Boosted Jet Tagging and Calibration in CMS

Oz Amram

On Behalf of the CMS Collaboration

July 31st, 2023

BOOST 2023

Overview

- Latest boosted jet **tagger** in CMS
- Performance **calibration** in data
- **New** method for calibrating **high-prong** jets
- Won’t cover heavy flavor jets

  → See Congqaio’s talk!

CMS Jet Tagging

Jets : Anti-kt R=0.8, PUPPI

Up to 100 jet constituents (42 feats. per)

Up to 7 secondary vertices (15 feats per.)

CMS Jet Tagging

arXiv:1902.08570
CMS: DP-2020-002

Jets : Anti-kt R=0.8, PUPPI
Up to 100 jet constituents (42 feats. per)
Up to 7 secondary vertices (15 feats per.)

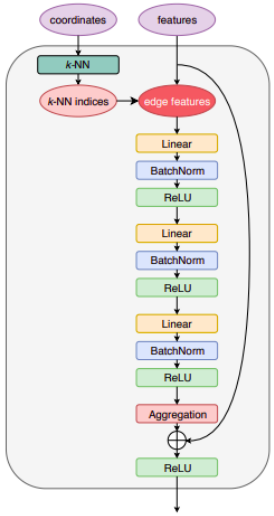
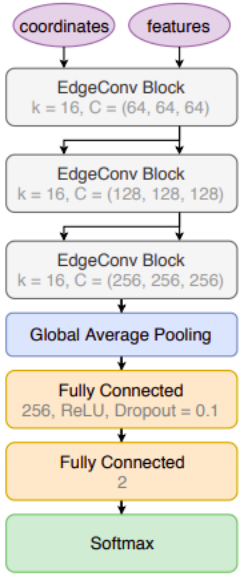


FIG. 1: The structure of the EdgeConv block.



(a) ParticleNet

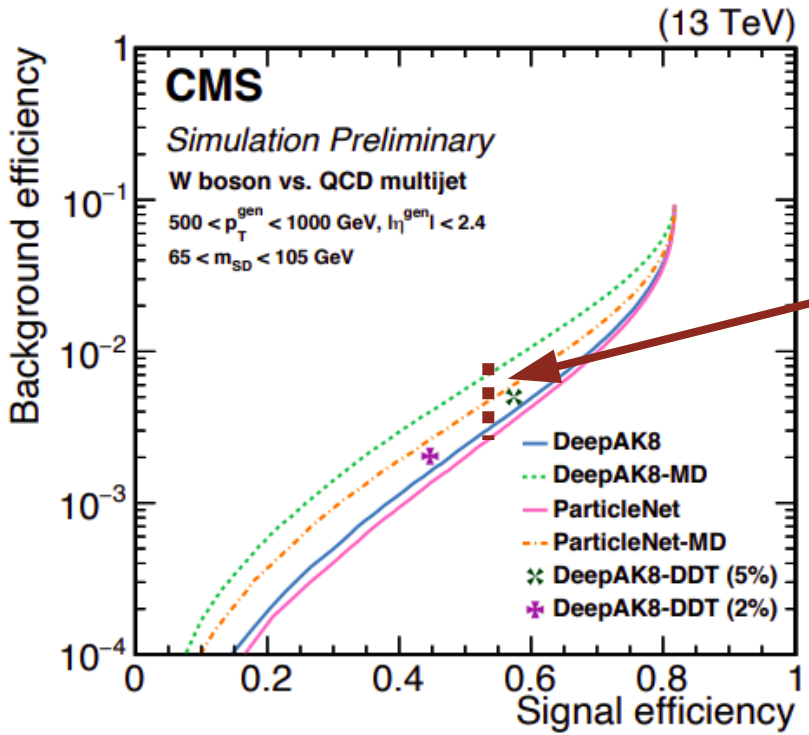
ParticleNet

- Architecture : **Graph based**
  - Processes inputs in permutation invariant way
  - Based on EdgeConv blocks
- Output: binary classification scores
  - X vs QCD

CMS Jet Tagging

arXiv:1902.08570
CMS: DP-2020-002

Jets : Anti-kt R=0.8, PUPPI
Up to 100 jet constituents (42 feats. per)
Up to 7 secondary vertices (15 feats per.)



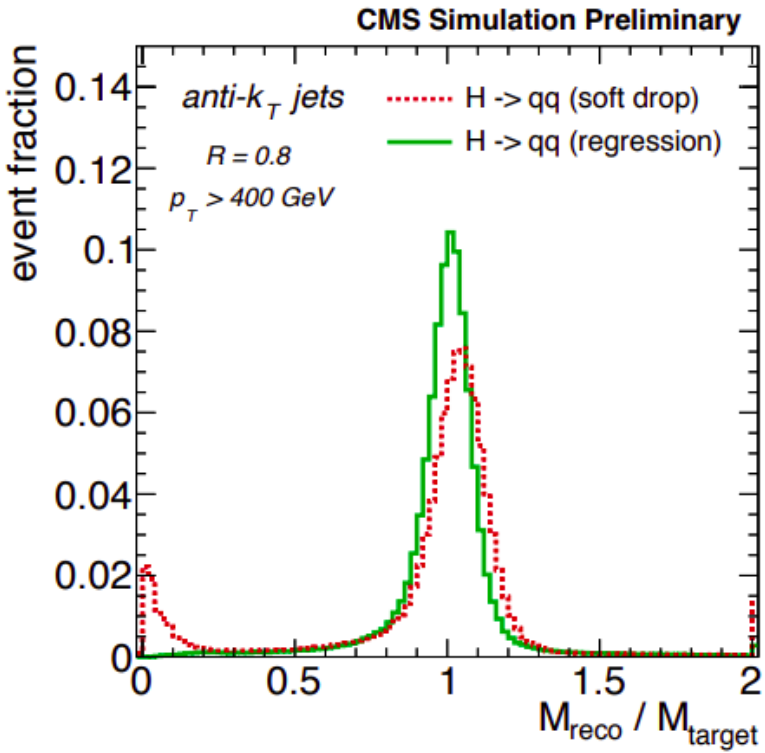
ParticleNet ~2x bkg rejection wrt prev. tagger (**DeepAK8**)

Jet Mass Regression

CMS-DP-2021-017

Training objective :

$$M_{\text{target}} = \begin{cases} M_{\text{SD}}^{\text{gen}} & \text{if jet is QCD} \\ m_{\text{H}} \in [15, 250] & \text{otherwise} \end{cases}$$



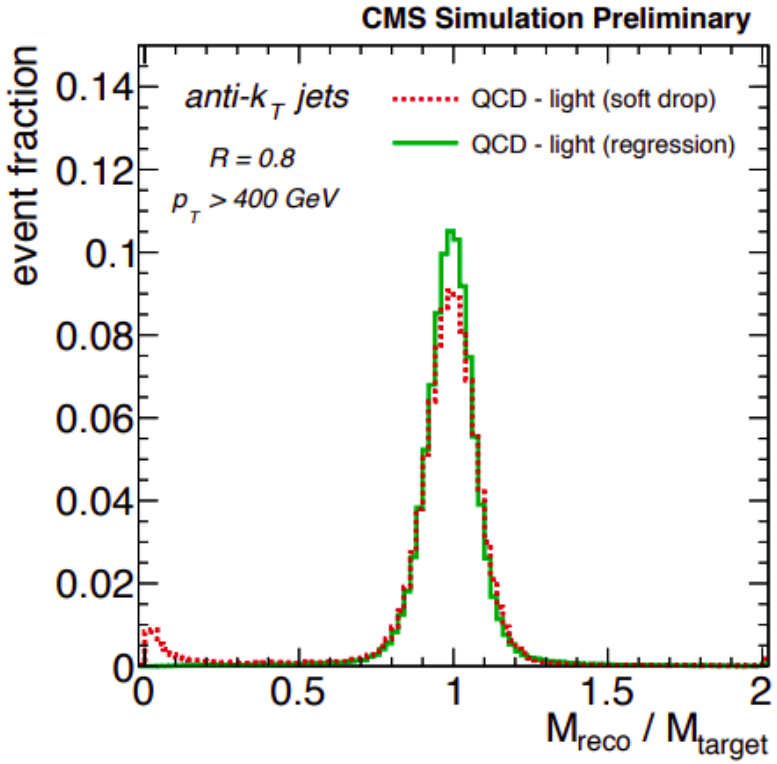
- Same **ParticleNet** architecture used to regress jet mass
- Significant improvement wrt **soft drop** on signal jets

Jet Mass Regression

CMS-DP-2021-017

Training objective :

$$M_{\text{target}} = \begin{cases} M_{\text{SD}}^{\text{gen}} & \text{if jet is QCD} \\ m_{\text{H}} \in [15, 250] & \text{otherwise} \end{cases}$$

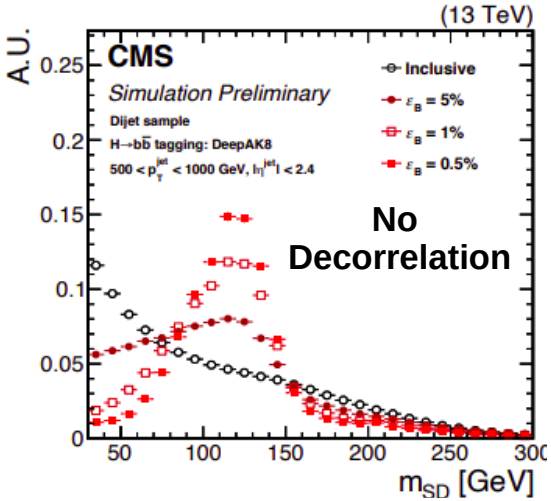


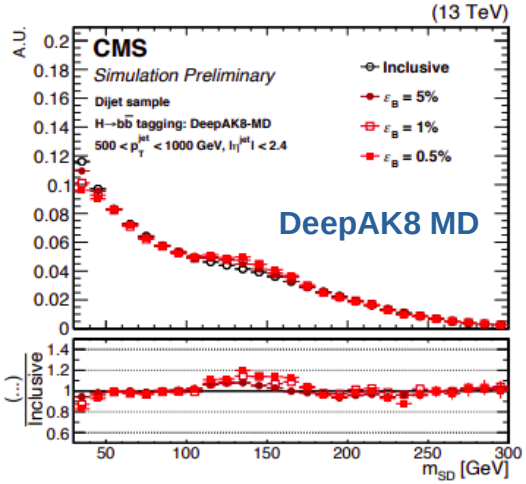
- Same **ParticleNet** architecture used to regress jet mass
- Significant improvement wrt **soft drop** on signal jets
- Doesn't distort shape for QCD jets

