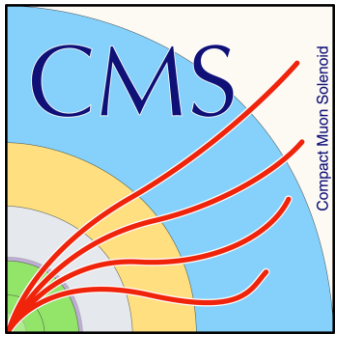


Boosted jet tagging at the CMS experiment

Donato Troiano[1,2]

October 8, 2024

CMS Italia 2024

1 INFN of Bari

2 University of Bari

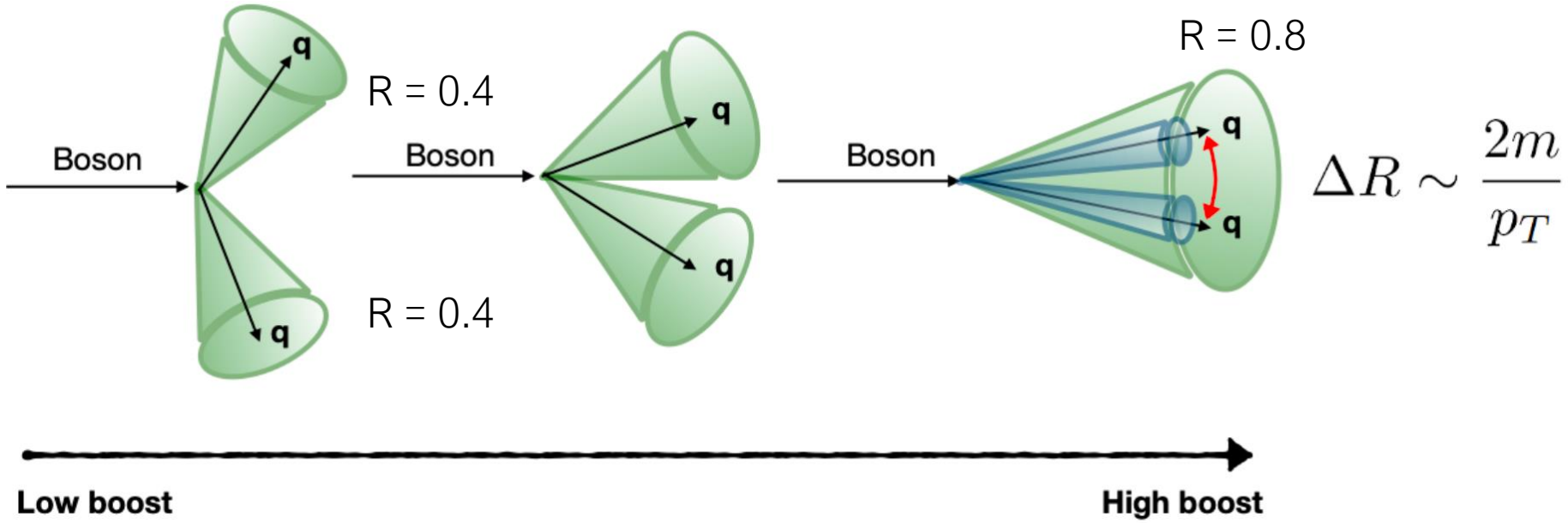
Boosted jets tagging

Hadronic decay products of highly boosted particles are collimated.

- Particles merged in one anti-k_T jet (radius R = 0.8 ⟶ AK8).

Identifying the particle generating the jet is a suitable task for ML:

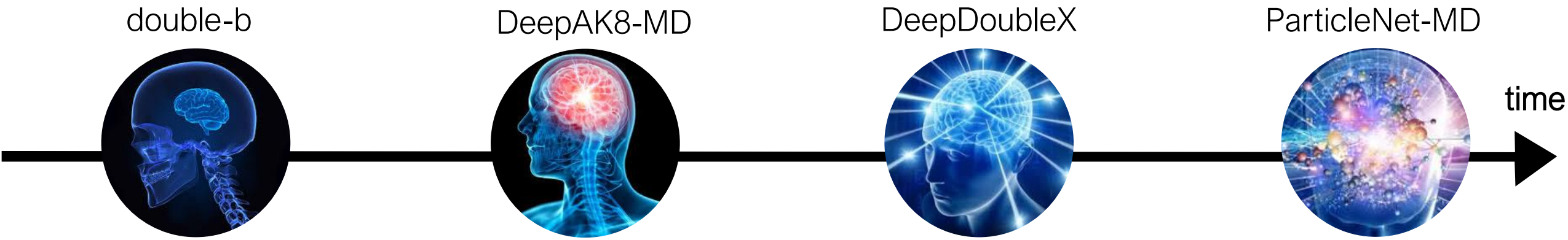
- large number of particles;
- diverse input variables for each particle.



Evolution of boosted jets tagging at CMS

early Run 2

state of art for Run 3 CMS



Model: Boosted decision Tree	1D Convolutional Neural Network (CNN)	1D CNN + Recursive NN	Dynamic Graph CNN (DGCNN)
Inputs: tracks and secondary vertices (SVs)	particle flow (PF) candidates and SVs	PF candidates and SVs	PF candidates, SVs and tracks
Outputs: • $H \rightarrow b\bar{b}$ vs QCD	• $X \rightarrow b\bar{b}$ vs QCD • $X \rightarrow c\bar{c}$ vs QCD	• $X \rightarrow b\bar{b}$ vs QCD • $X \rightarrow c\bar{c}$ vs QCD • $X \rightarrow b\bar{b}$ vs $X \rightarrow c\bar{c}$	• $p(X \rightarrow b\bar{b})$, $p(X \rightarrow c\bar{c})$, • $p(X \rightarrow$ light quarks), • $p(X \rightarrow \tau\bar{\tau})$, p(QCD)

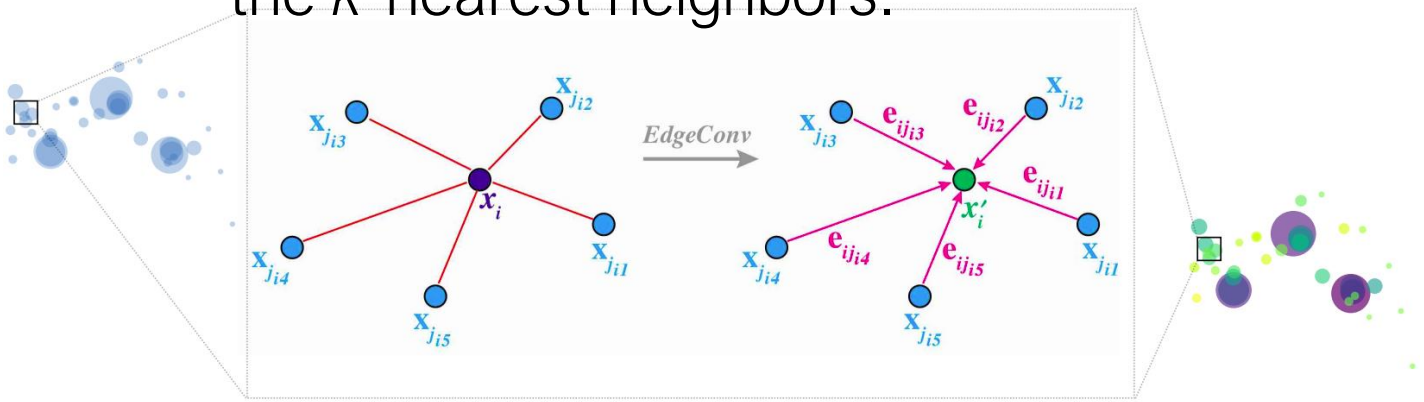
ParticleNet-MD

Architecture

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

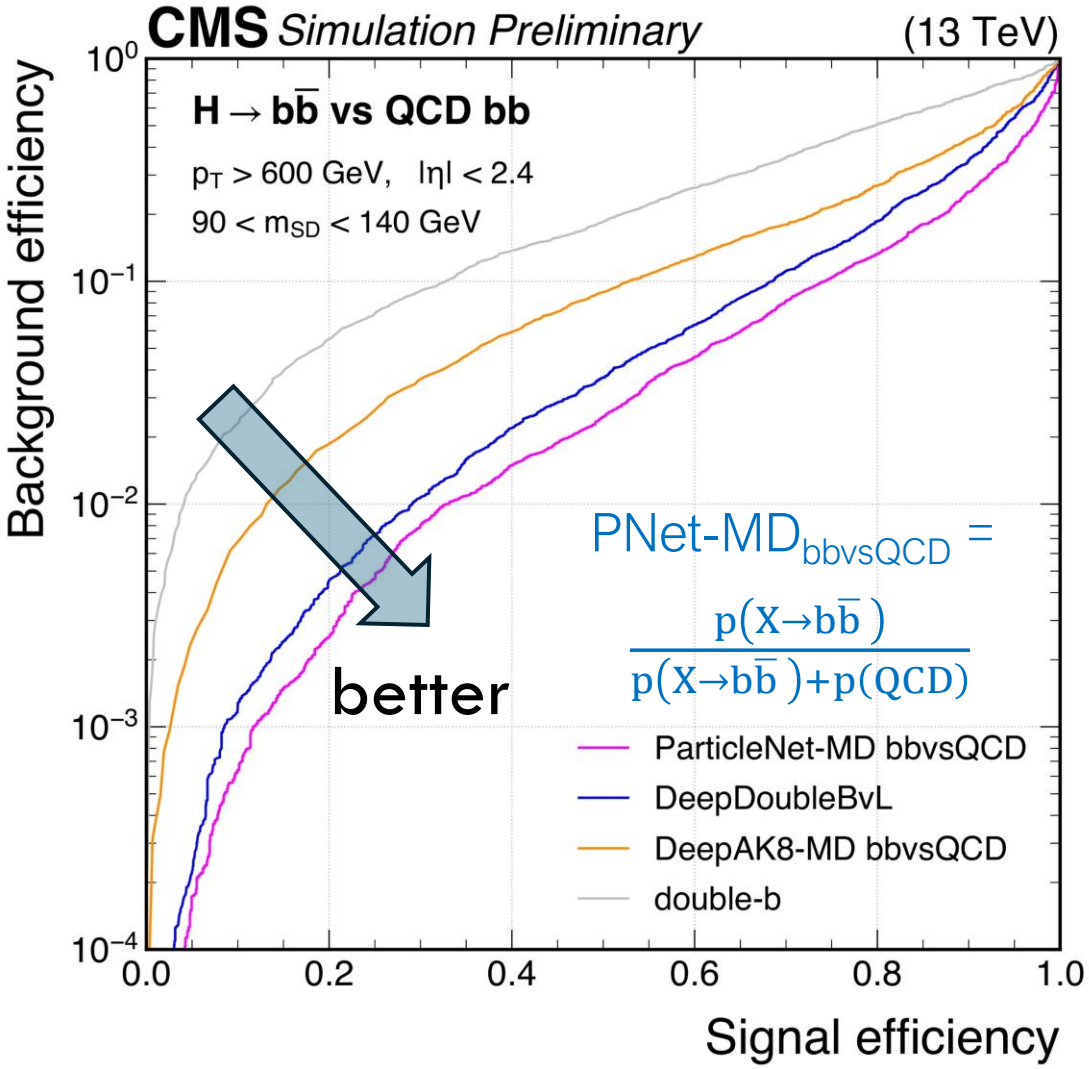
EdgeConv block:

- 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.

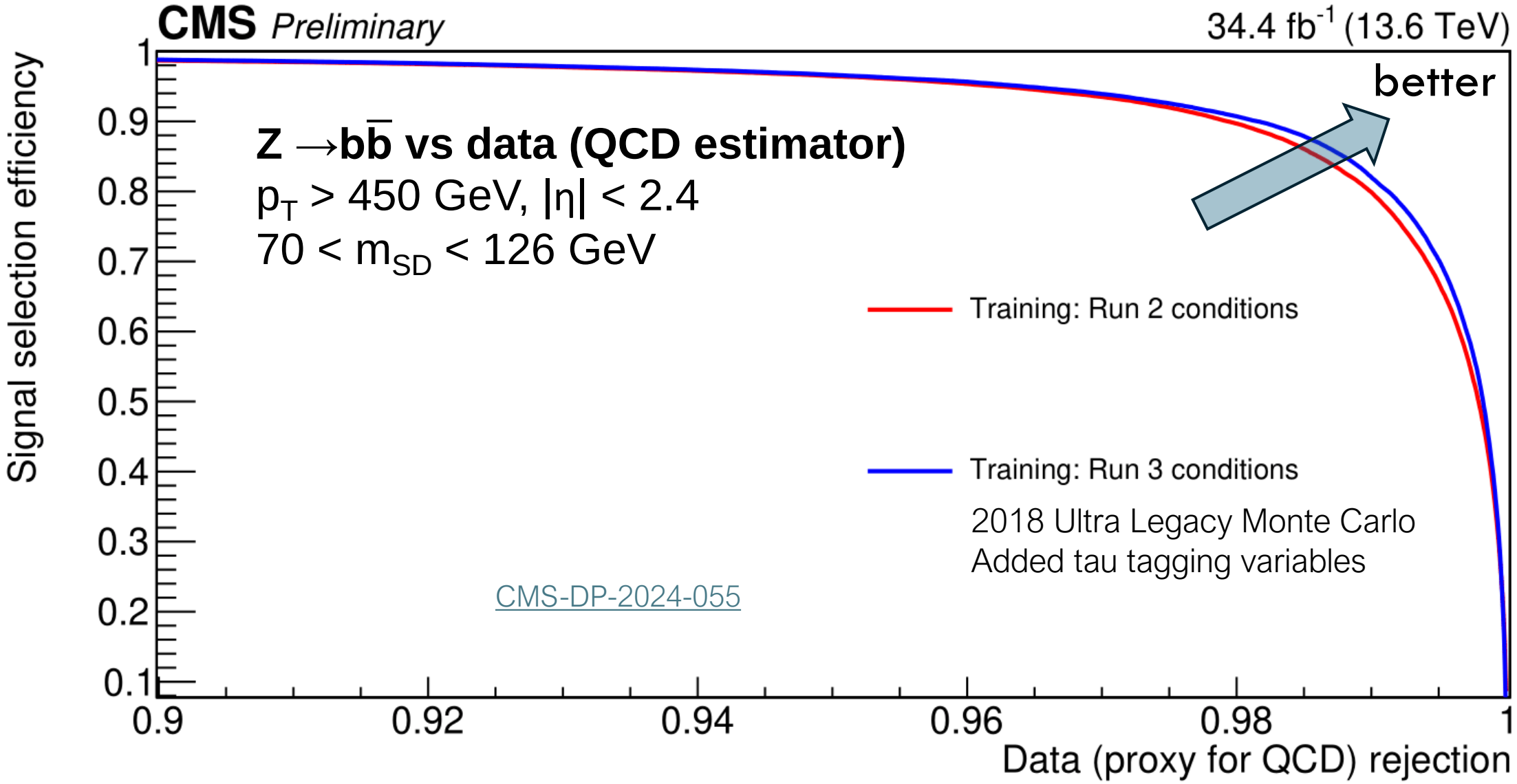


H. Qu, L. Gouskos, PRD 101, 056019 (2020)

CMS-PAS-BTV-22-001



ParticleNet-MD new training performance



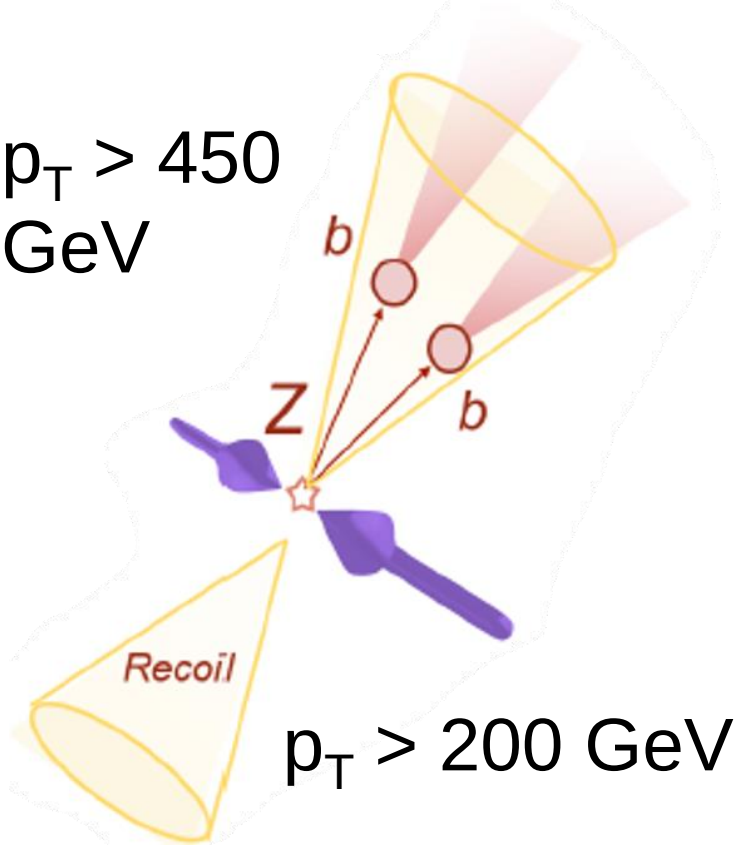
My work: First Run3 ParticleNet-MD validation

- Validation done on 2022 data events containing highly boosted $Z \rightarrow b\bar{b}$ + jets (CMS-DP-2024-055).

