

Jet substructure measurements in CMS

Suman Chatterjee *
for the CMS Collaboration

*Institute of High Energy Physics (HEPHY), Austrian Academy of Sciences (ÖAW),
Nikolsdorfer Gasse 18, 1050 Vienna, Austria*

E-mail: suman.chatterjee@cern.ch

A number of measurements related to the study of hadronic jet substructure in proton-proton collisions at a center-of-mass energy of 13 TeV with the CMS experiment are presented. The measurement of jet substructure observables describing the distribution of particles within quark- and gluon-initiated jets is carried out with dijet and Z+jet event samples. The calibration of algorithms developed for the identification of hadronically decaying heavy particles is also reported. The results improve the understanding of jet composition and extend the horizon of the LHC physics program.

*** *The European Physical Society Conference on High Energy Physics (EPS-HEP2021), ****

*** *26-30 July 2021 ****

*** *Online conference, jointly organized by Universität Hamburg and the research center DESY ****

*Speaker

1. Introduction

Quantum chromodynamics (QCD) is a gauge theory describing the strong interaction between partons (quarks and gluons). Jets are reconstructed using hadrons produced by the fragmentation of partons in collisions [1]. The study of the distribution of particles within a jet referred to as jet substructure is very useful to improve the understanding of QCD [2–4]. For example, the application of jet grooming techniques, developed based on the understanding of jet substructure, greatly enhances the region of the applicability of perturbative QCD [5, 6]. Jet substructure techniques have also found applications in measuring the parameters of the standard model of particle physics, e.g., top quark mass [7], the evolution of strong coupling with energy [8], among others. Several techniques have been developed to distinguish the jets capturing decay products of heavy particles, for example, top quark, W, Z, and Higgs bosons, produced at high transverse momentum (p_T) from jets originating from a single parton [9–11], and are used both in precision measurements and new physics searches at LHC. Dedicated measurements of jet substructure observables help to understand and model the heavy particle identification algorithms better. Recently, experimental measurements have used jet substructure variables to sense the emergent phenomena in QCD, for example, the formation of quark-gluon plasma [12], dead-cone effects due to heavy-quark mass [13], among others. In the following, a few recent jet substructure measurements performed by the CMS experiment are reported.

2. Study of quark and gluon jet substructure

The difference in Casimir color factor between quarks and gluon leads to differences in particle composition in jets originated from quarks and gluons: gluon jets are expected to be wide and characterized by a large number of soft constituents, whereas quark jets are likely to contain a small number of hard particles inside. A class of observables, known as jet angularities (λ_β^κ) as defined in Eq. (1), is sensitive to the differences between quark- and gluon-initiated jets [14].

$$\lambda_\beta^\kappa = \sum_{i \in \text{jet}} z_i^\kappa \left(\frac{\Delta R_i}{R} \right)^\beta \quad (1)$$

In Eq. (1), R is the jet distance parameter, and z_i and ΔR_i represent the fractional transverse momentum carried by the i -th jet constituent and its displacement from the jet axis, respectively. Measurements of λ_β^κ with $(\kappa, \beta) = (0, 0), (1, 0.5), (1, 1), (1, 2), (2, 0)$, which correspond to physical quantities: multiplicity, Les Houches angularity, width, thrust, and $p_T^{D^2}$, used previously in quark-gluon tagging [15, 16], are performed with jets reconstructed using anti- k_T algorithm [17] in events with $Z(\mu\mu)$ +jet and dijet production in back-to-back topology [18], selected according to the conditions mentioned in Table 1. At low p_T , the jet from the Z +jet sample is likely to have originated from a quark, whereas the central jet in the dijet sample is more likely to have been initiated by a gluon. Dijet forward sample is enriched by jets originated by gluons at low p_T and those initiated by quarks at high p_T .

A two-dimensional unfolding in $(p_T, \lambda_\beta^\kappa)$ is performed using TUNFOLD package [19] to remove the detector effects as modeled by simulation. Several experimental and theoretical sources of systematic uncertainty are considered. Among them, the dominant contribution comes from modeling

Table 1: Summary of the selection criteria for the Z+jet and dijet event samples. Here, y and η denote rapidity and pseudorapidity, respectively.