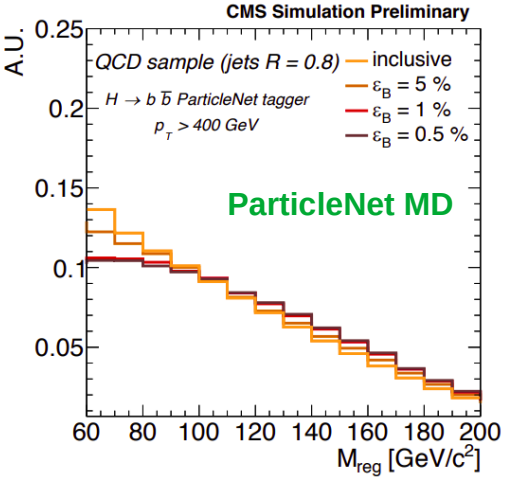
Mass Decorrelation

DP-2020-002
CMS-DP-2021-017

- **Crucial** for analyses doing bump-hunts in jet mass
- **ParticleNet MD** trains on samples generated & reweighted to be flat in m_{SD} & p_T
  - **DeepAK8 MD** used adversarial network
- **ParticleNet MD** achieves slightly better decorrelation on Higgs peak

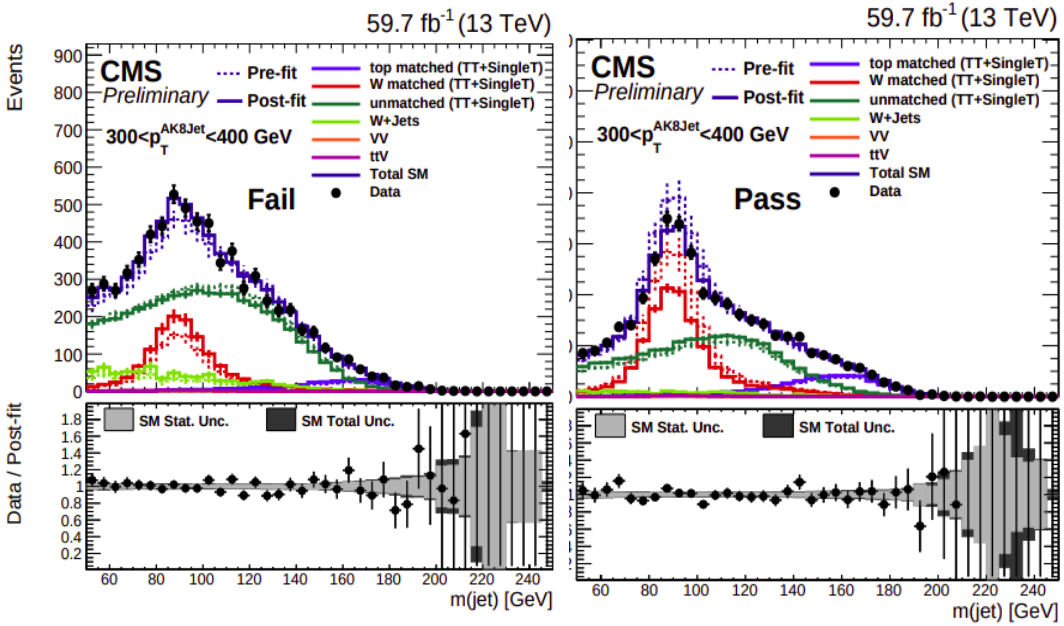






Tagging Efficiency Calibration

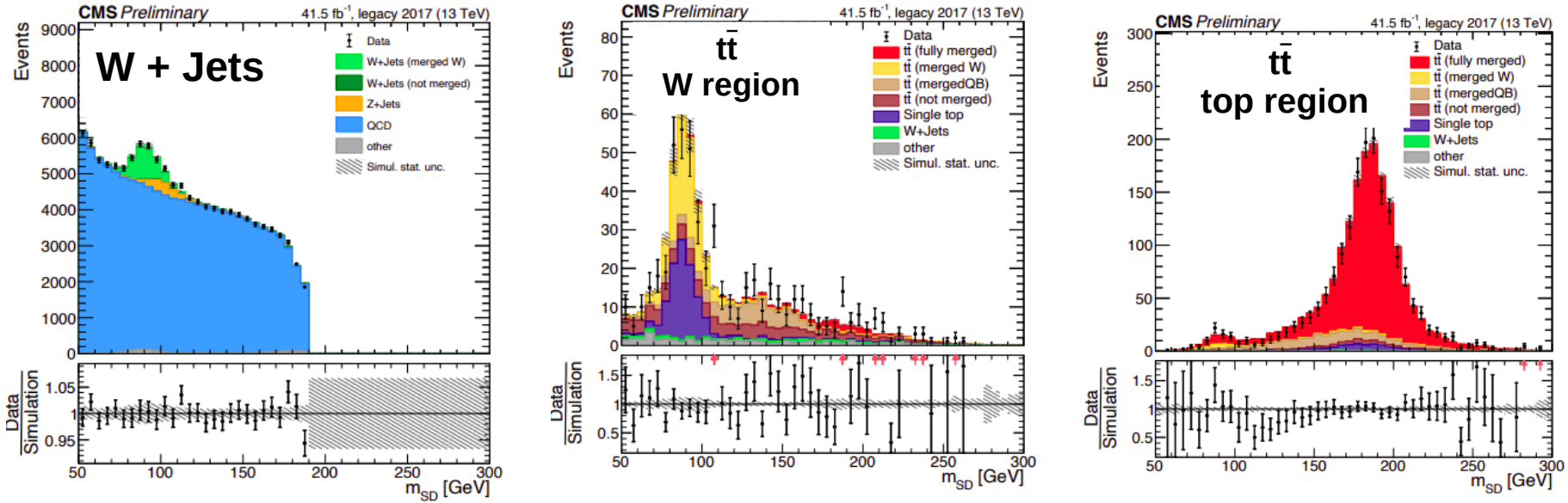
CMS-DP-2020-025



- Calibrating tagging efficiency is crucial for usage by analyzers
- Pick a tagging cut → create pass/fail regions → fit for relative normalization
- Often done in semi-lep. $t\bar{t}$ for clean sample of **W**'s and **top**'s

Jet Mass Scale Calibration

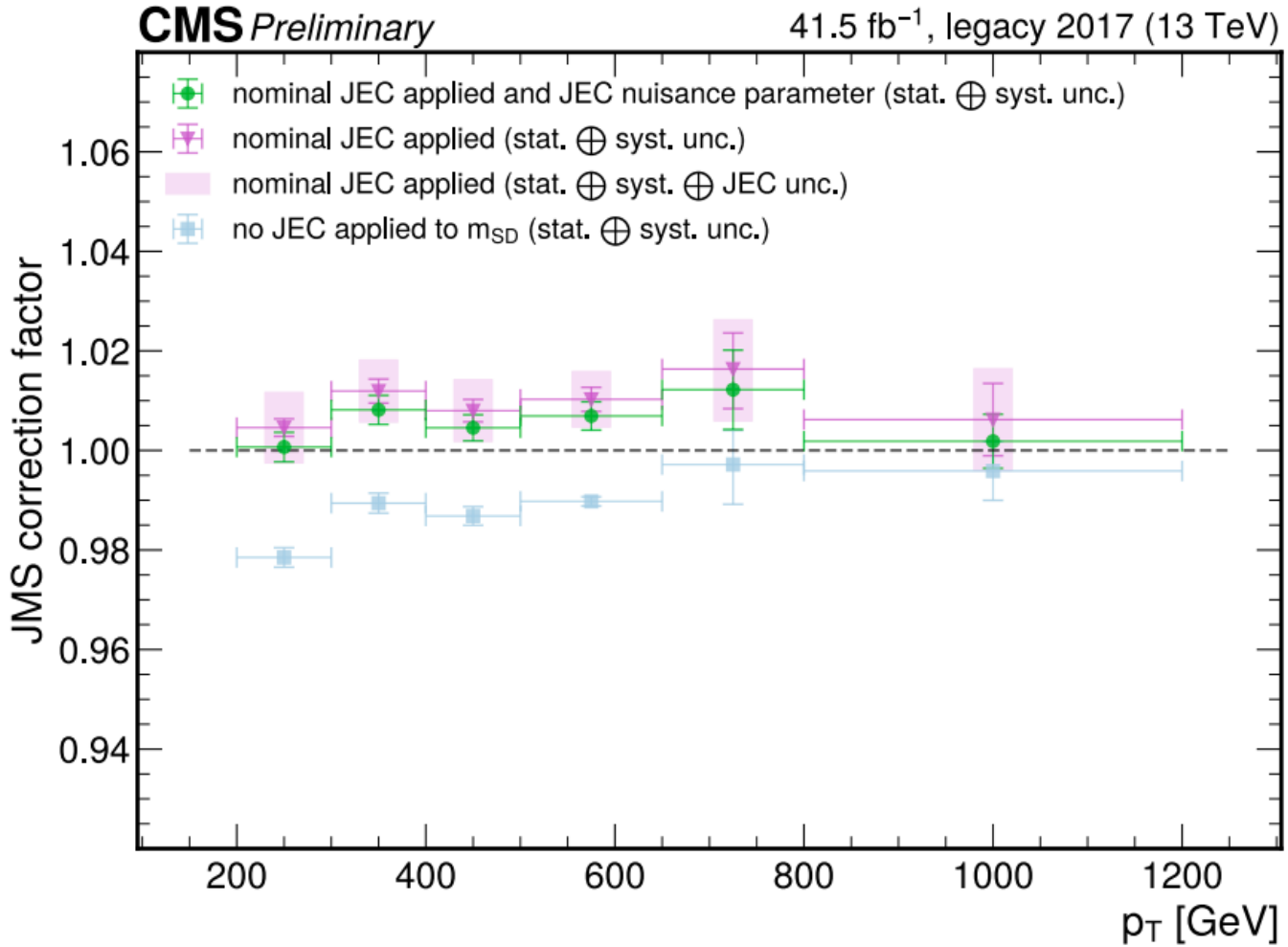
CMS-DP-2023-044



- Use substructure cuts (or ParticleNet) to construct pure samples of W's & top's
- Fit for jet mass scale in data vs sim.