- Likelihood fit on the soft-drop mass (m_{SD}) to match data and SM prediction ($Z \rightarrow q\bar{q}$, $W \rightarrow q\bar{q}'$ and QCD).
  - m_{SD}: jet mass after removing soft and wide-angle radiation.

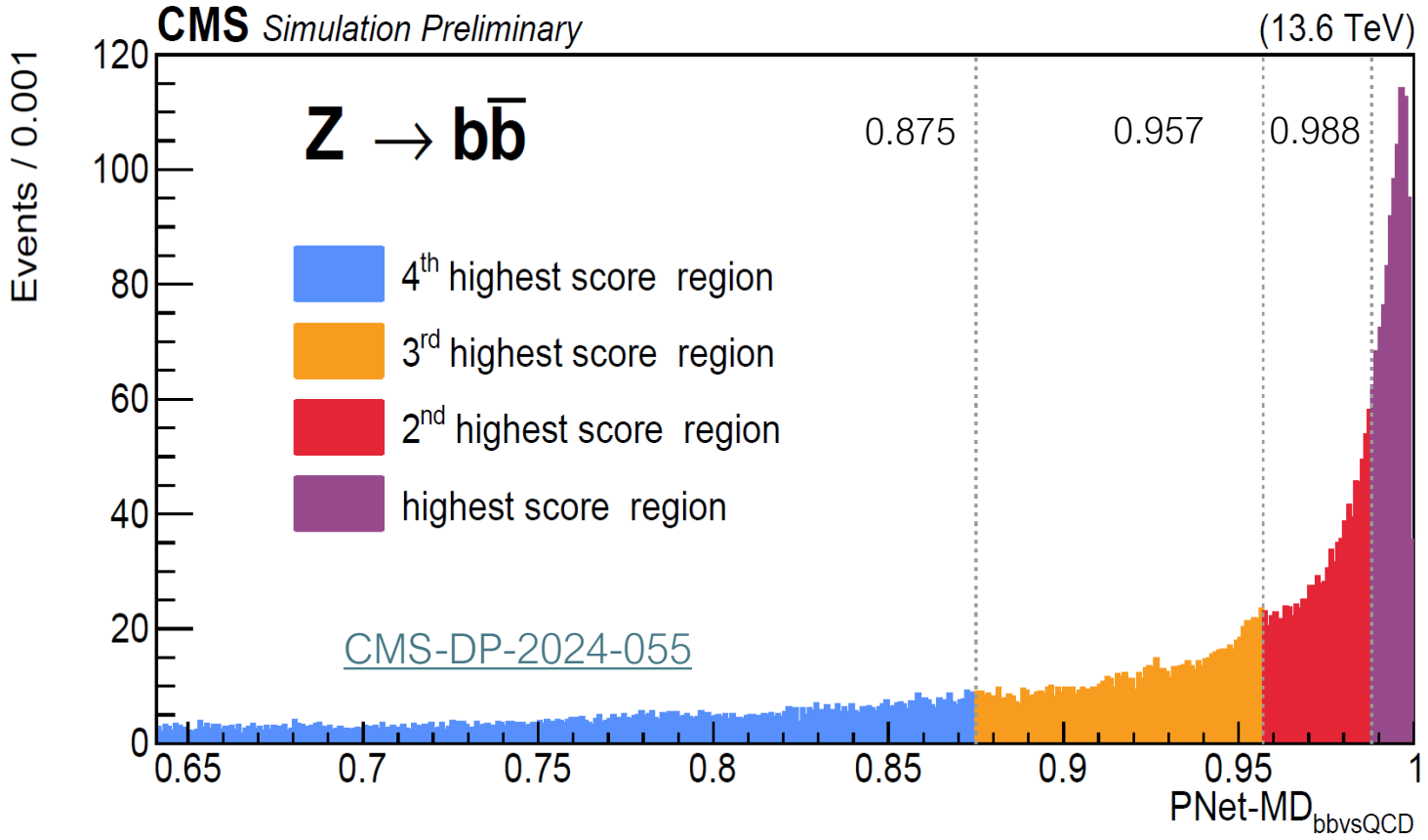
- Data-driven estimation of QCD multijet background:
  - average of the fits of the Z-candidate m_{SD} distributions in nine mass sidebands.

Z-candidate (leading-p_T AK8 jet)



ParticleNet-MD categorization

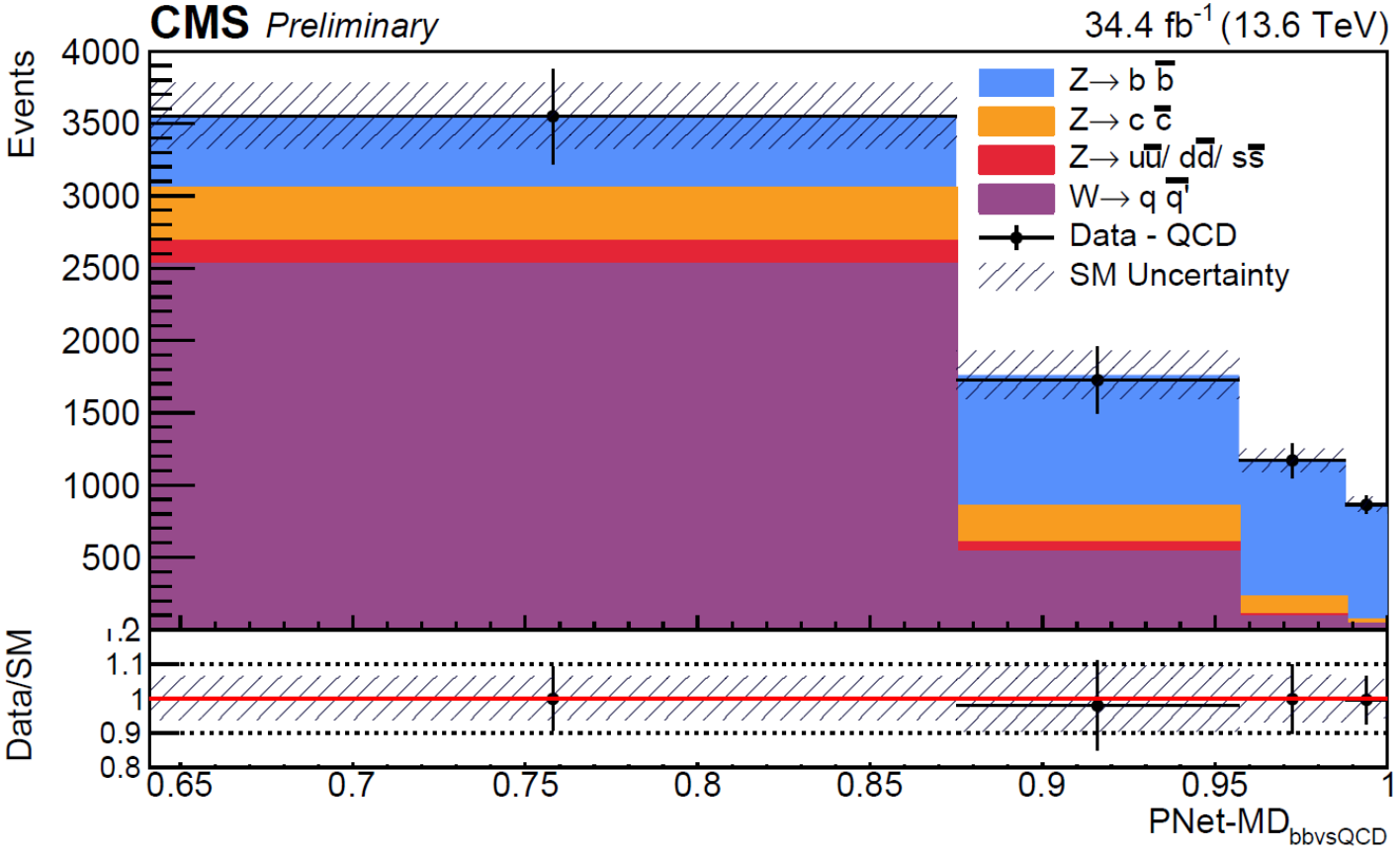
- Events categorized in five region of PNet-$MD_{bbvsQCD}$.
  - 20% of signal events in each PNet-$MD_{bbvsQCD}$ region.
  - Lowest score region (<0.641) almost completely populated by QCD events and not included in the fit.
  - Independent Likelihood fit performed in each of the four highest score regions.

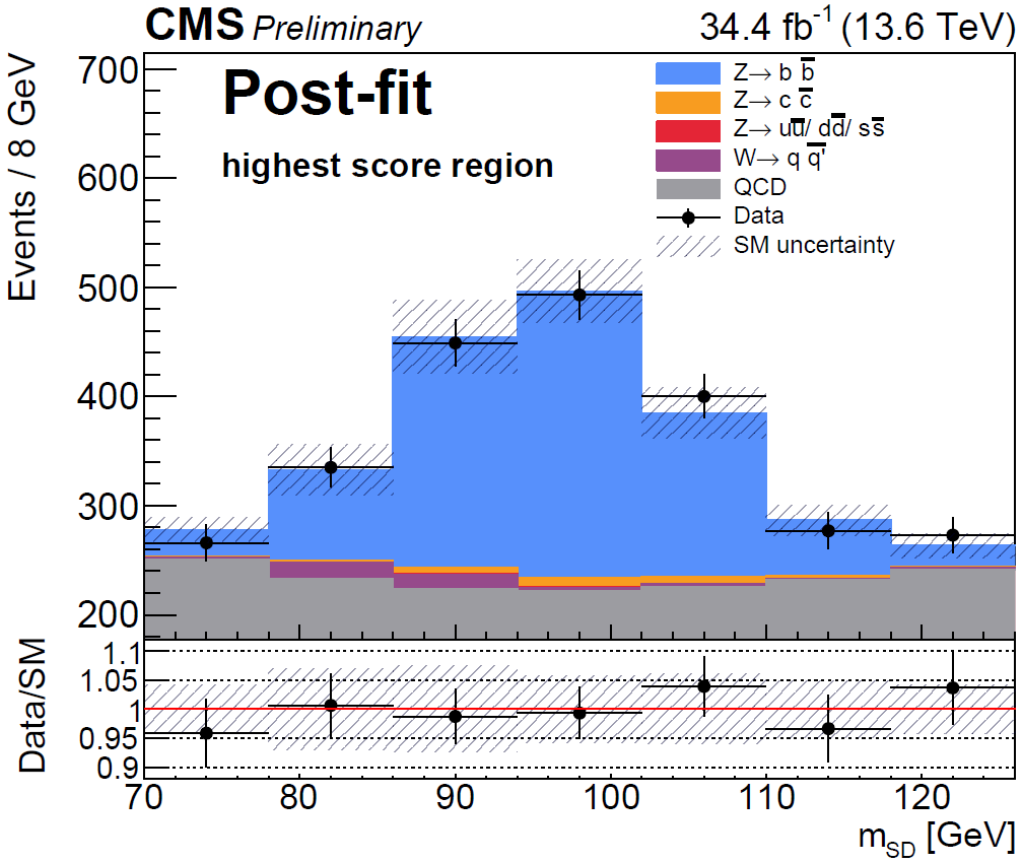


Boosted $Z \to b\bar{b}$ validation

- Good data-prediction agreement.
- Clear Z-peak in the data distribution.
  - Good signal discrimination from background.

CMS-DP-2024-055





Future plans for development of PNet at HLT

- ParticleNet-MD in HLT paths not optimized since Run 2.

- My plan: ParticleNet-MD optimization at HLT

  ☑ ParticleNet-MD input variables implemented in the BTV ntuplizer.

  ☒ Produce ntuples with flat p_t and mass distributions for mass decorrelation (MC production ongoing).

  ☒ Train on HLT features using BTV framework b-hive (targeting December for first HLT optimization).

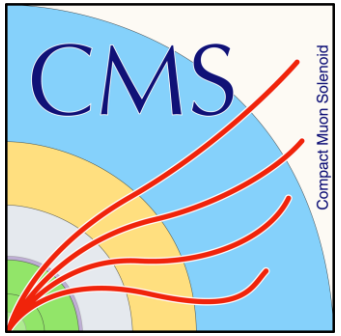
Conclusions

- Over the years, CMS has developed algorithms for boosted objects with increasing efficiency and performance.

- ParticleNet-MD best performing tagger for boosted jets, now optimized for Run 3.

- Validation on $Z \rightarrow b\bar{b}$-like events in 2022 data.
  - Stable performance, with some improvement with respect to Run 2.
  - Good data-prediction agreement.
  - Clear Z-peak in data distribution at high score.

- Currently working on boosted jet tagging at HLT.