Sample	Event selection
Z($\mu\mu$)+jet	A pair of oppositely charged muons each satisfying $ \eta < 2.4$ and $p_T^\mu > 26$ GeV $ m_{\mu\mu} - m_Z < 20$ GeV ≥ 1 jet with $ y < 1.7$ and $p_T^j > 30$ GeV not overlapping with muons from Z $\Delta\phi(j_1, Z) > 2$ $ p_T^{j_1} - p_T^Z / (p_T^{j_1} + p_T^Z) < 0.3$
Dijet (central and forward)	≥ 2 jets with $ y < 1.7$ and $p_T^j > 30$ GeV $\Delta\phi(j_1, j_2) > 2$ $ p_T^{j_1} - p_T^{j_2} / (p_T^{j_1} + p_T^{j_2}) < 0.3$
Dijet (central)	The jet with smaller $ y $ among two leading jets in the dijet sample
Dijet (forward)	The jet with larger $ y $ among two leading jets in the dijet sample

Table 2: Parameters varied in the measurement. Each column represents one dimension in the variation.

Sample	Kinematic region	Jet distance parameter	Constituents	Jet grooming
Z($\mu\mu$)+jet	[50, 65] GeV,	0.4	All particles	No grooming
Dijet (central)	...,	0.8	Charged particles only	Soft drop [20]
Dijet (forward)	> 1000 GeV			

of parton shower and hadronization, which is estimated taking the difference of unfolded results obtained using PYTHIA and HERWIG++ simulations. Results are presented at the particle level in multiple variants as summarized in Table 2.

Average values of λ_β^K in quark- and gluon-enriched jet samples are shown in Fig. 1, where the values in data are compared to those predicted by various simulations. Several important features are present in Fig. 1. The data-to-simulation comparison in the top panel reproduces the well-known fact that quark jets are better modeled, particularly by PYTHIA parton showering model. In contrast, predictions from PYTHIA and HERWIG++ parton showers for gluon jets are on two opposite sides of data. Predictions from modern generators SHERPA and HERWIG7, and the ones from PYTHIA using the latest underlying event (UE) tunes [21] model gluon jets better than quark jets, which implies that the latest LHC measurements used as inputs to the simulations have a significant effect in the modeling of quark-gluon responses. Fig. 1 also shows that angularities with $\kappa = 1$, which are infrared- and collinear-safe are not necessarily better modeled than the other two, which points to the fact that the mismodeling of data by simulations and also the variation between different predictions are not only due to non-perturbative effects but also occurs from the treatment of collinear radiation. Comparing PYTHIA predictions with different UE tunes, it is found that the prediction using CP2 tune, which uses a larger value of strong coupling and a smaller color reconnection range, describes data better than the one with the CP5 tune.

Quark-gluon differences are quantified using the ratios of average values of the angularities as shown in Fig. 2. The difference of the ratio of angularities from 1, as seen in Fig. 2, points to the quark-gluon separation power of the variables. However, this discrimination is overestimated by all the predictions as compared to data except at high p_T , where all of the samples have roughly equal

