Exploring the space of jets with CMS open data

Exploring the space of jets with CMS open data

Patrick T. Komiske,[1,2,*] Radha Mastandrea,[1,†] Eric M. Metodiev,[1,2,‡] Preksha Naik,[1,§] and Jesse Thaler[1,2,‖]

[1]*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[2]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*

(Received 7 September 2019; accepted 21 January 2020; published 11 February 2020)

We explore the metric space of jets using public collider data from the CMS experiment. Starting from 2.3 fb^{-1} of proton-proton collisions at $\sqrt{s} = 7$ TeV collected at the Large Hadron Collider in 2011, we isolate a sample of 1,690,984 central jets with transverse momentum above 375 GeV. To validate the performance of the CMS detector in reconstructing the energy flow of jets, we compare the CMS Open Data to corresponding simulated data samples for a variety of jet kinematic and substructure observables. Even without detector unfolding, we find very good agreement for track-based observables after using charged hadron subtraction to mitigate the impact of pileup. We perform a range of novel analyses, using the "energy mover's distance" (EMD) to measure the pairwise difference between jet energy flows. The EMD allows us to quantify the impact of detector effects, visualize the metric space of jets, extract correlation dimensions, and identify the most and least typical jet configurations. To facilitate future jet studies with CMS Open Data, we make our datasets and analysis code available, amounting to around two gigabytes of distilled data and one hundred gigabytes of simulation files.

DOI: 10.1103/PhysRevD.101.034009

I. INTRODUCTION

Ever since the first evidence for jet structure [1], the fragmentation of short-distance quarks and gluons into long-distance hadrons has been a rich area for experimental and theoretical investigations into quantum chromodynamics (QCD). A variety of observables have been proposed over the decades to probe the jet formation process [2–8], especially with recent advances in the field of jet substructure [9–20]. The stress-energy flow [21–23] is a particularly powerful probe of jets, since it in principle contains all the information about a jet that is infrared and collinear (IRC) safe [24–26]. A variety of observables have been built around the energy flow concept [27–31], including recent work on machine learning for jet substructure [32–34].

The unprecedented release of public collider data by the CMS experiment [35] starting in November 2014 [36] has enabled new exploratory studies of jets. The first such jet analyses [37,38] were performed using the CMS 2010 Open Data [39], corresponding to 31.8 pb^{-1} of 7 TeV data from Run 2010B at the Large Hadron Collider (LHC). Among other aspects of jets, these studies explored the groomed momentum fraction z_g [40], which has subsequently been measured in proton-proton and heavy-ion collisions by CMS [41], ALICE [42], and STAR [43]. The CMS Open Data release from LHC Run 2011A includes detector-simulated Monte Carlo (MC) samples, facilitating machine learning studies [44–46], an underlying event study [47], as well as a novel search for dimuon resonances [48]. CMS has also released data from Runs 2012B and 2012C, which have been used to search for nonstandard sources of parity violation in jets [49] and extract standard model cross sections [50]. Beyond CMS, archival ALEPH data [51] have been used by Ref. [52] to search for new physics and by Refs. [53–55] to perform QCD studies. While analyses using public collider data cannot match the sophistication or scope of official measurements by the experimental collaborations, they can enable proof-of-concept collider investigations and help stress-test archival data strategies.

In this paper, we perform the first exploratory study of the "space" of jets using the CMS 2011 Open Data. This data and MC release corresponds to 2.3 fb^{-1} of proton-proton collisions collected at a center-of-mass energy of $\sqrt{s} = 7$ TeV. The key idea, as proposed in Ref. [56], is to compute the pairwise distance between jet energy flows, and then use this information to construct a metric space.

*pkomiske@mit.edu
†rmastand@mit.edu
‡metodiev@mit.edu
§prekshan@mit.edu
‖jthaler@mit.edu

This enables a variety of distance-based jet analyses, including quantitative characterizations and qualitative visualizations. Because this is an exploratory study, we do not unfold for detector effects nor estimate systematic uncertainties, but the general agreement between the CMS Open Data and simulated MC samples provides evidence for the experimental robustness of these methods.

The metric we use is the "energy mover's distance" (EMD) [56], inspired by the famous earth mover's distance [57–61] sharing the same acronym. The EMD has units of energy (i.e., GeV) and quantifies the amount of "work" in energy times angle to make one jet radiation pattern look like another, including the cost of creating energy for jets with different p_T. While we focus on the EMD between pairs of jets in this study, the same concept could be applied to pairs of events as a whole. Crucially, the CMS Open Data contains full information about reconstructed particle flow candidates (PFCs) [62–64], which provide a robust proxy for the energy flow of a jet. It also contains information about primary vertices, allowing us to mitigate pileup (multiple proton-proton collisions per beam crossing) through charged hadron subtraction (CHS) [65]. Because of the improved resolution and pileup insensitivity of charged particles (i.e., tracks), we use a track-based variant of EMD for these exploratory studies.