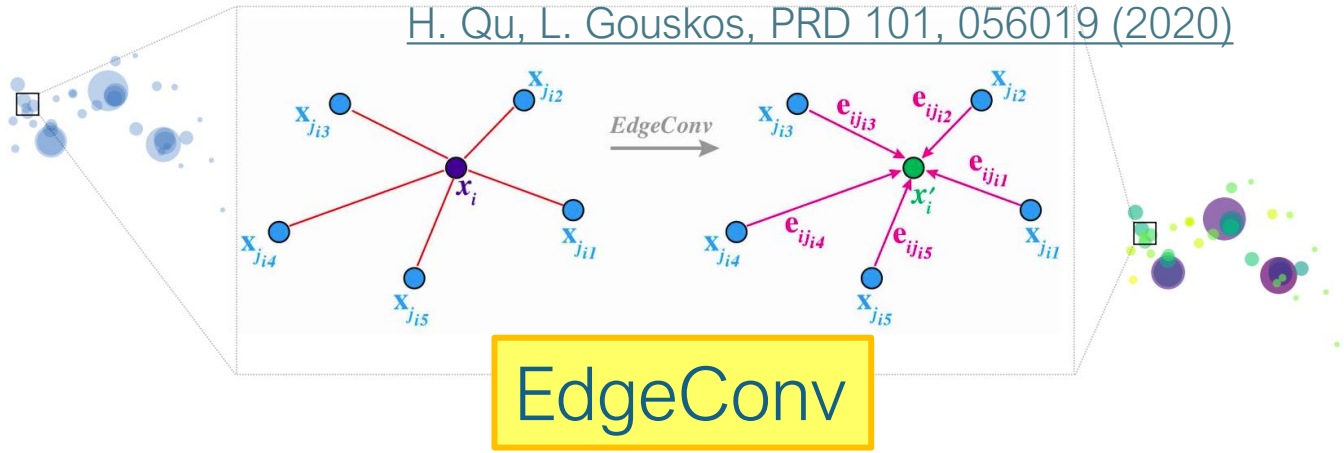


Backup

Jets as particle clouds: ParticleNet-MD

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

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



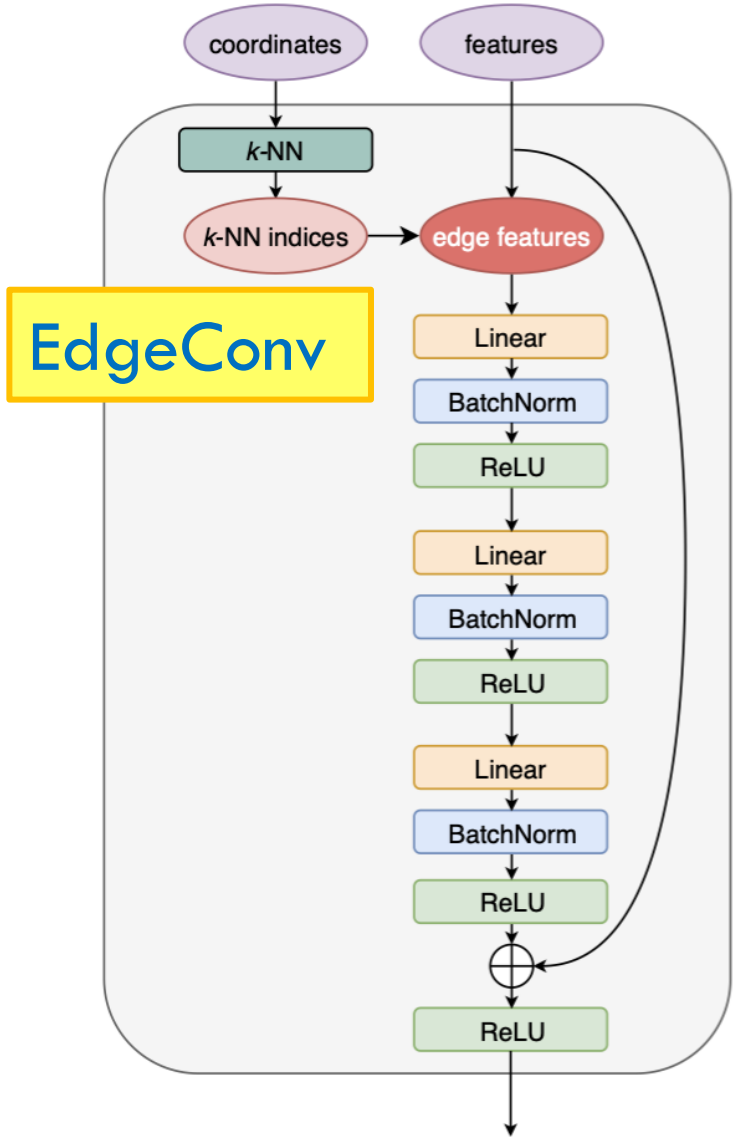
EdgeConv block:

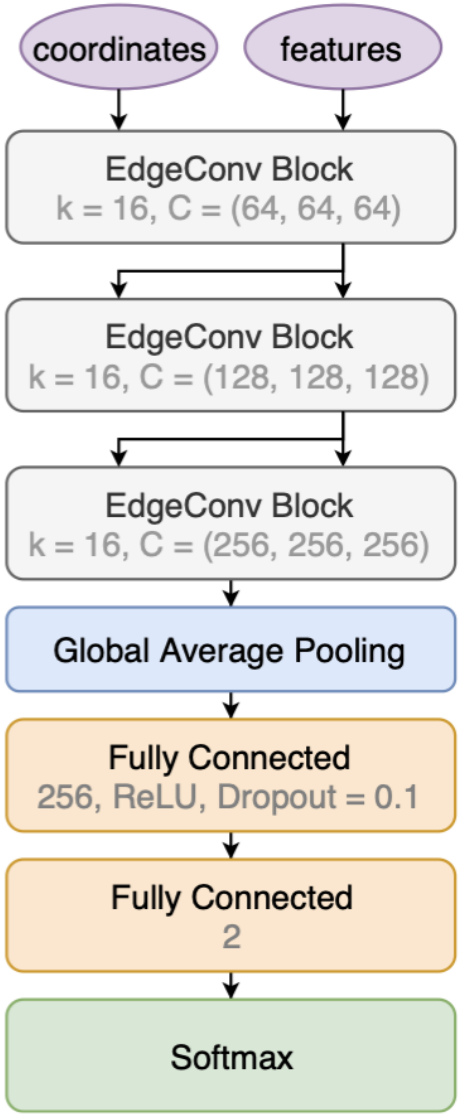
- NN module part of the ParticleNet architecture;
- New features vector associated to each jet constituent and based on the features of the *k*-nearest neighbors.

Mass decorrelation:

- Trained on Monte Carlo (MC) simulations containing boosted resonances (X), decaying in taus and quarks, 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.

ParticleNet architecture





ParticleNet

Binary outputs:

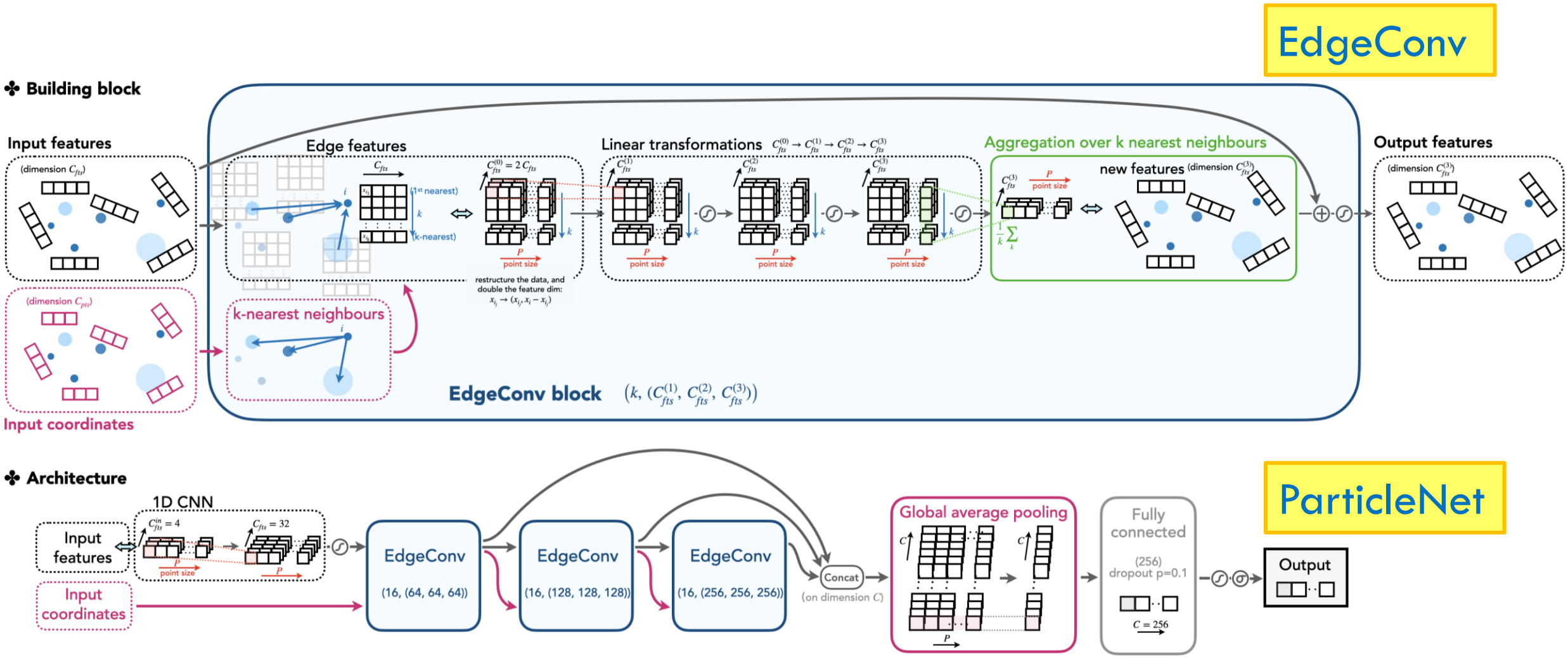
$$\text{PNet-MD}_{\text{bbvsQCD}} = \frac{p(X\to b\bar{b})}{p(X\to b\bar{b})+p(\text{QCD})}$$

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

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

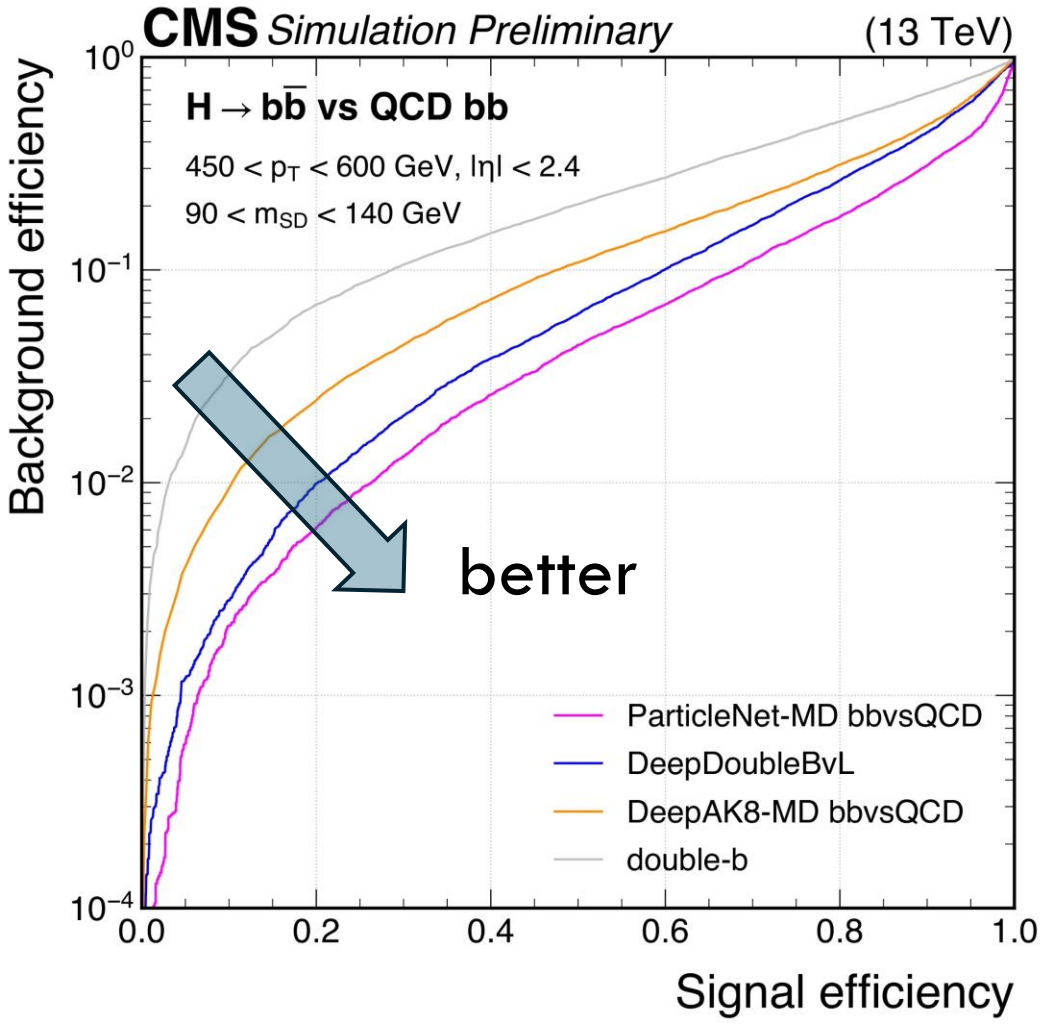
$$\text{PNet-MD}_{\tau\tau\text{vsQCD}} = \frac{p(X\to \tau\bar{\tau})}{p(X\to \tau\bar{\tau})+p(\text{QCD})}$$

ParticleNet architecture

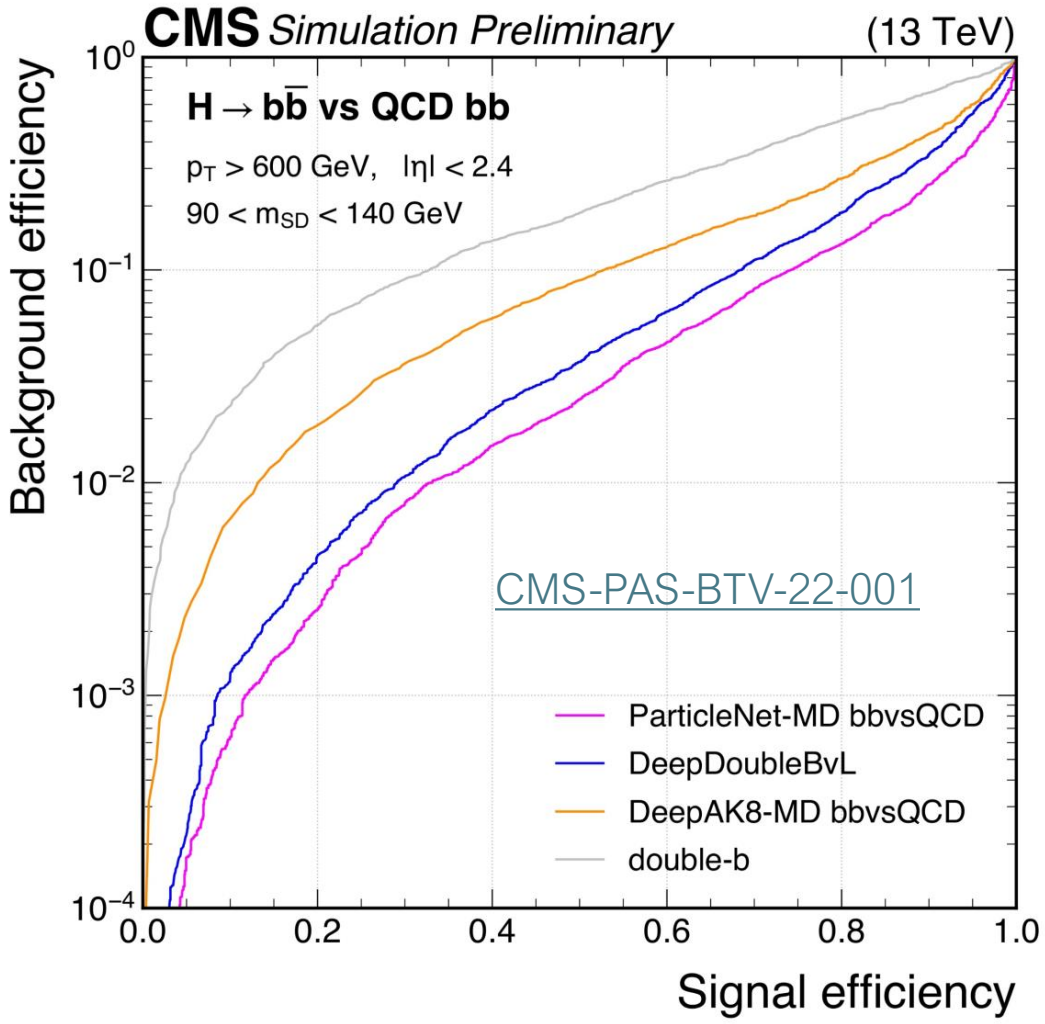


ROC curve $b\bar{b}$ tagging performances (Run 2)