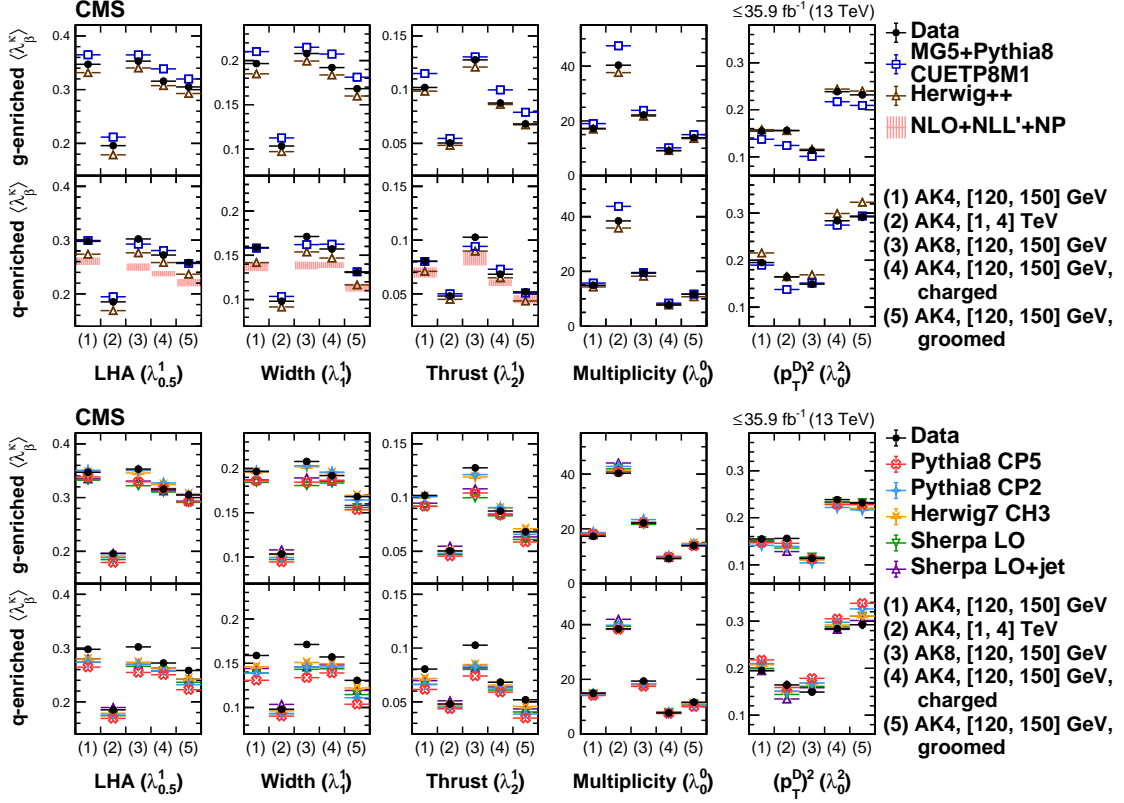


Figure 1: Mean of λ_β^K observables in regions with gluon-enriched and quark-enriched jets, for the following configurations: (1) ungroomed AK4 $120 < p_T < 150$ GeV, (2) ungroomed AK4 $1000 < p_T < 4000$ GeV, (3) ungroomed AK8 $120 < p_T < 150$ GeV, (4) ungroomed charged-only AK4 $120 < p_T < 150$ GeV, and (5) groomed AK4 $120 < p_T < 150$ GeV; shown for each of the five λ_β^K variables. The error bars on the data correspond to the total uncertainties of the experimental data. The error bars on the simulation correspond to the statistical uncertainties of the simulation [18].

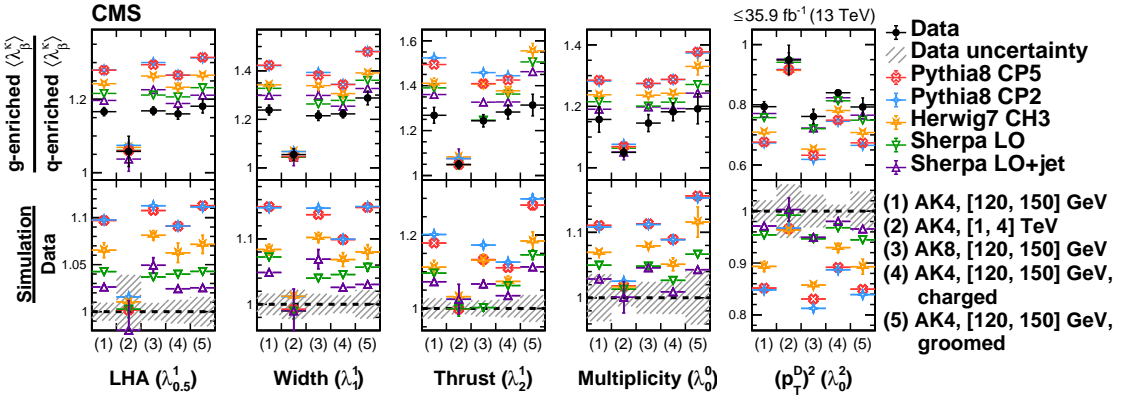


Figure 2: The ratio of the mean of λ_β^K in regions with gluon- and quark-enriched jets, for the same configurations as used in Fig. 1 [18].

amounts of quark and gluon jets, showing a clear need for improvements in the simulation. It is also evident that data-to-simulation comparison is not significantly affected by the variation of UE tune, jet size, composition, and grooming.

