Jet substructure studies with CMS open data

Jet substructure studies with CMS open data

Aashish Tripathee,[1,*] Wei Xue,[1,†] Andrew Larkoski,[2,‡] Simone Marzani,[3,§] and Jesse Thaler[1,‖]
[1]Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2]Physics Department, Reed College, Portland, Oregon 97202, USA
[3]University at Buffalo, The State University of New York, Buffalo, New York 14260-1500, USA
(Received 9 May 2017; published 3 October 2017)

We use public data from the CMS experiment to study the two-prong substructure of jets. The CMS open data are based on 31.8 pb^{-1} of 7 TeV proton-proton collisions recorded at the Large Hadron Collider in 2010, yielding a sample of 768,687 events containing a high-quality central jet with transverse momentum larger than 85 GeV. Using CMS's particle flow reconstruction algorithm to obtain jet constituents, we extract the two-prong substructure of the leading jet using soft-drop declustering. We find good agreement between results obtained from the CMS open data and those obtained from parton shower generators, and we also compare to analytic jet substructure calculations performed to modified leading-logarithmic accuracy. Although the 2010 CMS open data do not include simulated data to help estimate systematic uncertainties, we use track-only observables to validate these substructure studies.

DOI: 10.1103/PhysRevD.96.074003

I. INTRODUCTION

In November 2014, the CMS experiment at the Large Hadron Collider (LHC) announced the CMS Open Data project [1]. To our knowledge, this is the first time in the history of particle physics that research-grade collision data has been made publicly available for use outside of an official experimental collaboration. The CMS open data were reconstructed from 7 TeV proton-proton collisions in 2010, corresponding to a unique low-luminosity running environment where pileup contamination was minimal and trigger thresholds were relatively low. The CMS open data present an enormous opportunity to the particle physics community, both for performing physics studies that would be more difficult at higher luminosities and for demonstrating the scientific value of open data releases.

In this paper, we use the CMS open data to analyze the substructure of jets. Jets are collimated sprays of particles that are copiously produced in LHC collisions, and by studying the substructure of jets, one can gain valuable information about their parentage [2–10]. A key application of jet substructure is tagging boosted heavy objects like top quarks [11–31] and electroweak bosons [3,4,6,14,22,30–59]. To successfully tag such objects, though, one first has to understand the radiation patterns of ordinary quark and gluon jets [26,60–75], which are the main backgrounds to boosted objects. The CMS open data are a fantastic resource for performing these baseline quark/gluon studies. Using the Jet Primary Dataset [76], we perform initial investigations of the two-prong substructure of jets as well as present a general analysis framework to facilitate future studies. This effort is complementary to the growing catalog of jet substructure measurements performed within the ATLAS and CMS collaborations [77–199].[1]

The core of our analysis is based on soft-drop declustering [46], which is a jet grooming technique [6,200–202] that mitigates jet contamination from initial state radiation (ISR), underlying event (UE), and pileup. For the studies in this paper, we set the soft-drop parameter β equal to zero, such that soft drop behaves like the modified mass drop tagger (mMDT) [203,204].[2] After soft drop, a jet is composed of two well-defined subjets, which can then be used to derive various two-prong substructure observables. In addition to comparing the CMS open data to parton shower generators, we perform first-principles calculations of soft-dropped observables using recently developed analytic techniques [46,205,206]. In a companion paper, we use soft drop to expose the QCD splitting function using the CMS open data [207]; a similar strategy was used in preliminary CMS [167], STAR [208], and ALICE [209] heavy ion studies to test for possible modifications to the splitting function from the dense QCD medium [210,211].

For studying jet substructure, the key feature of the CMS open data is that they contain full information about particle

*aashisht@mit.edu
†weixue@mit.edu
‡larkoski@reed.edu
§smarzani@buffalo.edu
‖jthaler@mit.edu

[1]To highlight the vibrancy of the field, we have attempted to list all published jet substructure measurements from ATLAS and CMS. Please contact us if we missed a reference.

[2]The original mass drop tagger [6] was a pioneering technique in jet substructure; see also precursor work in Refs. [2–5].

flow candidates (PFCs). The particle flow algorithm [212,213] synthesizes information from multiple detector elements to create a unique particlelike interpretation of each collision event. Within CMS, these PFCs are used directly in jet reconstruction [214]. Here, we can exploit the PFC information to perform detailed jet substructure studies, using standard particle-based jet analysis tools.

The main limitation of the 2010 CMS open data release is that it only provides minimal calibration information, and therefore we cannot properly estimate systematic uncertainties from detector effects. Ideally, we would like a detector simulation or a smearing parametrization to account for finite resolution and granularity. Absent that, we cannot make a direct comparison of CMS open data to properly folded particle-level distributions. With that caveat in mind, our plots will overlay detector-level CMS open data (without further calibration) and particle-level theory distributions (without detector simulation). The overall agreement turns out to be rather good, highlighting the excellent performance of the CMS detector and CMS's particle flow reconstruction. One must always keep in mind, though, that our plots cannot be interpreted like standard LHC experimental plots, both because of the absence of detector (un)folding and the absence of systematic uncertainties in the error bars.

To gain confidence in the robustness of our substructure analysis, we perform cross-checks using track-based variants. Distributions using only charged particles are expected to exhibit better resolution than those using all particles, and we indeed find better qualitative agreement with parton showers using these track-based observables. We also attempted to estimate detector effects using the DELPHES fast simulation tool [215], but we found that the default CMS-like detector settings led to oversmearing of the distributions, so no DELPHES results will be shown in this paper. For the future, we plan to repeat these studies using the 2011 CMS open data [216], which do come accompanied by detector-simulated Monte Carlo files.

The remainder of this paper is organized as follows. In Sec. II, we give an overview of the CMS open data and corresponding analysis tools. In Sec. III, we present basic kinematic and substructure properties of the hardest jet in the event, comparing the CMS open data to parton shower generators. In Sec. IV, we review the soft-drop algorithm and compare analytic calculations of the two-prong substructure to open data and parton shower distributions. Based on our experience with the CMS open data, we provide recommendations to CMS and to the broader particle physics community in Sec. V. We conclude in Sec. VI, leaving additional details and plots to the Appendixes.

II. THE CMS OPEN DATA

The CMS open data are available from the CERN Open Data Portal [1], with the initial release corresponding to Run 2010B of the LHC. The primary data sets are in the form of analysis object data (AOD) files, which is a file format used internally within CMS based on the ROOT framework [217]. To process the CMS data, one first has to install a virtual machine (VM) with CERNVM running SCIENTIFIC LINUX CERN 5. Within the VM, one can then run the official CMS software framework (CMSSW), which provides access to the complete analysis tools needed to parse the AOD files.

Our jet substructure study is based on the Jet Primary Dataset [76], which is a subset of the full open data release with events that pass a predefined set of single-jet and multijet triggers. There are 1664 AOD files in the Jet Primary Dataset, corresponding to 20,022,826 events and 2.0 terabytes of disk space. Within CMSSW, it is possible to access the AOD files remotely through the XROOTD interface [218]. We found it more convenient to first download the AOD files and then process them locally, being careful to maintain the same directory structure as on the open data servers in order to ensure consistency of the workflow. We then converted AOD files into a text-based MIT Open Data (MOD) format to facilitate the use of external analysis tools.

A. The CMS software framework

CMSSW is a hybrid PYTHON/C++ analysis framework where event processing takes place through user-defined modules. The version provided with the CMS open data is 4.2.8, which was also used internally by CMS in 2010 (as of this writing, the current CMSSW version is 9.0.0). In principle, we could have used CMSSW directly to perform our jet substructure studies, but we found it more convenient to simply use CMSSW for data extraction and then use external tools for analysis, described in Sec. II D.

Within CMS, there are multiple tiers of data, but only AOD files are provided by the CMS open data. Starting from RAW detector-level data, CMS derives RECO (reconstructed) data which includes both low-level objects (like reconstructed tracks) and high-level objects (like clustered jets). For most CMS analyses, only a subset of the RECO data is required, and this is the basis for the AOD files. For our open data analysis, the AOD files contain far more information than needed, so we use CMSSW to isolate only the required physics objects and event information.

To use CMSSW for data extraction, we rely on a chain of user-defined modules. We use a `Source` module to read in events from the AOD files and an `EDProducer` called `MODProducer` to convert the AOD format into our own text-based MOD format (see Sec. II C).[3] The `MODProducer` software is available through a GITHUB repository [219].

[3]Here, `ED` refers to "event data." Strictly speaking, since we are not modifying the AOD files directly, we could have used an `EDAnalyzer` instead of an `EDProducer`. We decided to use `EDProducer` because the name aligns better with what the module is actually doing, namely "creating data," albeit in the MOD format. Also, CMS recommends using an `OutputModule` when writing to an external file, but we instead used the standard C++ libraries for output.

In order to maintain reasonable file sizes and enable easier data validation, we wanted `MODProducer` to generate a separate MOD file for each of the 1664 AOD files, rather than one monolithic MOD file. While we could have run `MODProducer` separately for each AOD file, it turns out that `MODProducer` has to load `FrontierConditions_GlobalTag_cff` and the appropriate global tag (`GR_R_42_V25::All`) in order to properly extract trigger information from the AOD file. Loading this information takes around 10 minutes at the beginning of a CMSSW run, so to save computing time, we wanted to process multiple AOD files in series in the same run. To the best of our knowledge, though, CMSSW does not allow an `EDProducer` to know which AOD file is being processed. To circumvent this limitation, we created a lightweight `FilenameMapProducer` that only runs on one file at a time and creates a map relating event and run numbers to the corresponding AOD filename. This filename map is then read in by `MODProducer`, along with a list provided by CMS of validated runs suitable for physics analyses.

From the AOD files, `MODProducer` extracts PFCs, jets clustered from these PFCs, associated jet calibration information, trigger information, luminosity information, and basic event identification information like event and run numbers. The PFCs provide a unique reference event interpretation in terms of reconstructed photons, electrons, muons, charged hadrons, and neutral hadrons [212,213]. Each PFC has a particle identification flag and a full Lorentz four-vector, with nonzero invariant mass when available. We use AK5 jets provided by CMS [214], corresponding to the anti-k_t jet clustering algorithm [220] with $R = 0.5$ and a minimum jet threshold of $p_T > 3.0$ GeV, and we later validate the anti-k_t clustering by running FastJet 3.1.3 [221] ourselves on the PFCs. The jet calibration information includes both jet quality criteria and jet energy corrections (JEC) factors, discussed further in Appendix A. We discuss trigger and luminosity information in more detail next.

B. The Jet Primary Dataset

The CMS open data is grouped into primary data sets, corresponding to the types of triggers that were used for event selection. Our analysis is based exclusively on the Jet Primary Dataset [76].[4] As listed in Table I, this data set has single-jet, dijet, quad-jet, and H_T triggers, though we only use single-jet triggers for our study. Each trigger has an associated prescale factor, which is the ratio of how often the triggering criteria are met compared to how many events the trigger actually records. A prescale factor of 1 indicates that all triggered events are kept, whereas larger prescale factors are assigned to frequently encountered event categories that would otherwise overwhelm data acquisition. The prescale factor used in the analysis is the product of the prescale factors from the underlying Level 1 trigger (based on low-level objects) and the final high level trigger (HLT). There are various versions of the triggers, indicated by suffixes like `_v2` and `_v3`, but we do not distinguish between the versions in our analysis.

TABLE I. Jet triggers provided in the Jet Primary Dataset [76], including the number of events for which the trigger was present and/or fired. Entries marked by * are used in this analysis (see Table II). `HNF` stands for `HcalNoiseFiltered`. We do not separate out the different versions of the same trigger in our analysis.

	Trigger	Present?	Fired?
Single-jet	`HLT_Jet15U`	16,341,190	1,342,155
	* `HLT_Jet15U_HNF`	16,341,190	1,341,930
	* `HLT_Jet30U`	16,341,190	604,287
	* `HLT_Jet50U`	16,341,190	870,649
	* `HLT_Jet70U`	16,341,190	5,257,339
	* `HLT_Jet100U`	16,341,190	3,689,951
	* `HLT_Jet140U`	5,989,945	1,898,874
	`HLT_Jet180U`	2,595,038	553,331
Dijet	`HLT_DiJetAve15U`	16,341,191	1,067,561
	`HLT_DiJetAve30U`	16,341,191	648,000
	`HLT_DiJetAve50U`	16,341,191	859,292
	`HLT_DiJetAve70U`	16,341,191	2,310,033
	`HLT_DiJetAve100U`	5,989,945	1,252,661
	`HLT_DiJetAve140U`	2,595,038	452,222
Quad-jet	`HLT_QuadJet20U`	10,351,245	677,451
	`HLT_QuadJet25U`	10,351,244	219,256
H_T	`HLT_HT100U`	10,351,245	7,369,985
	`HLT_HT120U`	10,351,245	4,090,218
	`HLT_HT140U`	10,351,245	2,430,208
	`HLT_EcalOnly_SumEt160`	10,351,246	208,718

The CMS single-jet triggers are designed to fire whenever *any* jet in the event is above a given p_T threshold. Since our substructure study is based only on the hardest jet in an event, we have to make sure that the correct "assigned" trigger fired for the hardest AK5 jet in an event. We also have to check that this trigger is nearly 100% efficient for jets of the given p_T.

In Fig. 1(a), we show the p_T spectrum of the hardest jet for the six triggers used in our analysis. All jets have passed a "loose" jet quality cut with appropriate JEC factors applied; see Table V and Fig. 15(a) in Appendix A. We further impose a pseudorapidity cut of $|\eta| < 2.4$ to ensure that jets are reconstructed in the central part of the CMS detector where tracking information is available. With prescale factors included, we see good overlap of the p_T spectra as desired, except for the `Jet140U` trigger which is systematically low. The reason is that the `Jet140U` trigger was not present for the entirety of Run 2010B, so we revert to the `Jet100U` trigger when needed.

Using `HLT_Jet15U_HcalNoiseFiltered` as the baseline, the trigger efficiencies of the five remaining

[4]In order to study lower p_T jets, we would have to incorporate the MinimumBias Primary Dataset [222]. Because primary data sets are overlapping, one has to be careful not to double count events when using multiple primary data sets.

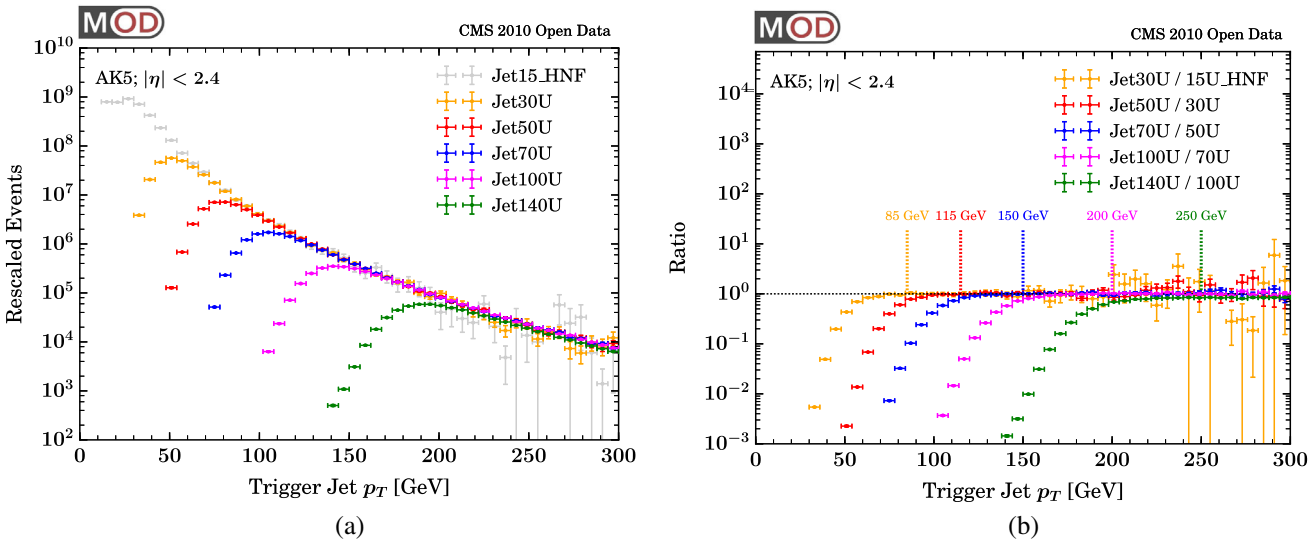

FIG. 1. (a) Hardest jet p_T spectrum in the CMS open data from the six triggers used in this analysis (see Table I). (b) Ratios of the jet p_T spectra from adjacent triggers used to determine when the triggers are nearly 100% efficient, which determine the jet trigger boundaries in Table II. Because the `Jet_140U` trigger was not present for the entirety of the run, it artificially appears systematically low in these plots.

triggers are shown in Fig. 1(b). For our analysis, we want to work with single triggers that are nearly 100% efficient when the hardest jet is in a given p_T range. Cross-checking Fig. 1(b) with Ref. [223], we define the trigger boundaries in Table II, where the $p_T > 250$ GeV bin uses either `Jet100U` or `Jet140U` depending on whether the latter is present. We see that lower p_T triggers have higher average prescale values as expected. Because each trigger selects a homogenous event sample, we can use the average prescale value for the assigned trigger when filling histograms, which is statistically preferable to using the individual event prescale values. For completeness, in Fig. 14 in Appendix A, we show the distribution of prescale values encountered for each trigger within their assigned p_T range.

Our event selection workflow is summarized in Table III. Starting from the 20 million events in the Jet Primary Dataset, we reduce the data set to about 82% by only including events that are in the official list of validated runs.

TABLE II. Assigned triggers for the hardest jet in a given p_T range, along with the average prescale value that determines subsequent histogram weights. Since the `Jet140U` trigger was not present for all of Run 2010B, we use `Jet100U` when needed for the highest p_T bin.

Hardest jet p_T	Trigger name	Events	⟨Prescale⟩
[85, 115] GeV	`HLT_Jet30U`	33,375	851.514
[115, 150] GeV	`HLT_Jet50U`	66,412	100.320
[150, 200] GeV	`HLT_Jet70U`	365,821	5.362
[200, 250] GeV	`HLT_Jet100U`	216,131	1.934
> 250 GeV	`HLT_Jet100U`	34,736	1.000
	`HLT_Jet140U`	177,891	1.000

Restricting to events that pass their assigned trigger in Table II drops the event sample to around 900,000 events, and this is used to define a skimmed data set. Requiring the loose jet quality criteria removes a small number of events, as does verifying that the AK5 jet provided by CMS matches those clustered by FastJet on the PFCs directly (see Secs. II C and II D). If the hardest jet passes $|\eta| < 2.4$, then it is used for substructure analyses (see Sec. II D). For later reference, Table III shows the number of events where the hardest jet has a valid two-prong substructure as determined by soft-drop declustering (see Sec. IV).

In the plots below, we always present normalized histograms in order to suppress fixed-order QCD corrections to the overall jet production rate. While knowledge of the total luminosity is therefore not needed for our study, it is still instructive to try to extract luminosity information from the CMS open data. The AOD files provide the integrated luminosities achieved during each luminosity block, such that the sum over blocks should give the total luminosity. Unfortunately, the AOD-extracted value of 309.5 pb^{-1} does not match the official recorded luminosity value of 31.79 pb^{-1} during Run 2010B [224,225].[5] This turns out to be a known limitation of the provided AOD files, though the AOD-extracted values do have the expected qualitative structures. Removing the overall vertical normalization to avoid confusion, the delivered and recorded integrated luminosities are shown in Fig. 2(a) and the cumulative distributions in Fig. 2(b). As expected, we see that Run 2010B occurred from September 22 to

[5]It is suspicious that the difference is very close to a factor of 10, but as far as we can tell, this is a coincidence.

TABLE III. Overall workflow to go from the events in the Jet Primary Dataset to the events used in our jet substructure analysis. The three steps above the first horizontal line indicate the steps included as part of event skimming. The next three steps are used for the `Hardest_Jet_Selection`. The final line is for events that pass the soft-drop requirement in Sec. IV.

	Events	Fraction
Jet Primary Dataset	20,022,826	1.000
Validated run	16,341,187	0.816
Assigned trigger fired (Table II)	894,366	0.045
Loose jet quality (Table V)	843,129	0.042
AK5 match	843,128	0.042
$\|\eta\| < 2.4$	768,687	0.038
Passes soft drop ($z_g > z_{\text{cut}}$)	760,055	0.038

October 29 in 2010, with a substantial ramp up of collected data over that two month period.