Low p_T: (450;600) GeV

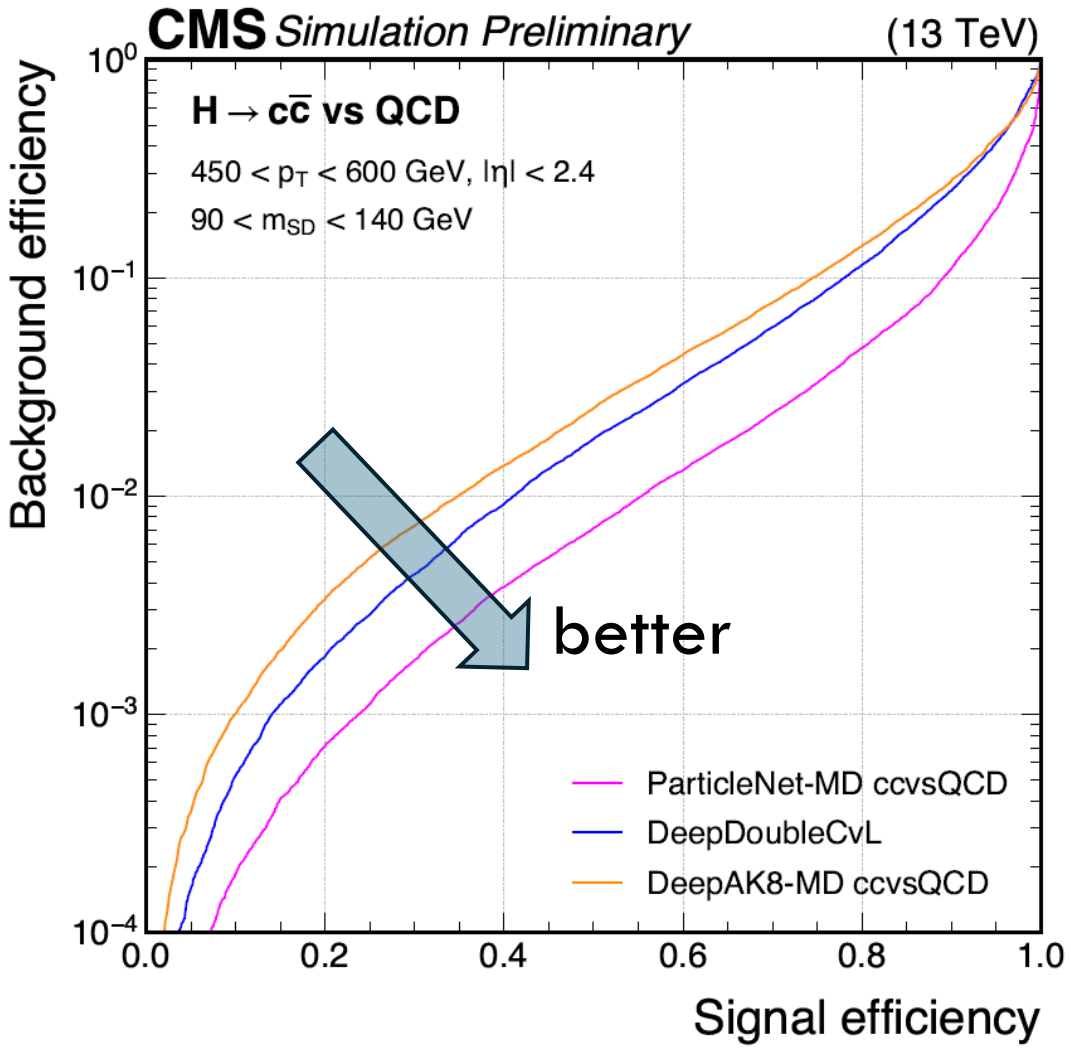


High p_T: (600;∞) GeV

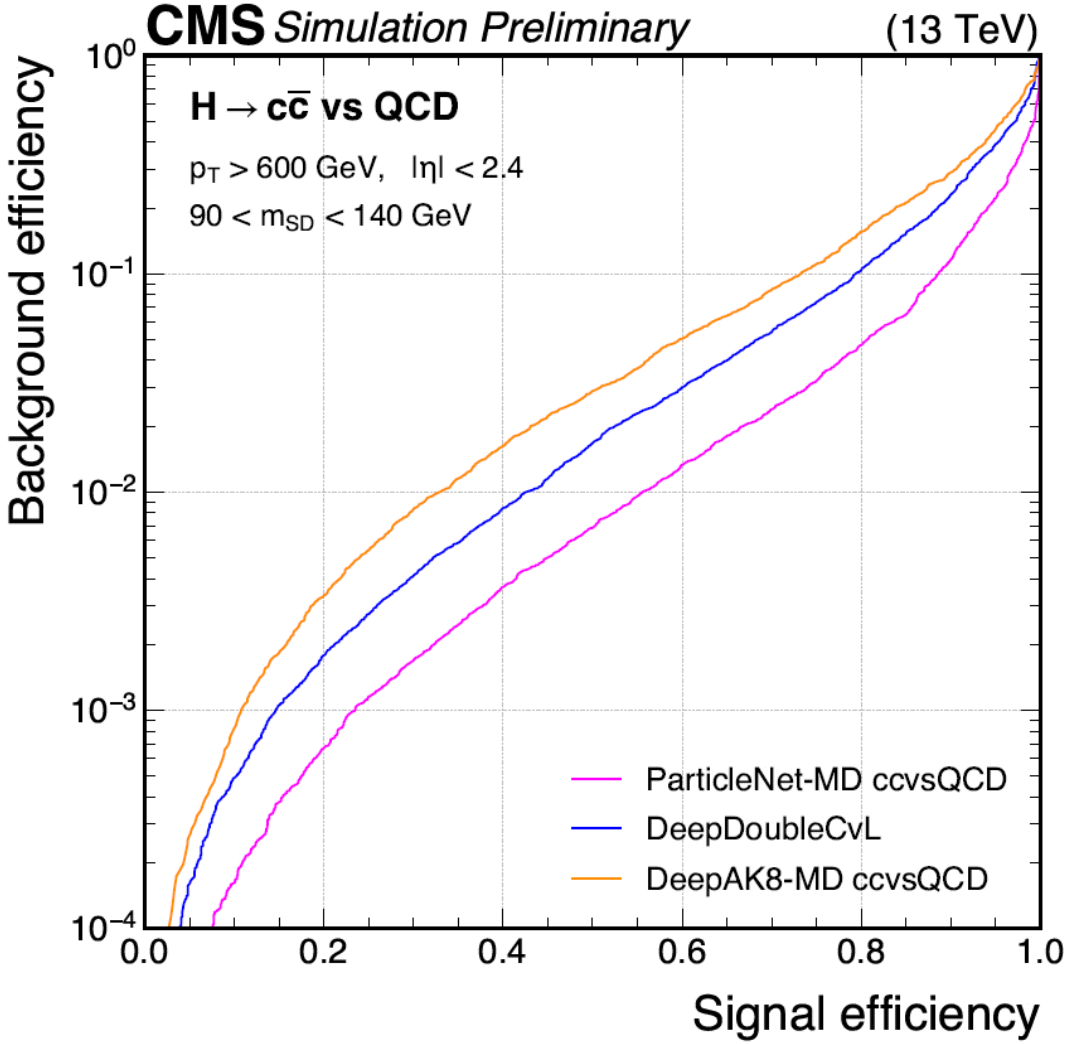


ROC curve $c\bar{c}$ tagging performances (Run 2)

Low p_t: (450;600) GeV



High p_t: (600;∞) GeV

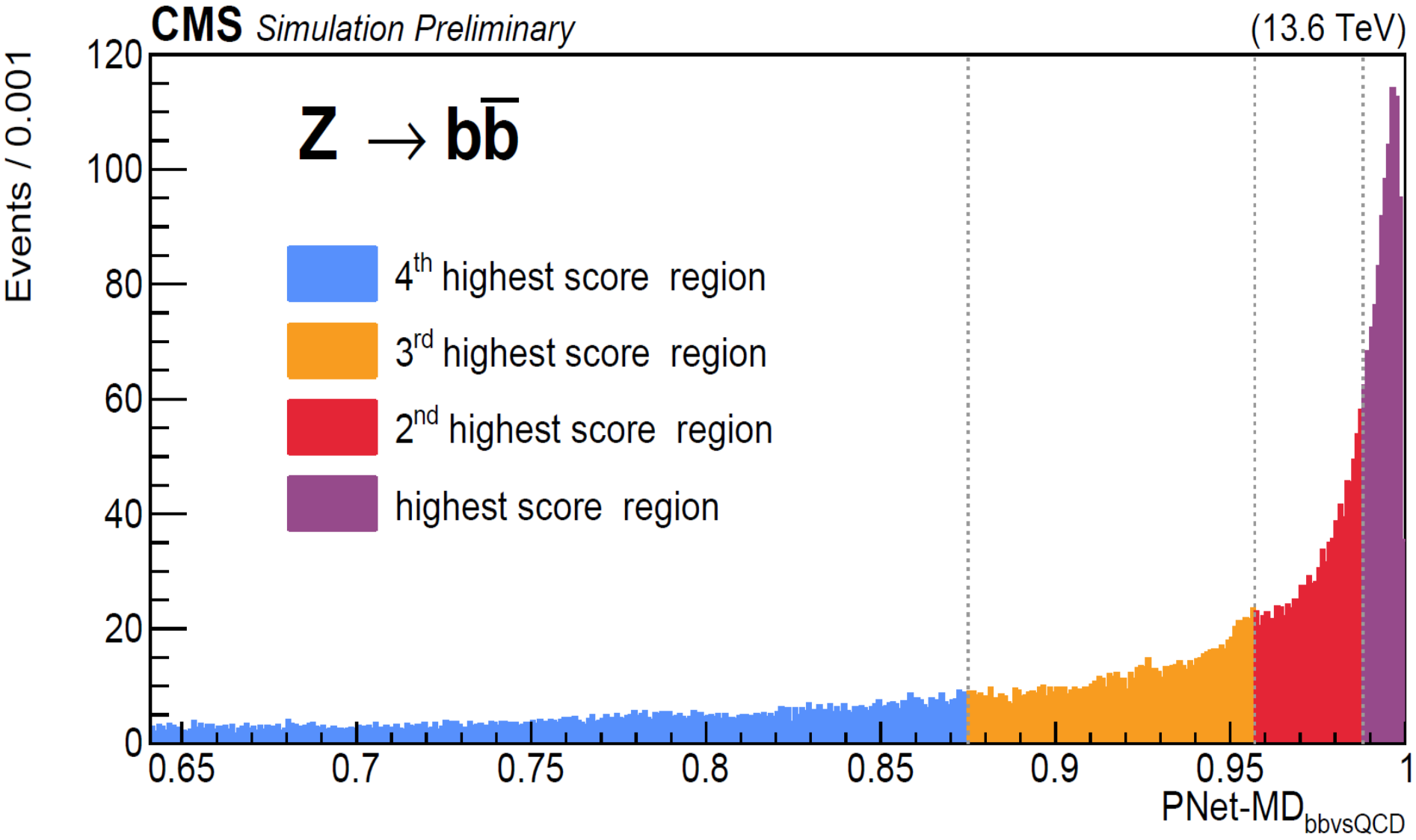


Boosted $Z \rightarrow b\bar{b}$ event selection

- **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→qq̄ after the event selection, normalized to 34.4 fb-1.
- Z-candidate compatible with a Z decaying to bb̄ at generation level.
- Each of the four highest score regions has the same amount (20%) of MC Z→bb̄ events .

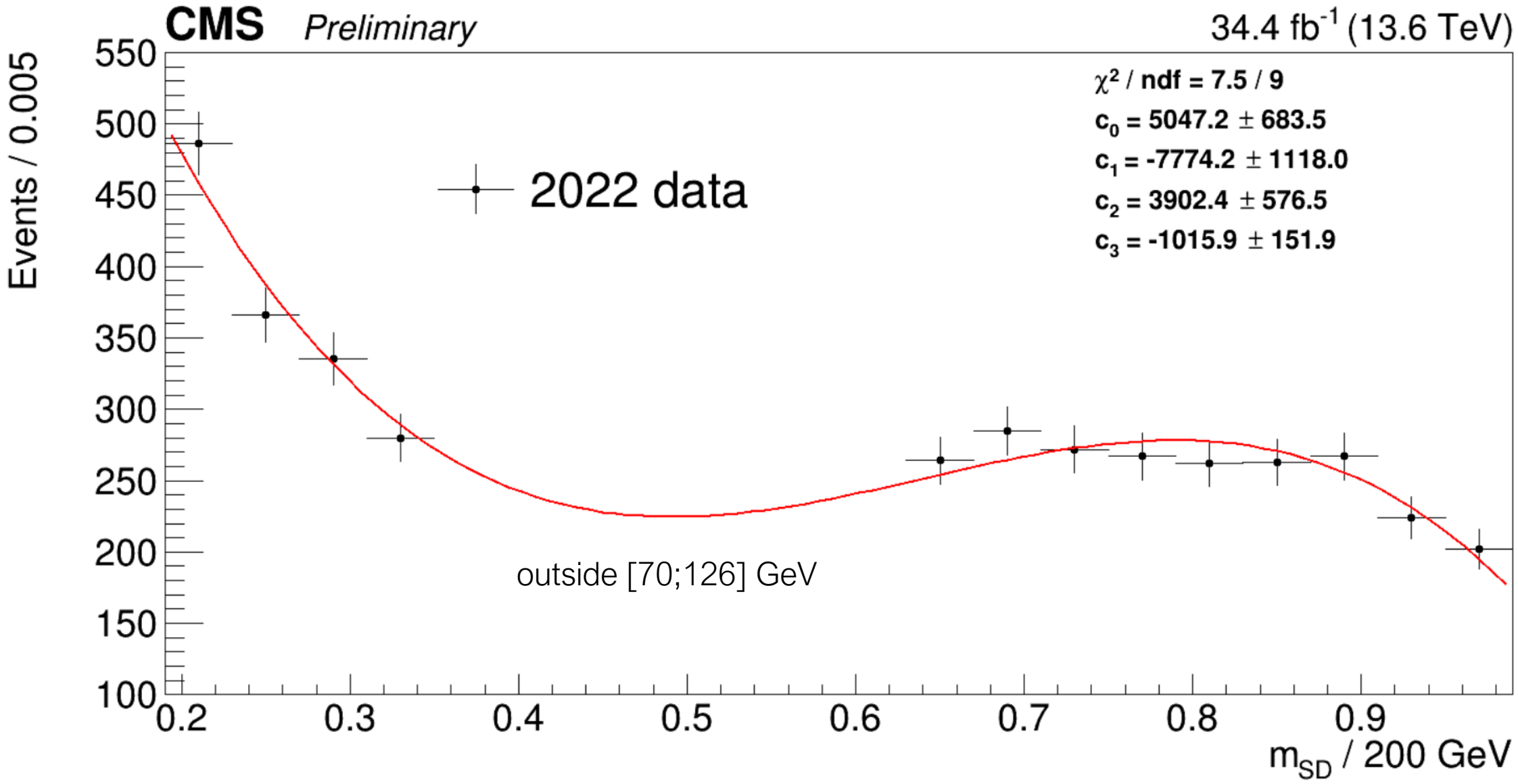
PNet-MD$_{bbvsQCD}$ validation: Likelihood fit

- 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→q$\bar{q}$ and W→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 →q$\bar{q}$ and W →q$\bar{q}$';
  - jet energy scale corrections for MC Z →q$\bar{q}$ and W →q$\bar{q}$';
  - the luminosity uncertainty.
- All the uncertainties, with the exception of the luminosity one, are assumed uncorrelated in the different score regions.

