Available on CMS information server

CMS DP -2023/046

The Compact Muon Solenoid Experiment

CMS Performance Note

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

24 July 2023 (v3, 25 July 2023)

Lund Plane Reweighting for Jet Substructure Correction

CMS Collaboration

Abstract

Many searches for new phenomena at the LHC utilize boosted large radius jets which contain the showers of multiple partons from the hard process. Such jets can be distinguished from those produced by the shower of a single parton due to their multiple prongs of energy. Due to limitations of the simulation of these jets, they are typically calibrated using similar standard model proxies in available in data. However, this approach cannot be straightforwardly applied to jets with a large number of prongs because no standard model process produces jets with more than three prongs in sufficient quantities. This note presents a new technique for correcting the substructure modeling of boosted, multi-prong, large radius jets which can be applied to jets with large numbers of prongs. Data corresponding to an integrated luminosity of 138 fb^{-1} collected by the CMS experiment between 2016-2018 at a center of mass energy of 13 TeV are analyzed to develop the technique and evaluate its performance. The technique is based on reclustering the jet's constituents into several subjets, such that each subjet captures a single prong, and separately correcting the radiation pattern of each subjet using a data-driven correction in the Lund jet plane. It is shown that the correction procedure improves the agreement between data and simulation in multiple different substructure observables of two and three pronged jets. This technique will enable future searches for high prong jets which were not possible with prior methods.

Lund Plane Reweighting for Jet Substructure Correction

The CMS Collaboration

cms-pog-conveners-jetmet@cern.ch

Glossary

- **AK4/AK8 jets**: jets clustered with the anti-kT algorithm [1] with a distance parameter of R=0.4/0.8. The AK4 jets use the charged hadron subtraction method for pileup mitigation and the AK8 jets use the PUPPI algorithm [12,13].
- **Lund Jet Plane (LJP)**: A two dimensional representation of the density of splittings inside the showering process of the jet. First introduced in [2]. This analysis makes use of the 'primary' LJP which includes the splittings along the hardest branch of the clustering history.
- **Prongs** : Cores of energy inside a large radius jet originating from the showering of a quark or gluon produced in the hard process.
- **n-subjettiness (τ_n):** A substructure tagging variable introduced in [3] that encodes how much of a jets momentum is aligned along n axes. Ratio's of these variables, $\tau_{n,n-1} = \tau_n/\tau_{n-1}$, are commonly used to tag jets with n prongs. For example, τ_{21} is used to tag two pronged jets.
- **p_T** : transverse momentum

Glossary

- **Soft drop mass (m_{SD}):** The groomed jet mass obtained from the "modified mass drop tagger" algorithm [4-6], also known as the "soft drop" (SD) algorithm, with angular exponent $\beta = 0$, soft cutoff threshold zcut = 0.1, and characteristic radius R of 0.8, is used to remove soft, wide-angle radiation from the jet.
- **Correction Factor (CF)**: The ratio of the efficiency of a selection in data to the efficiency of the same selection in simulation.
- **DeepAK8**: A machine learning based algorithm use in the CMS collaboration to tag various kinds of boosted large radius (AK8) jets [7]. Discriminants are constructed that reflect the likelihood of a jet originating from a boosted heavy resonance versus a standard quark or gluon jet. A version of the discriminant decorrolated with jet mass (-MD) is often used. For example, the DeepAK8-MD (W vs QCD) score is a typical variable used to tag W jets.
- **DeepJet**: A machine learning based algorithm used in the CMS collaboration to tag AK4 jets originating from b quark decays [8].

Introduction

- In the CMS collaboration many analyses make use of boosted large radius jets with substructure.
- The modeling of substructure observables in simulation is known to have deficiencies, so calibration of the tagging efficiency is generally based on the use of standard model (SM) proxy jets.
  - For example, boosted W jets are used for 2-prong tagging calibration, and boosted top jets for 3-prong tagging calibration. More details can be found in [7]
- However, some analyses may search for signals producing jets with large numbers of prongs (>3), for which there are no abundant SM proxy jets in data.
- In this note, a novel method to correct and calibrate the substructure of large radius multi-prong jets is presented.
- The method is based on reclustering the multi-prong jet, such that each prong is contained in a separate subjet. The simulated substructure of each subjet is then corrected via a data-driven reweighting of the subjet's splittings in the LJP.
- The data-driven correction is derived from a sample of jets enriched in hadronic W decays, and validated on a sample of hadronic top quark decays.
- This approach is generic, and therefore applicable to any boosted multi-prong jet provided that each prong originated from a SM quark.