C. The MIT Open Data format

The output of `MODProducer` is a text-based MOD file, which contains a subset of the AOD data, similar in spirit to the mini-AOD format being developed internally within CMS [226]. The MOD format is intended to be lightweight, easy to parse, and human readable, so it uses space-separated entries with keyword labels. While there are other text-based file formats used within high energy physics, such as HepMC [227] and LHEF [228,229], they are primarily intended for use with Monte Carlo generators and therefore do not have a standard way to incorporate CMS-specific information like triggers and JEC factors. Instead of trying to augment these existing file formats and risk breaking backward compatibility, we decided to develop our own MOD format. Ultimately, one could envision a standard file format for open collider data, since the MOD file already contains much of the information common to all collider analyses. In our analysis, we use the MOD format not only for experimental data but also for data generated from parton showers (see Sec. II E). As a cross-check of the results in this paper, we also performed an independent analysis using an internal ROOT-based framework.

A typical MOD event consists of the following six keywords:

(i) `BeginEvent`: A header that indicates the source of the event: CMS open data or parton shower generator.[6] It also includes the version number of the MOD format (currently version 5).

(ii) `Cond`: Basic information about the run and event conditions, including run and event numbers, a timestamp, the number of reconstructed primary vertices, and information about the luminosity block.

(iii) `Trig`: List of all triggers used in the Jet Primary Dataset, their associated prescale factors, and flags indicating whether a given trigger fired for that event.

(iv) `AK5`: List of anti-k_t $R = 0.5$ jets provided by CMS. In addition to the jet four-momentum, CMS provides a JEC factor, a jet area value [230], and information about jet quality.

(v) `PFC`: List of PFCs, with their four-momenta and particle identification codes.

(vi) `EndEvent`: A footer indicating the end of an event.

An example MOD event is included in the `arXiv` source files of this paper. For MOD files coming from parton shower generators, we replace `Cond` and `Trig` with event weight information and rename `PFC` to `Part` to indicate truth-level particles. The MOD format can be easily extended to accommodate additional information in the future.

The list of valid particle identification codes for the PFCs is given in Table IV, along with their prevalence in the hardest jet sample ($p_T > 85$ GeV, $|\eta| < 2.4$). These codes, determined by the CMS particle flow algorithm, are inspired by the Monte Carlo particle number scheme in Ref. [231]. For example, all charged hadron candidates are assigned a code of ± 211 corresponding to charged pions, which are more prevalent than charged kaons. Neutral pions, which decay as $\pi^0 \to \gamma\gamma$, are typically reconstructed as one or two photon candidates with code 22. Neutral hadron candidates are assigned code 130 corresponding to K-long. Electrons (± 11) and muons (± 13) are relatively rare in our jet sample.

Although the AK5 jets are derived from clustering the PFCs, we need to separately extract the AK5 jets provided by CMS in order to obtain JEC factors and impose jet quality cuts. Throughout our analysis, we impose the recommended "loose" jet quality cut; see Table V in Appendix A. Due to numerical rounding issues when outputting the MOD text file, the AK5 jets from CMS and ones we cluster ourselves from the PFCs can be subtly different, though if we restrict our attention to the hardest jet, this is a rare effect that has almost no impact in our analysis (see further discussion in Sec. II D).[7]

[6] We also generated samples using fast detector simulation. As already mentioned, because of apparent oversmearing by the default CMS-like Delphes configuration [215], we do not show any fast simulation results in this paper.

[7] Alternatively, we could have decided to directly identify the PFC constituents of the AK5 jet using CMSSW. This leads to a different numerical rounding issue where the jet is not the four-vector sum of its constituents. These issues could have been avoided by not relying on text-based output, at the expense of requiring ROOT dependencies in MODAnalyzer. Our internal ROOT-based analysis framework encounters no numerical rounding issues.

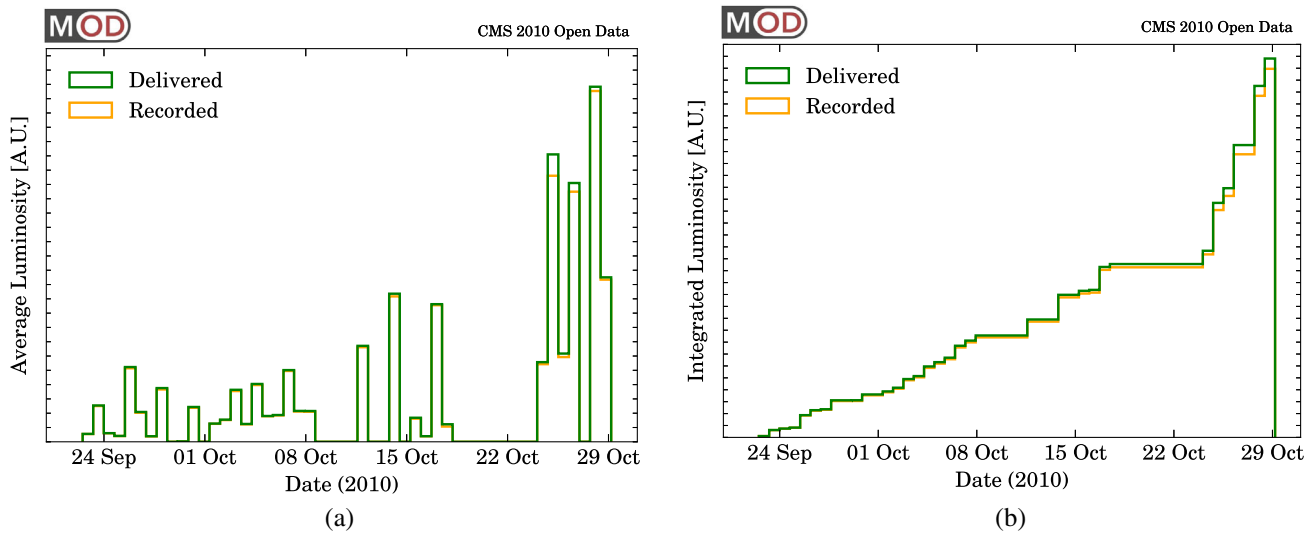

FIG. 2. Integrated luminosity collected by the CMS experiment during Run 2010B, plotted (a) per day and (b) cumulative. Because the luminosity information provided in the AOD files does not match the official recorded integrated luminosity of 31.8 pb^{-1}, we suppress the vertical normalization in these plots. The qualitative features shown here do agree with the official Run 2010B luminosity profile.

After running `gzip` for compression, the final MOD files are roughly 10 times smaller than the corresponding AOD files (which are already in a compressed ROOT format). Furthermore, if we restrict to a skimmed data set where the hardest jet has $p_T > 85$ GeV and the assigned trigger fired, we reduce the 198.8 gigabytes of compressed MOD files down to 11.6 gigabytes. This is small enough to easily fit on a flash drive.

D. Analysis tools

With the MOD files in hand, we are no longer tied to CMSSW. In order to leverage existing jet substructure tools, we built an external analysis framework in C++ based on the FastJet package [221]. This framework, called MODAnalyzer, is available from a GitHub repository [232], which also includes the Python histogramming and plotting tools used for this paper. For the soft-drop studies in Sec. IV, we use the RecursiveTools 1.0.0 package from FastJet contrib 1.019 [233].

TABLE IV. Valid particle identification codes for PFCs, with their most likely hadron interpretation. The total counts are taken from the sample of a hard central jet with $p_T > 85$ GeV and $|\eta| < 2.4$. In the forward region with $|\eta| > 2.4$, one also finds code 1 (for forward hadron candidate) and code 2 (for forward electron/photon candidate). The last column lists the counts after the $p_T^{\text{min}} = 1.0$ GeV cut derived in Fig. 3.

Code	Candidate	Total count	$p_T > 1$ GeV
11	Electron (e^-)	32,917	32,900
−11	Positron (e^+)	32,984	32,968
13	Muon (μ^-)	12,941	12,653
−13	Antimuon (μ^+)	13,437	13,110
211	Positive hadron (π^+)	6,908,914	5,183,048
−211	Negative hadron (π^-)	6,729,328	5,027,146
22	Photon (γ)	9,436,530	4,805,173
130	Neutral hadron (K_L^0)	2,214,385	1,658,892

The structure of MODAnalyzer mirrors the structure of the MOD files. The core class is `Event`, which is not only a container for all of the event information but also handles parsing of the MOD files and selecting the assigned trigger for the hardest jet. The `Cond` and `Trig` MOD entries are stored in `Condition` and `Trigger` classes. The `AK5` and `PFC` MOD entries are stored as FastJet `PseudoJet` objects. To amend these `PseudoJets` with additional MOD-specific information, we define `InfoCalibratedJet` and `InfoPFC` classes that inherit from FastJet's `UserInfoBase`. Apart from the `Event` class, the elements of MODAnalyzer are relatively lightweight, since much of the required functionality is already provided by FastJet.

The main complication in processing the MOD files is handling the duplicate jet information. Within MODAnalyzer, we have two types of jets: AK5 jets clustered by CMS and anti-k_t $R = 0.5$ jets clustered internally from the PFCs. Note that the AK5 jets are associated with JEC factors and jet quality criteria, whereas the internally clustered jets are not, so we cannot discard the AK5 jets completely. To define the hardest jet in the event (i.e. the "trigger jet"), we use the AK5 jet sample from CMS, rescaling the jet p_T values by the appropriate JEC factors and checking whether the assigned trigger fired. At this point, we remove events where the trigger jet fails the loose jet quality cut. We then find the internal PFC jet that is closest in rapidity-azimuth to the trigger jet. If this internal jet has the same number of constituents as the trigger jet and if the four-momenta agree up to 1 MeV

precision (after rescaling the internal jet by the same JEC factor), then we declare a match and perform all subsequent analyses on the internal jet. In rare cases where there is no match, we discard the event, though this only affects 1 event out of 843,129 in our analysis (see Table III).

Within MODANALYZER, we have a few ways to speed up the workflow. A large fraction of MOD events are unsuitable for analysis, mostly because the hardest jet was below the 85 GeV minimum p_T threshold set in Table II. We can therefore perform event skimming, where we read in each MOD file and generate a new MOD file with only events where the assigned trigger fired.[8] Similarly, because our analysis is only based on the hardest jet in the event, we can output MOD files with a `Hardest_Jet_Selection` header, where only the PFC constituents of the hardest jet are stored, and the minimally required `Trig`, `Cond`, and `AK5` information is consolidated under the `1JET` keyword.

After `gzip` compression, the `Hardest_Jet_Selection` MOD files only take 725 megabytes, which is small enough that we plan to make the files publicly available ourselves through DSPACE@MIT.[9] This reduced MOD file can be used directly with MODANALYZER, or one could build an alternative MOD analysis framework.

E. Parton shower generators

For the initial 2010 CMS open data release, no simulated Monte Carlo data sets were provided. In order to compare jet substructure results from open data with theoretical predictions, we use three parton shower generators: PYTHIA 8.219 [235], HERWIG 7.0.3 [236], and SHERPA 2.2.1 [237]. For each generator, we use the default settings for dijet production, since this is the process that dominates the single-jet triggers. To efficiently populate the full phase space, we use a p_T-weighted event generation strategy, which is highly efficient for jet production with $p_T > 85$ GeV, allowing us to use a single parton shower run to probe multiple p_T ranges. Our analyses are based on the raw output of the parton shower generators, without any detector simulation.[10]

Each generator outputs to HEPMC format [227], which we then convert to the same MOD file format used for the open data, suitably modified to eliminate CMS-specific information like triggers, luminosity, and JEC factors. After event skimming and applying `Hardest_Jet_Selection`, the MOD files from the open data and the parton shower generators look essentially identical, such that the same workflow can be used for all sources.

Because the parton showers do not include detector effects by default, we have to be careful in drawing conclusions about agreement or disagreement with the open data. For example, depending on the kinematics, the CMS particle flow reconstruction can sometimes reconstruct $\pi^0 \to \gamma\gamma$ as a single "photon" instead of two photons, which can affect jet substructure observables like constituent multiplicity.[11]

To partially account for the finite energy resolution of the CMS detector, we impose a restriction of $p_T^{\min} = 1.0$ GeV on each PFC (or truth-level particle in the case of the parton showers). This cut is motivated by Fig. 3, which suggests that PFCs below 1 GeV are subject to inefficiencies and mismeasurements. Crucially, this $p_T^{\min}$ restriction is only imposed for substructure observables; the original jet kinematics are given by all PFCs with the CMS-provided JEC factors. This universal $p_T^{\min}$ cut is similar in spirit to the SOFTKILLER approach to pileup mitigation [238].

Comparing Fig. 3(a) for neutral PFCs to Fig. 3(b) for charged PFCs, we see comparatively smaller differences between the CMS open data and the parton shower for charged PFCs; this will also be reflected in the substructure studies below. For this reason, we always perform cross-checks with track-based variants to address the finite granularity of the CMS calorimeter. Since the particle flow algorithm uses information from both the tracker and the calorimeter, the angular resolution of charged particles is much better than for neutral particles. This allows us to test whether there are large distortions to jet substructure observables from finite calorimeter cell size, especially for soft-dropped observables which probe the collinear core of the jet.[12] These track-based variants exploit the excellent track resolution of the CMS detector at the expense of losing neutral particle information, but since almost all of our substructure observables we study are dimensionless, the impact of switching to track-based variants is mild (see also [239,240]).

III. HARDEST JET PROPERTIES

We now present basic kinematic and substructure observables for the hardest p_T jet in an event, comparing CMS open data to parton shower generators. Unless otherwise stated, the jet p_T values always include the

[8]For the trigger and luminosity studies in Sec. II B, we of course had to use the unskimmed MOD files.

[9]The CMS open data are released under the Creative Commons CC0 waiver [234]. If you use the `Hardest_Jet_Selection` MOD files as part of an analysis, please cite the CMS Jet Primary Dataset [76] as well as this paper.

[10]As mentioned in the Introduction, we attempted to use the fast detector simulation tool DELPHES 3.3.2 [215], but the default CMS-like detector settings were intended to be used for jet studies, not jet substructure studies. In the future, since DELPHES does have a rudimentary version of particle flow reconstruction, it should be possible to tune DELPHES to match published CMS jet substructure results.

[11]One could partially mitigate this effect by forcing the π^0 to be stable within the generators, but this is not a replacement for a real detector simulation.

[12]We also tried preclustering the jet into small subjets as a way to mimic a finite angular resolution, but this simply led to increased smearing without improved agreement between data and parton showers.

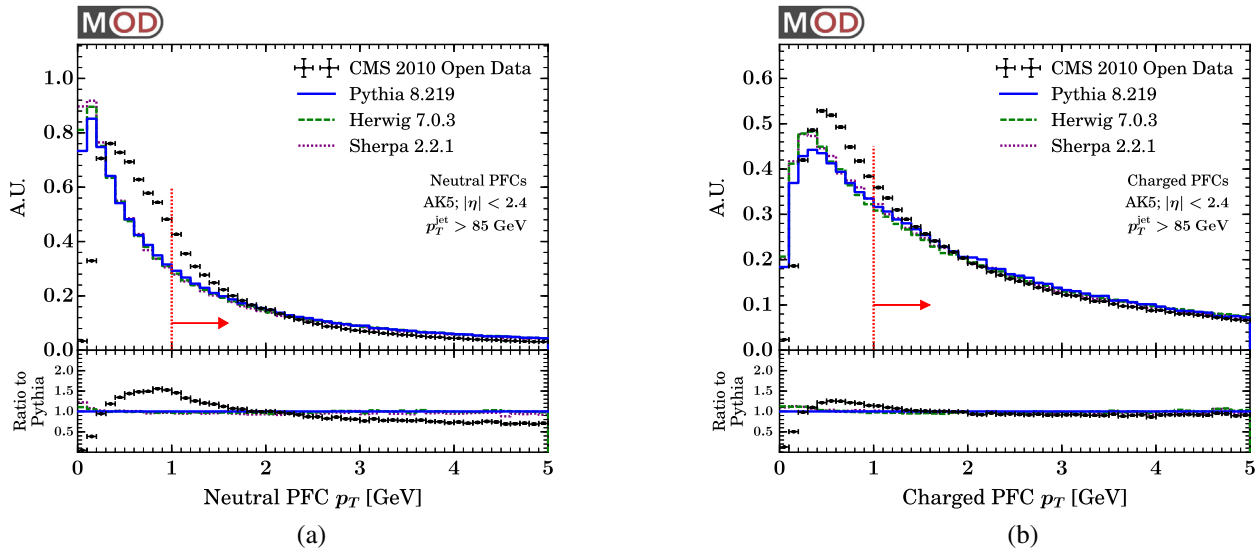

FIG. 3. Transverse momentum spectrum of raw PFCs, for (a) neutral candidates and (b) charged candidates. These histograms are populated only with PFCs from the hardest jet in the stated jet p_T range, comparing the CMS open data to parton shower generators. The cuts used in our jet substructure studies are $p_T^{\text{min}} = 1.0$ GeV, applied to both neutral and charged PFCs. For this and all remaining plots in this paper, one must keep in mind that the detector-level CMS open data and the particle-level parton showers are not directly comparable. See Fig. 16 in Appendix B for a version of this figure with an extended p_T range.

appropriate JEC factors, and we restrict our attention to jets with $|\eta| < 2.4$ and $p_T > 85$ GeV. Following the 2010 CMS default, the anti-k_t jet radius is always $R = 0.5$. In the text, we primarily show distribution for $p_T > 150$ GeV in order to avoid the large prescale values associated with the `HLT_Jet15U/Jet30U` triggers. In the `arXiv` source for this paper, each figure corresponds to a multipage file that has distributions for the full $p_T > 85$ GeV range, as well as for each of the p_T ranges defined in Table II.

A. Jet kinematics

The p_T spectrum for the hardest jet is shown in Fig. 4(a), going down to the 85 GeV threshold set by the lowest trigger in Table II. We see excellent agreement with parton shower predictions. As shown in Fig. 4(b), this good agreement is only possible because proper JEC factors were applied. Because we plot normalized histograms and because the p_T spectrum is steeply falling, the impact of the JEC factors is not so apparent at low p_T, but becomes increasingly visible going to higher p_T.

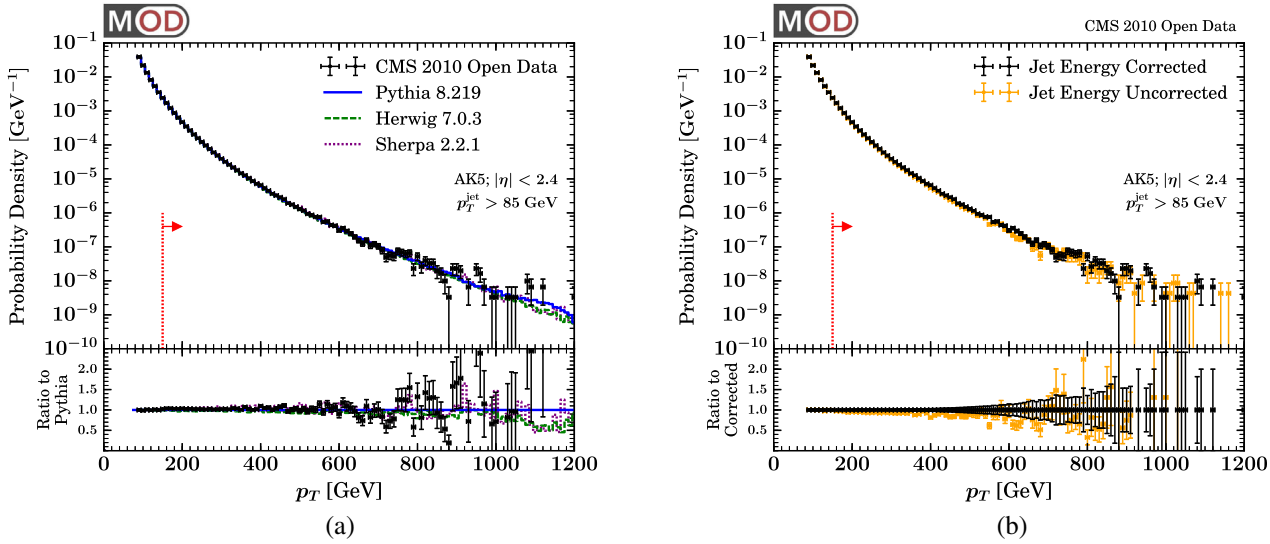

FIG. 4. (a) Hardest jet p_T spectrum, comparing the CMS open data with PYTHIA 8.219, HERWIG 7.0.3, and SHERPA 2.2.1. The maximum jet p_T in the Jet Primary Dataset is 1277 GeV. (b) Hardest jet p_T before and after applying the appropriate JEC factors. Because these are normalized histograms with the same $p_T > 85$ GeV cut, the mismatch in JEC values is only apparent at high p_T.