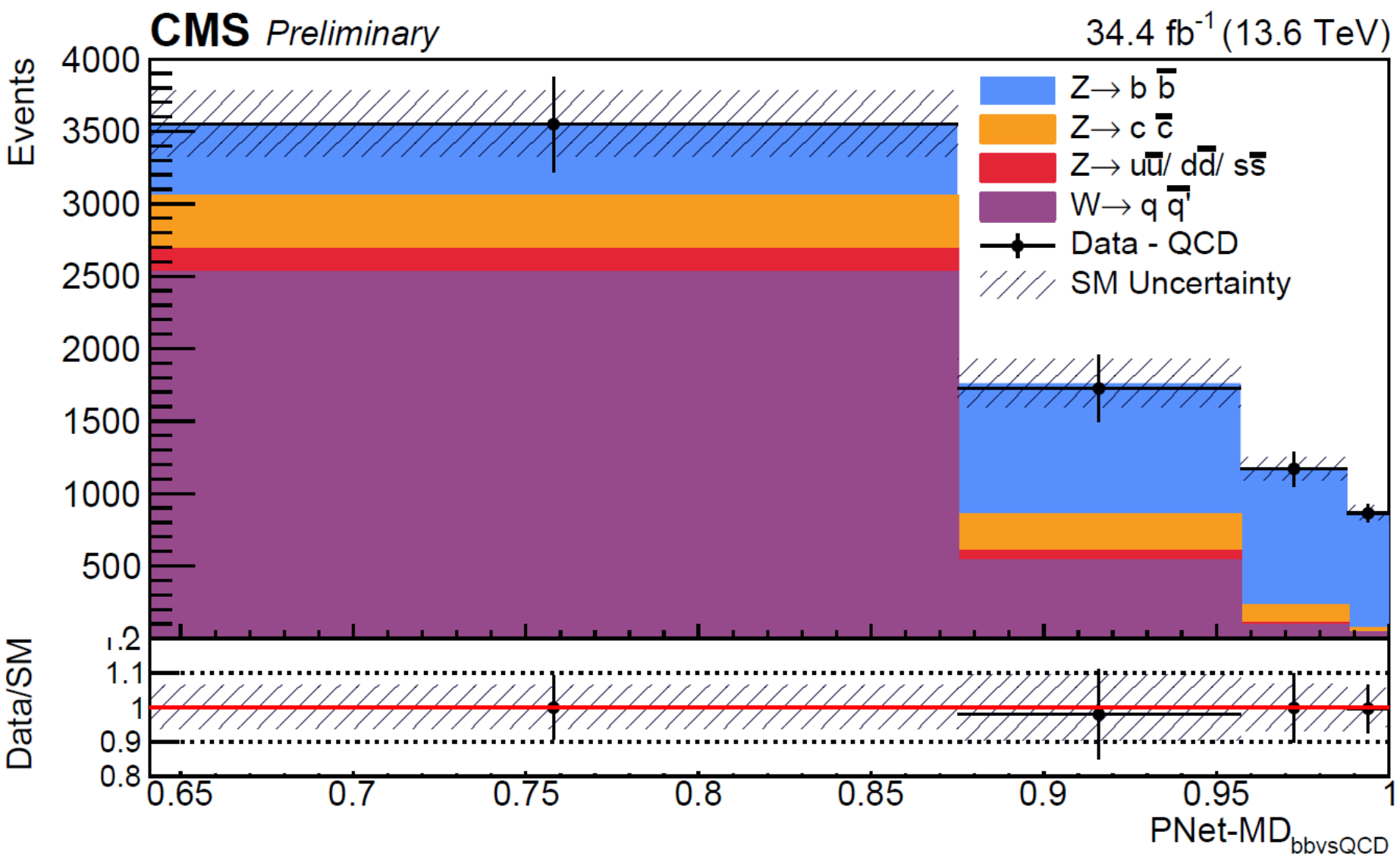
QCD evaluation

- QCD estimation Data-driven technique
  1. QCD estimated by fitting data m_{SD} distribution ([50-150] GeV) outside [70;126], [62;134], [78;118], [54;142], [86,110], [62;126], [78;126], [70,134] and [70;118] GeV windows.
     - QCD obtained as the average of the fits.
  2. Chebyshev polynomial functions have been used to fit Data.
     - Order established by means of the Fisher test (CL at 5%).
- QCD distributions assumed with the following uncertainties:
  - **fit:** symmetric error in each m_{SD} bin equal to the average of the fit function errors associated to each bin.
  - **window:** symmetric error in each m_{SD} bin equal to the root mean square deviation of the m_{SD} bin values obtained from the different fit functions.

QCD evaluation



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



- PNet-MD$_{bbvsQCD}$ distribution of the stack of W→qq̄' and Z→qq̄ (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→qq̄' and Z→qq̄ events, with q (light) or c quark, decreases increasing the score value.

Signal fraction

	Z→qq (q=b,c,u,d,s) percentage		
	Z→bb	Z→cc	Z→ uu/ dd/ ss
4th highest score region	48.8%	35.5%	15.7%
3rd highest score region	73.9%	20.5%	5.6%
2nd highest score region	87.7%	10.9%	1.4%
highest score region	96.6%	3.18%	0.26%

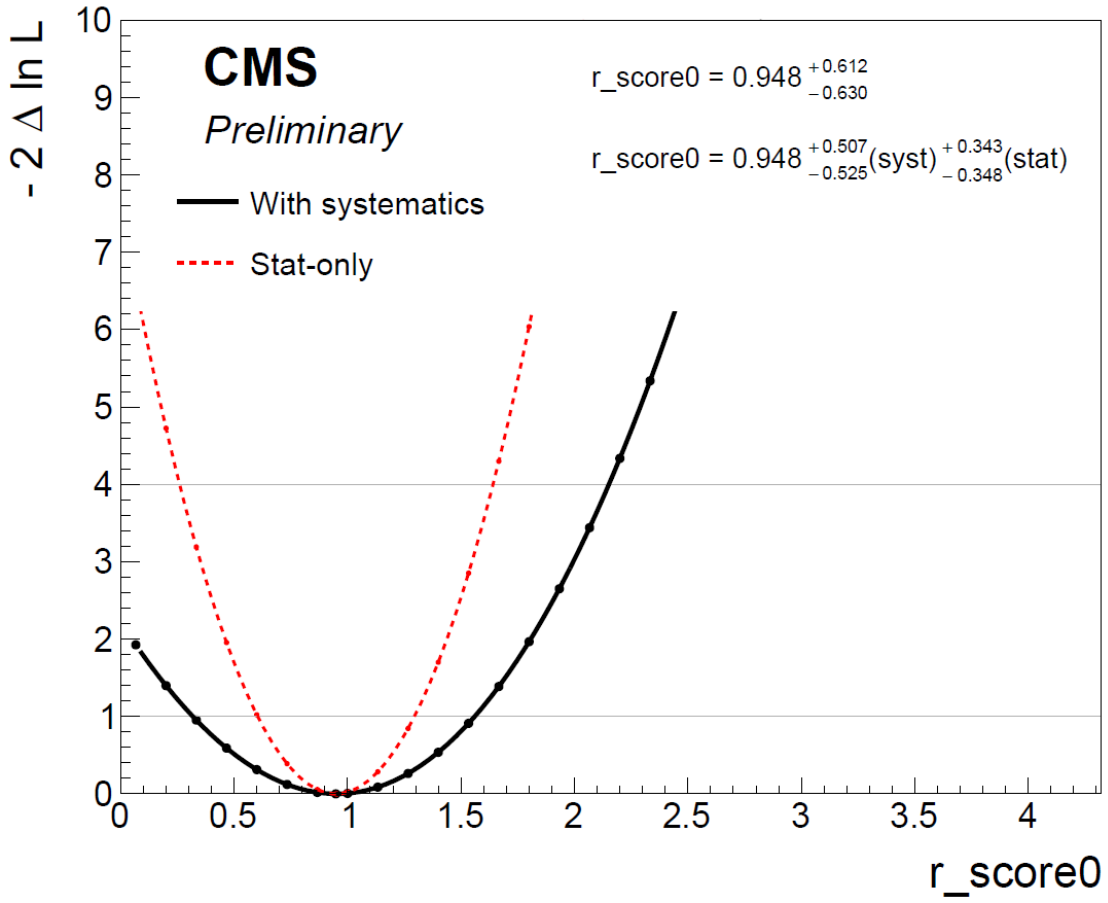
- MC mis-identified $Z \rightarrow b\bar{b}$ percentage less than 4% in highest score region.
- 20% of $Z \rightarrow b\bar{b}$ signal events lost moving from a score region to the next.

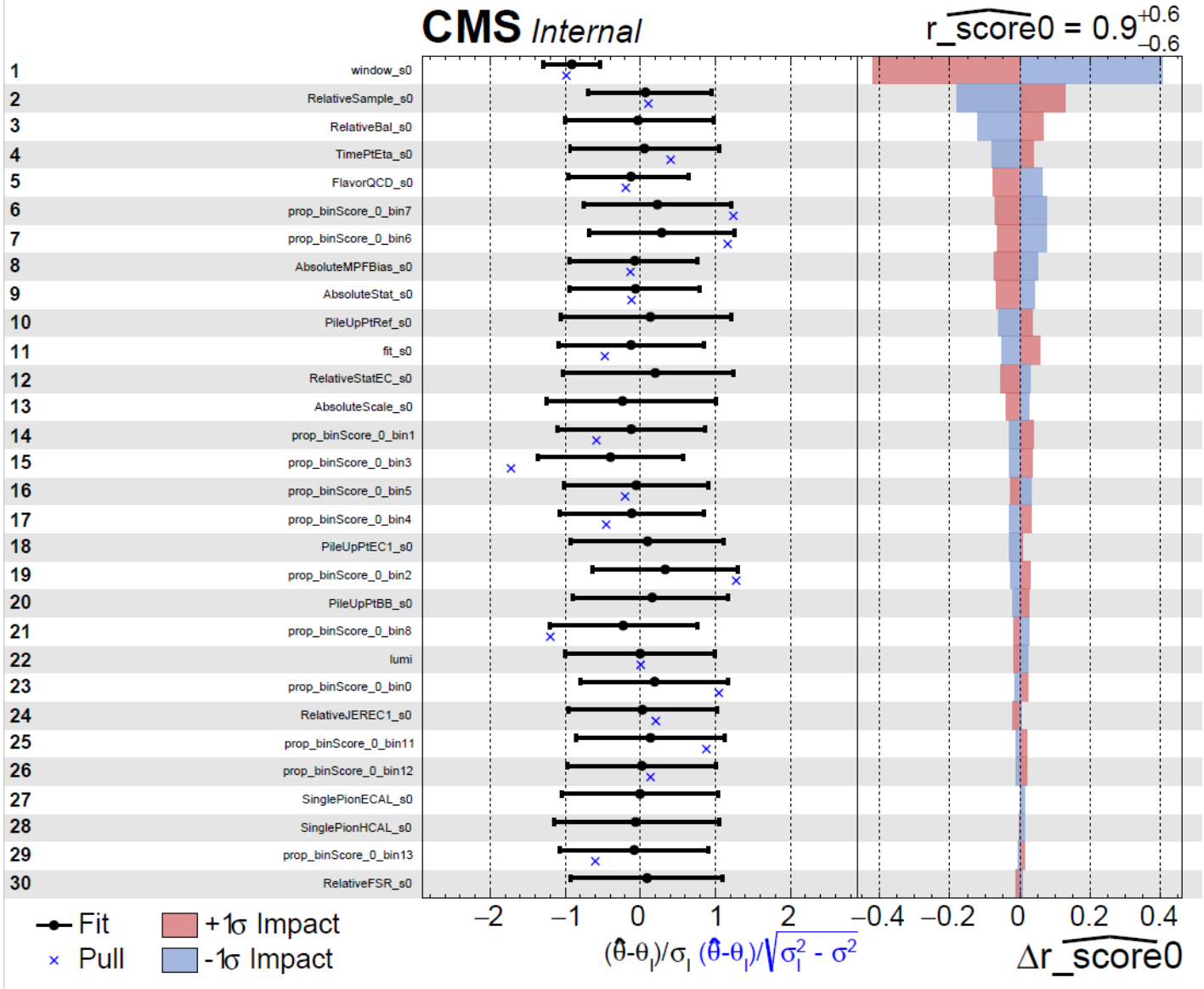
ParticleNet-MD validation results

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

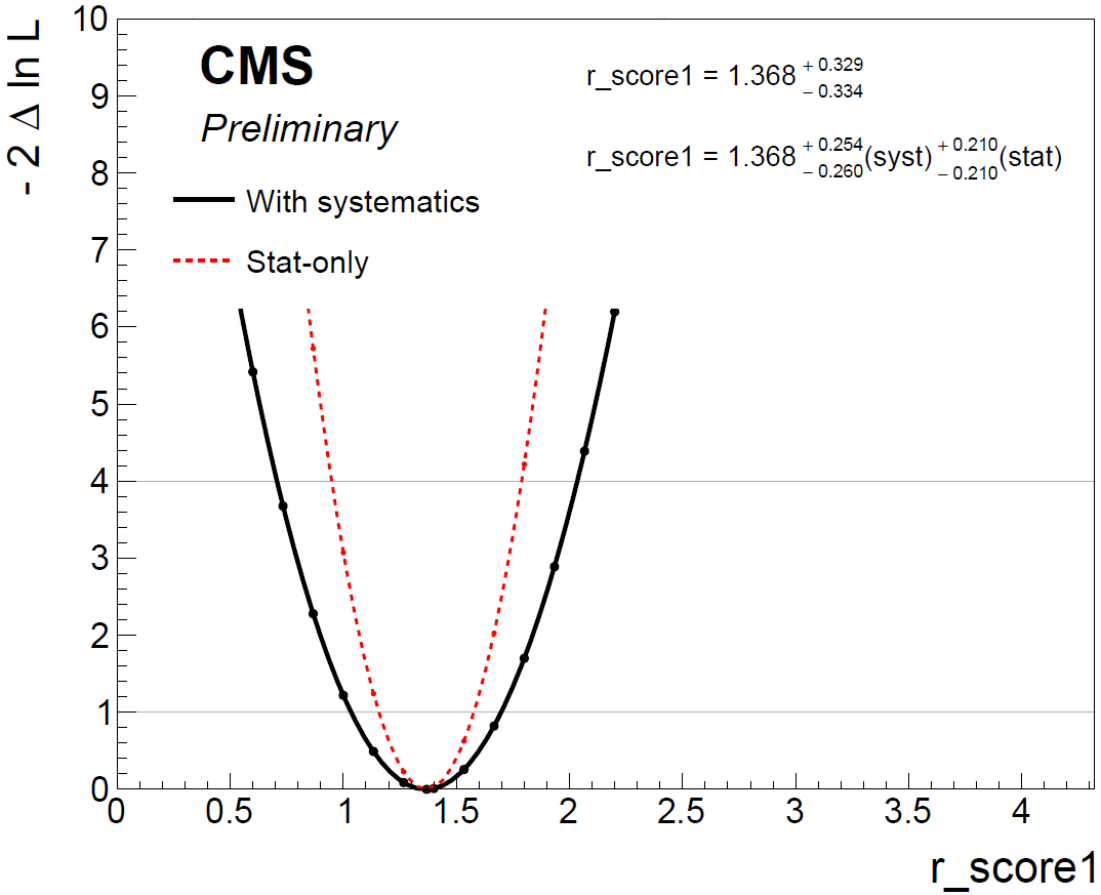
- $Z \rightarrow q\bar{q}$ (q = u, d, s, c, b) normalization factors (r_Z) with the corresponding error (split in statistical and systematic errors) in the four highest score regions.
- Normalization factors compatible with unity within uncertainties.