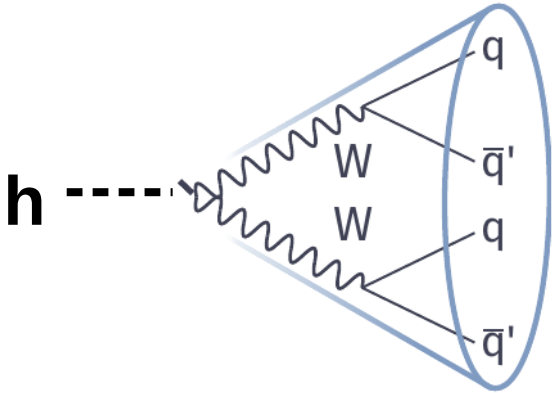
Jet Mass Scale Calibration

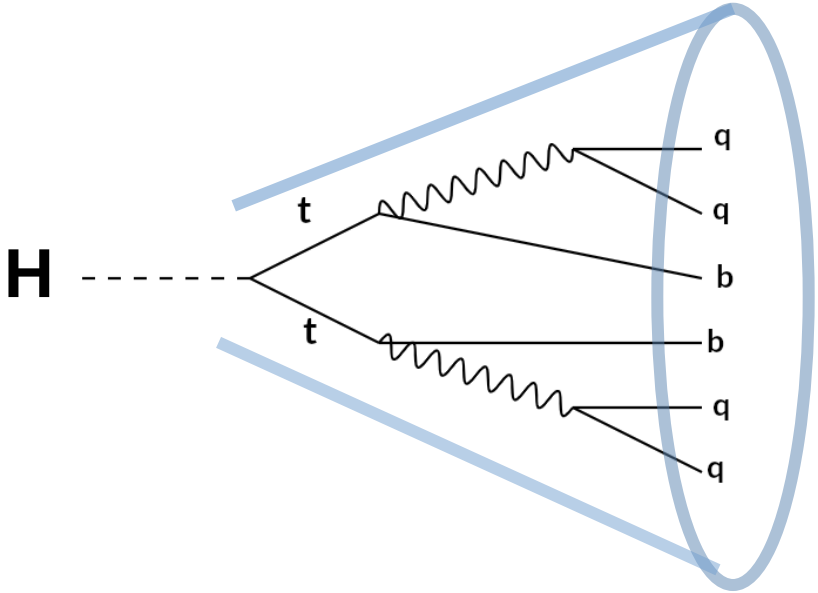
CMS-DP-2023-044



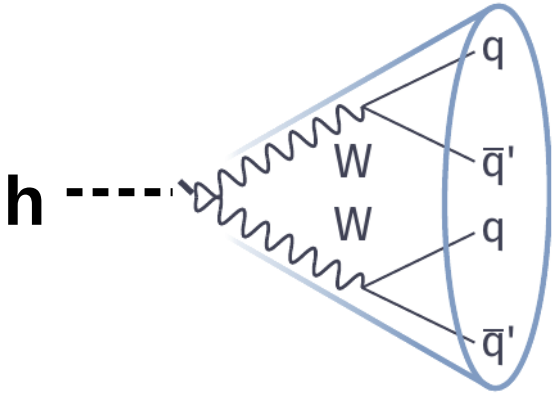
Data vs Sim. JMS agree within **2%** after JEC applied

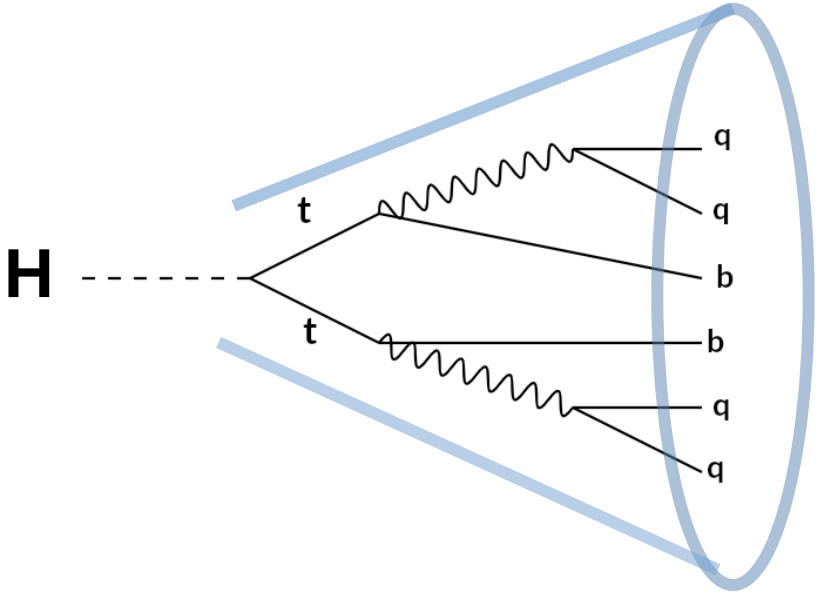
High Prong Jets?





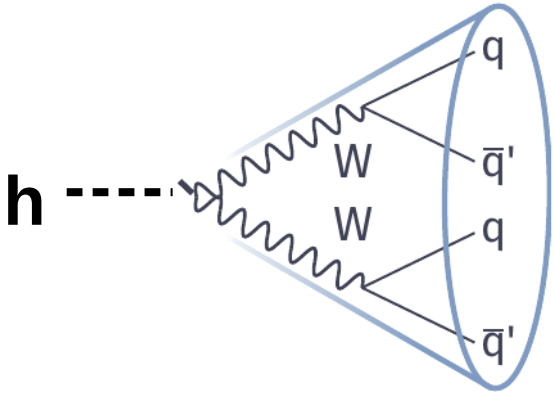
High Prong Jets?

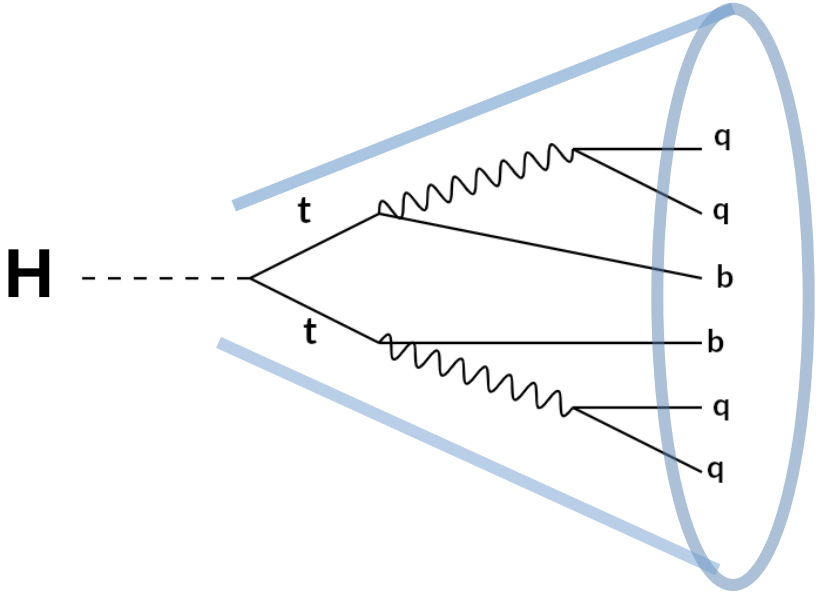




- To calibrate (B)SM jets typically use SM proxy with same number of prongs
- What to do for high prong (>3) jets ?
  - No abundant SM proxies in data...

High Prong Jets?



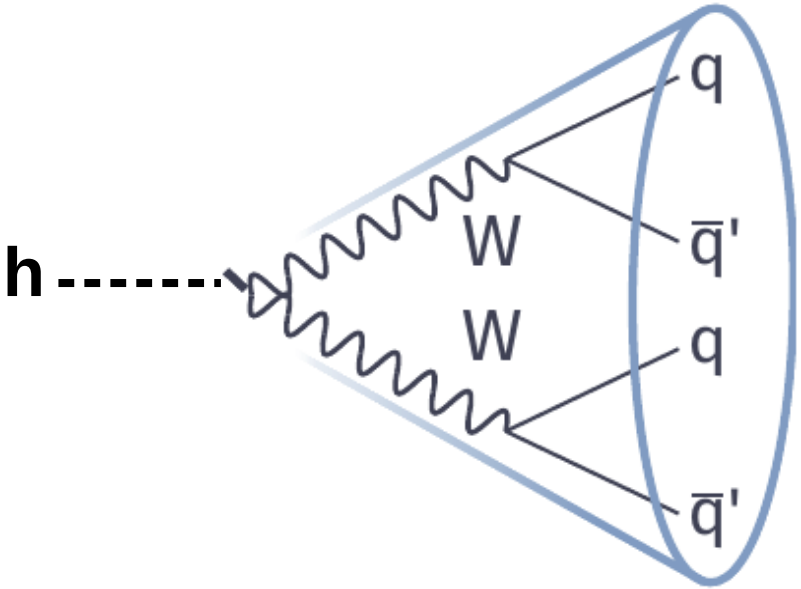


- To calibrate (B)SM jets typically use SM proxy with same number of prongs
- What to do for high prong (≥ 3) jets ?
  - No abundant SM proxies in data...