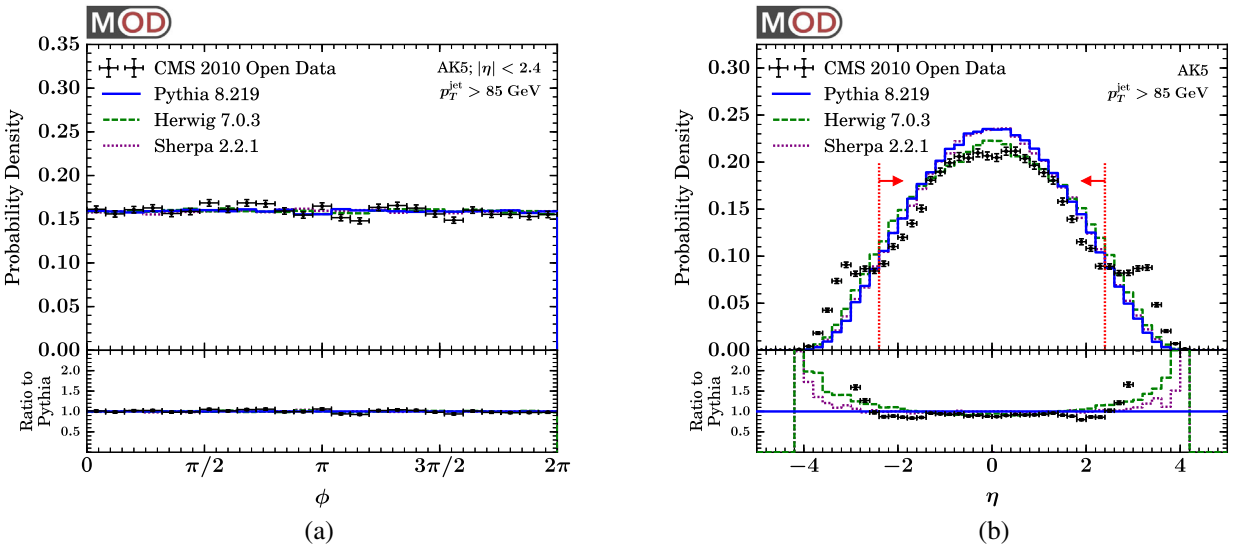

FIG. 5. (a) Azimuthal angle of the hardest jet, which is flat as desired. (b) Pseudorapidity spectrum for the hardest jet. Note the population of anomalous jets at $|\eta| > 2.4$, coming from the edge of tracking acceptance, which is why we enforce $|\eta| < 2.4$ in our analysis.

Turning to angular information, we show the jet azimuthal spectrum in Fig. 5(a), which is flat as expected. For the jet pseudorapidity distribution in Fig. 5(b), central jets with $|\eta| < 2.4$ match parton shower expectations within uncertainties. We see, however, a population of jets at $|\eta| > 2.4$ above parton shower expectations. These are most likely jet fakes that are able to erroneously pass the jet quality criteria due to the lack of tracking information at forward rapidities. For this reason, we restrict our attention to jets with $|\eta| < 2.4$ in our substructure studies.[13]

B. Basic substructure observables

The most basic jet substructure observable is the multiplicity of jet constituents, though this is very sensitive to the details of CMS's particle flow reconstruction. As mentioned in Sec. II E, we impose a cut of $p_T^{\min} = 1.0$ GeV on each PFC to avoid counting very soft particles that might not be efficiently reconstructed. That said, CMS cannot resolve arbitrarily small angles and therefore particles can be merged by the particle flow algorithm, especially for $\pi_0 \to \gamma\gamma$. For this reason, without a proper detector model, one has to be careful drawing conclusions from these substructure distributions. With that caveat in mind, we proceed to overlay the detector-level CMS open data with the particle-level parton shower generators.

In Fig. 6(a), we show the CMS open data constituent multiplicity distribution, which matches rather well to HERWIG and SHERPA. Once one restricts to charged particles in Fig. 6(b), however, the open data distribution shifts to lie closer to the PYTHIA distribution. We therefore conclude that the finite resolution of the calorimeter is an important detector effect that impacts jet substructure studies. Without a detector model, though, we cannot meaningfully comment on the correspondence between the open data and the parton showers, especially for distributions like multiplicity that are infrared and collinear (IRC) unsafe. The large differences between parton shower generators for charged particle multiplicity has been previously noted in e.g. Ref. [241], indicating that unfolded measurement of multiplicity should be used in parton shower tuning.

We can see the same sensitivity to detector effects for the observable p_T^D [86,242], defined as

$$p_T^D = \frac{\sqrt{\sum_{i\in\text{jet}} p_{Ti}^2}}{\sum_{i\in\text{jet}} p_{Ti}}. \quad (1)$$

This observable is soft safe but collinear unsafe and used in CMS's quark/gluon discrimination studies [97]. Using a logarithmic scale to emphasize the shape, we see in Fig. 6(c) that the CMS open data are at systematically higher values of p_T^D compared to parton shower predictions,

[13]Even if the jet axis satisfies $|\eta| < 2.4$, the jet constituents can extend to higher η values where the tracking degrades quickly. We explicitly checked that none of the jet substructure distributions studied below is substantially modified by taking the more conservative restriction of $|\eta| < 1.9$ (i.e. 2.4 minus the $R = 0.5$ jet radius). We further checked that there were no obvious pathologies for jets with $1.9 < |\eta| < 2.4$, even for observables like track multiplicity. For substructure studies, this tracking issue is subdominant to the choice of $p_T^{\min}$ in Fig. 3(b), in part because the jet cross section is falling with increasing $|\eta|$, so any tracking pathologies affect only a small portion of phase space. The CMS jet mass study in Ref. [87] considers $|y| < 2.5$ despite similar potential tracking issues.

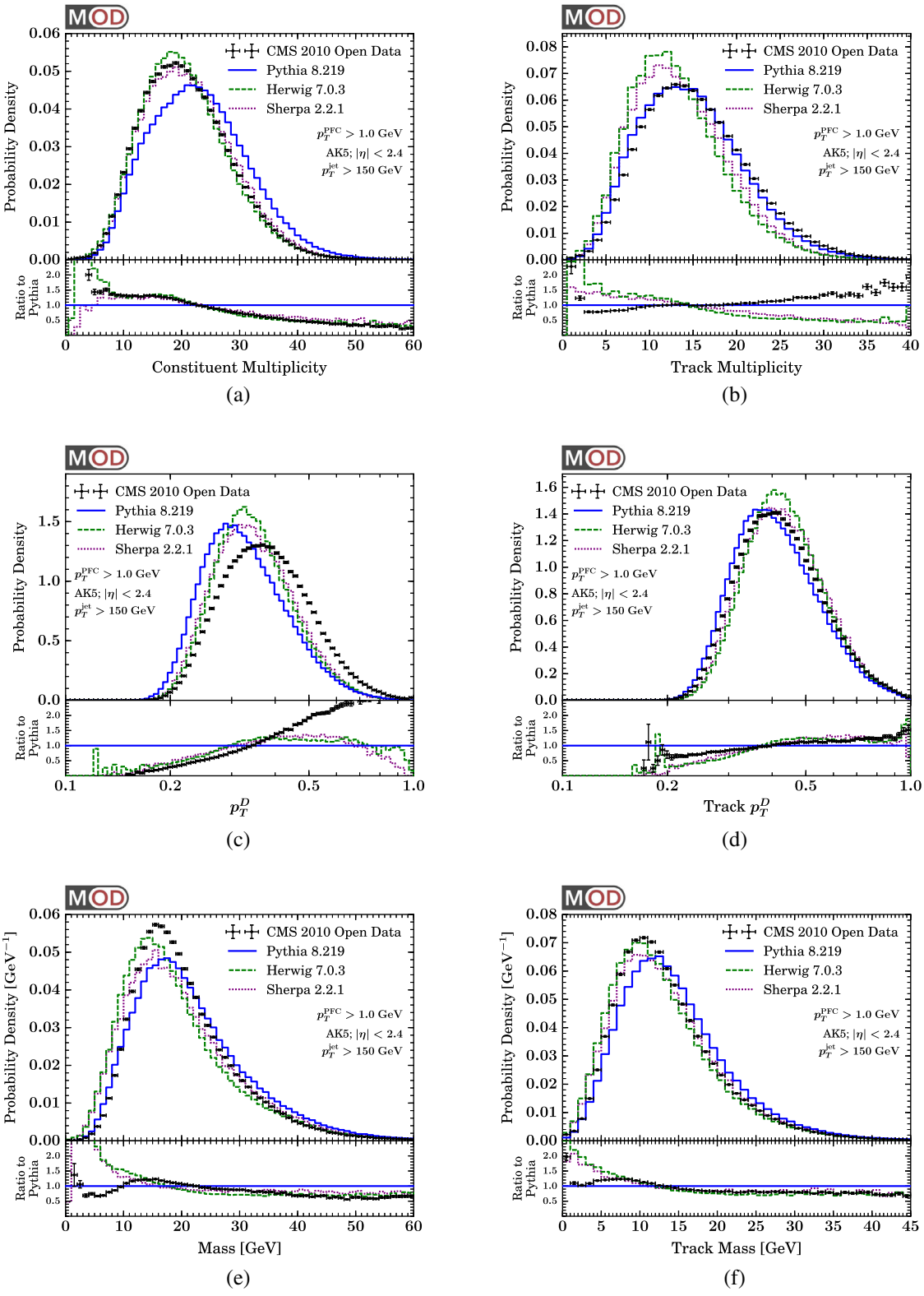

FIG. 6. Basic substructure observables for the hardest jet, using (left column) all PFCs and (right column) only charged PFCs, in both cases imposing a PFC cut of $p_T^{\mathrm{min}} = 1.0$ GeV. The observables are (a), (b) constituent multiplicity, (c), (d) p_T^D on a logarithmic scale and (e), (f) jet mass. We emphasize that in this and all subsequent figures, the distributions are not directly comparable, since the CMS open data have not been unfolded to account for detector effects and the parton shower generators have not been folded with detector effects. Similarly, only statistical uncertainties are shown for the open data.

again indicative of particle merging by the particle flow algorithm. Testing the track-based variant in Fig. 6(d), we see much better agreement between the open data and the parton shower generators, where the differences between detector-level and particle-level are comparable to the differences seen between generators.

For IRC-safe observables, we expect the impact of finite angular and energy resolution of the CMS detector to be less pronounced. In Fig. 6(e), the jet mass distribution agrees rather well between CMS open data and the parton showers, with differences again comparable to the differences between generators. Here, we have not applied the JEC factor to the mass distribution, since these are obtained after the PFC cut of $p_T^{\min} = 1.0$ GeV. In Fig. 6(f), we show the track-based variant (which is not corrected for the charged energy fraction), which shows similar agreement between the open data and the parton showers. While the lack of a detector model means that we cannot use the CMS open data to make quantitative statements about the jet mass distribution, we can say that the overall CMS detector performance is sufficient to draw qualitative conclusions about jet substructure distributions.

C. Jet angularities

A powerful way to study the radiation pattern of quark and gluon jets is to use jet angularities [14,68,243–245]. These are IRC-safe observables, defined as

$$e^{(\alpha)} = \sum_{i \in \text{jet}} z_i \theta_i^{\alpha}, \tag{2}$$

where

$$z_i = \frac{p_{Ti}}{\sum_{j \in \text{jet}} p_{Tj}}, \qquad \theta_i = \frac{R_i}{R}, \tag{3}$$

and R_i is the rapidity/azimuth distance to a recoil-free axis. Because the jet axis itself is sensitive to recoil [26,245–248], we use the winner-take-all axis [245,249,250] defined from Cambridge-Aachen (C/A) clustering [251,252].

By adjusting the value of α one can test radiation patterns mainly in the core ($\alpha < 1$) or periphery ($\alpha > 1$) of the jet. Three commonly used benchmarks are the Les Houches Angularity (LHA, $\alpha = 1/2$) [70,75]; jet width ($\alpha = 1$) [246,253,254]; and jet thrust ($\alpha = 2$) [255]. The corresponding distributions are shown in Fig. 7, plotted on a logarithmic scale to emphasize the behavior in the soft and collinear limit (i.e. small values of the angularities). Even though these are IRC-safe observables, we continue to place a cut of $p_T^{\min} = 1.0$ GeV on both the detector-level and particle-level constituents.

At large values of the angularities, the agreement between the CMS open data and the parton showers is rather good. At small values of the angularities where energy and angular resolution play an important role, the CMS open data are shifted to systematically higher values than the parton shower. Since the shift is less pronounced for the track-based variants, we suspect that the finite angular resolution of neutral PFCs is driving the bulk of the disagreement. For this reason, in the soft-drop study presented next, we have to be mindful of the challenge of resolving small angular scales using neutral particles.

IV. TWO-PRONG JET SUBSTRUCTURE

We now test the two-prong substructure of the hardest jet using soft-drop declustering [46]. This method has been used in both ATLAS [116] and CMS [105,136,142,148,149,154,155,160–169,175–179] jet studies, including a recent CMS heavy ion result [167]. There are also proposals to use soft drop to study the deadcone effect in top quarks [256] and gluon splitting to heavy flavor [257]. Here, we exploit the fact that soft drop is amenable to first-principle QCD calculations [46,206,258–260]. While there are a variety of different two-prong observables one could test on the CMS open data (e.g. N-subjettiness [22,23], energy correlation functions [26,49], and Qjet volatility [45,261]), soft drop has the advantage that it removes soft contamination from a jet, making it relatively robust to potential pileup and detector effects associated with soft particles.

As in the basic substructure analysis in Secs. III B and III C, we impose a restriction of $p_T^{\min} = 1.0$ GeV on all PFCs before passing them to the soft-drop algorithm. We again perform cross-checks with track-based variants which use only charged PFCs, which are expected to better resolve the small angular scales probed by soft drop.

A. Soft-drop declustering

The soft-drop algorithm reclusters the constituents of a jet using the C/A algorithm [251,252] to create an angular-ordered clustering tree. As shown in Fig. 8, soft drop then declusters the jet starting from the top of the tree, removing the softer p_T branch until a $1 \to 2$ branching is found that satisfies

$$z > z_{\text{cut}} \theta^{\beta}. \tag{4}$$

Here, z_{cut} is an energy fraction cut, β is an adjustable angular exponent, and the $1 \to 2$ kinematics are defined by

$$z = \frac{\min[p_{T1}, p_{T2}]}{p_{T1} + p_{T2}}, \qquad \theta \equiv \frac{R_{12}}{R}. \tag{5}$$

For the branching that passes the soft-drop condition, we denote the resulting kinematic observables by z_g and θ_g, which characterize the hard two-prong substructure of the jet. The g subscript is a reminder that these are groomed observables, subject to the soft-drop condition.

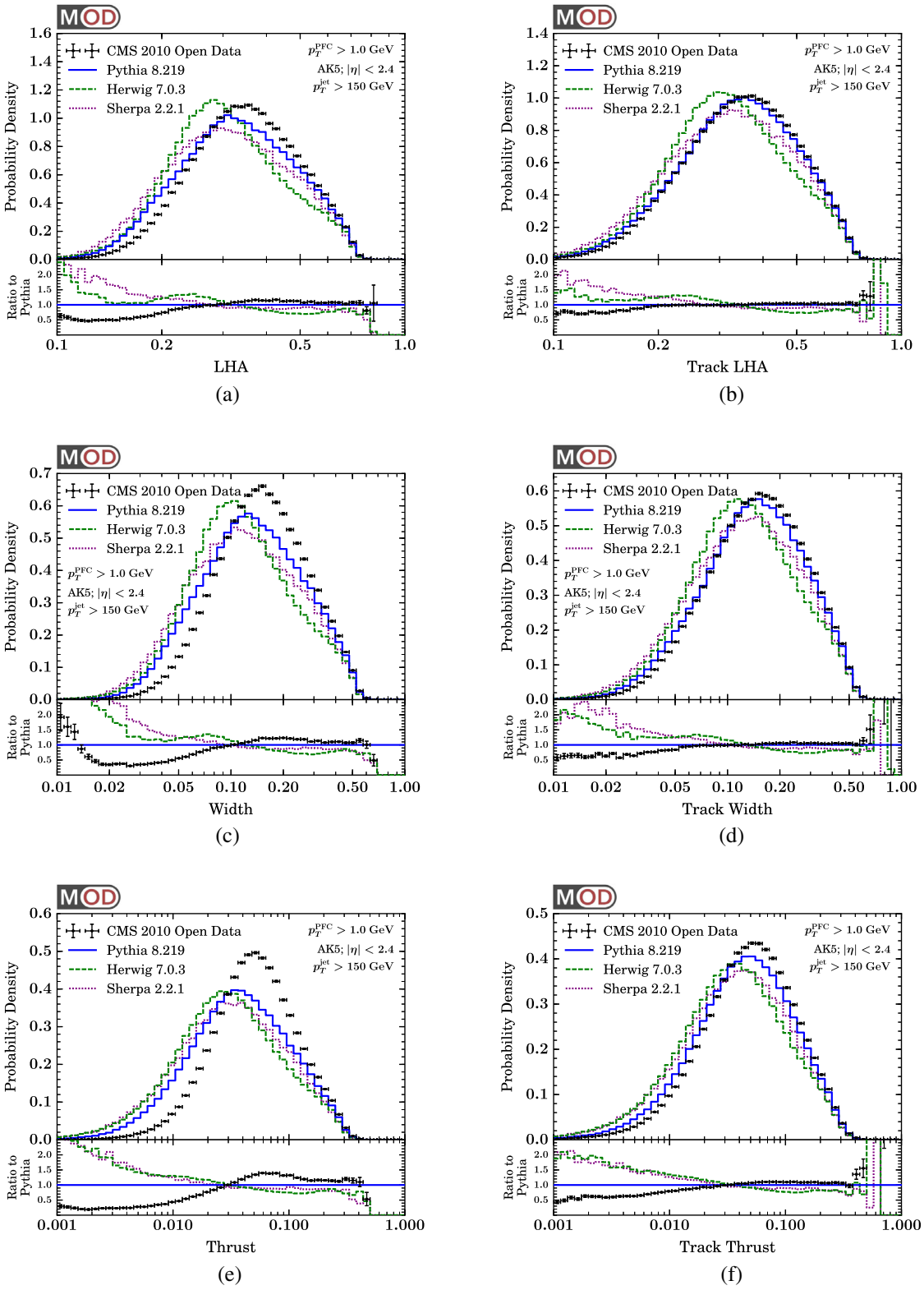

FIG. 7. Same as Fig. 6 but for the IRC-safe recoil-free jet angularities: (a), (b) LHA with $\alpha = 1/2$, (c), (d) jet width with $\alpha = 1$, and (e), (f) jet thrust with $\alpha = 2$. Once again we compare (left column) all particle distributions to (right column) track-only variants. Note the logarithmic scale of the distributions.

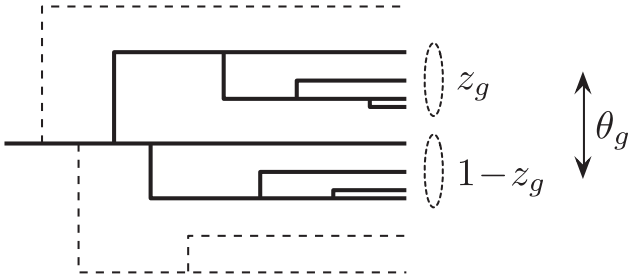

FIG. 8. Schematic of the soft-drop algorithm, which recursively removes branches from the C/A clustering tree if the momentum fraction z fails to satisfy $z > z_{\rm cut}\theta^{\beta}$. The g subscript indicates the final groomed kinematics.

In effect, soft drop simultaneously performs three tasks. First, it removes wide-angle soft contamination from jets, which helps mitigate the effect of jet contamination from ISR, UE, and pileup. Second, it dynamically changes the effective jet radius to match the size of the hard jet core. Third, it provides the two-prong kinematic observables z_g and θ_g, which can be used to perform foundational tests of QCD [167,206–209,257] as well as discriminate boosted W, Z, and Higgs bosons from ordinary quark/gluon jets [6,113]. In general, groomers like soft drop have an interesting interplay with discrimination variables [31,59,262].

In our study, we focus on the soft-drop parameters

$$z_{\rm cut} = 0.1, \qquad \beta = 0. \tag{6}$$

For this choice of β, the soft-drop condition reduces to $z > z_{\rm cut}$ and becomes independent of angular information. This then matches the behavior of the mMDT with $\mu = 1$ [203,204]. Without any explicit cut on θ_g, this enables us to probe rather small angular scales within the jet, though we need to be cognizant of the finite angular resolution of the CMS detector. We show distributions for five observables derived from soft drop:

$$z_g, \qquad \theta_g, \qquad e_g^{(1/2)}, \qquad e_g^{(1)}, \qquad e_g^{(2)}, \tag{7}$$

where

$$e_g^{(\alpha)} = z_g \theta_g^{\alpha} \tag{8}$$

is a single-emission groomed variant of the angularities introduced in (2).