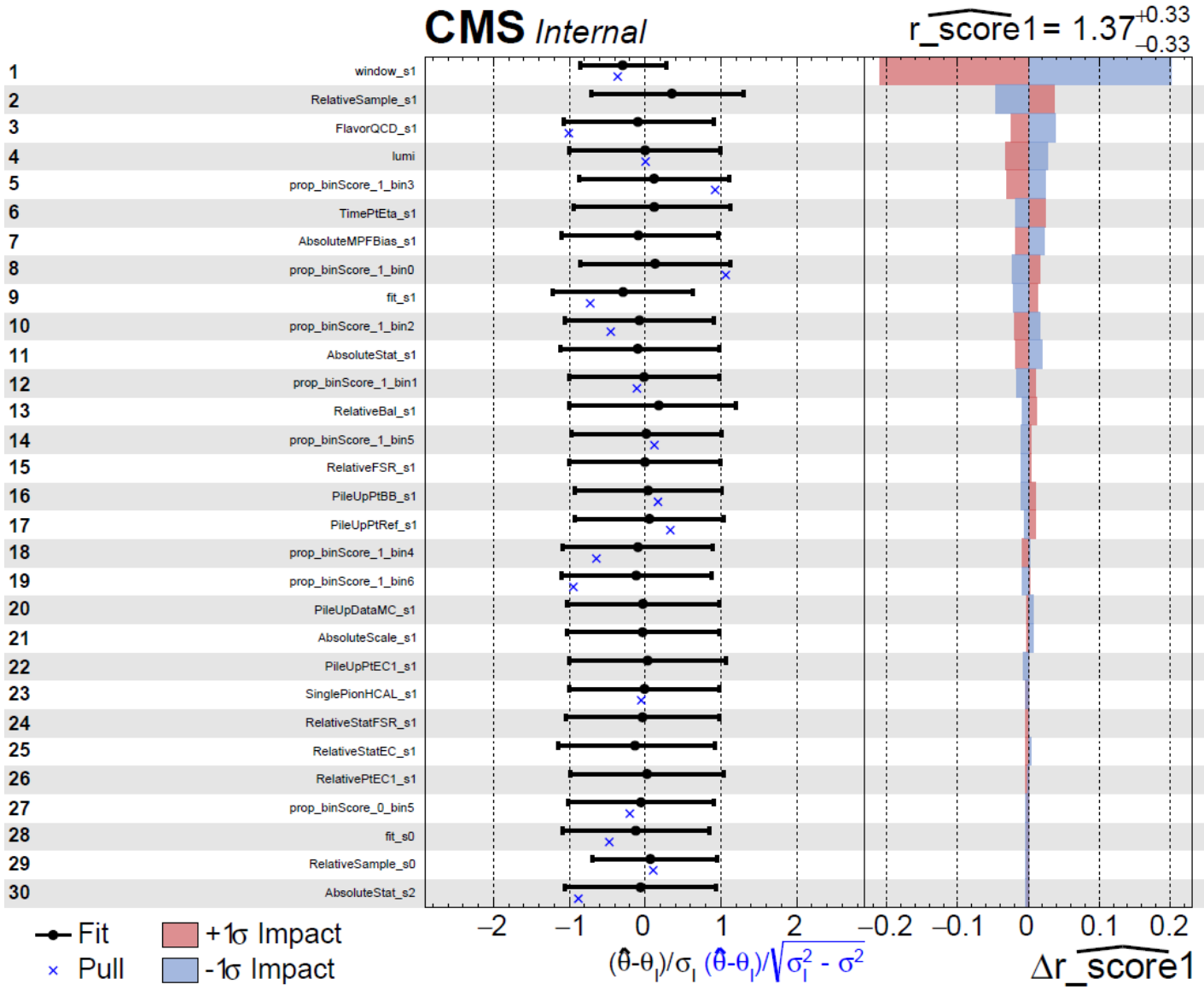
4th highest score region



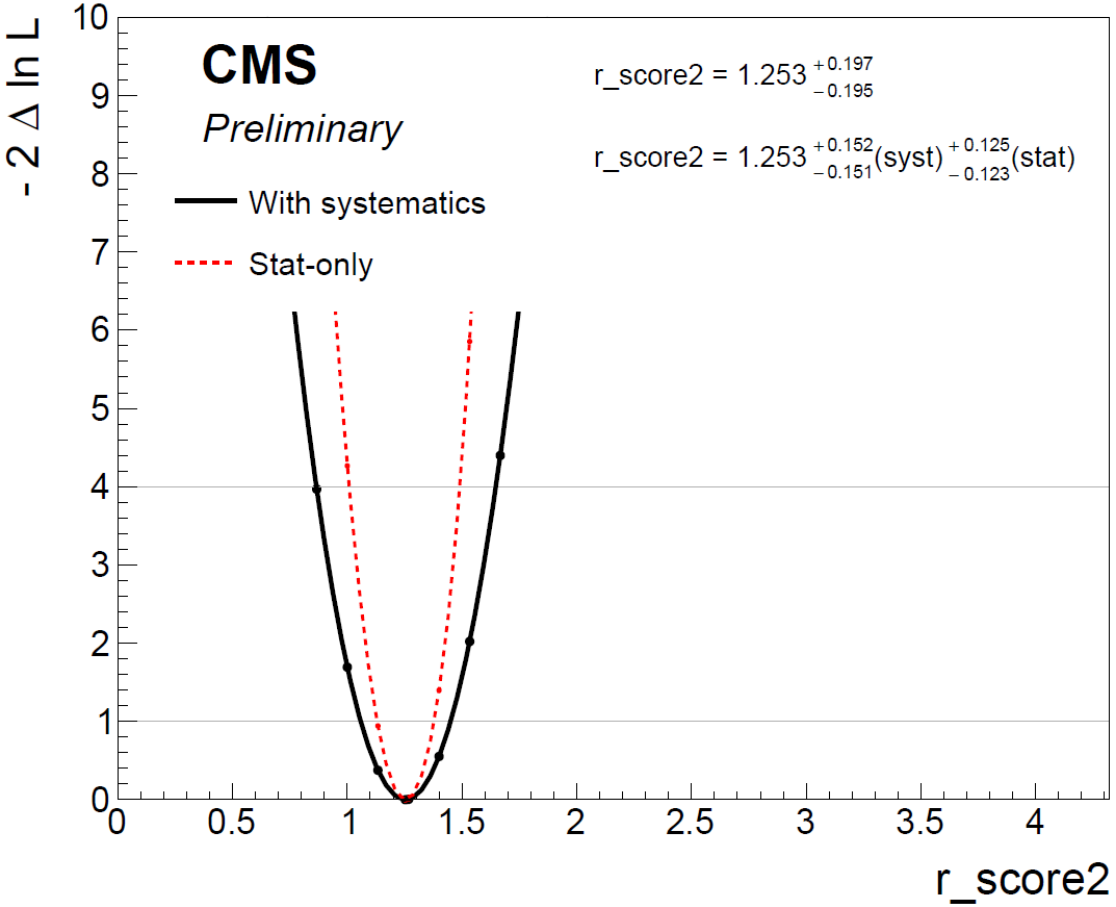


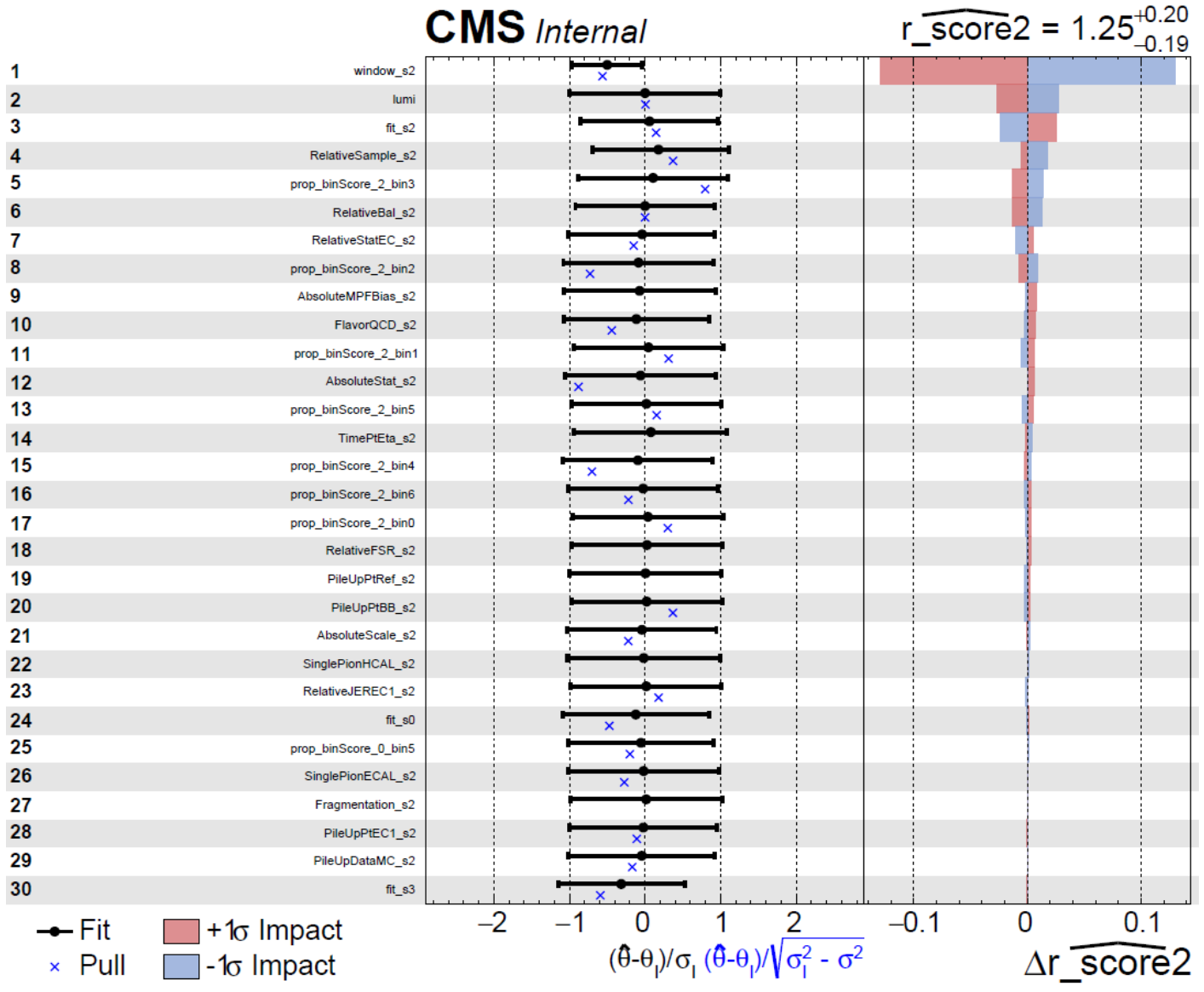
3rd highest score region



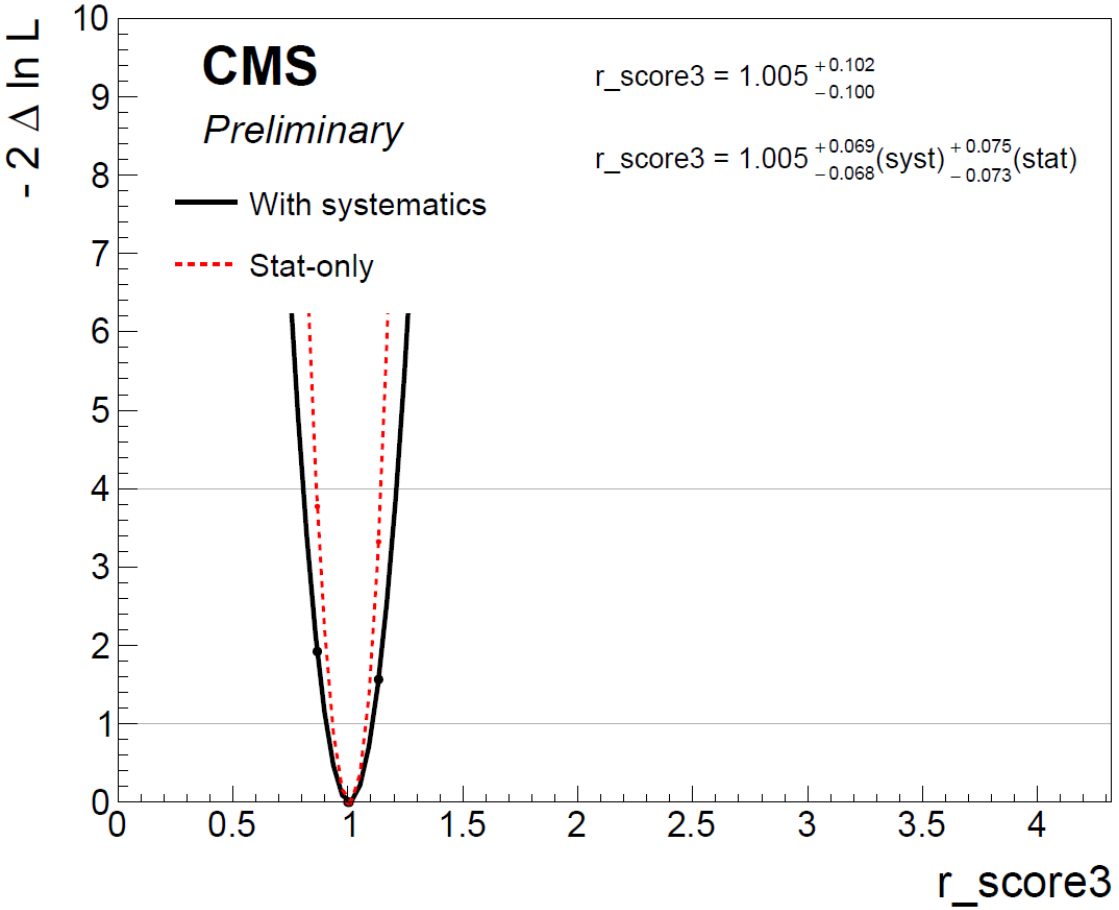


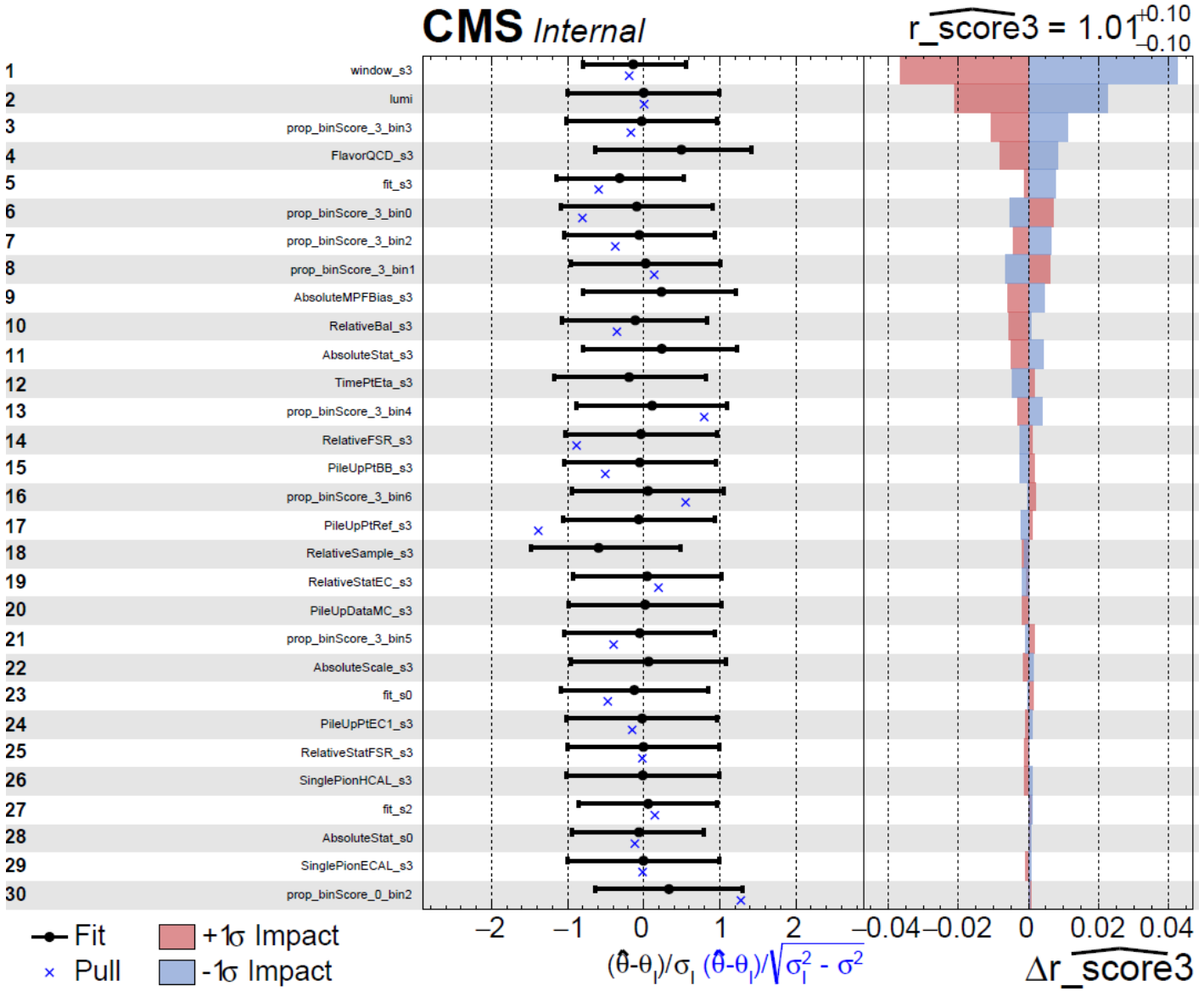
2nd highest score region





Highest score region





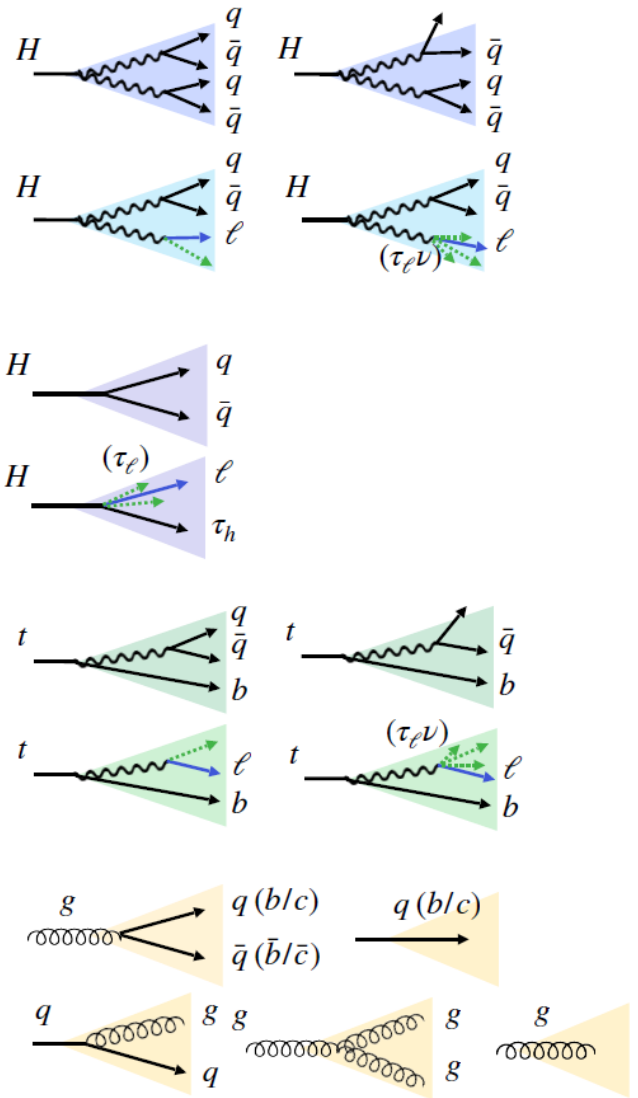
ParticleTranformer

- ParticleTransformer architecture shows superior performance than ParticleNet [arXiv:2202.03772].

- Next goal: extend tagging to a large set of final states including top-tagging (3 prong), X → VV with hadronic and leptonic decays, taus etc.

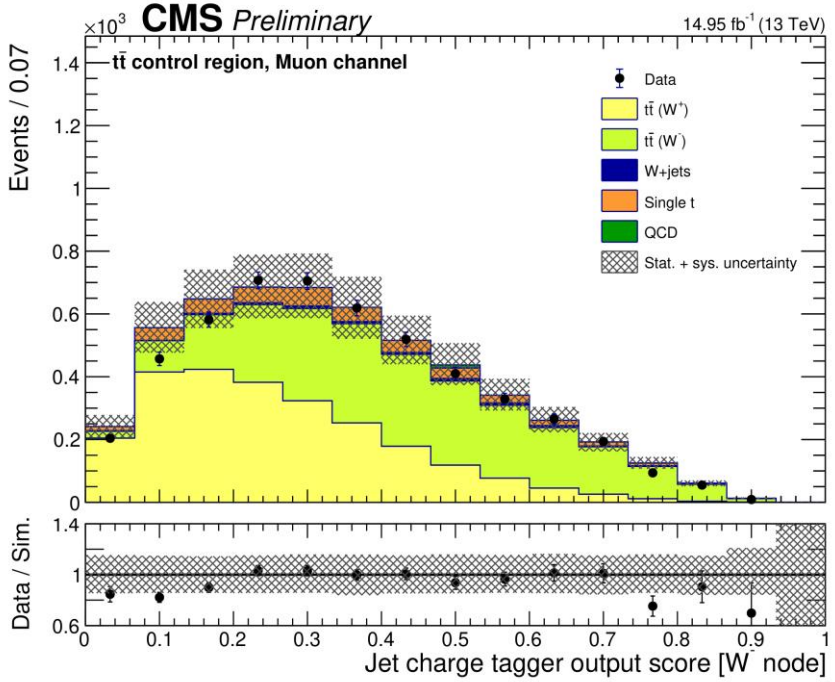
- ParticleTransformer used already in HH→VV→4b analysis [CMS-PAS-HIG-23-012]

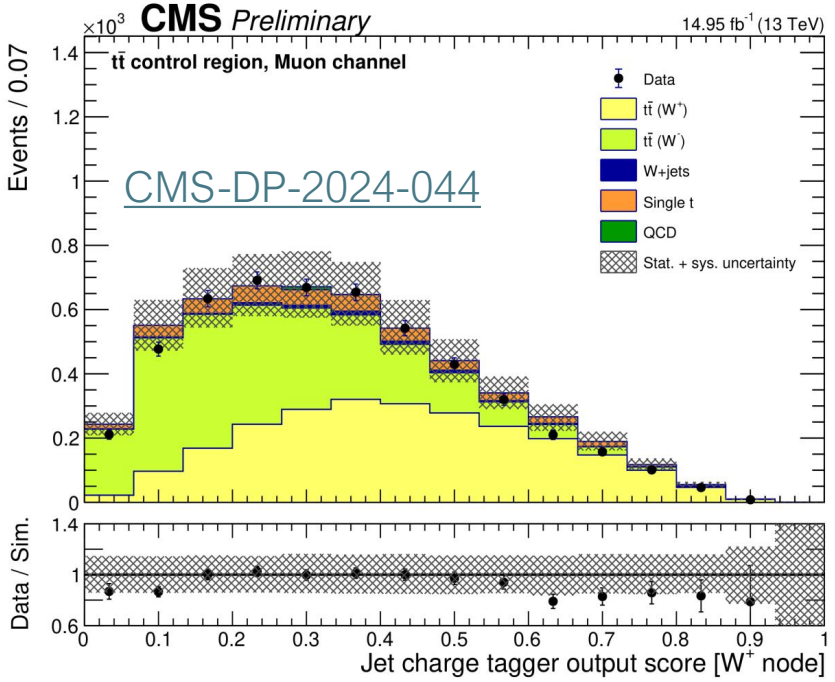
Process	Final state/ prongness	heavy flavour	# of classes