Semi Leptonic $t\bar{t}$ Region

- The LJP correction is defined and validated in a data region dominated by semi-leptonic $t\bar{t}$ events.
- Events containing one isolated muon with p_T > 60 GeV, missing transverse energy greater than 50 GeV, an AK4 jet tagged as a b-jet by the DeepJet algorithm, and AK8 jet are selected.
- Events with m_{SD} between 60 and 110 GeV (125 and 225 GeV) and the p_T of the AK8 jet > 225 (500) GeV are referred to as the W-region (top-region).
- Simulated $t\bar{t}$ events are categorized based on how many generator-level quarks are contained in the cone of the AK8 jet.
  - If the two quarks from the W decay but not the b quark from the top decay are inside, it is categorized as W-matched.
  - If all three quarks from the top decay are inside, it is categorized as t-matched.
  - The remaining events are categorized as unmatched.

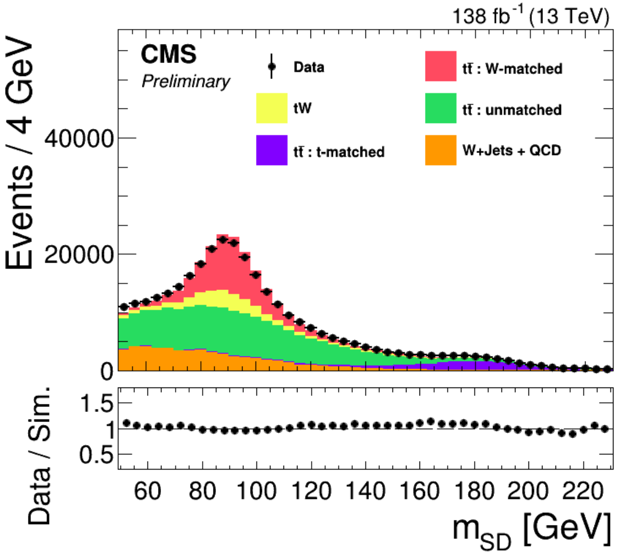

Figure 1: The softdrop mass distribution of events in semi-leptonic $t\bar{t}$ region. The W-region (top-region) covers the mass range 60 to 110 (125 to 225) GeV, and is enriched in W-matched (t-matched) events shown in the red (blue) histogram)

Lund Jet Plane Ratio

- We construct our LJPs using the subjets of the AK8 jet. The two axes of the LJP are the angle of the branching (Δ) and the transverse momentum of the emission relative to the emitter (k_t). We additionally bin the LJP in 6 bins of subjet p_T : 0, 50, 100, 175, 250, 350+ GeV.
- To obtain the subjets, the constituents inside the AK8 jets are reclustered with the exclusive kt algorithm into a number of subjets equal to the number of prongs. For example, 2 subjets are used for W jets, and 3 subjets are used for top jets.
- For each subjet, the constituents with transverse momentum greater than 1 GeV are then reclustered with the Cambridge/Aachen algorithm [9] to get the clustering history. For each splitting along the hardest branch of the clustering history, a point in the LJP is filled.
- Simulated and data AK8 jets in the W-region are reclustered into two subjets. For each subjet p_T bin, the LJP of data, W-matched $t\bar{t}$ simulation, and background simulation are constructed. The ratio of the data LJP minus the background LJP versus the W-matched $t\bar{t}$ LJP is taken.