B. MLL analytic predictions

In addition to parton shower predictions, we compare the CMS open data to first-principles QCD theory distributions made using the techniques of Refs. [46,205,206], working to modified leading-logarithmic (MLL) accuracy.

For the observable θ_g, it is convenient to express the probability distribution as

$$\frac{1}{\sigma}\frac{{\rm d}\sigma}{{\rm d}\theta_g} \equiv p(\theta_g) = \frac{\rm d}{{\rm d}\theta_g}\Sigma(\theta_g), \tag{9}$$

where the cumulative probability distribution $\Sigma(\theta_g)$ was calculated to MLL accuracy in Ref. [46]. For $\beta = 0$, the result for a parton of flavor i is

$$\Sigma_i^{\rm MLL}(\theta_g;\mu_\theta) = \exp\left[-\frac{2C_i}{\pi}\int_{\theta_g}^{1}\frac{{\rm d}\theta}{\theta}\int_{z_{\rm cut}}^{1/2}{\rm d}z\bar{\alpha}_s(z\theta\mu_\theta)\bar{P}_i(z)\right], \tag{10}$$

where C_i is the Casimir factor ($C_F = 4/3$ for quarks and $C_A = 3$ for gluons). At the lowest nontrivial order, the QCD splitting functions are

$$P_q(z) = \frac{1+(1-z)^2}{2z}, \tag{11}$$

$$P_g(z) = \frac{1-z}{z} + \frac{z(1-z)}{2} + \frac{n_f T_R}{2C_A}[z^2+(1-z)^2], \tag{12}$$

with $n_f = 5$ and $T_R = 1/2$; these appear in (10) in a symmetrized form,

$$\bar{P}_i(z) = P_i(z) + P_i(1-z). \tag{13}$$

The one-loop QCD running coupling is $\bar{\alpha}_s$, where the bar indicates that we have frozen the running below the IR scale $\mu_{\rm NP} \sim 1.0$ GeV,

$$\bar{\alpha}_s(\mu) = \alpha_s(\mu)\Theta(\mu-\mu_{\rm NP}) + \alpha_s(\mu_{\rm NP})\Theta(\mu_{\rm NP}-\mu). \tag{14}$$

The running coupling is evaluated at the canonical renormalization group scale

$$\mu_\theta = p_T R, \tag{15}$$

and we estimate uncertainties by varying both this scale and $\mu_{\rm NP}$ up and down by a factor of 2.

To get a physical distribution for θ_g, we need to determine the relative fraction of quark and gluon jets with our selection, such that the final cumulative distribution is

$$\Sigma^{\rm MLL} = f_q\Sigma_q^{\rm MLL} + f_g\Sigma_g^{\rm MLL}. \tag{16}$$

To determine the fractions f_q and f_g, we generate a leading-order (LO) sample of dijets using MadGraph5_aMC@NLO v 2.4.0 [263] with parton distribution functions (PDFs) given by NNPDF2.3 LO [264], extracting the average flavor composition from both jets. We set the renormalization and

factorization scales to the total transverse momentum of the dijet event,

$$\mu_h = p_{T1} + p_{T2}, \tag{17}$$

and vary this up and down by a factor of 2 to estimate uncertainties. Note that the renormalization scale does not affect the relative quark and gluon composition since it only rescales the total cross section by changing α_s. By contrast, the factorization scale does affect the flavor composition through the PDFs.

Strictly speaking, the above method for determining the quark/gluon fraction of the hardest jet is not IRC safe, since the flavor composition of the hardest jet at next-to-leading order (NLO) is no longer the same as the average flavor composition at LO. In practice, though, the hardest jet at NLO is more or less randomly determined from the two degenerate jets at LO, so the strategy used in this paper is sufficient for the current level of theoretical accuracy. There are various ways we could improve this procedure in a future analysis. Arguably the easiest method would be to study the inclusive jet spectrum instead of focusing on just the hardest jet in the event. While conceptually straightforward, it is technically more involved, since for dijet events close to a trigger boundary, the same event can have different assigned triggers for the two different jets. If we only wanted to study a single jet per event, we could use a dijet trigger for event selection but then only analyze the more central of the two jets, since that is a well-defined selection at LO.

To predict the probability distributions for z_g and $e_g^{(\alpha)}$, we use the strategy of Ref. [206]. Since $z_g = e_g^{(0)}$ (i.e. $\alpha = 0$), we can use the same method to calculate the remaining four observables in (7). We express the full probability distribution for $e_g^{(\alpha)}$ and θ_g,

$$p(e_g^{(\alpha)}, \theta_g) \equiv \frac{1}{\sigma} \frac{\mathrm{d}^2\sigma}{\mathrm{d}e_g^{(\alpha)} \mathrm{d}\theta_g}, \tag{18}$$

in terms of the probability for θ_g from (9) multiplied by the conditional probability for $e_g^{(\alpha)}$ given θ_g,

$$p(e_g^{(\alpha)}, \theta_g) = p(\theta_g) p(e_g^{(\alpha)} | \theta_g). \tag{19}$$

To obtain the probability for $e_g^{(\alpha)}$ alone, we simply integrate over all values of θ_g,

$$p(e_g^{(\alpha)}) = \int \mathrm{d}\theta_g p(\theta_g) p(e_g^{(\alpha)} | \theta_g). \tag{20}$$

To leading fixed order in the collinear limit, the conditional probability distribution is

$$p^{\text{LO-c}}(e_g^{(\alpha)} | \theta_g; \mu_z) = \frac{\bar{\alpha}_s(e_g^{(\alpha)} \theta_g^{1-\alpha} \mu_z) \theta_g^{-\alpha} \bar{P}_i(e_g^{(\alpha)} \theta_g^{-\alpha})}{\int_{z_{\text{cut}}}^{1/2} \mathrm{d}z \bar{\alpha}_s(z\theta_g \mu_z) \bar{P}_i(z)} \times \Theta(\theta_g^\alpha - 2e_g^{(\alpha)}) \Theta(e_g^{(\alpha)} - z_{\text{cut}} \theta_g^\alpha). \tag{21}$$

We note the dependence on a (different in principle) renormalization group scale,

$$\mu_z = p_T R, \tag{22}$$

which can be varied up and down by a factor of 2.

In summary, these theory distributions depend on four different scales,

$$\mu_{\text{NP}}, \quad \mu_\theta, \quad \mu_h, \quad \mu_z, \tag{23}$$

which can be varied to estimate theoretical uncertainties. As established, these variations do yield properly normalized distributions. To estimate perturbative uncertainties, we take the envelope of all scale variations, noting that the envelope will not, in general, be normalized.

There are two known effects which are not included in our theoretical uncertainty estimates. The first is genuine nonperturbative corrections. The above distributions are calculated perturbatively, with only the frozen coupling in (14) acknowledging the impact of nonperturbative physics. When z_g or θ_g are dominated by nonperturbative dynamics, though, these perturbative distributions can no longer be trusted. For double-differential distributions, this occurs when

$$z_g \theta_g \lesssim \frac{\Lambda}{p_T R}, \tag{24}$$

where $\Lambda \sim \mathcal{O}$ (GeV) and p_T is the lowest value in the plotted range. Projecting to the single observables, nonperturbative dynamics becomes relevant when

$$\theta_g \lesssim \frac{\Lambda}{z_{\text{cut}} p_T R}, \qquad e_g^{(\alpha)} \lesssim \max\{1, z_{\text{cut}}^{1-\alpha}\} \left(\frac{\Lambda}{p_T R}\right)^\alpha. \tag{25}$$

To indicate this in the plots below, we change the theory curves to a dashed style when nonperturbative modes dominate, using $\Lambda = 2$ GeV for concreteness. Note that z_g itself ($\alpha = 0$) is a collinear unsafe observable, so strictly speaking it is always sensitive to nonperturbative dynamics. Because z_g is a Sudakov safe [205,206] observable, though, the collinear singularity is regulated by the Sudakov form factor for θ_g. Also note that the theory calculations do not include the $p_T^{\text{min}} = 1.0$ GeV cut, which can be considered as part of the nonperturbative uncertainty.

The second missing effect is matching to fixed-order matrix elements. This is expected to have a small impact

because the jet radius is reasonably small and we are mostly focused on the $e_g^{(\alpha)} \ll 1$ limit. Nevertheless, there will be important fixed-order corrections to our theory predictions above the characteristic scale of $e_g^{(\alpha)} \simeq z_{\text{cut}}$, though we have not indicated that scale explicitly on the plots below. Indeed, there is a noticeable disagreement between our theory predictions and the open data/parton showers in the fixed-order regime, especially for $\theta_g \to 1$ (as illustrated in Fig. 12). A detailed study of fixed-order corrections is beyond the scope of this paper, and would anyway require a proper IRC-safe definition of the measured jet.

C. Open data results

We start in Fig. 9 with the full two-dimensional distributions for $p(z_g, \theta_g)$ from the open data, compared to the MLL analytic results and the three parton showers. All of the distributions show a peak at small values of z_g and θ_g, corresponding to the soft and collinear singularities of QCD. This structure is explained in more detail in a companion paper [207]. In principle, the θ_g distribution could extend all the way to $\theta_g \to 0$, but it is regulated by the perturbative form factor in (10), nonperturbative hadronization corrections, and the finite angular resolution of the CMS detector. Note the expected cut at $z_g = z_{\text{cut}}$ from the soft-drop condition. The $z_g = \theta_g = 0$ bin indicates jets which only have one constituent after soft drop.

Because of the logarithmic nature of the soft/collinear singularities of QCD, it is instructive to also plot $p(z_g, \theta_g)$ on a logarithmic scale, shown in Fig. 10. The overall qualitative structure is similar between the CMS open data and the theory distributions, but there are visible differences, especially when nonperturbative physics is important. Specifically, in the parton shower generators there is a strong peak around $\theta_g \simeq 0.1$, which is suppressed in the CMS open data. It would be interesting to know whether the parton shower is exhibiting a physical structure that is simply washed out in the open data or if there is a pathology in the parton shower generators in this kinematic regime. Because this feature appears exactly where nonperturbative physics is expected to matter, the perturbative MLL distribution is not a useful guide to answer this question.

To better compare the open data to theory predictions, we now consider the projected observables from (7). We show both all-particle and track-only observables to highlight the impact of angular resolution. Strictly speaking, the MLL distributions from Sec. IV B are only valid for all-particle observables, but we show dashed versions of the same curves on the track-only plots for ease of comparison. One could imagine using the track function formalism [239,240] to make sensible track-based MLL predictions, but that is beyond the scope of the present work.

We start with z_g in Fig. 11, which is also studied in Refs. [167,207–209]. Especially for the track-only measurement, the agreement between all five distributions is remarkable. For the all-particle distributions, there is a noticeable excess in the CMS open data compared to the theory distributions at $z_g \simeq z_{\text{cut}}$, as well as an excess of events that failed the soft-drop procedure; both of these features could be explained by the degraded angular resolution for neutral particles. On a logarithmic scale, one can see that the z_g distribution is roughly flat, as expected from the singularity structure of the splitting functions in Eqs. (11) and (12).

We can get a better understanding of angular effects by looking at θ_g directly in Fig. 12. Not surprisingly, the largest differences between the MLL distribution and the parton showers occur in the regime where nonperturbative dynamics matters. Especially on the logarithmic scale, the feature at $\theta_g \simeq 0.1$ is prominent in the parton shower generators. Note that the CMS heavy ion analysis in Ref. [167] placed a cut of $R_g > 0.1$ ($\theta_g > 0.2$) to avoid modeling issues in the small θ_g regime. Given the relatively good agreement between the CMS open data and the parton shower generators in the track-based distributions, we do not see an immediate reason to distrust small θ_g values, and measurements of θ_g could indeed be relevant for parton shower tuning.

Turning to the groomed single-emission angularities $e_g^{(\alpha)}$, in Fig. 13 we see reasonable agreement between the CMS open data and the parton shower generators, especially for the track-based observables. The MLL distributions exhibit the expected kinks at $e_g^{(\alpha)} = z_{\text{cut}}$, but the slope below this kink value differs noticeably. For the p_T range shown, though, the location of the kink is not so far from the scale where nonperturbative physics dominates, so measurements with more energetic jets are needed to test whether or not there is any tension with perturbative predictions.

The above plots are only a subset of the soft-dropped distributions we have made with the CMS open data. In the `arXiv` source files, the plots in Figs. 11, 12 and 13 are part of a multipage file that not only has multiple jet p_T ranges, but also $z_{\text{cut}} = 0.05$ and $z_{\text{cut}} = 0.2$ distributions. We leave a study of alternative β values to future work. For completeness, in Appendix B we show soft-dropped versions of all of the substructure distributions from Sec. III. We also show the fractional change in the jet p_T due to soft drop, which was shown in have interesting analytic properties in Ref. [46]. Additional soft-dropped observables can be provided to interested readers upon request (or derived using the publicly available MOD software framework).

V. ADVICE TO THE COMMUNITY

From a physics perspective, our experience with the CMS open data was fantastic. With PFCs, one can essentially perform the same kinds of four-vector-based analyses on real data as one would perform on collisions from parton shower generators. Using open data has the potential to accelerate scientific progress (pun intended) by

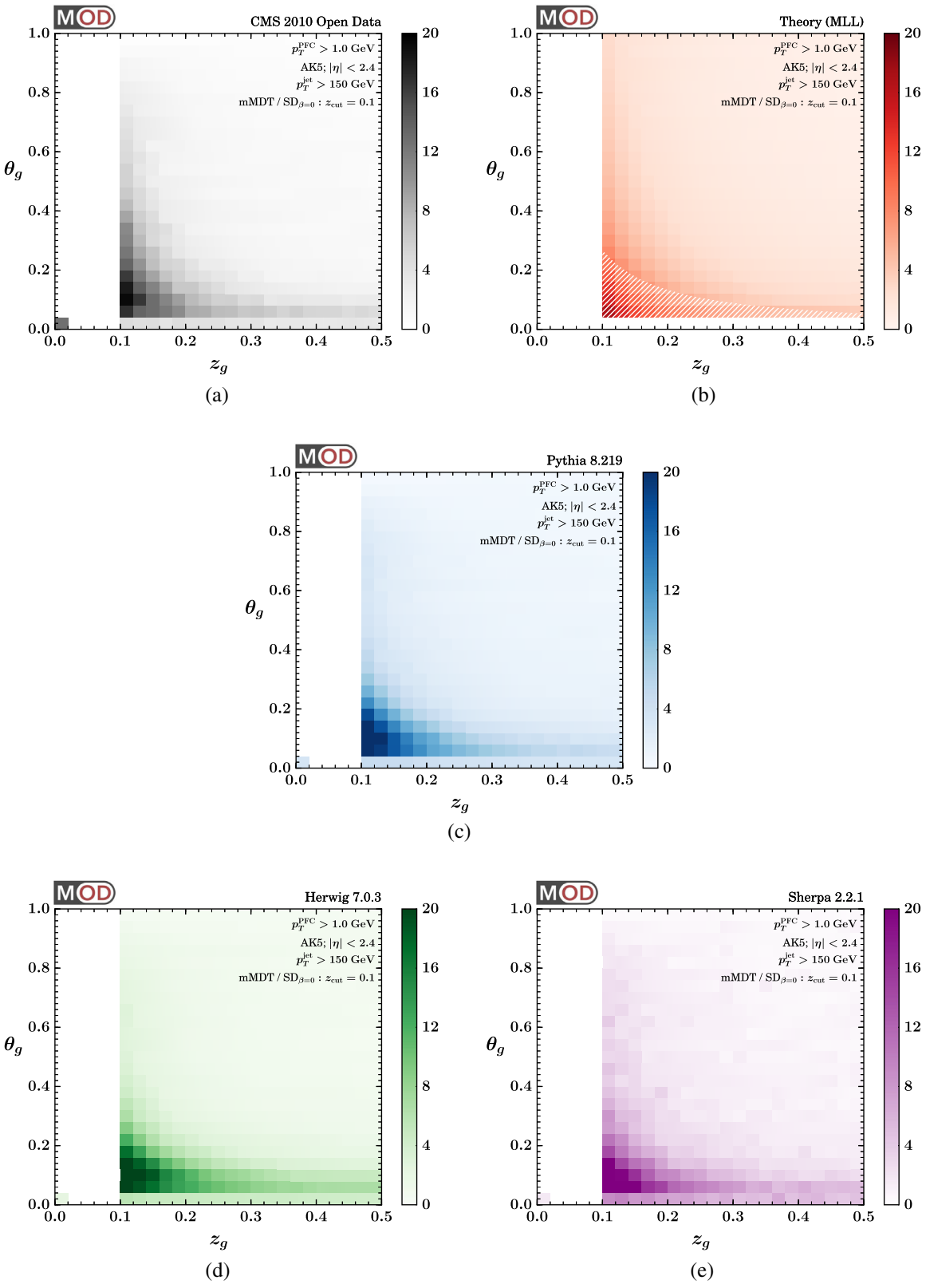

FIG. 9. Two-dimensional distributions of z_g versus θ_g from soft drop with $\beta = 0$ (i.e. mMDT with $\mu = 1$) in (a) CMS open data and (b) the MLL analytic prediction, compared to (c) PYTHIA, (d) HERWIG, and (e) SHERPA. Here, we are plotting the dimensionless probability density $p(z_g, \theta_g)$ whose integral is 1. The hard vertical cut corresponds to $z_g = z_{\mathrm{cut}}$, and the (0,0) entry corresponds to jets that fail the soft-drop procedure (not present for the analytic calculation). The white hashing in the MLL distribution corresponds to where nonperturbative physics dominates.

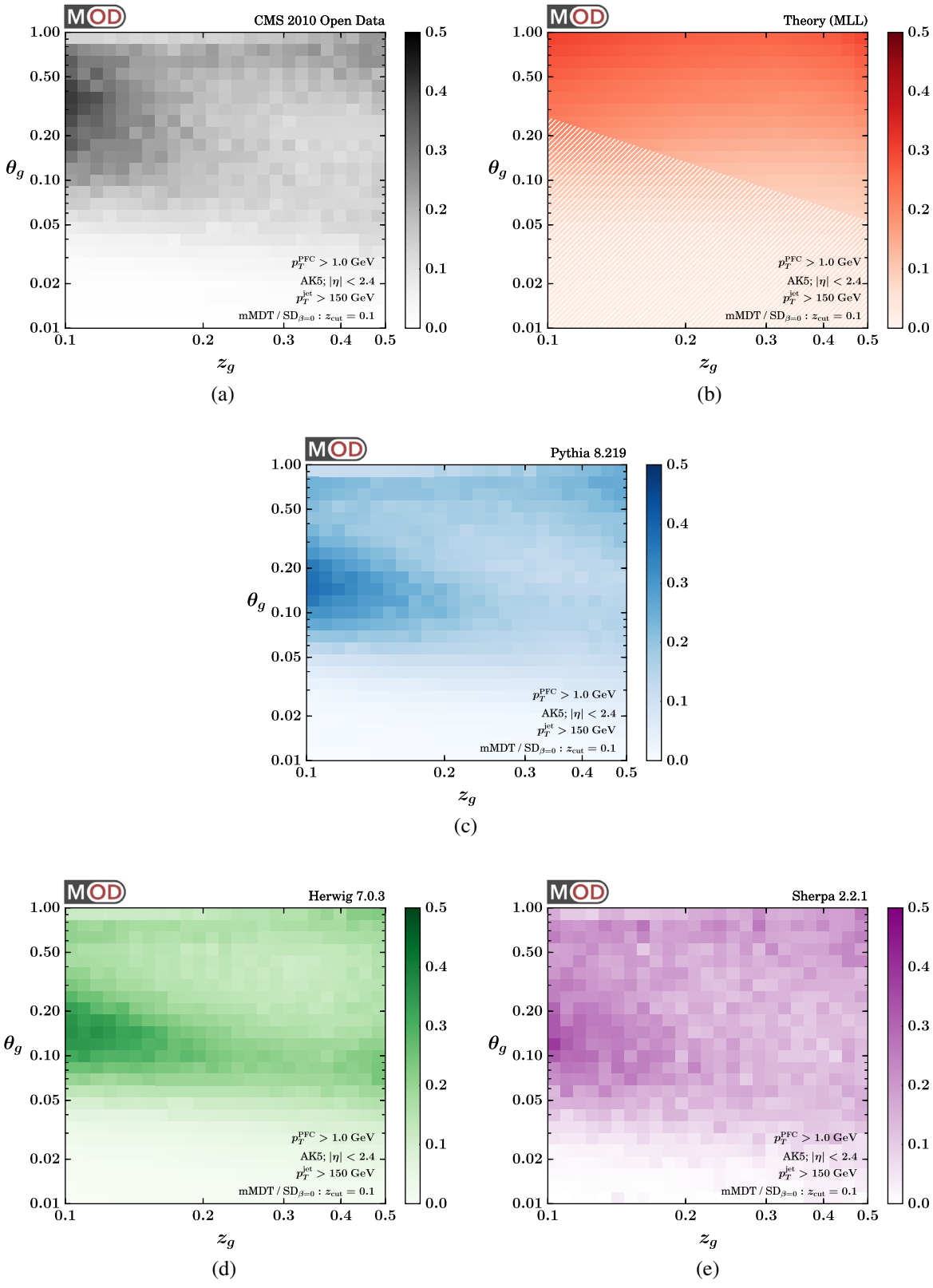

FIG. 10. Same as Fig. 9 but on a logarithmic scale to highlight the soft/collinear limit. Here, we are plotting the dimensionless probability density $p(\log z_g, \log \theta_g) = z_g \theta_g p(z_g, \theta_g)$ whose integral is 1 in logarithmic variables.