New Method!
Calibrate each prong separately
via Lund Jet Plane Reweighting

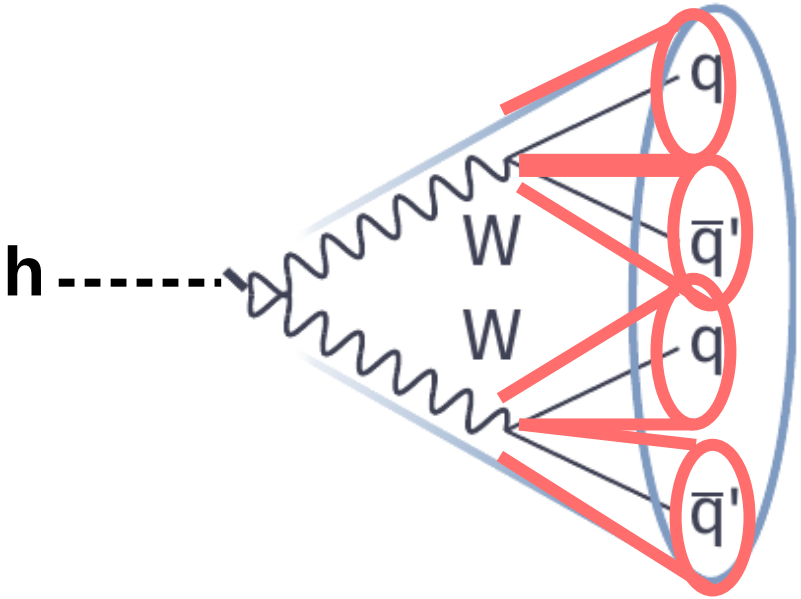
Multi-Prong Calibration Technique

CMS DP-2023/046



Multi-Prong Calibration Technique

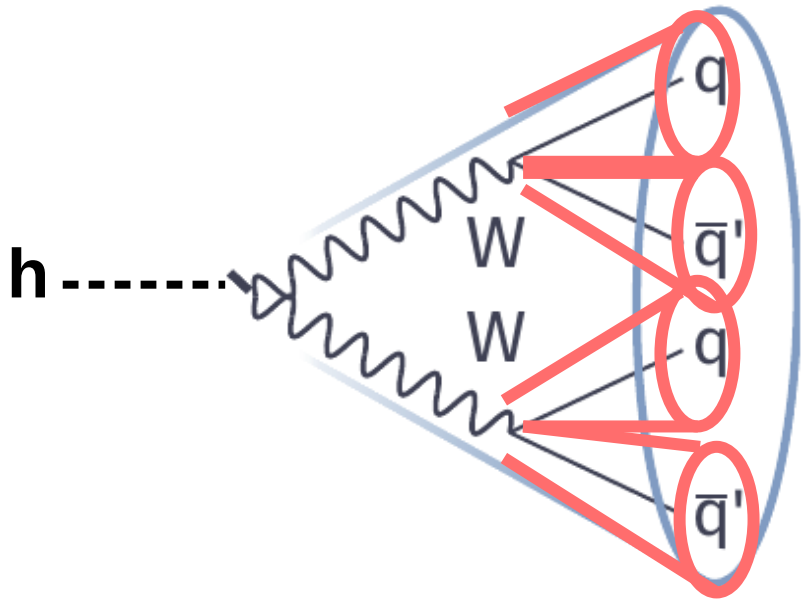
CMS DP-2023/046



- **Recluster** AK8 jet so each prong in a **separate subjet**

Multi-Prong Calibration Technique

CMS DP-2023/046

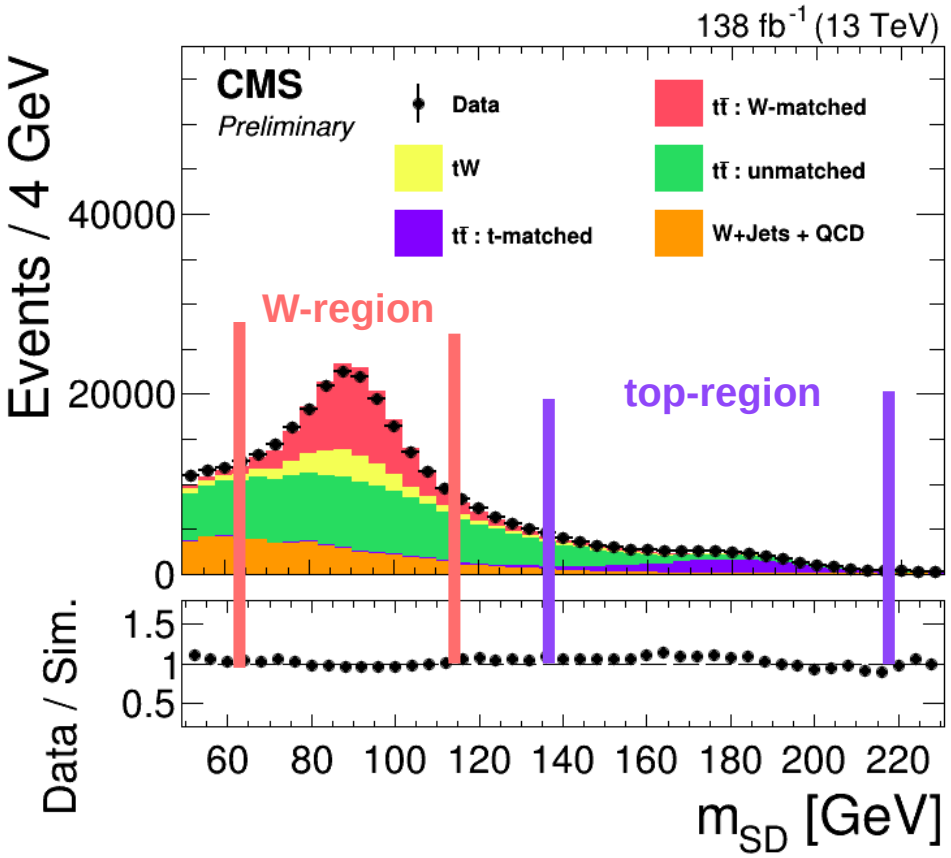


Key assumption : Each prong originates from a SM quark

- **Recluster** AK8 jet so each prong in a **separate subjet**
- Data-driven correction for each **subjet** using the **Lund Jet Plane**
- Correction is ‘per-prong’ so can extrapolate to higher-prong jets!

Semi-lep. $t\bar{t}$

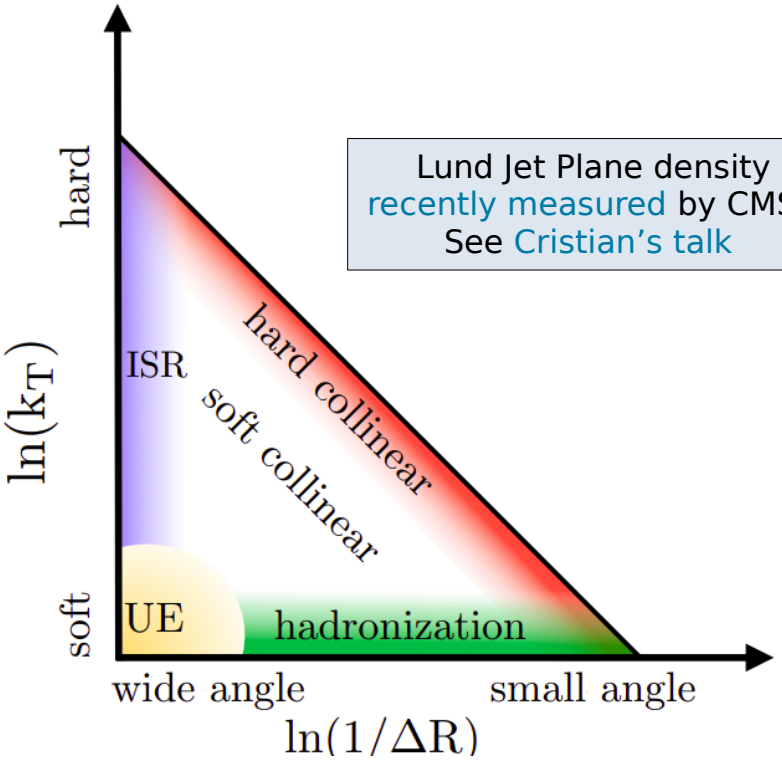
- Extract & test data-driven subjet correction using semi-leptonic $t\bar{t}$ events
- Derive data/sim. ratio of **Lund Jet Plane** from **boosted W's**
- Test calibration on **W's** and **top's**



The Lund Jet Plane (LJP)

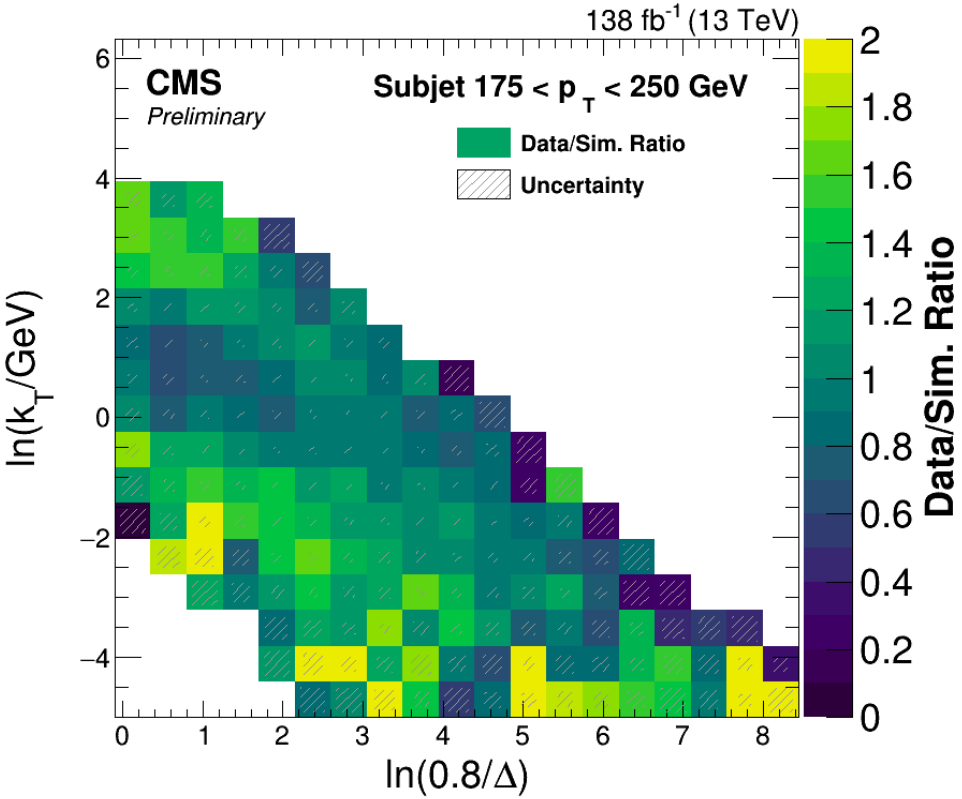
1807.04758

- A 2D representation of the density of splittings inside the jet
- To construct our **subjet Lund Jet Plane**
  - **Recluster** AK8 jet into #prongs using **exclusive kt** algorithm
  - Recluster each subjet using **Cambridge/Aachen** to get splitting history
  - **Fill** points based on **splittings** along hardest branch