Data/Simulation Lund Jet Plane Ratio

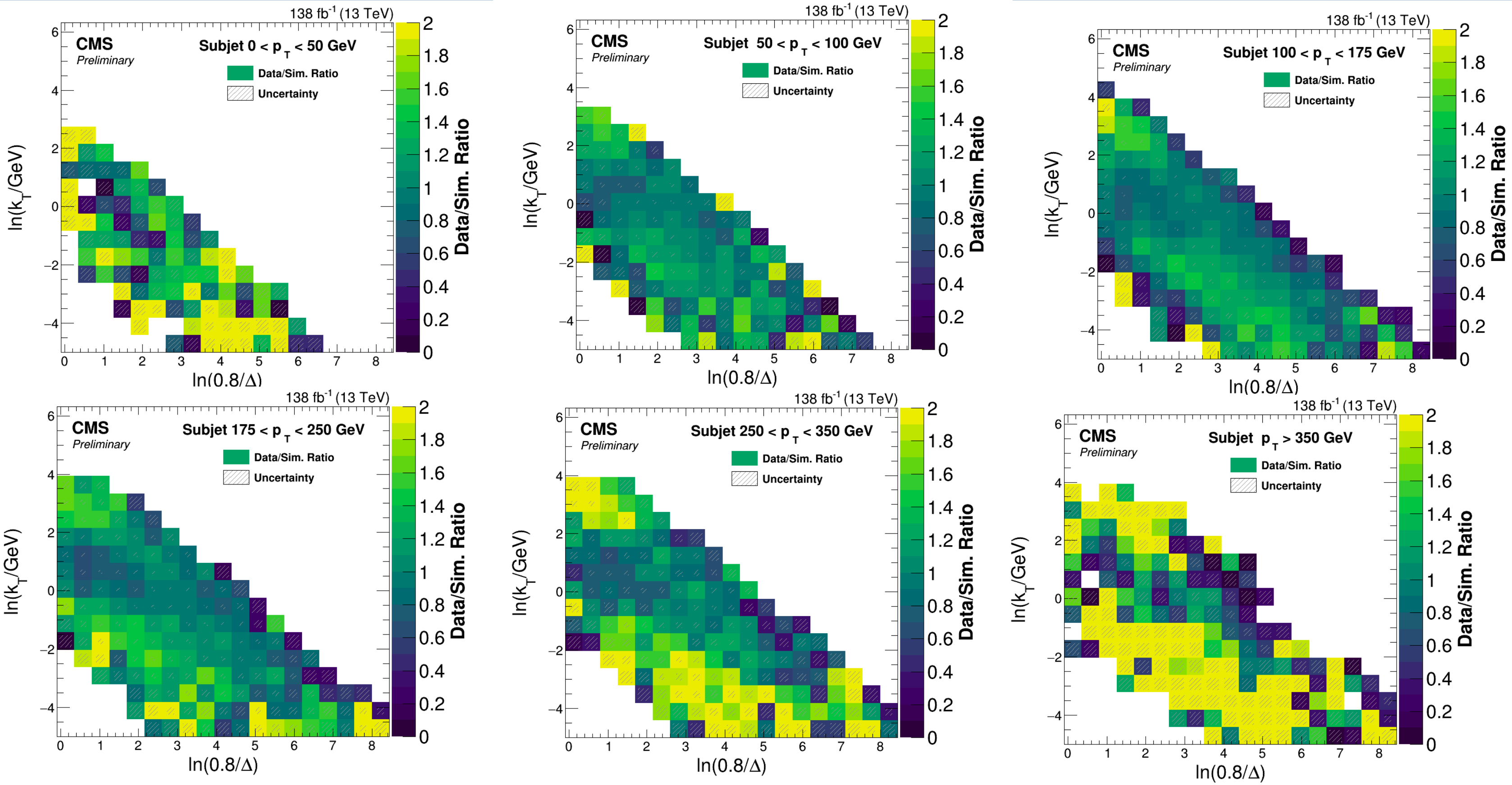

Figure 2: Ratios of the LJP in data and simulation in each subjet p_T bin. The combined statistical and systematic uncertainty on the ratio is represented by the area of the hatched region in each bin. Bins with no data or simulation events are shown as white, but assumed to have a ratio value of unity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

Correction Procedure

- The LJP data/simulation ratio is used to reweight simulated W-matched events in the W-region. For each splitting in the subjets, the corresponding correction factor from the ratios shown in Fig. 2. The total correction is the multiplication of the correction factor for each splitting.
- To validate the correction procedure is applicable to higher prong jets, it is performed on top jets. Simulated top-matched events in the top region are reclustered into three subjets and reweighted based on the same LJP data/simulation ratio.
- The data-simulation agreement of substructure observables before and after the correction are checked. It is seen that the correction significantly improves agreement for both W jets and top jets.
- Because the substructure of background processes are not corrected, it is expected that there are still some residual data-simulation discrepancies.