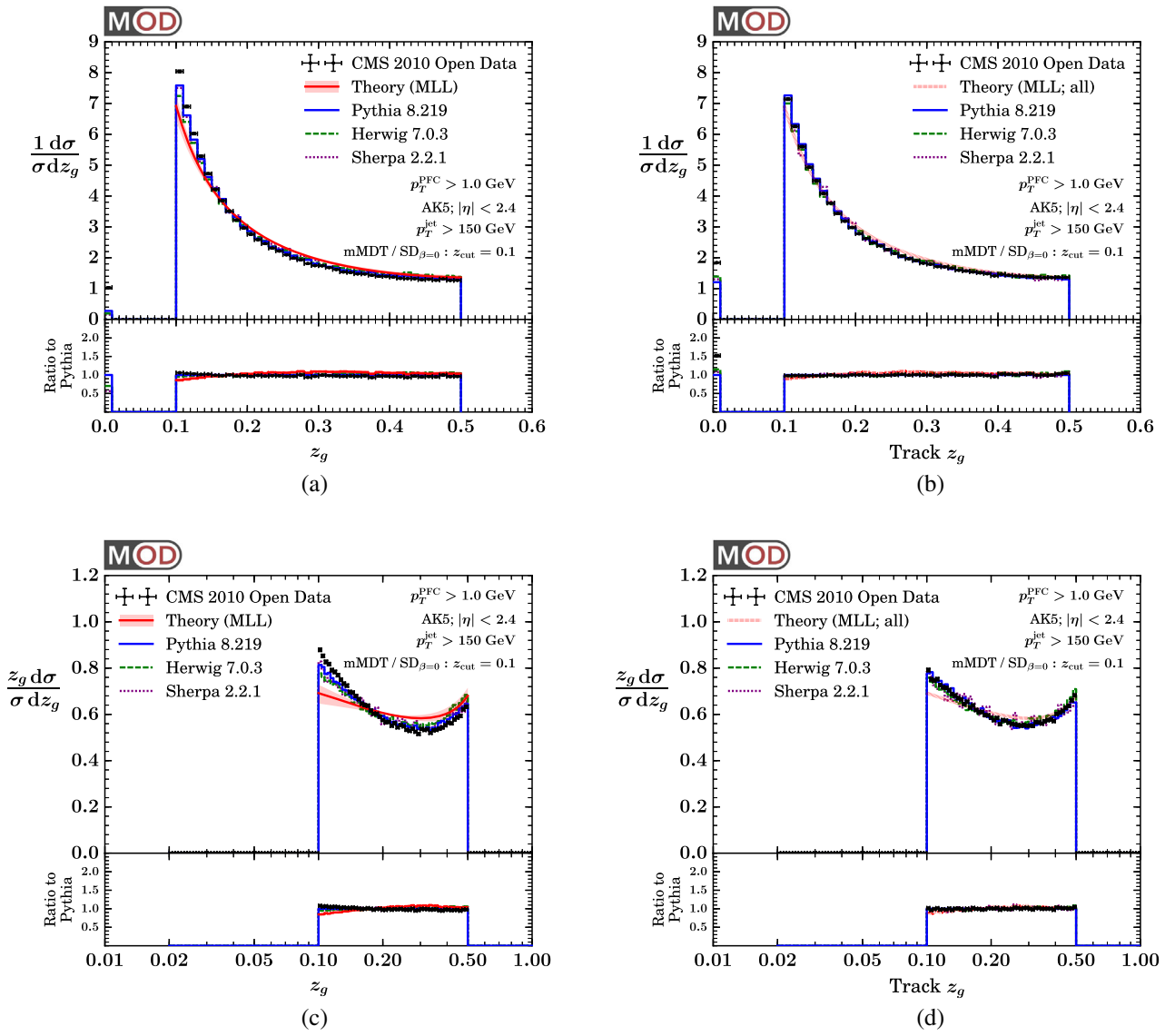

FIG. 11. Soft-dropped distributions for z_g using (left column) all particles and (right column) only charged particles. In this and subsequent plots, the MLL distributions are the same in both columns and do not account for the $p_T^{\mathrm{min}} = 1$ GeV cut on PFCs or the switch to charged particles (hence the dashed version on the right). The top row (a), (b) shows the linear distributions while the bottom row (c), (d) shows the logarithmic distributions.

allowing scientists outside of the official detector collaborations to pursue innovative analysis techniques. We hope that our jet substructure studies have demonstrated both the value in releasing public data and the enthusiasm of potential external users. We encourage other members of the particle physics community to take advantage of this unique data set.

From a technical perspective, though, we encountered a number of challenges. Some of these challenges were simply a result of our unfamiliarity with the CMSSW framework and the steep learning curve faced when trying to properly parse the AOD file format. Some of these challenges are faced every day by LHC experimentalists, and it is perhaps unreasonable to expect external users to have an easier time than collaboration members. Some of these challenges (particularly the issue of detector-simulated samples) have been partially addressed by the 2011A CMS open data release [216]. That said, we suspect that some issues were not anticipated by the CMS open data project, and we worry that they have deterred other analysis teams who might have otherwise found interesting uses for open data. Therefore, we think it is useful to highlight the primary challenges we faced, followed by specific recommendations for how potentially to address them.

A. Challenges

Here are the main issues that we faced in performing the analyses in this paper.

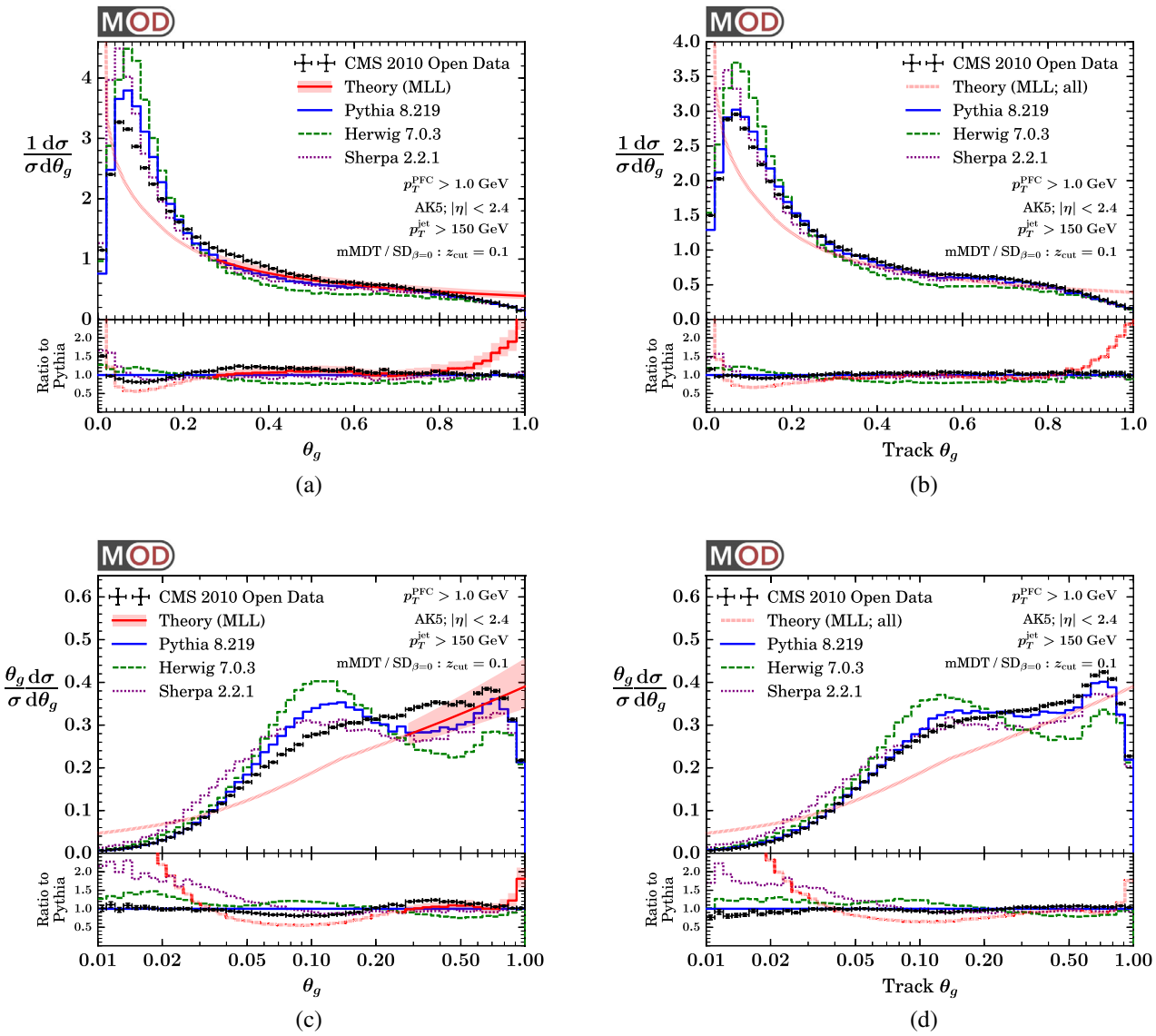

FIG. 12. Same as Fig. 11 but for θ_g. For the MLL distributions, the region where nonperturbative dynamics matters is indicated by the use of dashing. We do not indicate the regime where fixed-order corrections matter, since we have no first-principles estimate for the transition point.

(i) *Slow development cycle*. As CMSSW novices, we often needed to perform run-time debugging to figure out how specific functions worked. There were two elements of the CMSSW workflow that introduced a considerable lag between starting a job and getting debugging feedback. The first is that, when using the XRootD interface, one has to face the constant overhead (and inconstant network performance) of retrieving data remotely. The second is that, as a standard part of every CMS analysis, one has to load configuration files into memory. Loading `FrontierConditions_GlobalTag_cff` (which is necessary to get proper trigger prescale values) takes around 10 minutes at the start of a run. For most users, this delay alone would be too high of a barrier for using the CMS open data. By downloading the AOD files directly and building our own MOD file format, we were able to speed up the development cycle through a lightweight analysis framework. Still, creating the `MODProducer` in the first place required a fair amount of trial, error, and frustration.

(ii) *Scattered documentation*. Though the CMS open data use an old version of CMSSW (v4.2 compared to the latest v9.0), there is still plenty of relevant documentation available online. The main challenge is that it is scattered in multiple places, including online TWiki pages, masterclass lectures, thesis presentations, and GitHub repositories. Eventually, with help from CMS insiders, we were able to figure

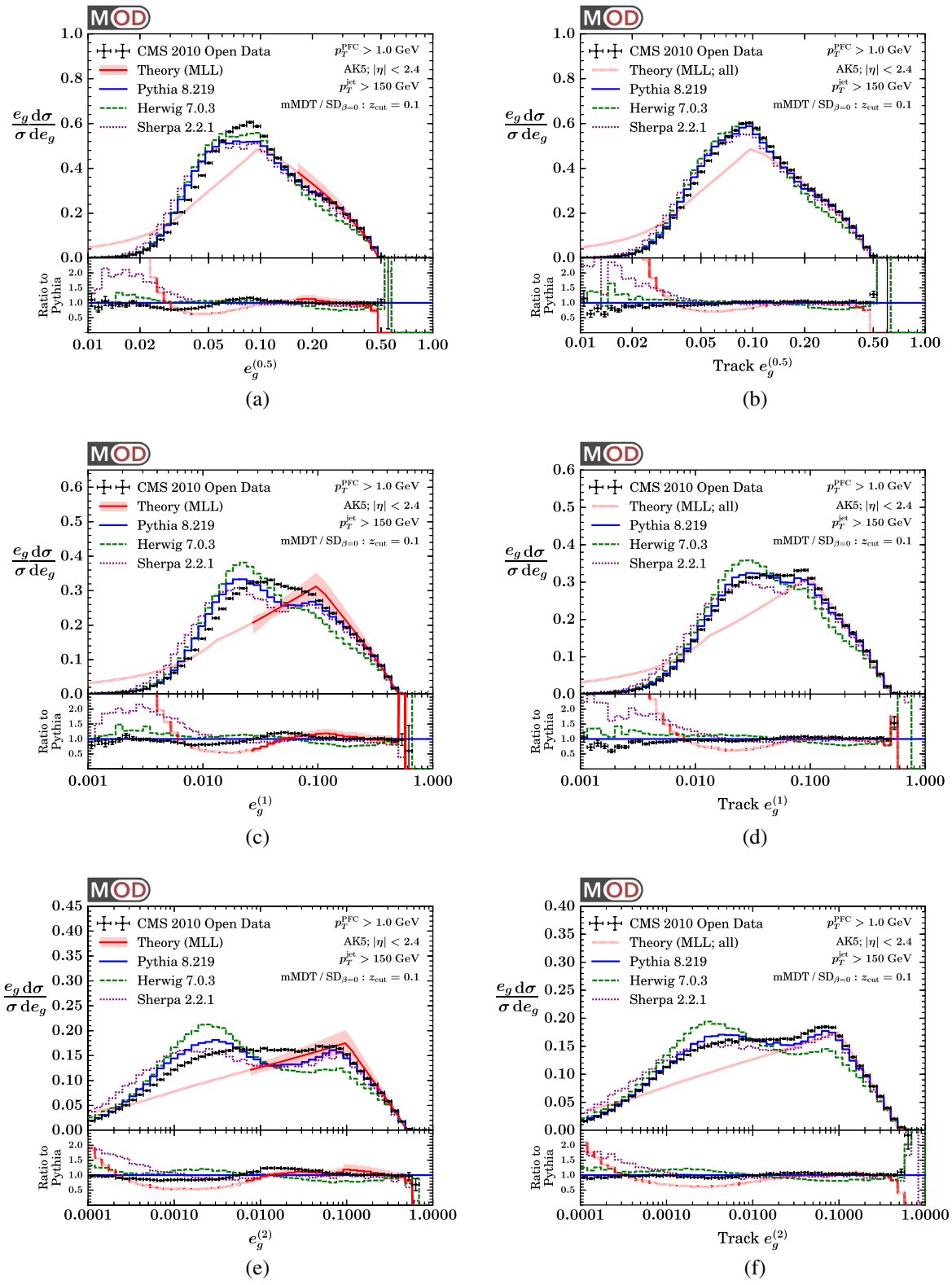

FIG. 13. Logarithmic distributions for (a), (b) $e_g^{(1/2)} = z_g\sqrt{\theta_g}$, (c), (d) $e_g^{(1)} = z_g\theta_g$, and (e), (f) $e_g^{(2)} = z_g\theta_g^2$, using (left column) all particles and (right column) only charged particles. As in Fig. 12, dashing indicates where nonperturbative physics dominates and we have not indicated the fixed-order regime.

out which information was relevant to a particular question, but we would have benefitted from more centralized documentation that highlighted the most important features of the CMS open data. Centralized documentation would undoubtedly help CMS Collaboration members as well, as would making more TWIKI pages accessible outside of the CERN authentication wall.

(iii) *Lack of validation examples*. When working with public data, one would like to validate that one is doing a sensible analysis by trying to match published results. While example files were provided, none of them (to our knowledge) involved the complications present in a real analysis, such as appropriate trigger selection, jet quality criteria, and jet energy corrections. Initially, we had hoped to reproduce the jet p_T spectrum measured by CMS on 2010 data [265], but that turned out to be surprisingly difficult, since very low p_T jet triggers are not contained in the Jet Primary Dataset, and we were not confident in our ability to merge information from the MinimumBias Primary Dataset. (In addition, the published CMS result is based on inclusive jet p_T spectra, while we restricted our analysis to the hardest jet in an event to simplify trigger assignment.) Ideally, one should be able to perform event-by-event validation with the CMS open data, especially if there are important calibration steps that could be missed.[14]

(iv) *Information overload*. The AOD files contain an incredible wealth of information, such that the majority of official CMS analyses can use the AOD format directly without requiring RAW or RECO information. While ideal for archival purposes, it is an overload of information for external users, especially because some information is effectively duplicated. The main reason we introduced the MOD file format was to restrict our access only to information that was essential for our analysis. This can be compared to the mini-AOD format currently being developed by CMS to address a similar problem [226].

(v) *Presence of superfluous data*. As described on the open data portal, one has to apply a cut to only select validated runs. This meant that of the initial 20 million events, only 16 million were actually usable. That said, this turns out to be a relatively small issue compared to trigger inefficiencies, which to our knowledge are not mentioned on the open data portal website. The Jet Primary Dataset includes any event where one of the jet-related triggers fired (see Table I). However, these triggers are not fully efficient down to the turn-on threshold, which is why we had to derive trigger efficiency curves in Fig. 1(b). Using just the triggers in Table II in the regime where they were nearly 100% efficient reduced the number of events for our analysis to less than 1 million, which is an order-of-magnitude smaller than the starting data set.

(vi) *No fast simulation or Monte Carlo samples*. While it is in principle possible to run the full CMS detector simulation on events from parton shower generators, we did not have the computing resources to do so. Without detector information, either in the form of CMS-approved fast simulation software or simulated Monte Carlo data sets, we cannot really say whether the good agreement seen between open data and parton showers is robust or merely accidental. Fast simulation tools like DELPHES can be used to some extent, but because they have not been optimized for jet substructure, we were not able to use them for this study. Official CMS Monte Carlo samples would have helped us greatly to estimate the size of detector corrections (and potentially even unfold distributions back to truth level). We are therefore encouraged by the inclusion of Monte Carlo samples in the 2011 CMS open data release [216].

Despite these above issues, though, we were able to perform a successful jet substructure analysis, in no small part due to the help of our CMS (and ATLAS) colleagues who generously offered their time and advice.

B. Recommendations

Given our experience, we would like to make the following recommendations to CERN and CMS about the continuation of the open data project. Many of these suggestions are also relevant for the 2012 ATLAS open data [266], though that effort is aimed more at education than research. Here are our recommendations, in rough order of priority.

(i) *Continue to release research-grade public data.* Particle physics experiments are expensive and, in many cases, unique. It is therefore incumbent on the particle physics community to extract as much useful information from collision data as possible. First priority for data analysis should of course go to members of the detector collaborations, especially since proper calibration can only be performed by physicists familiar with the detection technology.[15]

[14]In the one case where we thought it would be the most straightforward to cross-check results, namely the luminosity study in Fig. 2(a), it was frustrating to later learn that the AOD files contained insufficient information.

[15]There also needs to be a strong incentive for experimentalists to join collaborations in the first place. Outside access to (calibrated) data should not be used to bypass the stringent internal collaboration review process.

After an appropriate lag time—four years in the case of the 2010 CMS open data—outside scientists can play a useful role in data analysis, especially because collaboration members might not have the time or interest to revisit old data once new data are available. Techniques that perform well on open data can then be incorporated into the analysis strategies used internally by the collaborations, enhancing the already strong feedback cycle within the particle physics community.[16]

(ii) *Continue to provide a unique reference event interpretation.* A key feature of the CMS open data is the presence of PFCs, which provides a unique reference event interpretation with four-vector-like objects. From our experience, this seems to be the right level of information for an outside user. If the CMS open data were to consist only of high-level objects, like reconstructed jets, then we would not have been able to pursue these jet substructure studies. On the flip side, more low-level information (or multiple versions of the same information) could overwhelm the external user and cause confusion. Since it is unlikely that open data could support arbitrary physics studies, the aim of open data should be to facilitate particle-level studies that do not require detailed knowledge of the detector.

(iii) *Provide validation examples.* We mentioned above the potential value of having centralized documentation about open data. Even more important than documentation, though, is having example analyses performed using open data. Explicit code helps emphasize analysis steps that might be missed by novices, including trigger selection, prescale factors, jet calibration, and luminosity extraction. Where possible, these validation examples should reproduce official published analyses. We expect that these validation examples will become the templates for future open data analyses, and good validation examples could minimize incorrect use of the data. We intend to make the present analysis software public, in order to guide future open data studies.