Lund Jet Plane density recently measured by CMS! See Cristian's talk

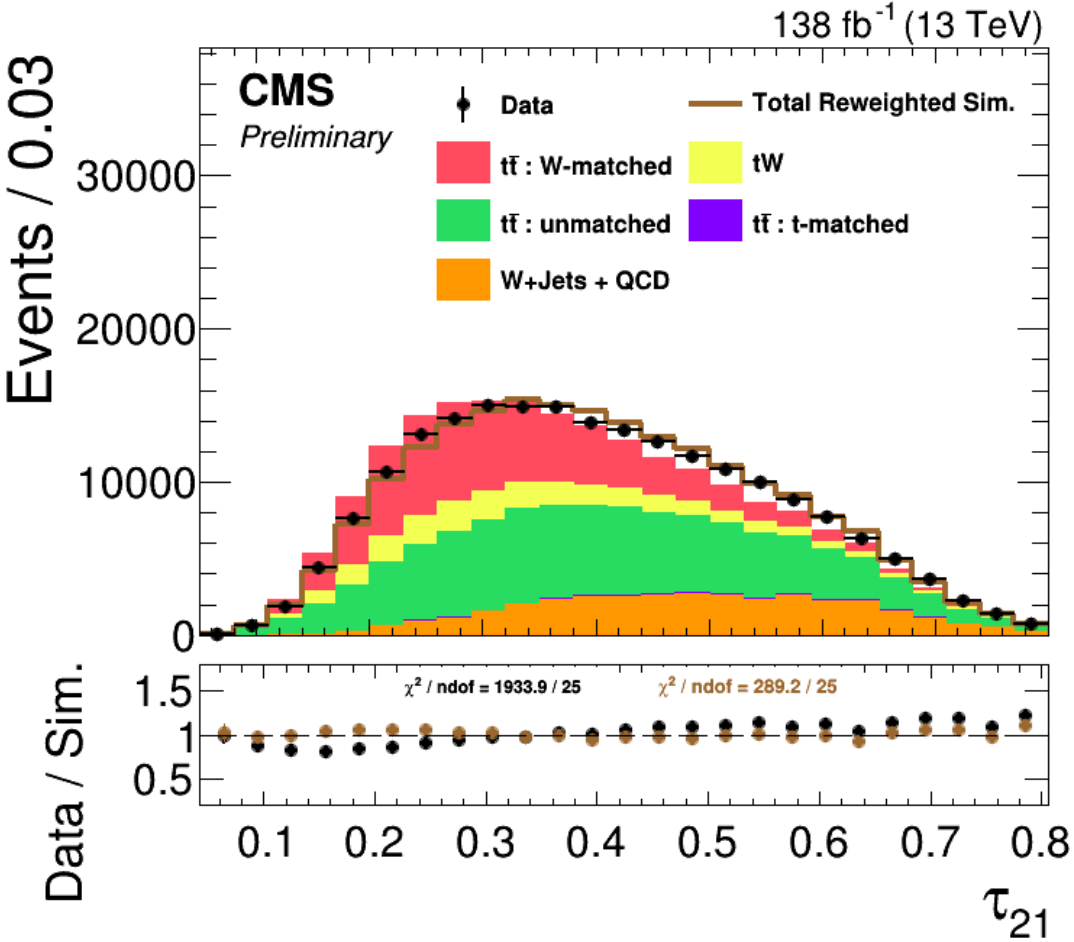
Derivation of Correction



- Recluster AK8 jets from **W-region** into 2 subjets
- Construct **LJP**'s of data and sim. → take ratio
  - Done in 6 bins of subjet p_T
- Use this ratio to **correct simulated jets**
- For each prong, reweight based on the multiplication of the **LJP** ratio of prong's splittings

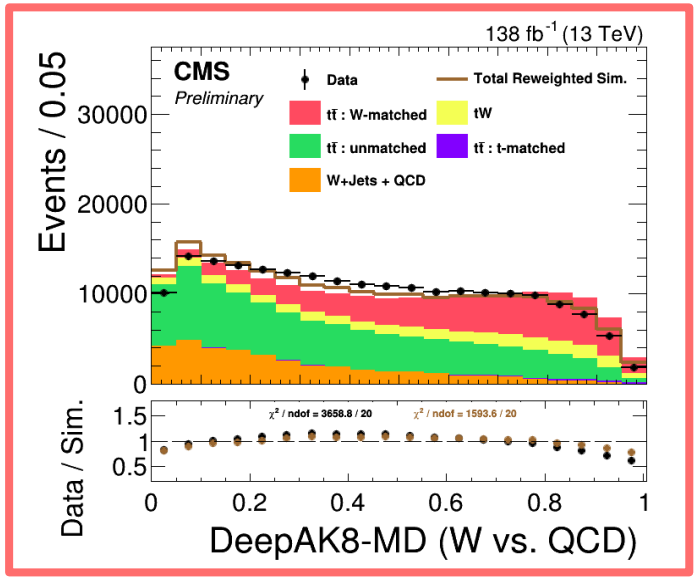
Application to W Jets

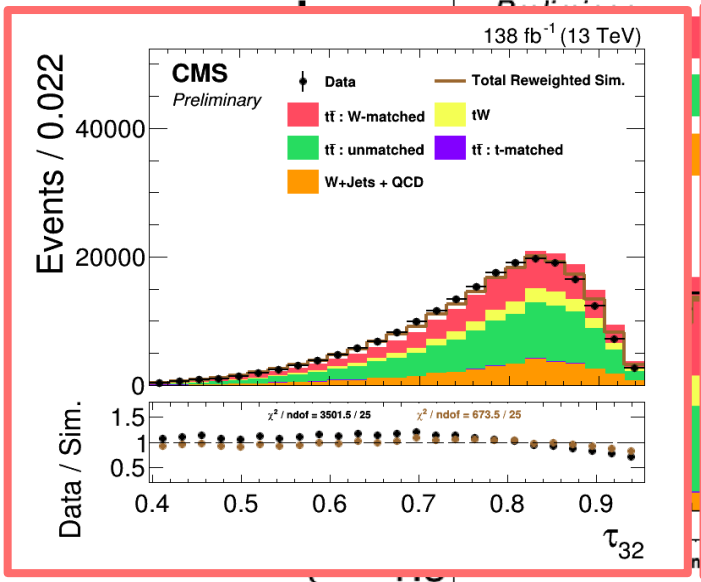
Application of **correction** to **W jets** significantly improves data/sim. agreement!