3. Calibration of heavy particle taggers

As already mentioned in Sec. 1, several observables are constructed to identify jets initiated by the decays of high- p_T heavy particles in comparison to those initiated by light quarks or gluons. Among those are n-subjettiness ratios [22], τ_{21} and τ_{32} , which measure the consistency of a jet having two and three clusters of energy, respectively. These variables are most commonly used in searches for new physics to identify large-radius jets originated from W boson and top quark decay [23]. Here, anti- k_T jets with distance parameter 0.8 (AK8 jets) are used after applying the pileup per particle identification (PUPPI) algorithm [24] for pileup mitigation. Events with μ +jets final state enriched by semileptonic $t\bar{t}$ production are used to check the modeling of n-subjettiness ratios of AK8 jets in data by simulation as shown in Fig. 3 (left) for τ_{32} , which shows that there is a need for improvement in the modeling. Distributions of soft-drop mass of jets associated to

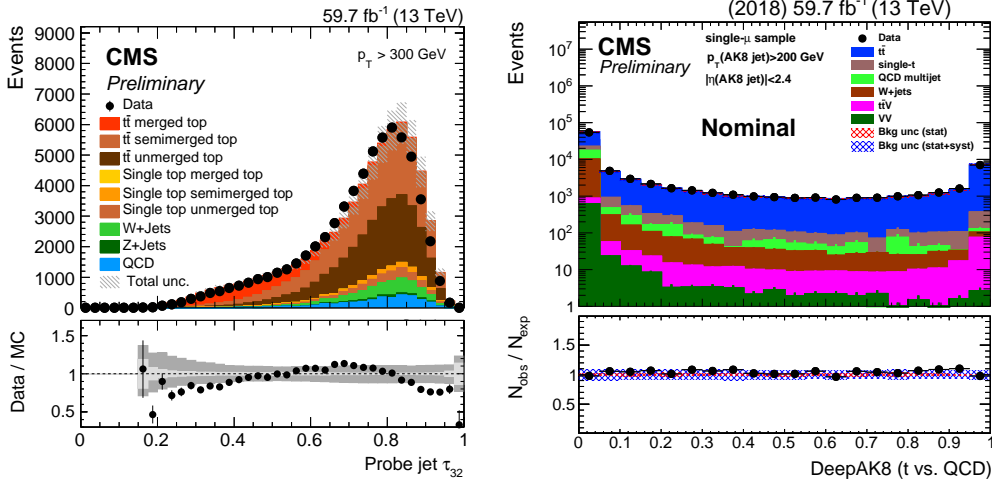


Figure 3: (Left) Distribution of n-subjettiness ratio τ_{32} of AK8 jets with $p_T > 300$ GeV. (Right) distribution of DEEPAK8 top tagger of AK8 jets with $p_T > 200$ GeV. Events with μ +jets in data and simulations corresponding to the 2018 run period are shown [25].

hadronic W or top, that pass or fail the criteria on n-subjettiness ratios, are fitted simultaneously to derive the identification efficiency in both data and simulation [25]. The ratio of identification efficiencies τ_{21} and τ_{32} variables in data and simulation, referred to as scale factor, is shown in Fig. 4. Systematic uncertainty on the scale factor includes the variations of jet energy scale, jet energy resolution, parton-showering model, and renormalization and factorization scales. Fig. 4 shows that the identification efficiencies with n-subjettiness ratios are very similar in data and simulation across different kinematic regions.