H→VV (full-hadronic)	qqqq	0c/1c/2c	3
	qqq		3
H→WW (semi-leptonic)	eνqq	0c/1c	2
	μνqq		2
	$\tau_e\nu$qq		2
	$\tau_\mu\nu$qq		2
	$\tau_h\nu$qq		2
H→qq		bb	1
		cc	1
		ss	1
		qq (q=u/d)	1
H→ττ	$\tau_e\tau_h$		1
	$\tau_\mu\tau_h$		1
	$\tau_h\tau_h$		1
t→bW (hadronic)	bqq	1b + 0c/1c	2
	bq		2
t→bW (leptonic)	beν	1b	1
	bμν		1
	b$\tau_e\nu$		1
	b$\tau_\mu\nu$		1
	b$\tau_h\nu$		1
QCD		b	1
		bb	1
		c	1
		cc	1
		others (light)	1

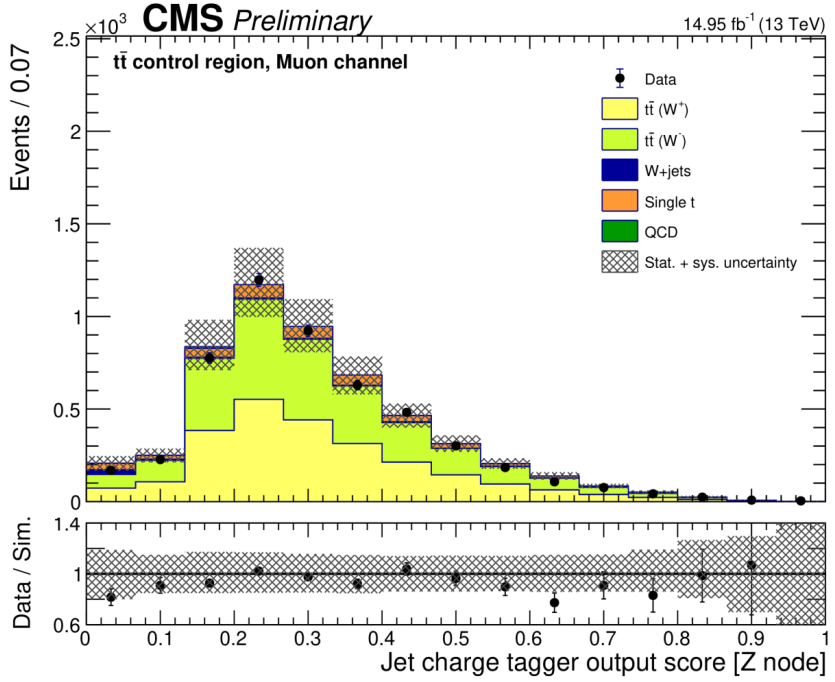


New Jet tagging developments (Jet charge tagger)

- DGCNN based on the ParticleNet architecture predicting the charge of the AK8 jet.
  - Discriminate W^+ and W^- bosons from Z boson.
- Observed good data-MC agreement selecting semileptonic $t\bar{t}$ events.



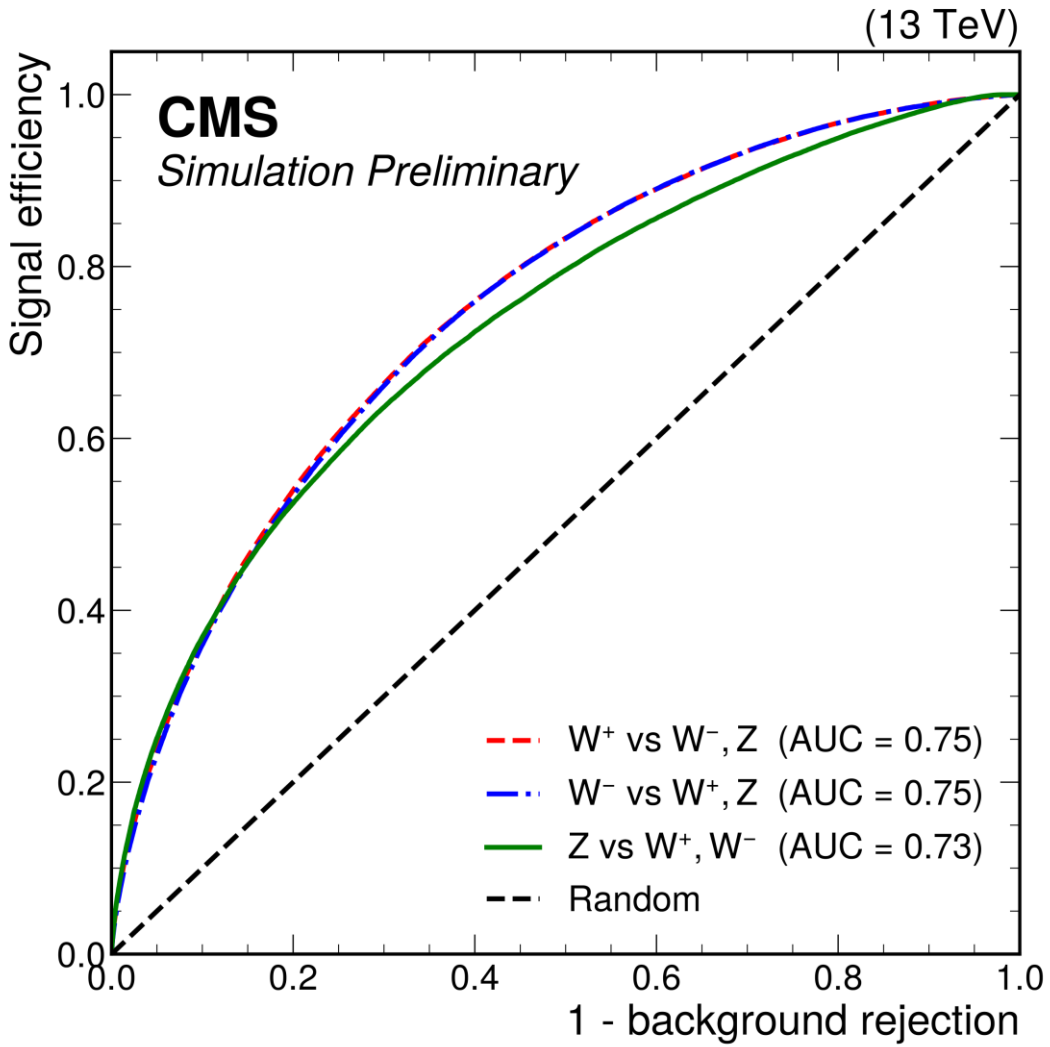




Jet charge tagger

- 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.