W Jets Substructure Observables

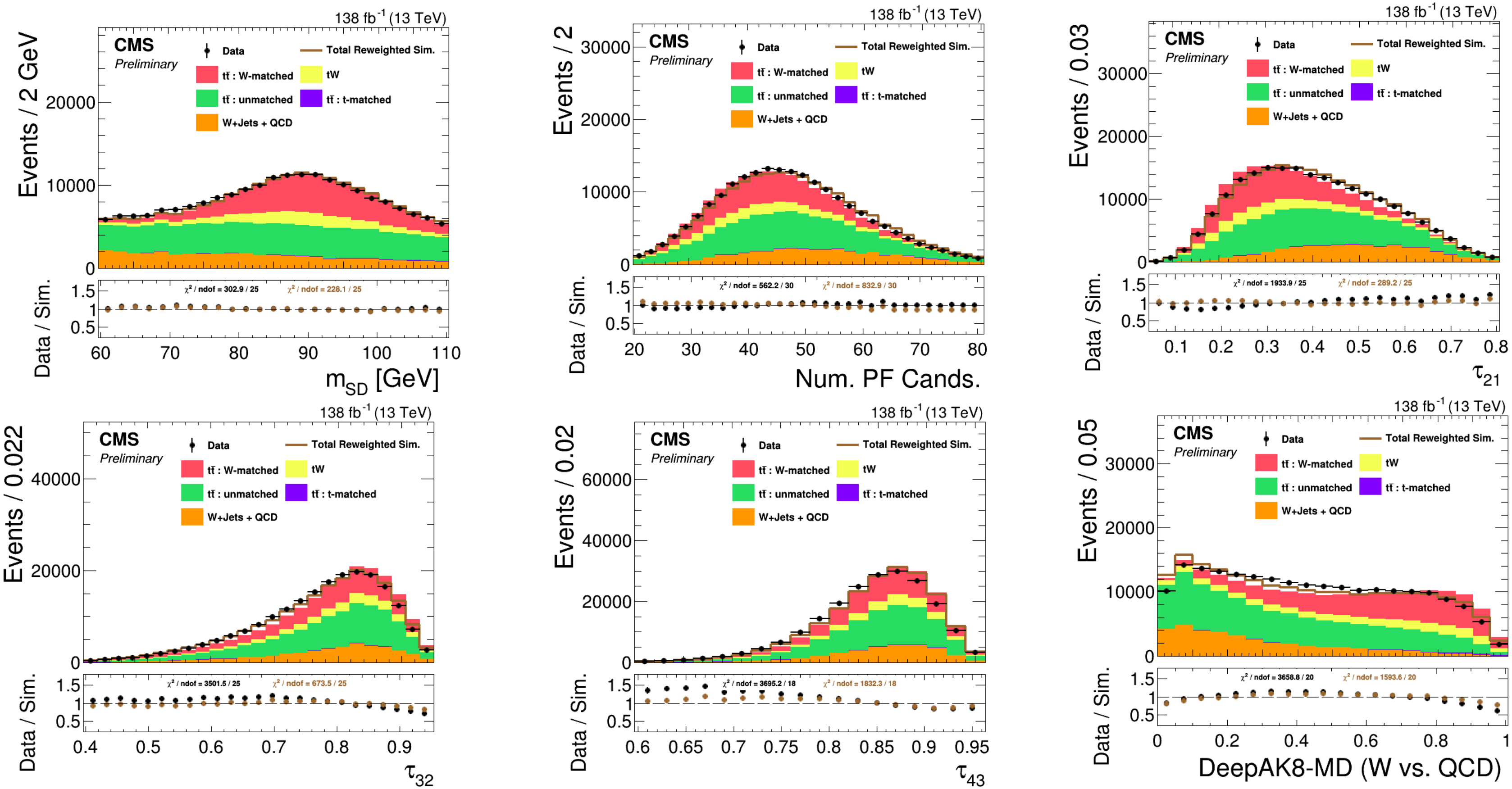

Figure 3: A comparison of the data/simulation agreement of various substructure observables in the W-region. The distribution of various simulated processes are shown in the colored histograms and observed data points are shown in black. The brown line shows the total simulated distribution after the LJP correction has been applied to the W-matched $t\bar{t}$ and tW simulations. Only statistical uncertainties are shown, and the computed χ^2 is based only on statistical uncertainties. The agreement between data and simulation improves with the correction, particularly in the two-pronged tagging variable τ_{21}.