More recently, multi-class particle identification algorithms, referred to as DEEPAK8 taggers, are developed to identify AK8 jets from the hadronic decays of top quark, W, Z, and Higgs boson. The DeepAK8 algorithm uses a deep one-dimensional convolutional neural network taking

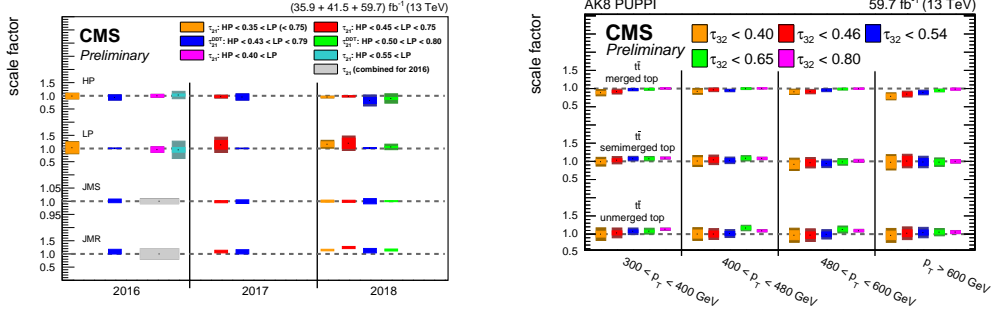


Figure 4: (Left) data-to-simulation scale factors for all three years in high purity (HP) and low purity (LP) categories for nominal τ_{21} and τ_{21} decorrelated with jet soft-drop mass. HP and LP categories refer to the cases when τ_{21} is required to be less and more than a threshold, respectively. (Right) data-to-simulation scale factors as a function of jet p_T for different thresholds on τ_{32} . The inner and outer areas indicate the statistical and the total uncertainties, respectively [25].

particles and secondary vertices associated with the jet as inputs. These taggers, well modeled by the simulation as shown Fig. 3 (right) for top tagging, have already been used in searches for new particles [26]. Scale factors for DEEPAK8 taggers are determined in a very similar manner as for n-subjettiness ratios and are shown in Fig. 5. Tagging efficiencies for DEEPAK8 taggers are well modeled in all three years for the identification of both W- and top-jets.

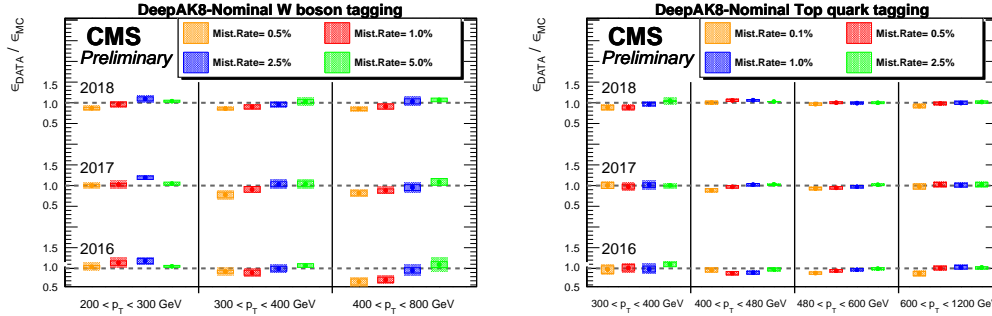


Figure 5: Data-to-simulation scale factors as a function of jet p_T in three years of data taking in case of DEEPAK8 algorithms for W-tagging (left) and top-tagging (right). Scale factors are shown for four misidentification rates. The solid open squares correspond to the statistical uncertainty, while the dashed squares to the total uncertainty obtained by adding statistical and systematic uncertainties in quadrature [25].

4. Conclusion

Measurements of jet substructure variables used for quark-gluon discrimination are presented. These are very useful to understand the details of parton-to-jet evolution, which results in jet composition. Calibration of algorithms for heavy particle tagging, based on both analytical expressions and machine learning, are presented for W boson and top quark identification. There is also a rich set of new measurements to appear in the near future.