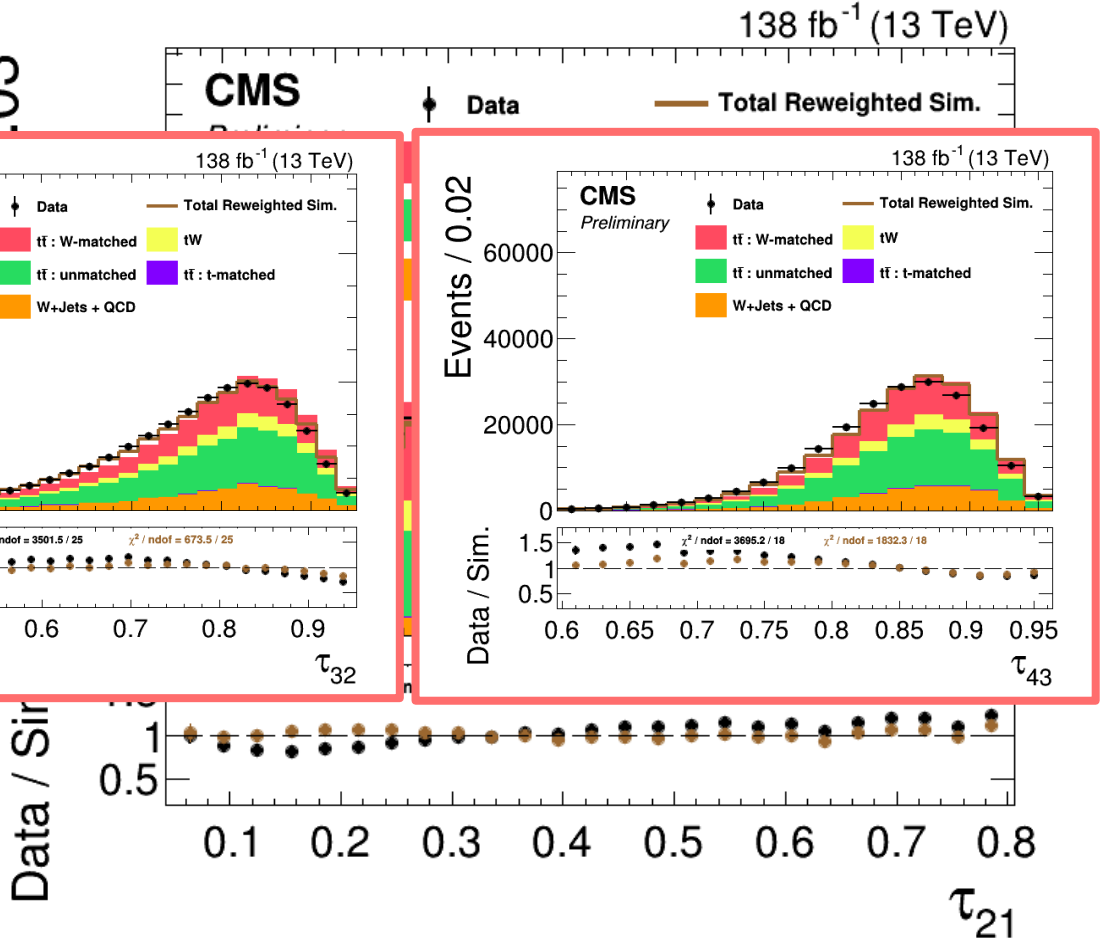


NB: Non-perfect closure b/c bkg processes are not corrected

Application to W Jets







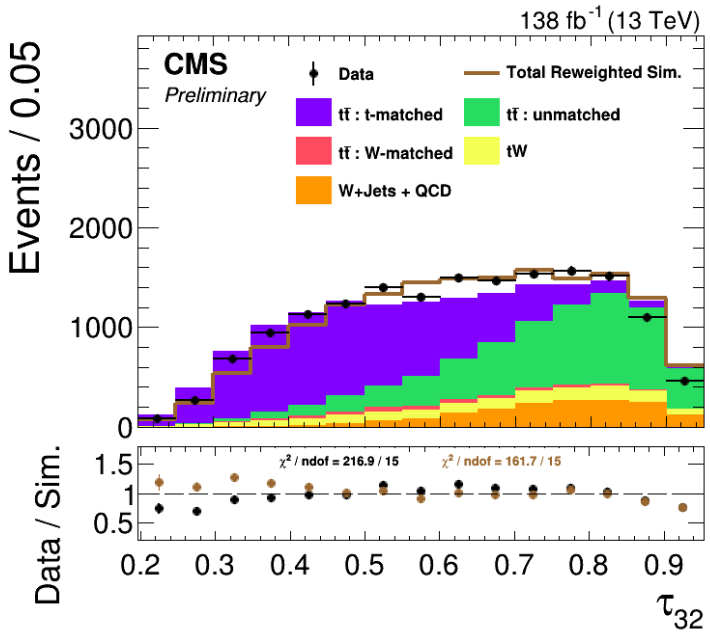
NB: Non-perfect closure b/c bkg processes are not corrected

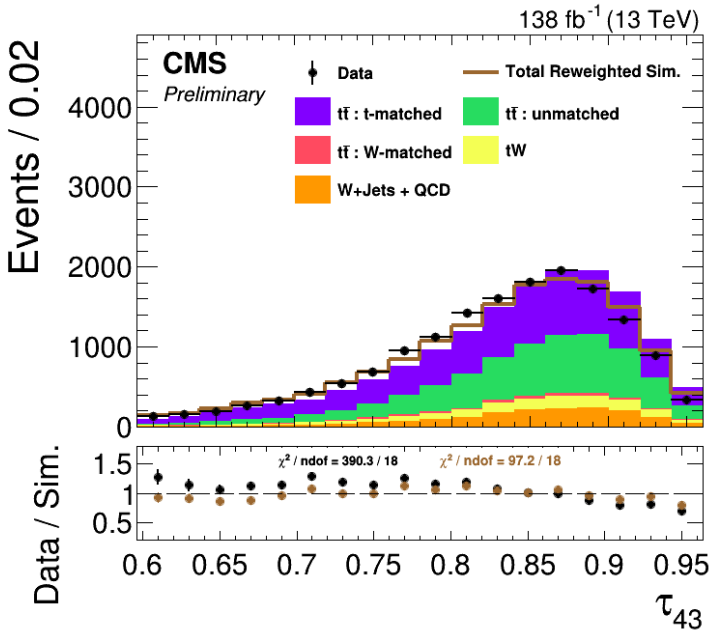
Application to Top Jets

- Recluster **top jets** into 3 subjets
- Apply data/sim **LJP** correction derived from **W**'s

Application to Top Jets

- Recluster **top jets** into 3 subjets
- Apply data/sim **LJP** correction derived from **W**'s





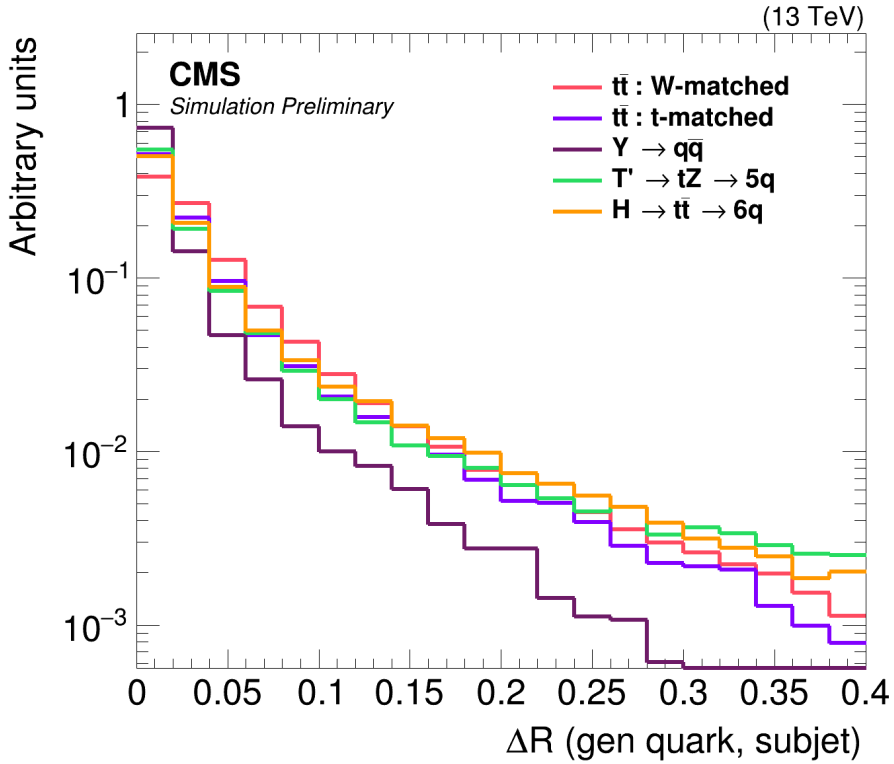
NB: Non-perfect closure b/c bkg processes are not corrected

Correction significantly improves agreement!

Uncertainties

- **Stat.** and **sys.** on extraction of data/sim. **LJP** ratio
- **Matching uncertainty** on how well the reclustered subjets correspond to the quarks from the hard process
  - Largest unc., grows with # of prongs
  - ~5% for 2-prong → 50% for 6-prong
- Minor uncertainties:
  - Extrapolation of correction in **subjet p_T**
  - Differences in showering of **bottom quarks** and light quarks

Reclustering works even for **6 prong** jets!

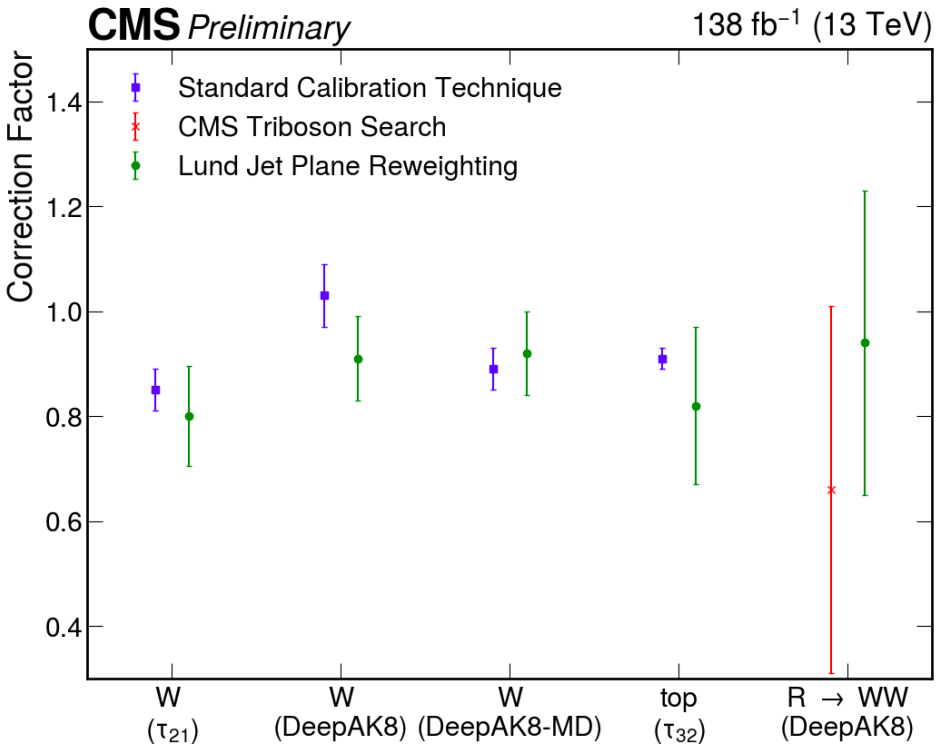


Correction Factor Comparison

- Use method to calibrate tagging efficiency
- Compare correction factor (ε_{data} / ε_{sim}) from **std technique** and **LJP reweighting**

  → Good agreement

- **LJP** has larger uncertainties b/c more general method
  - BUT enables calibration of high prong jets!



Conclusions

- **ParticleNet** tagger now standard in CMS
  - Also used for regressing jet mass
- Proper **calibration** of substructure tagging and jet mass scale crucial
- **New method** to calibrate high prong jets using the **Lund Jet Plane**
- Look out for the usage of these techniques in future CMS searches!