(iv) *Provide detector response information.* The biggest physics gap in our study was our limited ability to estimate detector corrections. Ideally, open data should be released with corresponding detector-simulated Monte Carlo samples, matched to the triggers of interest. Indeed, the 2011 CMS open data—released in April 2016—do provide these samples, which will make it possible to estimate (some) detector systematics.[17] Eventually, if open data are used to place (unofficial) bounds on physics beyond the standard model, an external user would also need access to a recommended fast simulation framework. While it is probably impossible for external users to assess systematic uncertainties with the same level of care as one can do within the collaboration, some understanding of detector effects is needed before concluding that an effect observed in the data is real and interesting.

(v) *Cull the data set.* Within the experimental collaborations, most studies are based on well-defined trigger paths with almost 100% trigger efficiencies and nearly constant prescale factors. These same requirements should be imposed on the open data such that only usable data are made available publicly. This would not only reduce the storage requirements for open data, but it would also help avoid some spurious features showing up in the data. Similarly, most official studies do not need the full information contained in the AOD file format, and a more restricted data format would help further shrink the data file sizes and reduce user errors. Of course, to maximize the archival value, it may still make sense to release the original AOD files for the expert users, along with the tools used to create the culled versions.

(vi) *Speed up the development cycle.* For archival purposes, it is valuable to have the full CMSSW framework operating in a VM environment. For the external user, though, it would be more efficient to have a simplified software framework that can run with minimal software dependencies.[18] We understand that developing an external software environment requires considerable effort by collaboration members, but a relatively small investment would greatly increase the usability of the CMS open data. Our MODANALYZER software (based heavily on FASTJET) might be a good starting point for such an analysis package, as would any of the existing private tools used internally by CMS analysis teams. It may also make sense for the collaborations to appoint an

[16]There are, of course, cases where a full open data analysis is not necessary to motivate the adoption of new techniques. Even in that context, though, it can still be valuable for the collaborations to release official Monte Carlo samples. At minimum, hadron-truth-level samples provide a standard benchmark to validate the performance of new techniques. More ambitiously, detector-simulated samples can be used to assess how a new technique might be affected by detector granularity, acceptance, and efficiency.

[17]The 2012 ATLAS open data [266] do provide detector-simulated samples, but not truth-level information, so it is not possible to derive detector response information.

[18]If the use of the CMSSW framework is essential, it would be helpful to have more centralized documentation for the core classes and methods of CMSSW.

official contact to answer questions from external users, possibly in the form of an open data "convenership."

While these recommendations are perhaps ambitious in their scope, we think that the enormous scientific value of particle physics data justifies this kind of investment in open data.

VI. CONCLUSION

As the LHC explores the frontiers of scientific knowledge, its primary legacy will be the measurements and discoveries made by the LHC detector collaborations. But there is another potential legacy from the LHC that could be just as important: granting future generations of physicists access to unique high-quality data sets from proton-proton collisions at 7, 8, 13, and 14 TeV.

In our view, the best way to build a legacy data set is to invest in open data initiatives right now, such that scientists outside of the LHC collaborations can stress-test archival data strategies. This paper represents the first such analysis made with 2010 CMS open data from 7 TeV collisions. We showed how to extract jet substructure observables with the help of CMS's particle flow algorithm, yielding results that are in good agreement with parton shower generators and first-principles QCD calculations. The recent release of the 2011 CMS open data is particularly exciting, since it now includes detector-simulated Monte Carlo samples, allowing one to properly estimate detector systematics. We hope our experience motivates the LHC collaborations to further their investment in public data releases and encourages the particle physics community to exploit the scientific potential of open data sets.

ACKNOWLEDGMENTS

We applaud CERN for the historic launch of the open data portal, and we congratulate the CMS Collaboration for the fantastic performance of their detector and the high quality of the resulting public data set. We thank Alexis Romero for collaboration in the early stages of this work. We are indebted to Salvatore Rappoccio and Kati Lassila-Perini for helping us navigate the CMS software framework. We benefitted from code and encouragement from Tim Andeen, Matt Bellis, Andy Buckley, Kyle Cranmer, Sarah Demers, Guenther Dissertori, Javier Duarte, Peter Fisher, Achim Geiser, Giacomo Govi, Phil Harris, Beate Heinemann, Harri Hirvonsalo, Markus Klute, Greg Landsberg, Yen-Jie Lee, Elliot Lipeles, Peter Loch, Marcello Maggi, David Miller, Ben Nachman, Christoph Paus, Alexx Perloff, Andreas Pfeiffer, Maurizio Pierini, Ana Rodriguez, Gunther Roland, Ariel Schwartzman, Liz Sexton-Kennedy, Maria Spiropulu, Nhan Tran, Ana Trisovic, Chris Tully, Marta Verweij, Mikko Voutilainen, and Mike Williams. This work is supported by the MIT Charles E. Reed Faculty Initiatives Fund. The work of J. T., A. T., and W. X. is supported by the U.S. Department of Energy (DOE) under Contracts No. DE-SC-00012567 and No. DE-SC-00015476. The work of A. L. was supported by the U.S. National Science Foundation, under Grant No. PHY-1419008, the LHC Theory Initiative. S. M. is supported by the U.S. National Science Foundation, under Grants No. PHY-0969510 (LHC Theory Initiative) and No. PHY-1619867. A. T. is also supported by the MIT Undergraduate Research Opportunities Program.

APPENDIX A: ADDITIONAL OPEN DATA INFORMATION

In this Appendix, we provide additional information about the overall CMS open data extraction from Sec. II. In Fig. 14, we show the distribution of prescale values obtained for the triggers in Table II. As expected, higher trigger thresholds have lower prescale values, but there is substantial variation in the prescale values which changed over the duration of the run. If we were to use the given prescale factors instead of the averages, we would have seen rather large statistical uncertainties in our distributions. Since we only ever use one trigger per p_T bin, it is valid to use the average prescale value instead.

To properly select the hardest jet, we have to impose jet quality criteria and apply JEC factors. The CMS-recommended jet quality criteria are shown in Fig. 5; we always use the "loose" selection in our analysis. In Fig. 15(a), we show the distribution of JEC factors encouraged for the hardest jet. These are multiplicative scaling factors that tend to give a 5–10% correction to the jet p_T. In addition to accounting for detector effects, the JEC factor accounts for pileup through area subtraction [230]. The distribution of jet areas for the hardest jet is shown in Fig. 15(b), which peaks at πR^2 for $R = 0.5$ as expected. Note that the impact of pileup was minimal in Run 2010B, since as shown in Table VI, the number of primary interactions per bunch crossing was less than 5

TABLE V. Recommended jet quality criteria provided by CMS for $|\eta| < 2.4$. For $|\eta| > 2.4$, where no tracking is available, the last three requirements are not applied, and all jet constituents are treated as neutral. For our analysis, we always impose the "loose" criteria.

	Loose	Medium	Tight
Neutral hadron fraction	<0.99	<0.95	<0.90
Neutral EM fraction	<0.99	<0.95	<0.90
Number of constituents	>1	>1	>1
Charged hadron fraction	>0.00	>0.00	>0.00
Charged EM fraction	<0.99	<0.99	<0.99
Charged multiplicity	>0	>0	>0

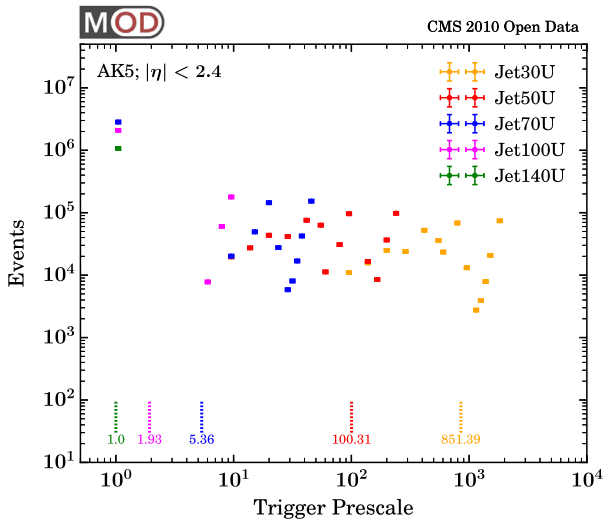

FIG. 14. Trigger prescale values for jets that pass the criteria in Table II. When filling histograms in this paper, we always use the average prescale values, not the individual ones.

TABLE VI. Number of primary interactions per bunch crossing. Since Run 2010B was a relatively low luminosity run, a large fraction of the event sample has $N_{\mathrm{PV}} = 1$, corresponding to no pileup contamination.

	Jet Primary Dataset		Hardest jet selection	
N_{PV}	Events	Fraction	Events	Fraction
1	4,716,494	0.289	190,277	0.248
2	4,814,495	0.295	246,387	0.321
3	3,630,413	0.222	180,021	0.234
4	1,933,832	0.118	93,587	0.122
5	819,835	0.050	38,598	0.050
6	294,612	0.018	13,805	0.018
7	93,714	0.006	4,318	0.006
8	27,550	0.002	1,242	0.002
9	7,481	0.000	330	0.000
10	2,041	0.000	91	0.000
11	540	0.000	21	0.000
12	125	0.000	6	0.000
13	41	0.000	3	0.000
14	9	0.000	1	0.000
≥ 15	5	0.000	0	0.000

(i.e. effectively no pileup) for over 90% of the events and never more than 15 for the selection used for our analysis (modest pileup).

To partially account for detector effects in our substructure analysis, we impose a PFC cut of $p_T^{\mathrm{min}} = 1.0$ GeV, motivated by Fig. 3. In Fig. 16, we plot the PFC p_T spectrum over an extended range, again restricting to PFCs within the hardest jet. For neutral particles, there is a growing difference between the CMS open data and the parton shower generators for constituents that carry a large fraction of the jet momentum, though this difference is reduced when considering only charged particles.

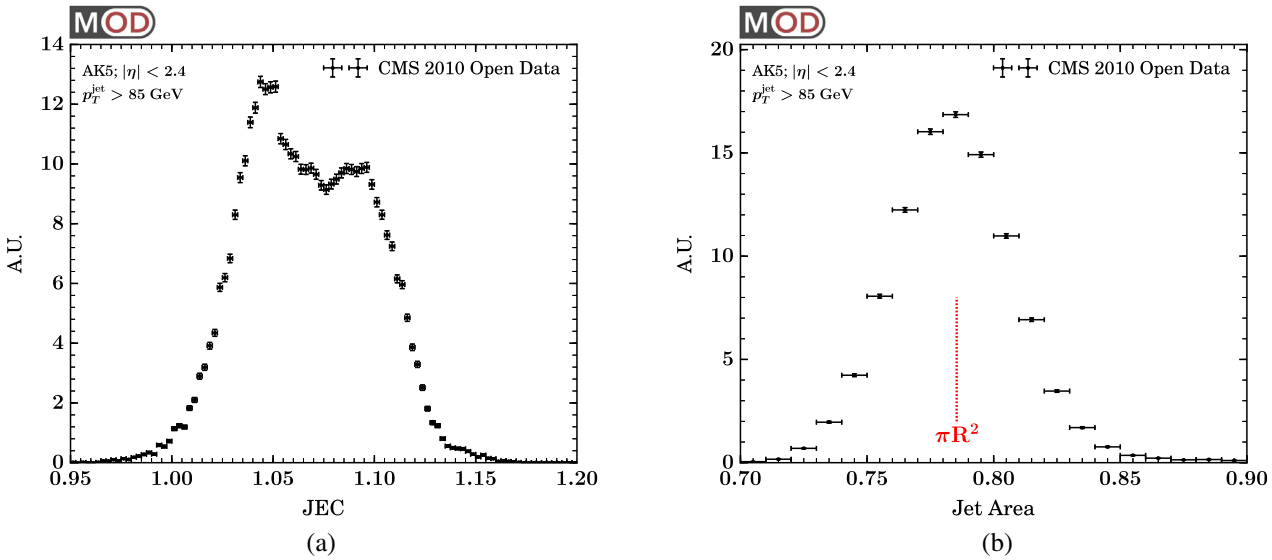

FIG. 15. Range of (a) JEC factors and (b) active jet areas [230] encountered for the hardest jet.

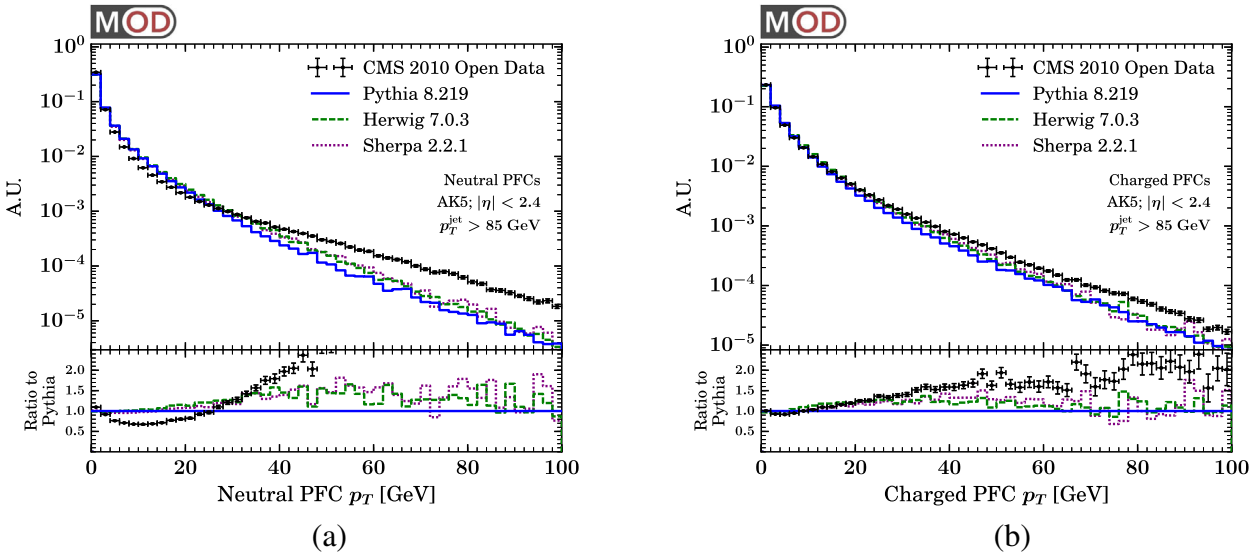

FIG. 16. Same as Fig. 3, for (a) neutral candidates and (b) charged candidates, but showing a wider range of PFC p_T values.

APPENDIX B: ADDITIONAL SOFT-DROPPED DISTRIBUTIONS

In this Appendix, we show additional distributions obtained from soft-drop declustering. In Fig. 17, we show the fraction of the original jet p_T discarded after soft drop, plotted logarithmically. This distribution was advocated in Ref. [46] as an interesting example of a Sudakov safe [205,206] observable, and we see good agreement between the CMS open data and parton showers.

The distributions in Sec. III were obtained prior to applying any jet grooming. In Fig. 18, we show the same basic substructure observables from Fig. 6, but now showing the impact of soft drop. Soft drop does not necessarily improve the agreement between the CMS open data and the PYTHIA parton shower, though it also does not make it any worse, and the track-based agreement is very good. We perform a similar study in Fig. 19 for the jet angularities from Fig. 7. There is good qualitative agreement between the open data and PYTHIA, but the track-only version has much better quantitative agreement as expected.

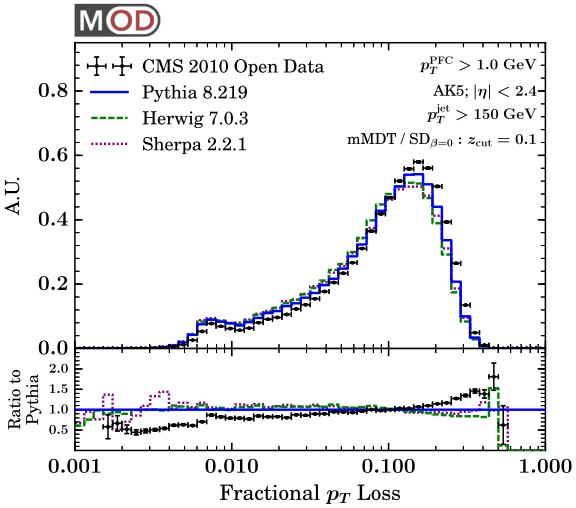

FIG. 17. Fraction of the original jet p_T lost after performing soft-drop declustering. Because this is a fraction, no JEC factors are applied.

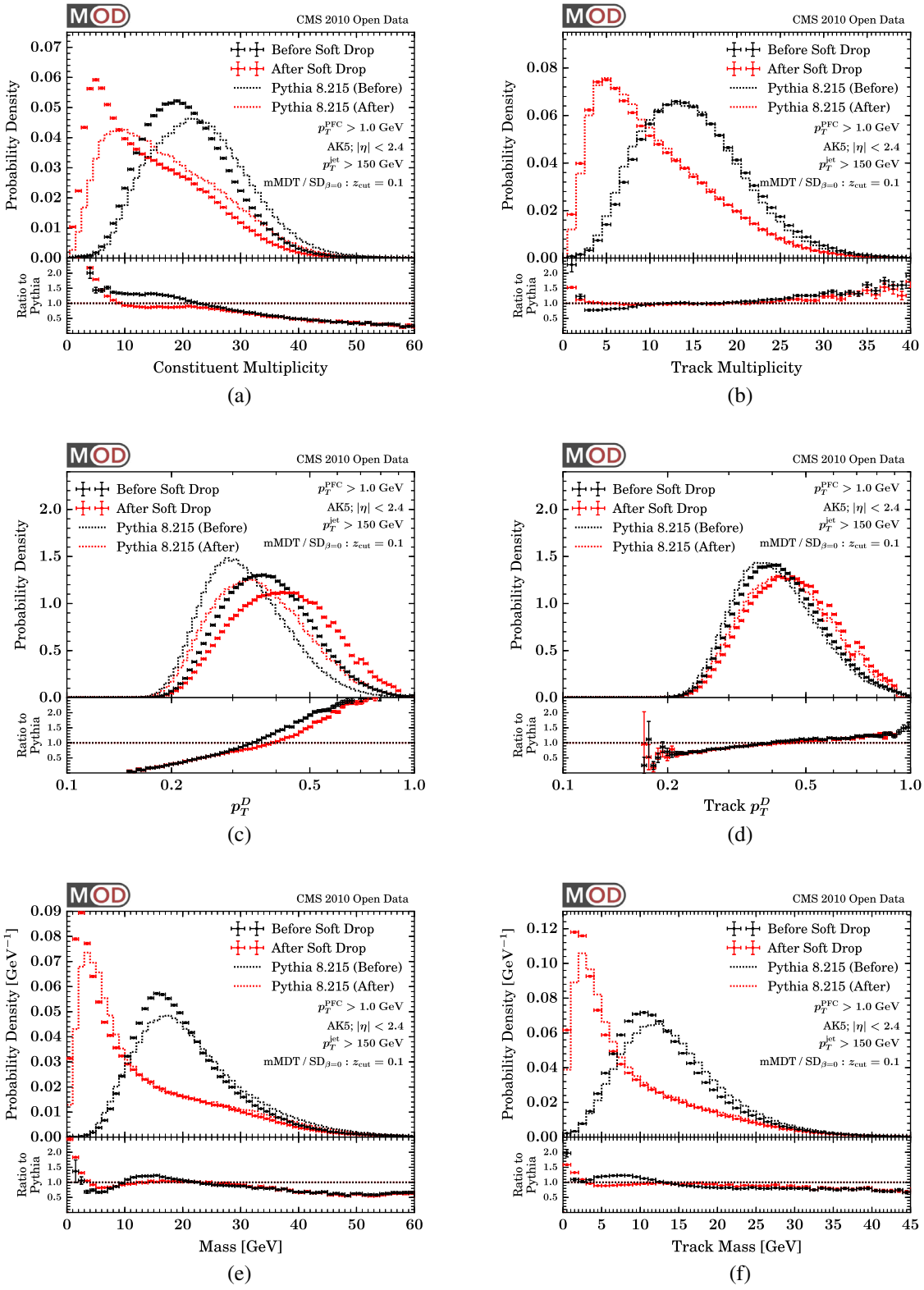

FIG. 18. Same observables as in Fig. 6, for (a), (b) constituent multiplicity, (c), (d) p_T^D on a logarithmic scale and (e), (f) jet mass, but now showing the original distributions (black) compared to those obtained after soft-drop declustering (red).