References

- [1] G. Sterman and S. Weinberg, “Jets from quantum chromodynamics”, *Phys. Rev. Lett.* **39** (1977) 1436, doi:[10.1103/PhysRevLett.39.1436](https://doi.org/10.1103/PhysRevLett.39.1436).
- [2] M. Dasgupta, A. Fregoso, S. Marzani, and A. Powling, “Jet substructure with analytical methods”, *Eur. Phys. J. C* **73** (2013), no. 11, 2623, doi:[10.1140/epjc/s10052-013-2623-3](https://doi.org/10.1140/epjc/s10052-013-2623-3), arXiv:[1307.0013](https://arxiv.org/abs/1307.0013).
- [3] 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](https://doi.org/10.1007/JHEP09(2013)029), arXiv:[1307.0007](https://arxiv.org/abs/1307.0007).
- [4] F. A. Dreyer, G. P. Salam, and G. Soyez, “The Lund Jet Plane”, *JHEP* **12** (2018) 064, doi:[10.1007/JHEP12\(2018\)064](https://doi.org/10.1007/JHEP12(2018)064), arXiv:[1807.04758](https://arxiv.org/abs/1807.04758).
- [5] ATLAS Collaboration, “Measurement of the Soft-Drop Jet Mass in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector”, *Phys. Rev. Lett.* **121** (2018) 092001, doi:[10.1103/PhysRevLett.121.092001](https://doi.org/10.1103/PhysRevLett.121.092001), arXiv:[1711.08341](https://arxiv.org/abs/1711.08341).
- [6] CMS Collaboration, “Measurements of the differential jet cross section as a function of the jet mass in dijet events from proton-proton collisions at $\sqrt{s} = 13$ TeV”, *JHEP* **11** (2018) 113, doi:[10.1007/JHEP11\(2018\)113](https://doi.org/10.1007/JHEP11(2018)113), arXiv:[1807.05974](https://arxiv.org/abs/1807.05974).
- [7] CMS Collaboration, “Measurement of the Jet Mass Distribution and Top Quark Mass in Hadronic Decays of Boosted Top Quarks in pp Collisions at $\sqrt{s} = \text{TeV}$ ”, *Phys. Rev. Lett.* **124** (2020) 202001, doi:[10.1103/PhysRevLett.124.202001](https://doi.org/10.1103/PhysRevLett.124.202001), arXiv:[1911.03800](https://arxiv.org/abs/1911.03800).
- [8] CMS Collaboration, “Measurement of jet substructure observables in $t\bar{t}$ events from proton-proton collisions at $\sqrt{s} = 13\text{TeV}$ ”, *Phys. Rev. D* **98** (2018), no. 9, 092014, doi:[10.1103/PhysRevD.98.092014](https://doi.org/10.1103/PhysRevD.98.092014), arXiv:[1808.07340](https://arxiv.org/abs/1808.07340).
- [9] A. J. Larkoski, I. Moult, and B. Nachman, “Jet substructure at the Large Hadron Collider: a review of recent advances in theory and machine learning”, *Phys. Rept.* **841** (2020) 1, doi:[10.1016/j.physrep.2019.11.001](https://doi.org/10.1016/j.physrep.2019.11.001), arXiv:[1709.04464](https://arxiv.org/abs/1709.04464).
- [10] R. Kogler et al., “Jet substructure at the Large Hadron Collider: Experimental review”, *Rev. Mod. Phys.* **91** (2019) 045003, doi:[10.1103/RevModPhys.91.045003](https://doi.org/10.1103/RevModPhys.91.045003), arXiv:[1803.06991](https://arxiv.org/abs/1803.06991).
- [11] S. Chatterjee, R. Godbole, and T. S. Roy, “Jets with electrons from boosted top quarks”, *JHEP* **01** (2020) 170, doi:[10.1007/JHEP01\(2020\)170](https://doi.org/10.1007/JHEP01(2020)170), arXiv:[1909.11041](https://arxiv.org/abs/1909.11041).
- [12] CMS Collaboration, “Measurement of the Splitting Function in pp and Pb-Pb Collisions at $\sqrt{s_{\text{NN}}} = 5.02$ TeV”, *Phys. Rev. Lett.* **120** (2018), no. 14, 142302, doi:[10.1103/PhysRevLett.120.142302](https://doi.org/10.1103/PhysRevLett.120.142302), arXiv:[1708.09429](https://arxiv.org/abs/1708.09429).
- [13] ALICE Collaboration, “Direct observation of the dead-cone effect in QCD”, arXiv:[2106.05713](https://arxiv.org/abs/2106.05713).
- [14] P. Gras et al., “Systematics of quark/gluon tagging”, *JHEP* **07** (2017) 091, doi:[10.1007/JHEP07\(2017\)091](https://doi.org/10.1007/JHEP07(2017)091), arXiv:[1704.03878](https://arxiv.org/abs/1704.03878).
- [15] ATLAS Collaboration, “Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector”, *Eur. Phys. J. C* **74** (2014) 3023, doi:[10.1140/epjc/s10052-014-3023-z](https://doi.org/10.1140/epjc/s10052-014-3023-z), arXiv:[1405.6583](https://arxiv.org/abs/1405.6583).
- [16] CMS Collaboration, “Jet algorithms performance in 13 TeV data”, CMS