Backup

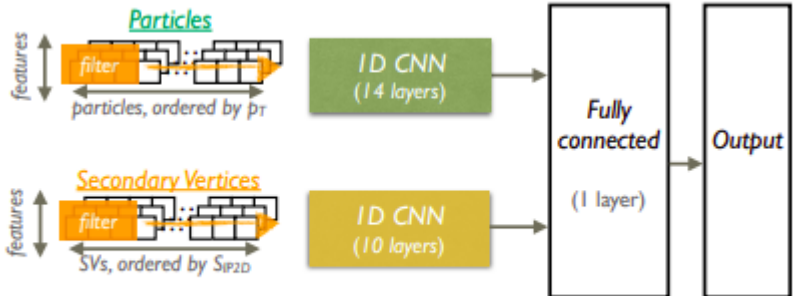
arXiv:2004.08262

Taggers : Deep AK8

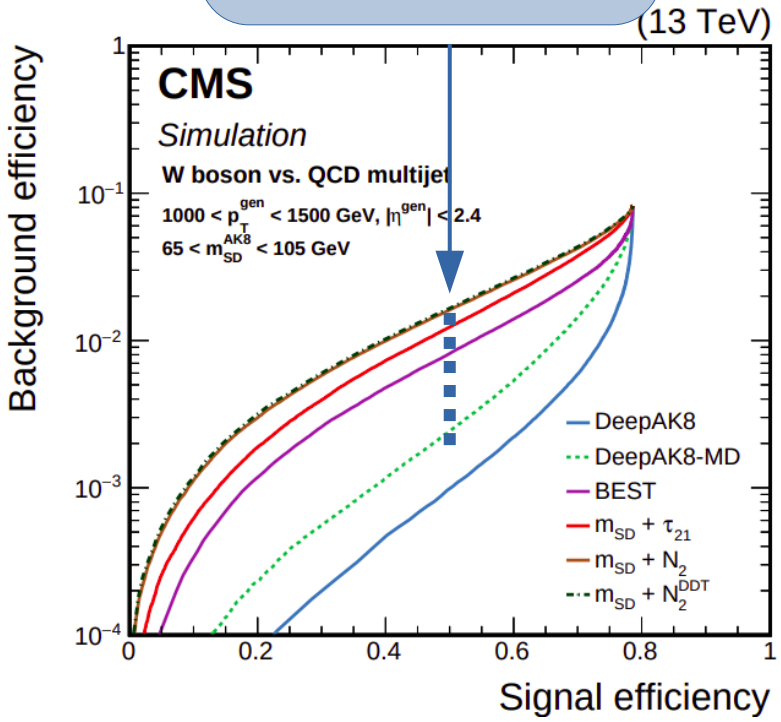
Jets : Anti-kt R=0.8, PUPPI
Up to 100 jet constituents (42 feats. per)
Up to 7 secondary vertices (15 feats per.)

DeepAK8

- Architecture : **1D CNN**'s
  - Order inputs by p_T & 2D IP
- Output: Multi-class scores
  - W/Z/t/H/other, split by decay modes (17 scores)
  - Build discriminants by taking ratios
- Mass-decorrelated version trained with an adversary

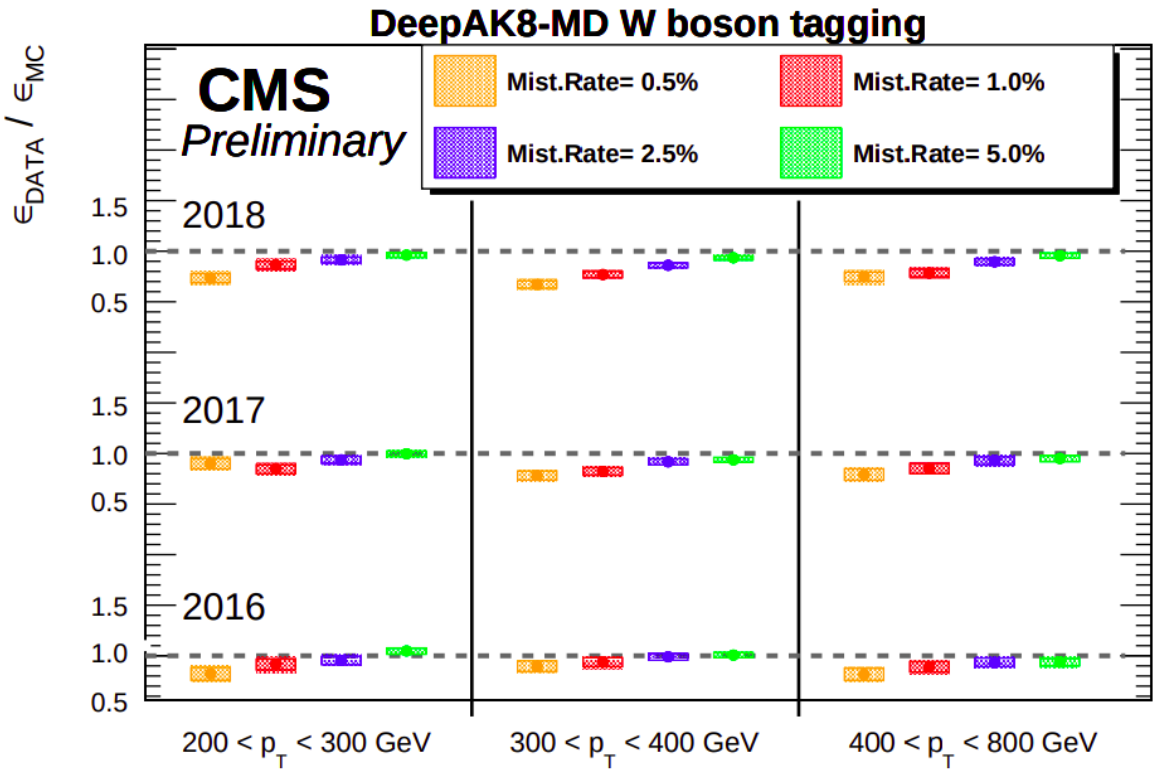


DeepAK8 gives **~10x** improvement in bkg rejection wrt traditional vars



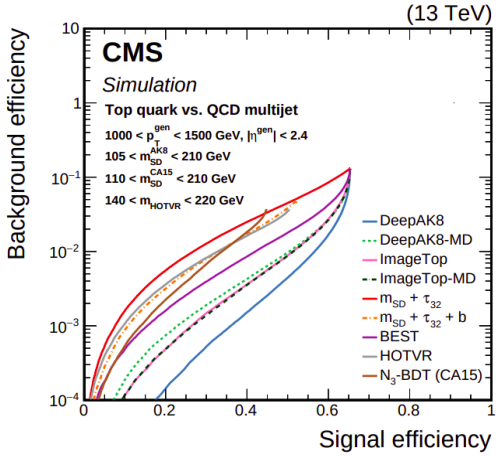
DeepAK8 W-Tagging Calibration

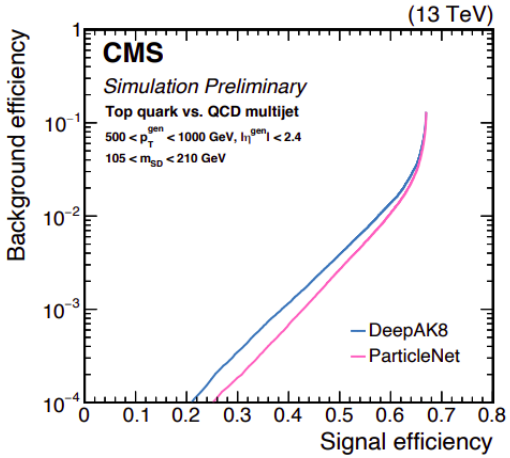
CMS-DP-2020-025

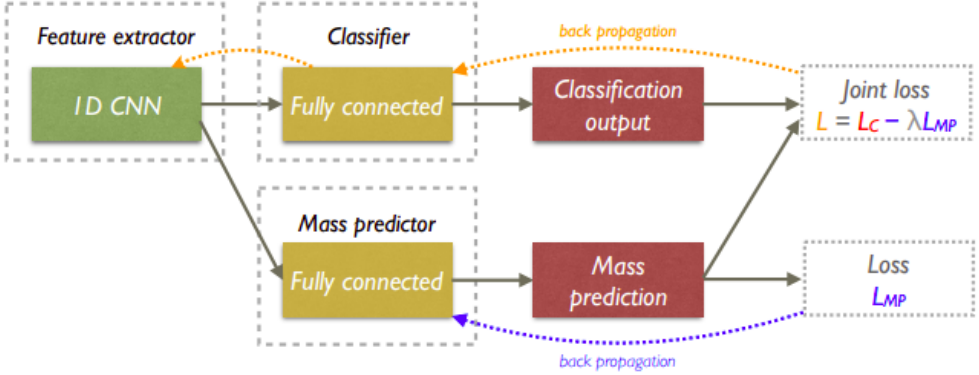


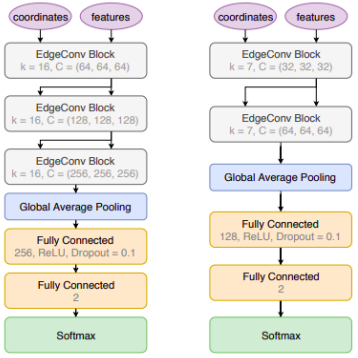
- Precise calibration of tagging efficiency, O(few %)
- Efficiency often significantly different in data vs. MC!

Tagger Backup









(a) ParticleNet (b) ParticleNet-Lite

FIG. 2: The architectures of the ParticleNet and the ParticleNet-Lite networks.