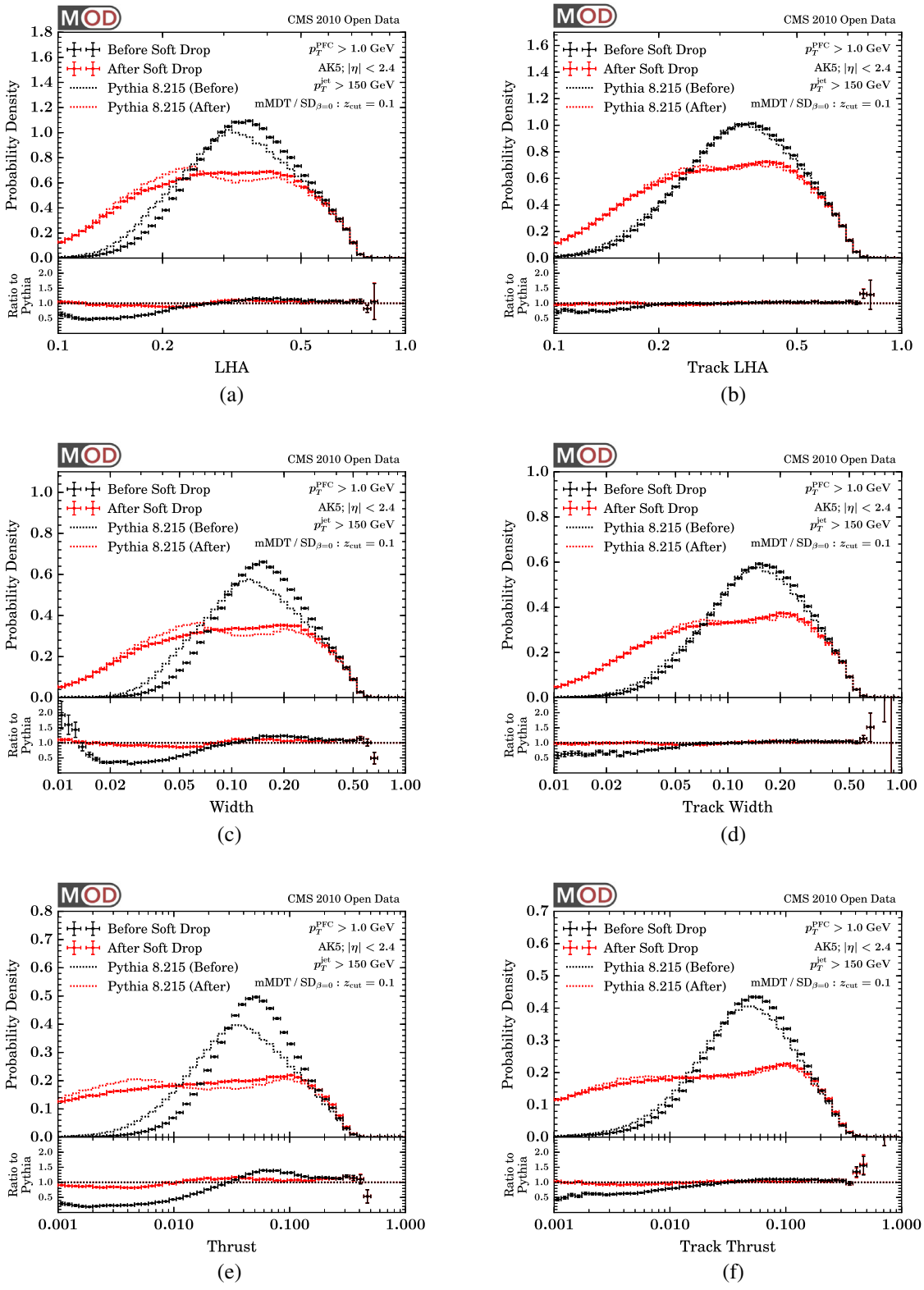

FIG. 19. Same observables as in Fig. 7, for (a), (b) LHA, (c), (d) jet width, and (e), (f) jet thrust, but now showing the original distributions (black) compared to those obtained after soft-drop declustering (red).

[1] CERN Open Data Portal, http://opendata.cern.ch.

[2] M. H. Seymour, Tagging a heavy Higgs boson, in *ECFA Large Hadron Collider (LHC) Workshop: Physics and Instrumentation Aachen, Germany, 1990* (CERN, Geneva, 1991).

[3] M. H. Seymour, Searches for new particles using cone and cluster jet algorithms: A comparative study, Z. Phys. C **62**, 127 (1994).

[4] J. M. Butterworth, B. E. Cox, and J. R. Forshaw, *WW* scattering at the CERN LHC, Phys. Rev. D **65**, 096014 (2002).

[5] J. M. Butterworth, J. R. Ellis, and A. R. Raklev, Reconstructing sparticle mass spectra using hadronic decays, J. High Energy Phys. 05 (2007) 033.

[6] 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**, 242001 (2008).

[7] A. Abdesselam, E. Bergeaas Kuutmann, U. Bitenc, G. Brooijmans, J. Butterworth *et al.*, Boosted objects: A probe of beyond the Standard Model physics, Eur. Phys. J. C **71**, 1661 (2011).

[8] A. Altheimer, S. Arora, L. Asquith, G. Brooijmans, J. Butterworth *et al.*, Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks, J. Phys. G **39**, 063001 (2012).

[9] A. Altheimer *et al.*, Boosted objects and jet substructure at the LHC, Report of BOOST2012, Eur. Phys. J. C **74**, 2792 (2014).

[10] D. Adams *et al.*, Towards an understanding of the correlations in jet substructure, Eur. Phys. J. C **75**, 409 (2015).

[11] G. Brooijmans, Report No. ATL-PHYS-CONF-2008-008 and ATL-COM-PHYS-2008-001, 2008.

[12] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks, Phys. Rev. Lett. **101**, 142001 (2008).

[13] J. Thaler and L.-T.Wang, Strategies to identify boosted tops, J. High Energy Phys. 07 (2008) 092.

[14] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung *et al.*, Substructure of high-p_T jets at the LHC, Phys. Rev. D **79**, 074017 (2009).

[15] L. G. Almeida, S. J. Lee, G. Perez, I. Sung, and J. Virzi, Top jets at the LHC, Phys. Rev. D **79**, 074012 (2009).

[16] CMS Collaboration, Technical Report CMS-PAS-JME-09-001, 2009.

[17] CMS Collaboration, Technical Report CMS-PAS-EXO-09-002, 2009.

[18] ATLAS Collaboration, Technical Report ATL-PHYS-PUB-2009-081, ATL-COM-PHYS-2009-255, 2009.

[19] T. Plehn, G. P. Salam, and M. Spannowsky, Fat Jets for a Light Higgs, Phys. Rev. Lett. **104**, 111801 (2010).

[20] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, Stop reconstruction with tagged tops, J. High Energy Phys. 10 (2010) 078.

[21] L. G. Almeida, S. J. Lee, G. Perez, G. Sterman, and I. Sung, Template overlap method for massive jets, Phys. Rev. D **82**, 054034 (2010).

[22] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.

[23] J. Thaler and K. Van Tilburg, Maximizing Boosted top identification by minimizing N-subjettiness, J. High Energy Phys. 02 (2012) 093.

[24] M. Jankowiak and A. J. Larkoski, Jet substructure without trees, J. High Energy Phys. 06 (2011) 057.

[25] D. E. Soper and M. Spannowsky, Finding top quarks with shower deconstruction, Phys. Rev. D **87**, 054012 (2013).

[26] A. J. Larkoski, G. P. Salam, and J. Thaler, Energy correlation functions for jet substructure, J. High Energy Phys. 06 (2013) 108.

[27] C. Anders, C. Bernaciak, G. Kasieczka, T. Plehn, and T. Schell, Benchmarking an even better top tagger algorithm, Phys. Rev. D **89**, 074047 (2014).

[28] M. Freytsis, T. Volansky, and J. R. Walsh, Tagging partially reconstructed objects with jet substructure, Phys. Lett. B **769**, 333 (2017).

[29] G. Kasieczka, T. Plehn, T. Schell, T. Strebler, and G. P. Salam, Resonance searches with an updated top tagger, J. High Energy Phys. 06 (2015) 203.

[30] T. Lapsien, R. Kogler, and J. Haller, A new tagger for hadronically decaying heavy particles at the LHC, Eur. Phys. J. C **76**, 600 (2016).

[31] I. Moult, L. Necib, and J. Thaler, New angles on energy correlation functions, J. High Energy Phys. 12 (2016) 153.

[32] G. D. Kribs, A. Martin, T. S. Roy, and M. Spannowsky, Discovering the Higgs boson in new physics events using jet substructure, Phys. Rev. D **81**, 111501 (2010).

[33] G. D. Kribs, A. Martin, T. S. Roy, and M. Spannowsky, Discovering Higgs bosons of the MSSM using jet substructure, Phys. Rev. D **82**, 095012 (2010).

[34] C.-R. Chen, M. M. Nojiri, and W. Sreethawong, Search for the elusive Higgs boson using jet structure at LHC, J. High Energy Phys. 11 (2010) 012.

[35] C. Hackstein and M. Spannowsky, Boosting Higgs discovery: The forgotten channel, Phys. Rev. D **82**, 113012 (2010).

[36] A. Falkowski, D. Krohn, L.-T. Wang, J. Shelton, and A. Thalapillil, Unburied Higgs boson: Jet substructure techniques for searching for Higgs' decay into gluons, Phys. Rev. D **84**, 074022 (2011).

[37] A. Katz, M. Son, and B. Tweedie, Jet substructure and the search for neutral spin-one resonances in electroweak boson channels, J. High Energy Phys. 03 (2011) 011.

[38] Y. Cui, Z. Han, and M. D. Schwartz, W-jet tagging: Optimizing the identification of boosted hadronically-decaying W bosons, Phys. Rev. D **83**, 074023 (2011).

[39] J.-H. Kim, Rest frame subjet algorithm with SISCone jet for fully hadronic decaying Higgs search, Phys. Rev. D **83**, 011502 (2011).

[40] J. Gallicchio, J. Huth, M. Kagan, M. D. Schwartz, K. Black, and B. Tweedie, Multivariate discrimination and the Higgs + W/Z search, J. High Energy Phys. 04 (2011) 069.

[41] J. Gallicchio and M. D. Schwartz, Seeing in Color: Jet Superstructure, Phys. Rev. Lett. **105**, 022001 (2010).

[42] A. Hook, M. Jankowiak, and J. G. Wacker, Jet dipolarity: Top tagging with color flow, J. High Energy Phys. 04 (2012) 007.

[43] D. E. Soper and M. Spannowsky, Finding physics signals with shower deconstruction, Phys. Rev. D **84**, 074002 (2011).

[44] L. G. Almeida, O. Erdogan, J. Juknevich, S. J. Lee, G. Perez, and G. Sterman, Three-particle templates for a boosted Higgs boson, Phys. Rev. D **85**, 114046 (2012).
[45] S. D. Ellis, A. Hornig, T. S. Roy, D. Krohn, and M. D. Schwartz, Qjets: A Non-Deterministic Approach to Tree-Based Jet Substructure, Phys. Rev. Lett. **108**, 182003 (2012).
[46] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft Drop, J. High Energy Phys. 05 (2014) 146.
[47] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, Jet-images: Computer vision inspired techniques for jet tagging, J. High Energy Phys. 02 (2015) 118.
[48] E. Izaguirre, B. Shuve, and I. Yavin, Improving Identification of Dijet Resonances at Hadron Colliders, Phys. Rev. Lett. **114**, 041802 (2015).
[49] A. J. Larkoski, I. Moult, and D. Neill, Power counting to better jet observables, J. High Energy Phys. 12 (2014) 009.
[50] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-images? Deep learning edition, J. High Energy Phys. 07 (2016) 069.
[51] M. Dasgupta, A. Powling, and A. Siodmok, On jet substructure methods for signal jets, J. High Energy Phys. 08 (2015) 079.
[52] A. J. Larkoski, I. Moult, and D. Neill, Analytic boosted boson discrimination, J. High Energy Phys. 05 (2016) 117.
[53] I. W. Stewart, F. J. Tackmann, J. Thaler, C. K. Vermilion, and T. F. Wilkason, XCone: N-jettiness as an exclusive cone jet algorithm, J. High Energy Phys. 11 (2015) 072.
[54] M. Dasgupta, L. Schunk, and G. Soyez, Jet shapes for boosted jet two-prong decays from first-principles, J. High Energy Phys. 04 (2016) 166.
[55] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, Jet substructure classification in high-energy physics with deep neural networks, Phys. Rev. D **93**, 094034 (2016).
[56] J. S. Conway, R. Bhaskar, R. D. Erbacher, and J. Pilot, Identification of high-momentum top quarks, Higgs bosons, and W and Z bosons using boosted event shapes, Phys. Rev. D **94**, 094027 (2016).
[57] J. Barnard, E. N. Dawe, M. J. Dolan, and N. Rajcic, Parton shower uncertainties in jet substructure analyses with deep neural networks, Phys. Rev. D **95**, 014018 (2017).
[58] T. S. Roy and A. M. Thalapillil, Augmenting collider searches and enhancing discovery potentials through stochastic jet grooming, Phys. Rev. D **95**, 075002 (2017).
[59] M. Dasgupta, A. Powling, L. Schunk, and G. Soyez, Improved jet substructure methods: Y-splitter and variants with grooming, J. High Energy Phys. 12 (2016) 079.
[60] H. P. Nilles and K. H. Streng, Quark-gluon separation in three jet events, Phys. Rev. D **23**, 1944 (1981).
[61] L. M. Jones, Tests for determining the parton ancestor of a hadron jet, Phys. Rev. D **39**, 2550 (1989).
[62] Z. Fodor, How to see the differences between quark and gluon jets, Phys. Rev. D **41**, 1726 (1990).
[63] L. Jones, Towards a systematic jet classification, Phys. Rev. D **42**, 811 (1990).
[64] L. Lönnblad, C. Peterson, and T. Rognvaldsson, Using neural networks to identify jets, Nucl. Phys. **B349**, 675 (1991).
[65] J. Pumplin, How to tell quark jets from gluon jets, Phys. Rev. D **44**, 2025 (1991).
[66] J. Gallicchio and M. D. Schwartz, Quark and Gluon Tagging at the LHC, Phys. Rev. Lett. **107**, 172001 (2011).
[67] J. Gallicchio and M. D. Schwartz, Quark and gluon jet substructure, J. High Energy Phys. 04 (2013) 090.
[68] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, Gaining (mutual) information about quark/gluon discrimination, J. High Energy Phys. 11 (2014) 129.
[69] B. Bhattacherjee, S. Mukhopadhyay, M. M. Nojiri, Y. Sakaki, and B. R. Webber, Associated jet and subjet rates in light-quark and gluon jet discrimination, J. High Energy Phys. 04 (2015) 131.
[70] J. R. Andersen *et al.*, Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report, in 9th Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2015) Les Houches, France, 2015 (2016), arXiv:1605.04692.
[71] D. Ferreira de Lima, P. Petrov, D. Soper, and M. Spannowsky, Quark-gluon tagging with shower deconstruction: Unearthing dark matter and Higgs couplings, Phys. Rev. D **95**, 034001 (2017).
[72] B. Bhattacherjee, S. Mukhopadhyay, M. M. Nojiri, Y. Sakaki, and B. R. Webber, Quark-gluon discrimination in the search for gluino pair production at the LHC, J. High Energy Phys. 01 (2017) 044.
[73] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, Deep learning in color: Towards automated quark/gluon jet discrimination, J. High Energy Phys. 01 (2017) 110.
[74] J. Davighi and P. Harris, Fractal based observables to probe jet substructure of quarks and gluons, arXiv:1703.00914.
[75] P. Gras, S. Hoeche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer, A. Siódmok, P. Skands, G. Soyez, and J. Thaler, Systematics of quark/gluon tagging, J. High Energy Phys. 07 (2017) 091.
[76] CMS Collaboration, Jet Primary Dataset in AOD format from RunB of 2010 (/Jet/Run2010B-Apr21ReReco-v1/AOD), CERN Open Data Portal, http://dx.doi.org/10.7483/OPENDATA.CMS.3S7F.2E9W.
[77] CMS Collaboration, Technical Report CMS-PAS-EXO-11-006, 2011.
[78] CMS Collaboration, Technical Report CMS-PAS-JME-10-013, 2011.
[79] D. W. Miller (ATLAS Collaboration), Technical Report ATL-PHYS-PROC-2011-142, 2011.
[80] S. Chatrchyan *et al.* (CMS Collaboration), Shape, transverse size, and charged hadron multiplicity of jets in pp collisions at 7 TeV, J. High Energy Phys. 06 (2012) 160.
[81] ATLAS Collaboration, Technical Report ATLAS-CONF-2012-066, ATLAS-COM-CONF-2012-097, 2012.
[82] G. Aad *et al.* (ATLAS Collaboration), ATLAS measurements of the properties of jets for boosted particle searches, Phys. Rev. D **86**, 072006 (2012).
[83] ATLAS Collaboration, Technical Report ATLAS-CONF-2012-065, ATLAS-COM-CONF-2012-095, 2012.
[84] G. Aad *et al.* (ATLAS Collaboration), Jet mass and substructure of inclusive jets in $\sqrt{s} = 7$ TeV pp collisions with the ATLAS experiment, J. High Energy Phys. 05 (2012) 128.
[85] S. Chatrchyan *et al.* (CMS Collaboration), Search for anomalous $t\bar{t}$ production in the highly-boosted all-hadronic final state, J. High Energy Phys. 09 (2012) 029.

[86] S. Chatrchyan *et al.* (CMS Collaboration), Search for a Higgs boson in the decay channel H to ZZ(*) to q qbar ℓ^- l+ in pp collisions at $\sqrt{s} = 7$ TeV, J. High Energy Phys. 04 (2012) 036.

[87] S. Chatrchyan *et al.* (CMS Collaboration), Studies of jet mass in dijet and $W/Z +$ jet events, J. High Energy Phys. 05 (2013) 090.

[88] G. Aad *et al.* (ATLAS Collaboration), Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector, J. High Energy Phys. 09 (2013) 076.

[89] G. Aad *et al.* (ATLAS Collaboration), Measurement of jet shapes in top-quark pair events at $\sqrt{s} = 7$ TeV using the ATLAS detector, Eur. Phys. J. C **73**, 2676 (2013).

[90] ATLAS Collaboration, Technical Report ATLAS-CONF-2013-087, ATLAS-COM-CONF-2013-099, 2013.

[91] ATLAS Collaboration, Technical Report ATLAS-CONF-2013-086, ATLAS-COM-CONF-2013-101, 2013.

[92] ATLAS Collaboration, Technical Report ATLAS-CONF-2013-085, ATLAS-COM-CONF-2013-100, 2013.

[93] ATLAS Collaboration, Technical Report ATLAS-CONF-2013-083, ATLAS-COM-CONF-2013-097, 2013.

[94] ATLAS Collaboration, Technical Report ATLAS-CONF-2013-084, ATLAS-COM-CONF-2013-074, 2013.

[95] CMS Collaboration, Technical Report CMS-PAS-HIG-13-008, 2013.

[96] CMS Collaboration, Technical Report CMS-PAS-JME-13-006, 2013.

[97] CMS Collaboration, Technical Report CMS-PAS-JME-13-002, 2013.

[98] CMS Collaboration, Technical Report CMS-PAS-JME-13-005, 2013.

[99] S. Fleischmann (ATLAS, CMS Collaborations), Boosted top quark techniques and searches for $t\bar{t}$ resonances at the LHC, J. Phys. Conf. Ser. **452**, 012034 (2013).

[100] J. Pilot (ATLAS, CMS Collaborations), Boosted top quarks, top pair resonances, and top partner searches at the LHC, Eur. Phys. J. Web Conf. **60**, 09003 (2013).

[101] CMS Collaboration, Technical Report CMS-PAS-QCD-10-041, CERN, 2010.

[102] G. Aad *et al.* (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**, 3023 (2014).

[103] P. Loch (ATLAS Collaboration), Studies of jet shapes and jet substructure in proton-proton collisions at $\sqrt{s} = 7$ TeV with ATLAS, Proc. Sci., EPS-HEP2013 (**2013**) 442.

[104] CMS Collaboration, Technical Report CMS-PAS-JME-13-007, 2014.

[105] CMS Collaboration, Technical Report CMS-PAS-JME-14-002, 2014.

[106] G. Aad *et al.* (ATLAS Collaboration), Measurement of the cross-section of high transverse momentum vector bosons reconstructed as single jets and studies of jet substructure in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector, New J. Phys. **16**, 113013 (2014).

[107] CMS Collaboration, Technical Report CMS-PAS-B2G-14-001, 2014.

[108] CMS Collaboration, Technical Report CMS-PAS-B2G-14-002, 2014.