Top Jets Substructure Observables

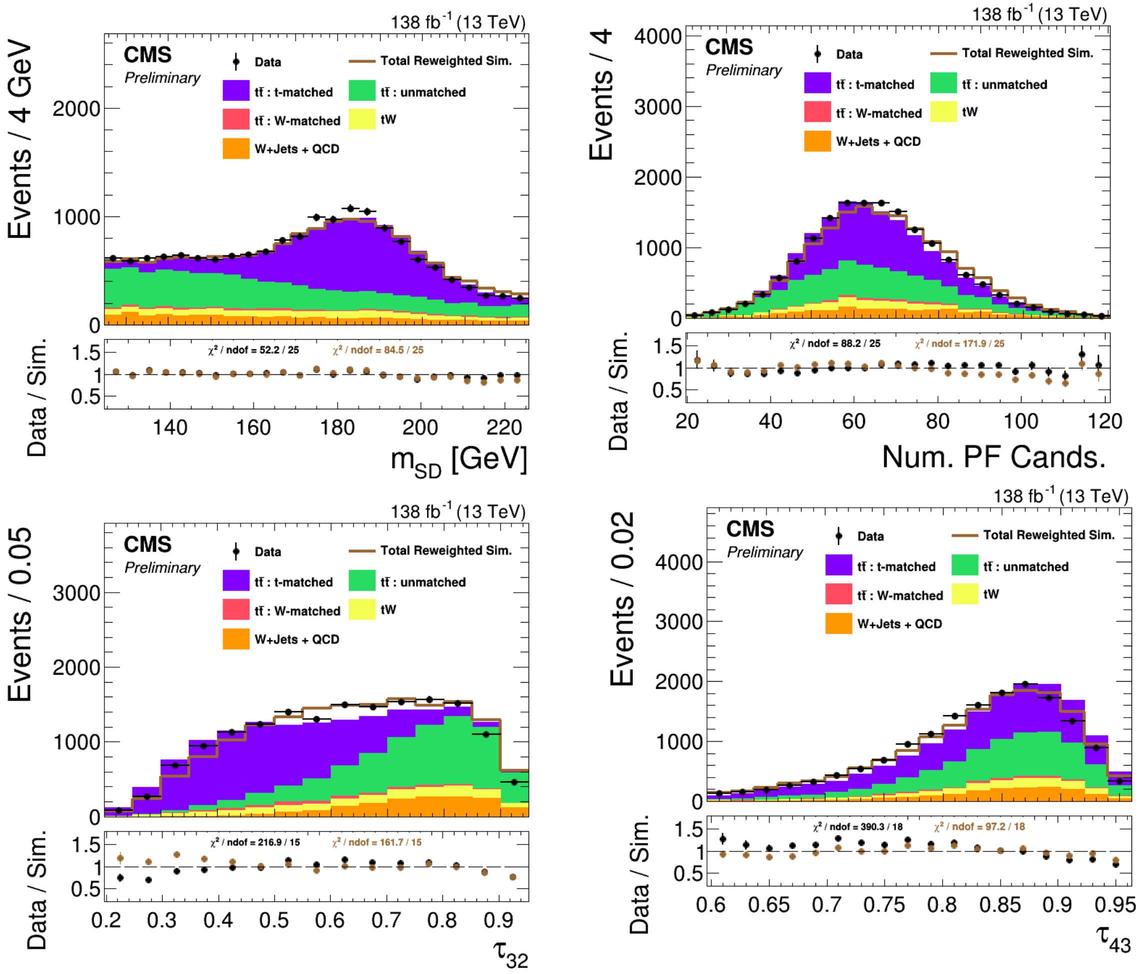

Figure 4: A comparison of the data/simulation agreement of various substructure observables in the top-region. The distribution of various simulated processes are shown in the colored histograms and observed data points are shown in black. The brown line shows the total simulated distribution after the Lund Plane correction has been applied to the top-matched $t\bar{t}$ simulation. Only statistical uncertainties are shown, and the computed χ^2 is based only on statistical uncertainties. The agreement between data and simulation improves with the correction, particularly in the three-pronged tagging variable τ_{32}.