All Data/Sim. Lund Jet Plane Ratios

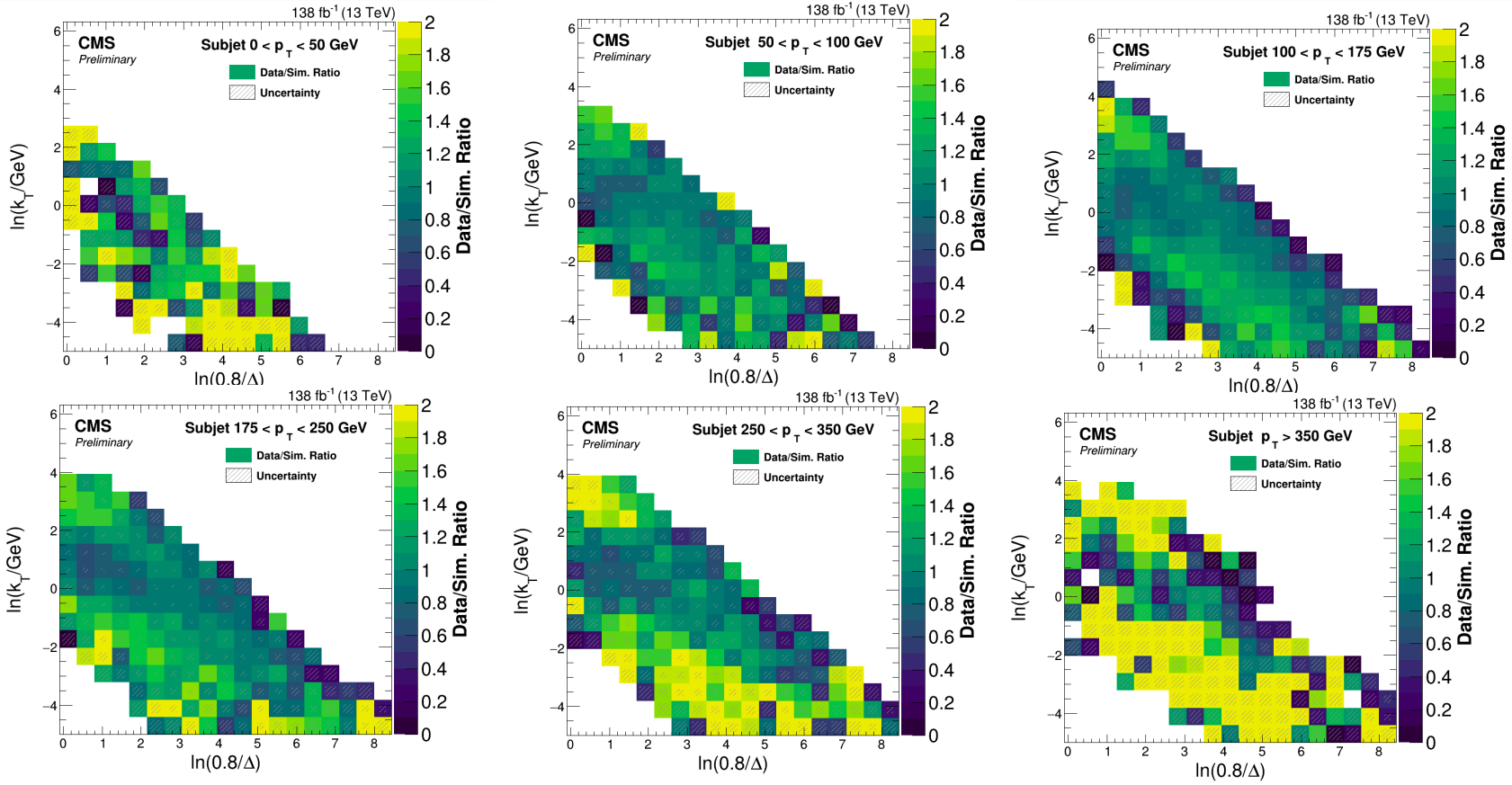


Figure 2: Ratios of the LJP in data and simulation in each subjet p_T bin. The combined statistical and systematic uncertainty on the ratio is represented by the area of the hatched region in each bin. Bins with no data or simulation events are shown as white, but assumed to have a ratio value of unity and 100% uncertainty.

Table of Uncertainties

Jet Type (# Prongs)	Selection Var.	Ratio Stat. Unc.	Ratio Sys. Unc.	p_T Extrap. Unc.	b Unc.	Matching Unc.	Tot. Unc.
W (2)	τ_{21}	0.03	0.04	-	-	0.07	0.09
W (2)	DeepAK8	0.01	0.02	-	-	0.08	0.08
W (2)	DeepAK8_MD	0.01	0.02	-	-	0.08	0.08
top (3)	τ_{32}	0.03	0.06	0.01	0.01	0.14	0.16
$Y \to qq$ (2)	τ_{21}	0.01	0.12	0.04	-	0.04	0.13
$R \to WW$ (4)	Deep-WH	0.02	0.02	0.06	-	0.28	0.29
$T' \to tZ$ (5)	τ_{43}	0.02	0.09	0.02	0.01	0.37	0.38
$H \to tt$ (6)	τ_{43}	0.02	0.05	0.01	0.01	0.50	0.50

Table 1: Uncertainties on the LJP calibration of the efficiency for tagging jets of various kinds using various substructure variables. For the tagging of W jets, both τ_{21} and two versions of the DeepAK8 discriminant (mass decorrelated and not) are evaluated. The matching uncertainty is seen to be dominant, and grows for jets with higher numbers of prongs due to denser environment leading to less well separated subjets.

Correction Factors

- To further validate the method, the LJP correction procedure is applied to various jet types and used to compute a correction factor. The CF is defined as the ratio of the efficiency in the corrected simulation divided by the efficiency of the nominal simulation.
- For tagging of W and top jets, the LJP CFs are compared to those measured in data using standard methods [7,10]. The LJP CFs are found to agree with the standard ones within uncertainties.
- CFs are also derived for higher prong jets from the decays of beyond standard model particles as example use cases.

Correction Factor Table

Jet Type (# Prongs)	Selection Variable	Lund Jet Plane	Comparison
W (2)	τ_{21}	$0.80 \pm 0.03 \pm 0.08$	0.85 ± 0.04
W (2)	DeepAK8 W-score	$0.91 \pm 0.01 \pm 0.08$	1.03 ± 0.06
W (2)	DeepAK8_MD W-score	$0.92 \pm 0.01 \pm 0.08$	0.89 ± 0.04
top (3)	τ_{32}	$0.82 \pm 0.03 \pm 0.15$	0.91 ± 0.02
$Y \to qq$ (2)	τ_{21}	$0.81 \pm 0.01 \pm 0.13$	-
R $\to$ WW (4)	Deep-WH	$0.94 \pm 0.02 \pm 0.29$	0.66 ± 0.35
$T' \to tZ$ (5)	τ_{43}	$0.82 \pm 0.02 \pm 0.37$	-
$H \to tt$ (6)	τ_{43}	$0.84 \pm 0.02 \pm 0.50$	-

Table 2: A comparison of CFs derived using the LJP correction procedure and other sources. The first uncertainty listed with the LJP value is statistical and the second is systematic. For W and top tagging, CFs derived with the LJP have larger uncertainties but agree well with those from traditional methods. For the R$\to$WW SF, the LJP correction factor is compared to the one obtained in a recent CMS search [11] which utilized top quarks with a hard gluon emission as a 4-prong proxy. No comparison correction factors exist for the higher prong signals.

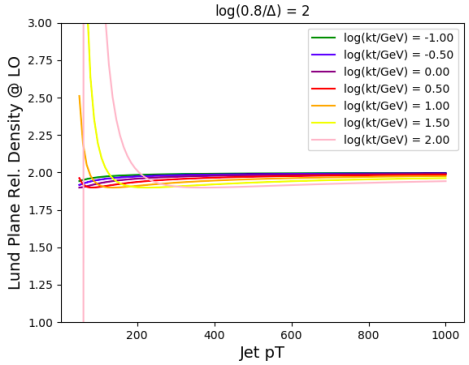
Uncertainties

- **Statistical uncertainty**: the limited number of data events in each bin of the LJP ratio. This uncertainty is propagated to the final correction.
- **Systematic uncertainty**: uncertainties in the efficiencies of the muon and jet selections used, and the modeling of the background processes in the W-region.
- **Matching uncertainty**: How well the reclustered subjets correspond to the quarks from the hard process. This uncertainty is evaluated in simulation based on how often the reclustered subjets fail a $\Delta R < 0.2$ match to a quark from the hard process, how often two quarks are matched to the same subjet and how often a quark is only partially contained in the AK8 jet.
- **High p_T extrapolation**: Events in the W-region contained limited numbers of high p_T subjets. The application of the correction to a higher p_T subjet therefore requires an extrapolation to the correction as a function of subjet p_T. The correction factor in each bin of the LJP is fit as a function of subjet p_T. This extrapolation is used for subjets with $p_T >$ 350 GeV, and the uncertainty of th fit is propagated to the correction.
- **b-jet uncertainty**: Subjets originating from bottom quarks may shower differently than those of light quarks due to the larger bottom quark mass. This difference is assessed by comparing the difference between the LJP of b quarks and light quarks in simulation. This difference is taken as an uncertainty when the correction procedure is applied to b-quark initiated subjets.

Subjet p_T Extrapolation

- W-jets we derive LPR from are in a limited range of p_T → need to extrapolate for some signals
- For splittings with $k_T << p_T$, expect LP density is indep. of pt
  - Corrections scale as $z=k_T/(\Delta\ p_T)$, ie $1/\ p_T$
- For each bin of LPR, fit dependence vs **1/subjet p_T**
- F-test to determine order of fit
  - Almost all bins prefer constant order, few prefer linear
- For subjets with pt > 350, fitted functions are used instead of directly meaured LPR
  - Any bins without a measured data/MC ratio, assume a LPR value of 1 with 100% uncertainty

Theoretical Scaling
(LO QCD)



P_T Extrapolation Theory

- From theory, we expect the extrapolation in subjet pt to be well behaved
- At LO in QCD, the density in a Lund Plane bin (Δ, k_t) is given by :

$$\bar{z} = \frac{k_t}{p_{t,\text{jet}}\Delta}$$

(source)

$$\rho(\Delta, k_t) = \frac{1}{N_{\text{jet}}} \frac{dn_{\text{emission}}}{d\ln(k_t)d\ln(1/\Delta)} = \frac{\alpha_s(k_T)C_F}{\pi}\bar{z}(p_{gq}(\bar{z}) + p_{gq}(1-\bar{z}))$$

$$p_{gq}(\bar{z}) = \frac{1+(1-\bar{z})^2}{\bar{z}}$$

- All the dependence on jet pt comes from the $\bar{z}$ part
- For $\bar{z} << 1$, ie soft/large angle splittings and/or high jet p_T, the $\bar{z}$ dependence is : $2 - z + 2z^2 + 2z^3 + O(z^4)$
  - Ie it asymptotes to 2