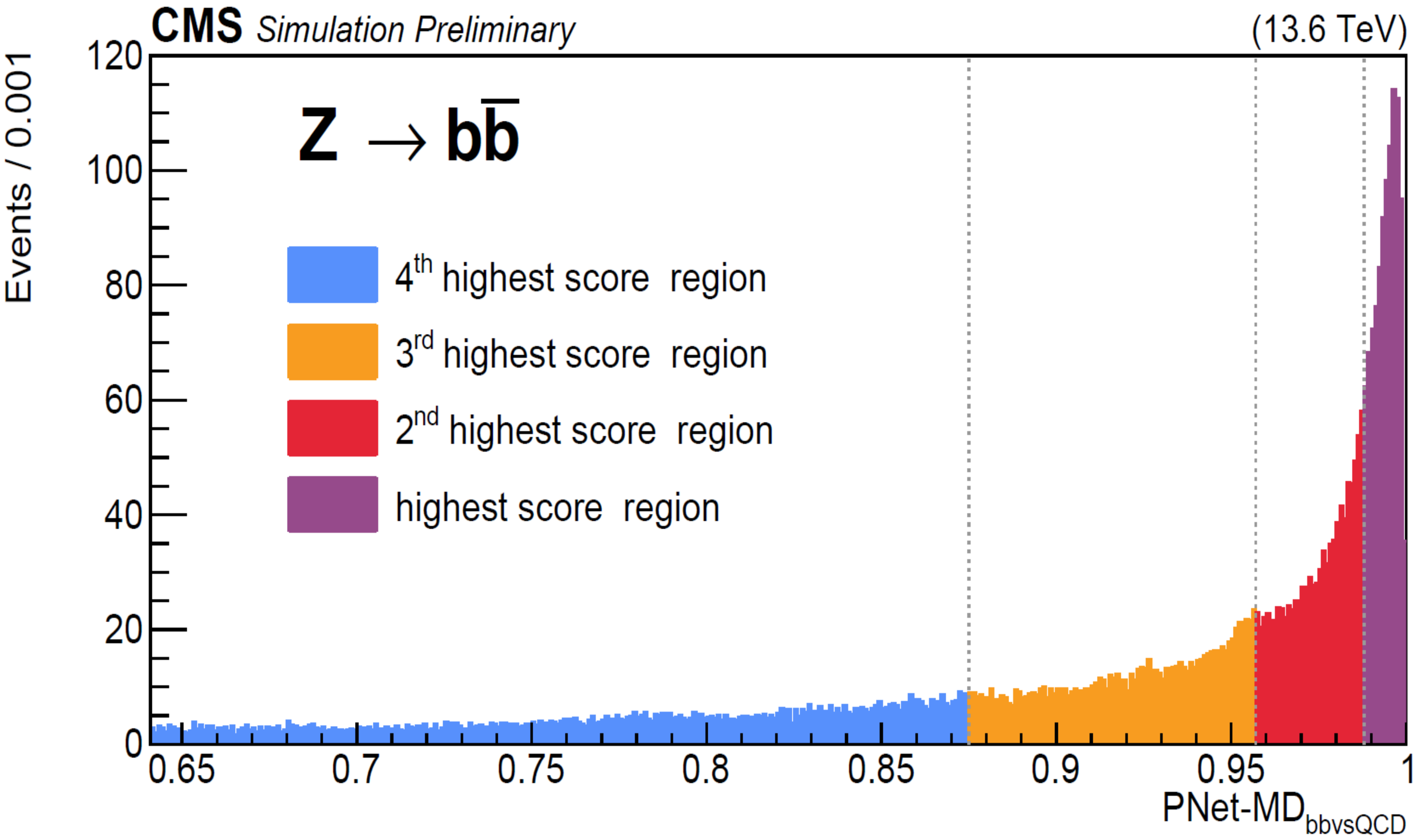


Boosted $Z \to b\bar{b}$ event selection

- **High-Level-Trigger paths :** PFHT1050, PFJet500, AK8PFJet500, AK8PFJet400_TrimMass30, AK8PFJet420_TrimMass30, AK8PFHT800_TrimMass50

- Leading-p_T AK8 jet (Z-boson candidate): $p_T > 450$ GeV and $|\eta| < 2.4$
- Sub-leading-p_T AK8 jet: $p_T > 200$ GeV and $|\eta| < 2.4$
- Veto events with at least one electron or muon with $p_T > 20$ GeV, $|\eta| < 2.4$, and satisfying the loosest identification and isolation working point
- Veto events with a b-tagged AK4 jet having $p_T > 30$ GeV, $|\eta| < 2.4$ and a distance ΔR from the leading AK8 jet greater than 0.8
    - The DeepJet medium working point is used to tag AK4 jets as originating from b-quark

PNet-MD$_{bbvsQCD}$ categorization



- Z-boson candidate PNet-MD$_{bbvsQCD}$ distribution from 0.641 to 1 for the MC $Z \rightarrow q\bar{q}$ after the event selection, normalized to 34.4 fb-1.
- Z-candidate compatible with a Z decaying to $b\bar{b}$ at generation level.
- Each of the four highest score regions has the same amount (20%) of MC $Z \rightarrow b\bar{b}$ events .

PNet-MD$_{bbvsQCD}$ validation: Likelihood fit

- The likelihood fit is performed within the signal mass window in the four highest score regions defined in slide 7.
  - The parameters of interest of the fit (the MC $Z \rightarrow q\bar{q}$ and $W \rightarrow q\bar{q}'$ normalization factors) are obtained independently in each score region.
- The background from QCD multijet events is estimated using the average of the fits of the Z-candidate m_{SD} distributions outside one of nine alternative mass windows.
- The following uncertainties are considered:
  - uncertainty on QCD estimate due to the Z-candidate m_{SD} distribution fit functions;
  - uncertainty on QCD estimate due to the use of the nine mass windows;
  - statistical uncertainties for MC $Z \rightarrow q\bar{q}$ and $W \rightarrow q\bar{q}'$;
  - jet energy scale corrections for MC $Z \rightarrow q\bar{q}$ and $W \rightarrow q\bar{q}'$;
  - the luminosity uncertainty.
- All the uncertainties, with the exception of the luminosity one, are assumed uncorrelated in the different score regions.

Results

	r_Z
$0.641 < \text{PNet-MD}_{bbvsQCD} \leq 0.875$	0.9 ± 0.5 (stat) ± 0.3 (syst)
$0.875 < \text{PNet-MD}_{bbvsQCD} \leq 0.957$	1.37 ± 0.26 (stat) ± 0.21 (syst)
$0.957 < \text{PNet-MD}_{bbvsQCD} \leq 0.988$	1.25 ± 0.15 (stat) ± 0.12 (syst)
$0.988 < \text{PNet-MD}_{bbvsQCD} \leq 1$	1.01 ± 0.07 (stat) ± 0.07 (syst)

- $Z \rightarrow q\bar{q}$ (q = u, d, s, c, b) normalization factors (r_Z) with the corresponding error (split in statistical and systematic errors) in the four highest score regions.
- Normalization factors compatible with unity within uncertainties.