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 – Quark Matching

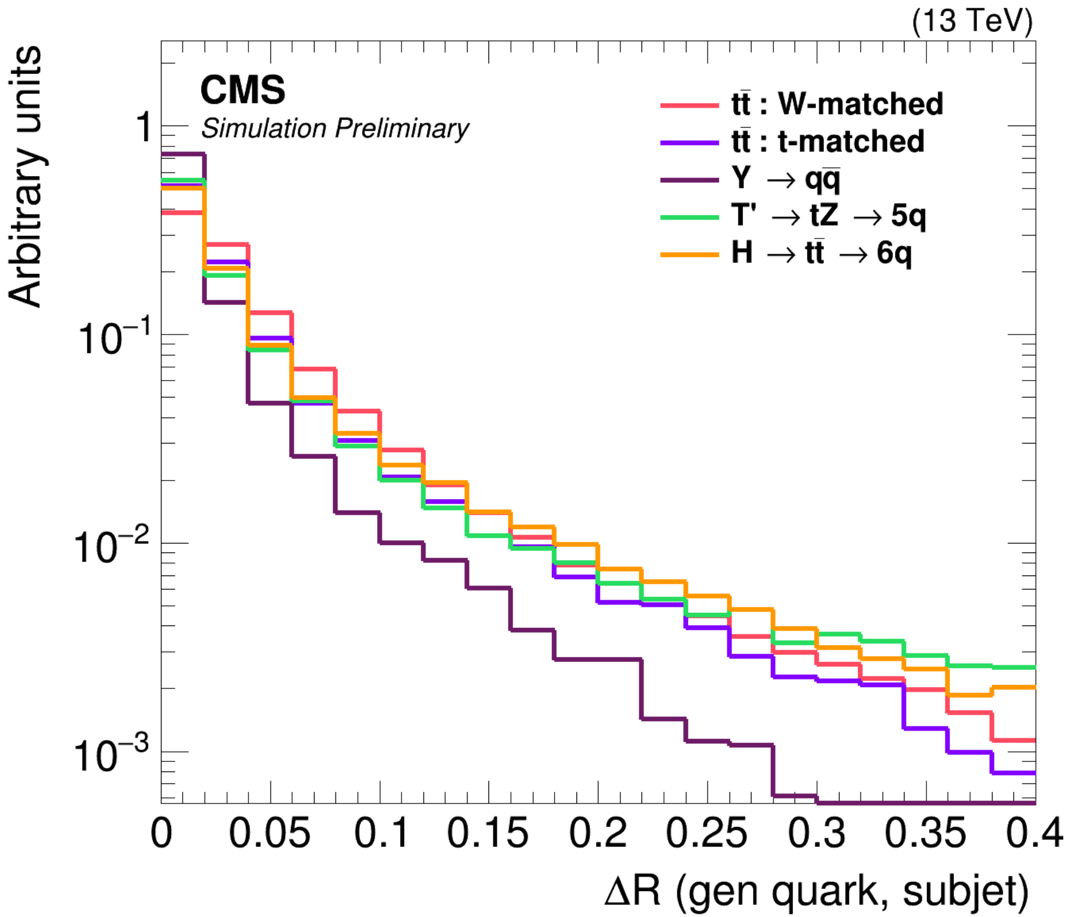

Figure 5: Distributions of the ΔR between subjets found by the reclustering with the exclusive k_t algorithm and gen-level quark of the heavy resonance decay. The distribution is shown for various different jet types simulated using CMS Run-2 conditions. Even for jets with a high number of prongs, the ΔR distribution remains steeply falling. But at high numbers of prongs, the showers from the multiple quarks begin to overlap, leading to increased 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
$Y \rightarrow qq$ (2)	τ_{21}	0.01	0.12	0.04	-	0.04	0.13
$R \rightarrow WW$ (4)	Deep-WH	0.02	0.02	0.06	-	0.28	0.29
$T' \rightarrow tZ$ (5)	τ_{43}	0.02	0.09	0.02	0.01	0.37	0.38
$H \rightarrow tt$ (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→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.