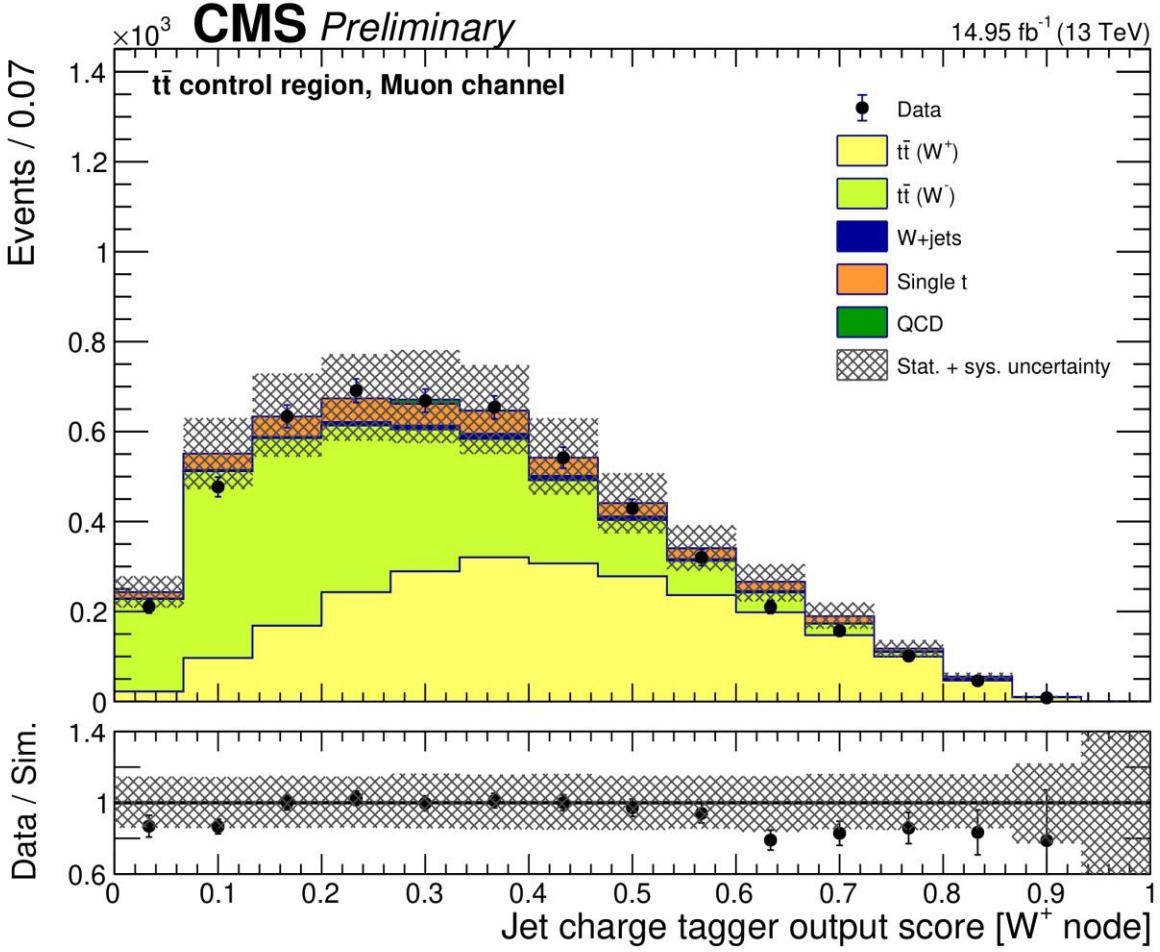
CMS-DP-2024-044



Jet charge tagger

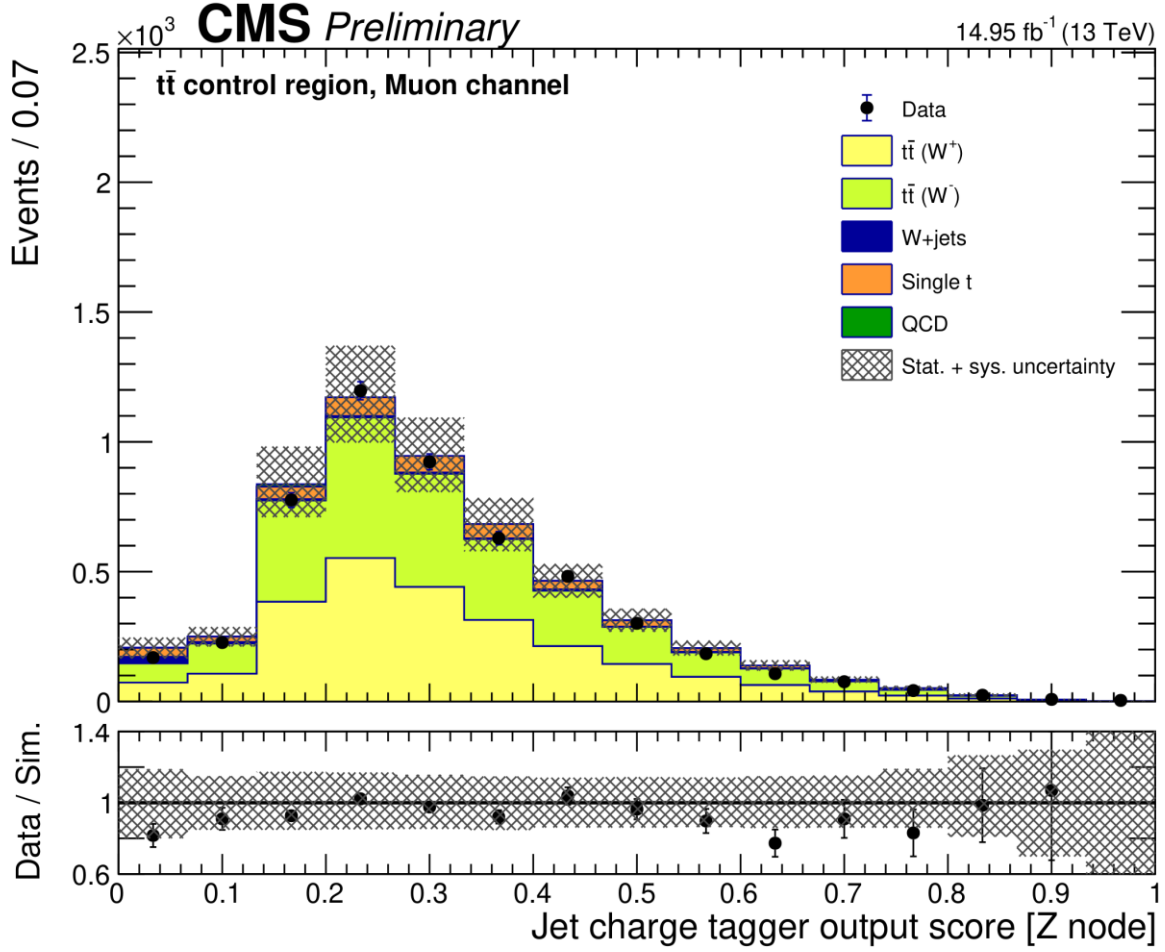
semileptonic $t\bar{t}$ enriched region

output score: W^+



semileptonic $t\bar{t}$ enriched region

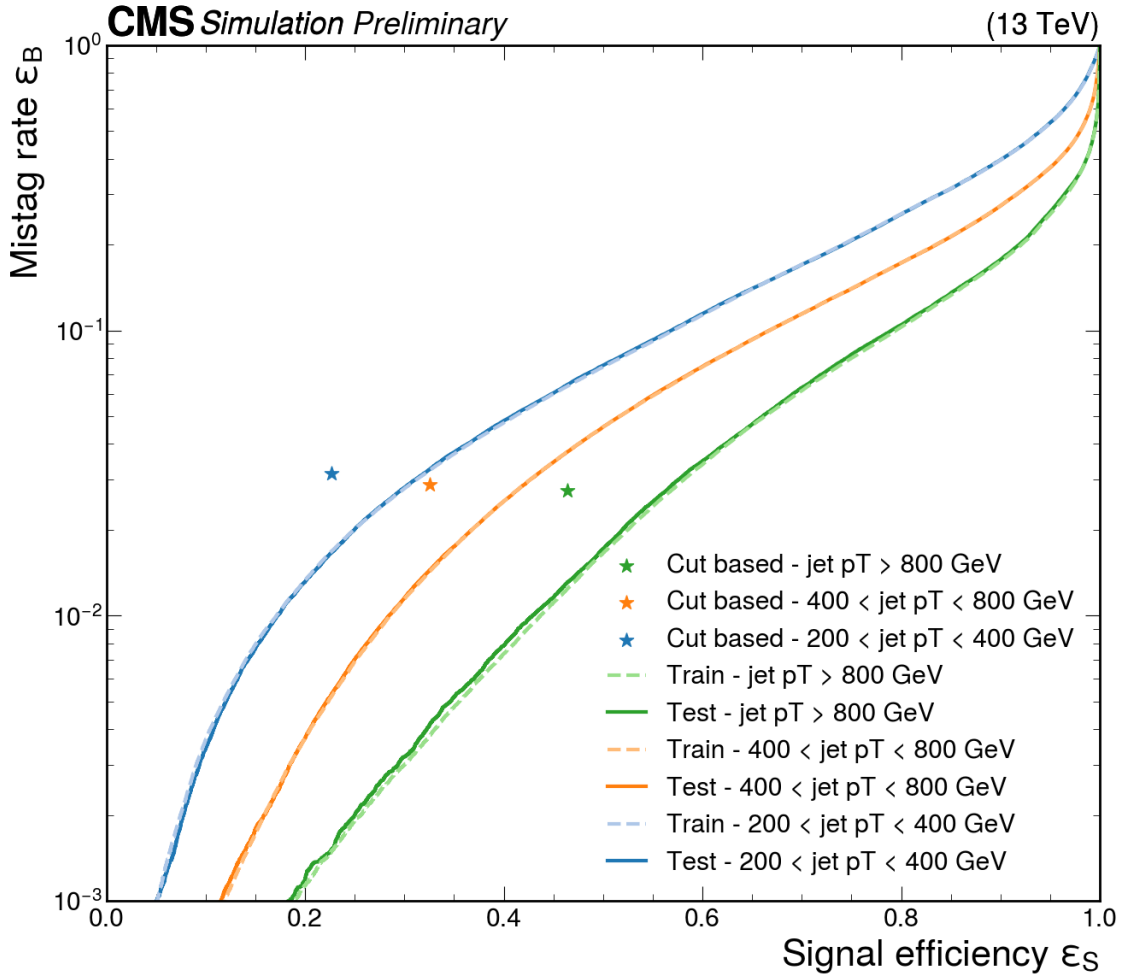
output score: Z (not expected Z bosons)



New Jet tagging developments (HOTVR)

- Heavy Object Tagger with Variable Radius.
  - R = 600 / p_T (R min 0.1).
  - Useful for 4 top final states in the intermediate region where the top quark is neither resolved (p_t < 200 GeV) nor boosted (p_t > 800 GeV).
  - For Run 2 top tagging was performed using a cut-based approach.
  - Developed a BDT, using as inputs the cut variables of the Run 2 analysis, to distinguish top quarks from QCD.
    - Observed improved with respect to the top tagging cut-based.

CMS-DP-2024-038



Top tagging with variable sized jets

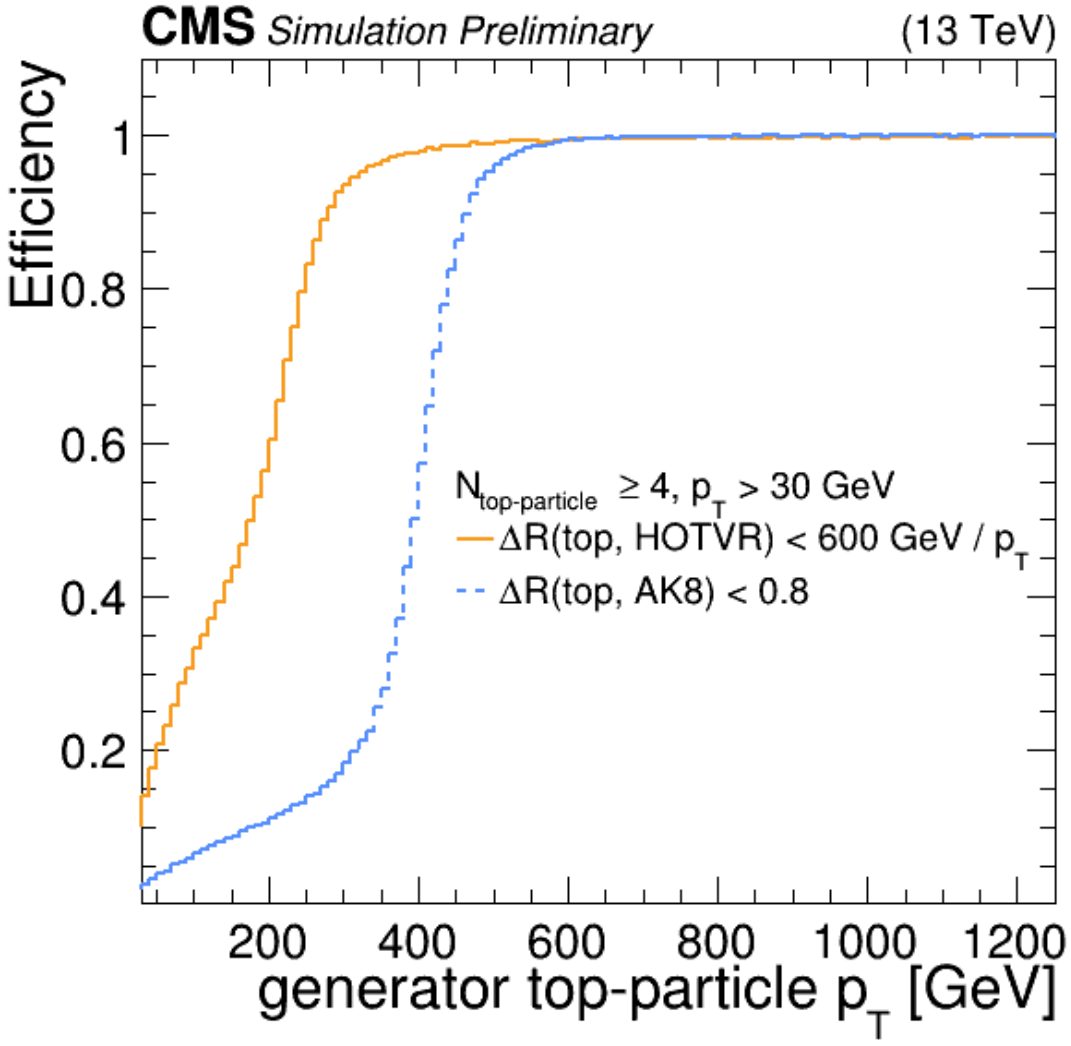
Heavy Object Tagger with Variable Radius

- 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

CMS-DP-2024-038



Top tagging with variable sized jets