[109] V. Khachatryan *et al.* (CMS Collaboration), Identification techniques for highly boosted W bosons that decay into hadrons, J. High Energy Phys. 12 (2014) 017.

[110] V. Khachatryan *et al.* (CMS Collaboration), Search for vector-like T quarks decaying to top quarks and Higgs bosons in the all-hadronic channel using jet substructure, J. High Energy Phys. 06 (2015) 080.

[111] CMS Collaboration, Technical Report CMS-PAS-B2G-12-013, 2012.

[112] V. Khachatryan *et al.* (CMS Collaboration), Search for a massive resonance decaying into a Higgs boson and a W or Z boson in hadronic final states in proton-proton collisions at $\sqrt{s} = 8$ TeV, J. High Energy Phys. 02 (2016) 145.

[113] G. Aad *et al.* (ATLAS Collaboration), Search for high-mass diboson resonances with boson-tagged jets in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector, J. High Energy Phys. 12 (2015) 055.

[114] G. Aad *et al.* (ATLAS Collaboration), Measurement of jet charge in dijet events from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector, Phys. Rev. D **93**, 052003 (2016).

[115] G. Aad *et al.* (ATLAS Collaboration), Measurement of colour flow with the jet pull angle in $t\bar{t}$ events using the ATLAS detector at $\sqrt{s} = 8$ TeV, Phys. Lett. B **750**, 475 (2015).

[116] G. Aad *et al.* (ATLAS Collaboration), Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV, Eur. Phys. J. C **76**, 154 (2016).

[117] G. Aad *et al.* (ATLAS Collaboration), Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector, Eur. Phys. J. C **76**, 581 (2016).

[118] G. Aad *et al.* (ATLAS Collaboration), Search for charged Higgs bosons in the $H^{\pm} \to tb$ decay channel in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector, J. High Energy Phys. 03 (2016) 127.

[119] ATLAS Collaboration, Technical Report ATLAS-CONF-2015-035, 2015.

[120] ATLAS Collaboration, Technical Report ATLAS-CONF-2015-037, 2015.

[121] ATLAS Collaboration, Technical Report ATLAS-CONF-2015-071, 2015.

[122] ATLAS Collaboration, Technical Report ATLAS-CONF-2015-073, 2015.

[123] V. Khachatryan *et al.* (CMS Collaboration), Search for excited leptons in proton-proton collisions at $\sqrt{s} = 8$ TeV, J. High Energy Phys. 03 (2016) 125.

[124] G. Aad *et al.* (ATLAS Collaboration), A new method to distinguish hadronically decaying boosted Z bosons from W bosons using the ATLAS detector, Eur. Phys. J. C **76**, 238 (2016).

[125] ATLAS Collaboration, Technical Report ATLAS-CONF-2015-036, 2015.

[126] V. Khachatryan *et al.* (CMS Collaboration), Search for pair-produced vectorlike B quarks in proton-proton collisions at $\sqrt{s} = 8$ TeV, Phys. Rev. D **93**, 112009 (2016).

[127] CMS Collaboration, Technical Report CMS-PAS-EXO-15-002, 2015.

[128] CMS Collaboration, Technical Report CMS-PAS-HIG-15-012, 2015.

[129] CMS Collaboration, Technical Report CMS-PAS-EXO-12-055, 2015.
[130] G. Aad *et al.* (ATLAS Collaboration), Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector, J. High Energy Phys. 06 (2016) 093.
[131] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-008, 2016.
[132] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-034, 2016.
[133] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-035, 2016.
[134] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-039, 2016.
[135] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-102, 2016.
[136] CMS Collaboration, Technical Report CMS-PAS-B2G-16-013, 2016.
[137] CMS Collaboration, Technical Report CMS-PAS-TOP-15-015, 2016.
[138] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-100, 2016.
[139] CMS Collaboration, Technical Report CMS-PAS-SMP-16-012, 2016.
[140] CMS Collaboration, Technical Report CMS-PAS-B2G-16-020, 2016.
[141] CMS Collaboration, Technical Report CMS-PAS-EXO-16-037, 2016.
[142] CMS Collaboration, Technical Report CMS-PAS-EXO-16-040, 2016.
[143] CMS Collaboration, Technical Report CMS-PAS-HIG-16-016, 2016.
[144] V. Khachatryan *et al.* (CMS Collaboration), Search for dark matter in proton-proton collisions at 8 TeV with missing transverse momentum and vector boson tagged jets, J. High Energy Phys. 12 (2016) 083; Erratum, J. High Energy Phys. 08 (2017) 35.
[145] CMS Collaboration, Technical Report CMS-PAS-BTV-15-002, 2016.
[146] CMS Collaboration, Technical Report CMS-PAS-SMP-15-003, 2016.
[147] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-016, 2016.
[148] CMS Collaboration, Technical Report CMS-PAS-TOP-16-013, 2016.
[149] CMS Collaboration, Technical Report CMS-PAS-B2G-16-005, 2016.
[150] V. Khachatryan *et al.* (CMS Collaboration), Search for supersymmetry in pp collisions at $\sqrt{s} = 8$ TeV in final states with boosted W bosons and b jets using razor variables, Phys. Rev. D **93**, 092009 (2016).
[151] V. Khachatryan *et al.* (CMS Collaboration), Search for heavy resonances decaying to two Higgs bosons in final states containing four b quarks, Eur. Phys. J. C **76**, 371 (2016).
[152] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-014, 2016.
[153] V. Khachatryan *et al.* (CMS Collaboration), Search for direct pair production of supersymmetric top quarks decaying to all-hadronic final states in pp collisions at $\sqrt{s} = 8$ TeV, Eur. Phys. J. C **76**, 460 (2016).
[154] CMS Collaboration, Technical Report CMS-PAS-EXO-16-017, 2016.
[155] CMS Collaboration, Technical Report CMS-PAS-B2G-15-003, 2016.
[156] CMS Collaboration, Technical Report CMS-PAS-B2G-16-007, 2016.
[157] CMS Collaboration, Technical Report CMS-PAS-EXO-16-013, 2016.
[158] CMS Collaboration, Technical Report CMS-PAS-B2G-16-004, 2016.
[159] CMS Collaboration, Technical Report CMS-PAS-HIG-16-004, 2016.
[160] CMS Collaboration, Technical Report CMS-PAS-BTV-15-001, 2016.
[161] CMS Collaboration, Technical Report CMS-PAS-B2G-15-002, 2016.
[162] CMS Collaboration, Technical Report CMS-PAS-B2G-16-009, 2016.
[163] CMS Collaboration, Technical Report CMS-PAS-SUS-16-029, 2016.
[164] CMS Collaboration, Technical Report CMS-PAS-B2G-16-008, 2016.
[165] CMS Collaboration, Technical Report CMS-PAS-EXO-16-012, 2016.
[166] CMS Collaboration, Technical Report CMS-PAS-EXO-16-030, 2016.
[167] CMS Collaboration, Technical Report CMS-PAS-HIN-16-006, 2016.
[168] CMS Collaboration, Technical Report CMS-PAS-B2G-16-001, 2016.
[169] CMS Collaboration, Technical Report CMS-PAS-B2G-15-008, 2016.
[170] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-083, 2016.
[171] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-082, 2016.
[172] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-055, 2016.
[173] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-104, 2016.
[174] ATLAS Collaboration, Technical Report ATLAS-CONF-2016-002, 2016.
[175] CMS Collaboration, Technical Report CMS-PAS-B2G-16-011, 2016.
[176] CMS Collaboration, Technical Report CMS-PAS-HIG-16-029, 2016.
[177] CMS Collaboration, Technical Report CMS-PAS-B2G-16-002, 2016.
[178] CMS Collaboration, Technical Report CMS-PAS-JME-15-002, 2016.
[179] CMS Collaboration, Technical Report CMS-PAS-JME-14-001, 2014.
[180] V. Khachatryan *et al.* (CMS Collaboration), Search for heavy resonances decaying into a vector boson and a Higgs boson in final states with charged leptons, neutrinos, and b quarks, Phys. Lett. B **768**, 137 (2017).
[181] V. Khachatryan *et al.* (CMS Collaboration), Search for single production of a heavy vector-like T quark decaying to a Higgs boson and a top quark with a lepton and jets in the final state, Phys. Lett. B **771**, 80 (2017).

[182] CMS Collaboration, Technical Report CMS-PAS-B2G-16-021, 2016.
[183] V. Khachatryan *et al.* (CMS Collaboration), Searches for invisible decays of the Higgs boson in pp collisions at $\sqrt{s} = 7$, 8, and 13 TeV, J. High Energy Phys. 02 (2017) 135.
[184] A. M. Sirunyan *et al.* (CMS Collaboration), Search for electroweak production of a vector-like quark decaying to a top quark and a Higgs boson using boosted topologies in fully hadronic final states, J. High Energy Phys. 04 (2017) 136.
[185] A. M. Sirunyan *et al.* (CMS Collaboration), Search for massive resonances decaying into WW, WZ or ZZ bosons in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 03 (2017) 162.
[186] A. M. Sirunyan *et al.* (CMS Collaboration), Search for high-mass $Z\gamma$ resonances in proton-proton collisions at $\sqrt{s} = 8$ and 13 TeV using jet substructure techniques, Phys. Lett. B **772**, 363 (2017).
[187] CMS Collaboration, Technical Report CMS-PAS-B2G-12-016, 2016.
[188] A. M. Sirunyan *et al.* (CMS Collaboration), Search for single production of vector-like quarks decaying to a Z boson and a top or a bottom quark in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 05 (2017) 029.
[189] CMS Collaboration, Technical Report CMS-PAS-B2G-16-022, 2017.
[190] A. M. Sirunyan *et al.* (CMS Collaboration), Search for associated production of dark matter with a Higgs boson decaying to b-bbar or gamma-gamma at $\sqrt{s} = 13$ TeV, arXiv:1703.05236.
[191] A. M. Sirunyan *et al.* (CMS Collaboration), Search for a heavy resonance decaying to a top quark and a vector-like top quark at $\sqrt{s} = 13$ TeV, arXiv:1703.06352.
[192] A. M. Sirunyan *et al.* (CMS Collaboration), Measurement of the jet mass in highly boosted t-tbar events from pp collisions at $\sqrt{s} = 8$ TeV, Eur. Phys. J. C **77**, 467 (2017).
[193] A. M. Sirunyan *et al.* (CMS Collaboration), Search for anomalous couplings in semileptonic WW/WZ to l nu qqbar production in proton-proton collisions at $\sqrt{s} =$ 8 TeV, Phys. Lett. B **772**, 21 (2017).
[194] A. M. Sirunyan *et al.* (CMS Collaboration), Search for dark matter produced with an energetic jet or a hadronically decaying W or Z boson at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 07 (2017) 014.
[195] CMS Collaboration, Technical Report CMS-PAS-HIG-17-002, 2017.
[196] CMS Collaboration, Technical Report CMS-PAS-SUS-16-049, 2017.
[197] CMS Collaboration, Technical Report CMS-PAS-B2G-17-001, 2017.
[198] CMS Collaboration, Technical Report CMS-PAS-B2G-17-002, 2017.
[199] CMS Collaboration, Technical Report CMS-PAS-B2G-17-007, 2017.
[200] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Techniques for improved heavy particle searches with jet substructure, Phys. Rev. D **80**, 051501 (2009).
[201] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches, Phys. Rev. D **81**, 094023 (2010).
[202] D. Krohn, J. Thaler, and L.-T. Wang, Jet trimming, J. High Energy Phys. 02 (2010) 084.
[203] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, Towards an understanding of jet substructure, J. High Energy Phys. 09 (2013) 029.
[204] M. Dasgupta, A. Fregoso, S. Marzani, and A. Powling, Jet substructure with analytical methods, Eur. Phys. J. C **73**, 2623 (2013).
[205] A. J. Larkoski and J. Thaler, Unsafe but calculable: Ratios of angularities in perturbative QCD, J. High Energy Phys. 09 (2013) 137.
[206] A. J. Larkoski, S. Marzani, and J. Thaler, Sudakov safety in perturbative QCD, Phys. Rev. D **91**, 111501 (2015).
[207] A. Larkoski, S. Marzani, J. Thaler, A. Tripathee, and W. Xue, Exposing the QCD Splitting Function with CMS Open Data, Phys. Rev. Lett. **119**, 132003 (2017).
[208] K. Kauder (STAR Collaboration), Measurement of the shared momentum fraction zg using jet reconstruction in $p + p$ and au + au collisions with star, in *Hard Probes* (2016).
[209] K. Lapidus (ALICE Collaboration), Hard substructure of jets probed in p-pb collisions, in *Quark Matter* 2017.
[210] Y.-T. Chien and I. Vitev, Probing the hardest branching of jets in heavy ion collisions, arXiv:1608.07283.
[211] Y. Mehtar-Tani and K. Tywoniuk, Groomed jets in heavy-ion collisions: Sensitivity to medium-induced bremsstrahlung, J. High Energy Phys. 04 (2017) 125.
[212] CMS Collaboration, Technical Report CMS-PAS-PFT-09-001, CERN, 2009.
[213] CMS Collaboration, Technical Report CMS-PAS-PFT-10-001, 2010.
[214] V. Khachatryan *et al.* (CMS Collaboration), Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV, J. Instrum. **12**, P02014 (2017).
[215] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.
[216] CMS Collaboration, Jet Primary Dataset in AOD format from RunA of 2011 (/Jet/Run2011A-12Oct2013-v1/AOD), CERN Open Data Portal, http://dx.doi.org/10.7483/OPENDATA.CMS.UP77.P6PQ.
[217] R. Brun and F. Rademakers, ROOT: An object oriented data analysis framework, Nucl. Instrum. Methods Phys. Res., Sect. A **389**, 81 (1997).
[218] XRootD Project, http://xrootd.org.
[219] MIT Open Data Producer, https://github.com/tripatheea/MODProducer.
[220] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k(t) jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[221] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, Eur. Phys. J. C **72**, 1896 (2012).
[222] CMS Collaboration, MinimumBias primary dataset in AOD format from RunB of 2010 (/MinimumBias/Run2010B-Apr21ReReco-v1/AOD), CERN Open Data Portal, http://dx.doi.org/10.7483/OPENDATA.CMS.6BPY.XFRQ.

[223] CMS Collaboration, Technical Report CMS-PAS-QCD-10-011, CERN, 2010.
[224] M. De Gruttola, The $Z \to \mu^+\mu^-$ decay channel in the CMS experiment at LHC: From cross-section measurement with the 2010 7 TeV collision dataset to offline machine luminosity monitor, Ph.D. thesis, Università degli Studi di Napoli Federico II, 2010.
[225] CMS Collaboration, Technical Report CMS-PAS-EWK-10-004, CERN, 1900.
[226] G. Petrucciani, A. Rizzi, and C. Vuosalo, Mini-AOD: A new analysis data format for CMS, J. Phys. Conf. Ser. **664**, 072052 (2015).
[227] M. Dobbs and J. B. Hansen, The HepMC C++ Monte Carlo event record for high energy physics, Comput. Phys. Commun. **134**, 41 (2001).
[228] E. Boos *et al.*, Generic user process interface for event generators, in *Physics at TeV colliders. Proceedings, Euro Summer School, Les Houches, France, 2001* (2001), http://lss.fnal.gov/archive/preprint/fermilab-conf-01-496-t.shtml.
[229] J. Alwall *et al.*, A standard format for Les Houches event files, Comput. Phys. Commun. **176**, 300 (2007).
[230] M. Cacciari, G. P. Salam, and G. Soyez, The catchment area of jets, J. High Energy Phys. 04 (2008) 005.
[231] C. Patrignani *et al.* (Particle Data Group), Review of particle physics, Chin. Phys. C **40**, 100001 (2016).
[232] MIT Open Data Analyzer, https://github.com/tripatheea/MODAnalyzer.
[233] Fastjet contrib, http://fastjet.hepforge.org/contrib/.
[234] Creative Commons CC0 Waiver, http://creativecommons.org/publicdomain/zero/1.0.
[235] T. Sjostrand, S. Mrenna, and P. Skands, A brief introduction to PYTHIA 8.1, Comput. Phys. Commun. **178**, 852 (2008).
[236] J. Bellm *et al.*, Herwig 7.0/Herwig++ 3.0 release note, Eur. Phys. J. C **76**, 196 (2016).
[237] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter, Event generation with SHERPA 1.1, J. High Energy Phys. 02 (2009) 007.
[238] M. Cacciari, G. P. Salam, and G. Soyez, SoftKiller: A particle-level pileup removal method, Eur. Phys. J. C **75**, 59 (2015).
[239] H.-M. Chang, M. Procura, J. Thaler, and W. J. Waalewijn, Calculating Track-Based Observables for the LHC, Phys. Rev. Lett. **111**, 102002 (2013).
[240] H.-M. Chang, M. Procura, J. Thaler, and W. J. Waalewijn, Calculating track thrust with track functions, Phys. Rev. D **88**, 034030 (2013).
[241] G. Aad *et al.* (ATLAS Collaboration), Measurement of the charged-particle multiplicity inside jets from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector, Eur. Phys. J. C **76**, 322 (2016).
[242] F. Pandolfi and D. Del Re, Search for the Standard Model Higgs boson in the $H \to ZZ \to l^+l^-q\bar{}q$ decay channel at CMS, Ph.D. thesis, Zurich, 2013.
[243] C. F. Berger, T. Kúcs, and G. Sterman, Event shape/energy flow correlations, Phys. Rev. D **68**, 014012 (2003).
[244] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, and C. Lee, Jet shapes and jet algorithms in SCET, J. High Energy Phys. 11 (2010) 101.
[245] A. J. Larkoski, D. Neill, and J. Thaler, Jet shapes with the broadening axis, J. High Energy Phys. 04 (2014) 017.
[246] S. Catani, G. Turnock, and B. R. Webber, Jet broadening measures in e^+e^- annihilation, Phys. Lett. B **295**, 269 (1992).
[247] Y. L. Dokshitzer, A. Lucenti, G. Marchesini, and G. P. Salam, On the QCD analysis of jet broadening, J. High Energy Phys. 01 (1998) 011.
[248] A. Banfi, G. P. Salam, and G. Zanderighi, Principles of general final-state resummation and automated implementation, J. High Energy Phys. 03 (2005) 073.
[249] D. Bertolini, T. Chan, and J. Thaler, Jet observables without jet algorithms, J. High Energy Phys. 04 (2014) 013.
[250] G. Salam, E_t^∞ scheme (unpublished).
[251] M. Wobisch and T. Wengler, Hadronization corrections to jet cross-sections in deep inelastic scattering, arXiv:hep-ph/9907280.
[252] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, Better jet clustering algorithms, J. High Energy Phys. 08 (1997) 001.
[253] P. E. L. Rakow and B. R. Webber, Transverse momentum moments of hadron distributions in QCD jets, Nucl. Phys. **B191**, 63 (1981).
[254] *Physics of the Superconducting Supercollider: Proceedings, 1986 Summer Study, 1986, Snowmass, Colorado*, edited by R. Donaldson and J. N. Marx (American Institute of Physics, New York, 1988).
[255] E. Farhi, A QCD Test for Jets, Phys. Rev. Lett. **39**, 1587 (1977).
[256] F. Maltoni, M. Selvaggi, and J. Thaler, Exposing the dead cone effect with jet substructure techniques, Phys. Rev. D **94**, 054015 (2016).
[257] P. Ilten, N. L. Rodd, J. Thaler, and M. Williams, Disentangling heavy flavor at colliders, arXiv:1702.02947.
[258] C. Frye, A. J. Larkoski, M. D. Schwartz, and K. Yan, Precision physics with pile-up insensitive observables, arXiv:1603.06375.
[259] C. Frye, A. J. Larkoski, M. D. Schwartz, and K. Yan, Factorization for groomed jet substructure beyond the next-to-leading logarithm, J. High Energy Phys. 07 (2016) 064.
[260] S. Marzani, L. Schunk, and G. Soyez, A study of jet mass distributions with grooming, J. High Energy Phys. 07 (2017) 132.
[261] S. D. Ellis, A. Hornig, D. Krohn, and T. S. Roy, On statistical aspects of Qjets, J. High Energy Phys. 01 (2015) 022.
[262] G. P. Salam, L. Schunk, and G. Soyez, Dichroic subjettiness ratios to distinguish colour flows in boosted boson tagging, J. High Energy Phys. 03 (2017) 022.
[263] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.
[264] R. D. Ball *et al.*, Parton distributions with LHC data, Nucl. Phys. **B867**, 244 (2013).
[265] S. Chatrchyan *et al.* (CMS Collaboration), Measurement of the Inclusive Jet Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV, Phys. Rev. Lett. **107**, 132001 (2011).
[266] ATLAS Open Data Portal, http://opendata.atlas.cern/.