- Physics Analysis Summary CMS-PAS-JME-16-003, 2017.
- [17] M. Cacciari, G. P. Salam, and G. Soyez, “The anti- k_T jet clustering algorithm”, *JHEP* **04** (2008) 063, doi:[10.1088/1126-6708/2008/04/063](https://doi.org/10.1088/1126-6708/2008/04/063), arXiv:[0802.1189](https://arxiv.org/abs/0802.1189).
 - [18] CMS Collaboration, “Study of quark and gluon jet substructure in Z+jet and dijet events from pp collisions”, arXiv:[2109.03340](https://arxiv.org/abs/2109.03340).
 - [19] S. Schmitt, “TUnfold: an algorithm for correcting migration effects in high energy physics”, *JINST* **7** (2012) T10003, doi:[10.1088/1748-0221/7/10/T10003](https://doi.org/10.1088/1748-0221/7/10/T10003), arXiv:[1205.6201](https://arxiv.org/abs/1205.6201).
 - [20] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, “Soft Drop”, *JHEP* **05** (2014) 146, doi:[10.1007/JHEP05\(2014\)146](https://doi.org/10.1007/JHEP05(2014)146), arXiv:[1402.2657](https://arxiv.org/abs/1402.2657).
 - [21] CMS Collaboration, “Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements”, *Eur. Phys. J. C* **80** (2020) 4, doi:[10.1140/epjc/s10052-019-7499-4](https://doi.org/10.1140/epjc/s10052-019-7499-4), arXiv:[1903.12179](https://arxiv.org/abs/1903.12179).
 - [22] J. Thaler and K. Van Tilburg, “Identifying Boosted Objects with N-subjettiness”, *JHEP* **03** (2011) 015, doi:[10.1007/JHEP03\(2011\)015](https://doi.org/10.1007/JHEP03(2011)015), arXiv:[1011.2268](https://arxiv.org/abs/1011.2268).
 - [23] CMS Collaboration, “Search for resonant $t\bar{t}$ production in proton-proton collisions at $\sqrt{s} = 13$ TeV”, *JHEP* **04** (2019) 031, doi:[10.1007/JHEP04\(2019\)031](https://doi.org/10.1007/JHEP04(2019)031), arXiv:[1810.05905](https://arxiv.org/abs/1810.05905).
 - [24] D. Bertolini, P. Harris, M. Low, and N. Tran, “Pileup per particle identification”, *JHEP* **10** (2014) 059, doi:[10.1007/JHEP10\(2014\)059](https://doi.org/10.1007/JHEP10(2014)059), arXiv:[1407.6013](https://arxiv.org/abs/1407.6013).
 - [25] CMS Collaboration, “W and top tagging scale factors for Run 2 data”, CMS Detector Performance Report CMS-DP-2020-025, 2020.
 - [26] CMS Collaboration, “Search for W bosons decaying to a top and a bottom quark at $\sqrt{s} = 13$ TeV in the hadronic final state”, *Phys. Lett. B* **820** (2021) 136535, doi:[10.1016/j.physletb.2021.136535](https://doi.org/10.1016/j.physletb.2021.136535), arXiv:[2104.04831](https://arxiv.org/abs/2104.04831).