Correction Factor Comparison Plot

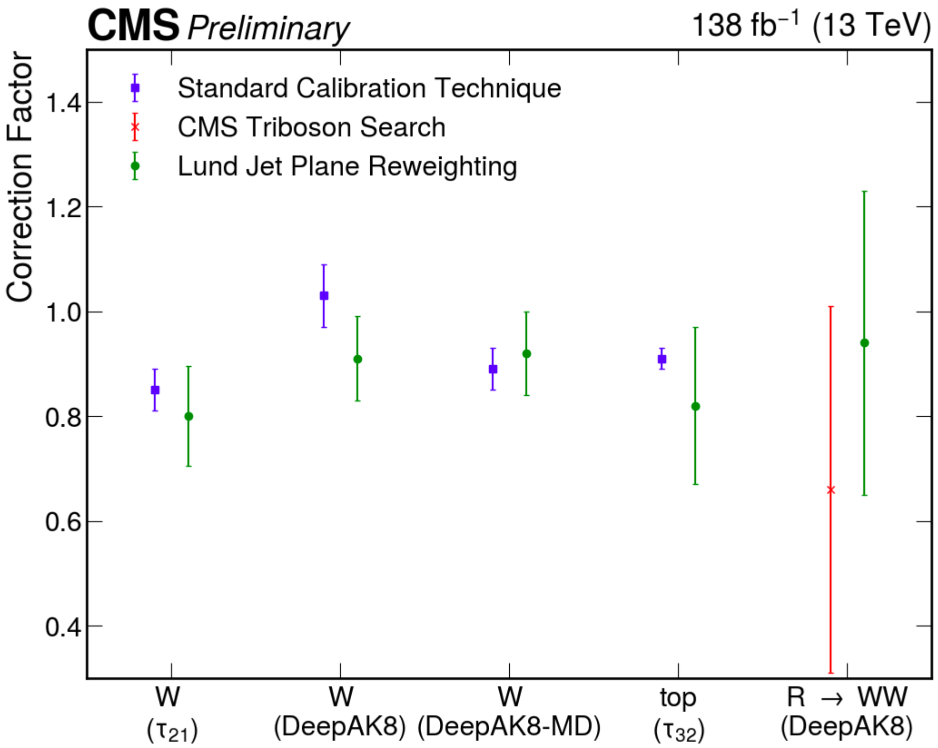

Figure 6: A comparison of CFs derived using the LJP correction procedure and other sources. For W and top tagging, CFs derived with the LJP have larger uncertainties but agree well with those from the standard calibration method. For the R→WW SF, the LJP correction factor is compared to the one obtained in a recent CMS search [11] which utilized top quarks with a hard gluon emission as a 4-prong proxy. The LJP CF agrees well with this value as well, with a ~20% reduction in uncertainty.

Summary

- A new technique to correct and calibrate the substructure of multi prong jets has been presented.
- The technique is based on a reclustering the multiprong jet into subjets, and applying correction to splittings of each subjet. The correction to the splittings is based on a data/simulation ratio of the LJP derived from boosted hadronic W decays.
- The technique was shown to improve the substructure modeling of W jets and was extrapolated to improve the modeling to higher pronged top jets.
- The technique will enable the calibration of high prong jets which were not previously possible. This calibration will enable future searches for such jets by the CMS collaboration.

References

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

[2] F. A. Dreyer, G. P. Salam, and G. Soyez, "The Lund Jet Plane", JHEP 12 (2018) 064, doi:10.1007/JHEP12(2018)064, arXiv:1807.04758.

[3] J. Thaler and K. Van Tilburg, "Identifying Boosted Objects with N-subjettiness", JHEP 03 (2011) 015, doi:10.1007/JHEP03(2011)015, arXiv:1011.2268.

[4] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, "Towards an understanding of jet substructure", JHEP 09 (2013) 029, doi:10.1007/JHEP09(2013)029, arXiv:1307.0007.

[5] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, "Jet substructure as a new Higgs search channel at the LHC", Phys. Rev. Lett. 100 (2008) 242001, doi:10.1103/PhysRevLett.100.242001, arXiv:0802.2470.

[6] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft drop", JHEP 05 (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.

References

[7] CMS Collaboration, "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques", JINST 15 (2020) P06005

[8] E. Bols et al., "Jet flavour classification using DeepJet", JINST 15 (2020) P12012, doi:10.1088/1748-0221/15/12/P12012, arXiv:2008.10519.

[9] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, "Better jet clustering algorithms", JHEP 08 (1997) 001, doi:10.1088/1126-6708/1997/08/001, arXiv:hep-ph/9707323.

[10] CMS Collaboration, "W and top tagging scale factors for Run 2 data", CMS DP 2020/025

[11] CMS Collaboration, "Search for resonances decaying to three W bosons in the hadronic final state in proton-proton collisions at s =13 TeV", Phys. Rev. D 106 (2022), no. 1, 012002, doi:10.1103/PhysRevD.106.012002, arXiv:2112.13090.

[12] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup per particle identification", JHEP 10

(2014) 059, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013.

[13] CMS Collaboration, "Pileup mitigation at CMS in 13 TeV data", JINST 15 (2020) P09018,

doi:10.1088/1748-0221/15/09/p09018, arXiv:2003.00503.