We base our study on the CMS 2011 `Jet` primary dataset [66] and focus on the `HLT_Jet300` single-jet trigger, which we show is fully efficient to reconstruct jets with transverse momentum (p_T) above 375 GeV. We also use dijet MC samples [67–81], generated with PYTHIA 6 [82] and simulated using GEANT 4 [83], to understand the performance of the CMS detector in reconstructing the jet energy flow. In order to facilitate future jet studies on the CMS Open Data, we make our MIT Open Data (MOD) software framework available [84,85], along with the distilled data [86] and MC [87–94] files needed to recreate the majority of our studies.

The remainder of this paper is organized as follows. We begin in Sec. II by describing the CMS Open Data and the MOD software framework used for our analysis. In Sec. III, we validate the `Jet` primary dataset by comparing the basic kinematic and substructure properties of jets between the CMS data and MC samples. The core of our analysis is in Sec. IV, where we perform a variety of exploratory studies using the EMD. We conclude in Sec. V with a discussion of future jet studies on public collider data.

II. PROCESSING THE CMS OPEN DATA

In this section, we describe the main steps for processing the CMS Open Data. Our eventual analyses will be based on a single unprescaled trigger above its turn-on threshold, but we include additional details here about the analysis pipeline in order to demonstrate the general capabilities of our framework. The reader already familiar with how CMS data is processed can safely skip to Sec. II E, where we describe the baseline jet selection criteria used for our substructure and EMD studies.

A. Jet primary dataset

The CMS Open Data is available on the CERN Open Data Portal [36], which currently hosts data collected by CMS in 2010 [95], 2011 [96], and 2012 [97], as well as specialized samples for machine learning studies [98]. It also contains limited datasets from ALICE [99], ATLAS [100], and LHCb [101], as well as data from the OPERA neutrino experiment [102]. Accompanying the CMS 2011 Open Data is a virtual machine which runs version 5.3.32 of the CMS software (CMSSW) framework. This open data initiative complements efforts like HEPDATA [103], RIVET [104], and REANA [105] to preserve the results and workflows of official collider analyses (see further discussion in Ref. [106]).

The CMS Open Data is grouped into primary datasets that contain a subset of the triggers used for event selection [107]. There are 19 primary datasets included in the 2011 release, along with corresponding MC samples (see Sec. II D below). All of the primary datasets are provided by CMS in their analysis object data (AOD) format, which provides high-level reconstructed objects used for the bulk of official CMS analyses in Run 1. A subsample of some primary datasets (e.g., `Jet` [108] and `MinimumBias` [109]) are also provided in the RAW format, containing the full readout of the CMS detector.

Our analysis is based on the `Jet` primary dataset [66], which includes a variety of single jet and dijet triggers. This primary dataset contains 30,726,331 events spread across 1,223 AOD files, totaling 4.7 TB. The 2011A data-taking period is subdivided into 318 runs, and the runs are subdivided into 109,428 luminosity blocks (LBs) [110]. A luminosity block is the smallest unit of data-taking for which there is calibrated luminosity information, and during one block, the triggers are guaranteed to have consistent requirements and prescale factors (see Sec. II C below). Of the events in the `Jet` primary dataset, 26,275,768 are contained in "valid" LBs which are certified by CMS for use in physics analyses [111].

Each event in the AOD format has a complete list of PFCs, which are particle-like objects containing the reconstructed four-momentum and a probable particle identification (PID) code. In addition, the AOD format has AK5 jets, which are clusters of PFCs identified by the anti-k_t jet algorithm [112] with radius parameter $R = 0.5$. Jet energy correction (JEC) factors are obtained for the AK5 jets, including a correction for pileup using the area-median subtraction procedure [113]. The jets also have the information needed to impose jet quality criteria (JQC).

B. MIT open data framework

Because of the technical challenges involved in using CMSSW, we only use it to extract information from the

AOD files, performing the actual physics analyses outside of the virtual machine. Building on the MOD software framework introduced in Ref. [38], we use a custom `MODProducer` module in CMSSW to translate each AOD file into a plain text MOD file. We then use a custom framework called `MODAnalyzer` to read in each MOD file and perform various jet analysis tasks using FASTJET 3.3.1 [114]. Finally, we convert the MOD files into HDF5 [115] files for universal usability.

As described in more detail in Sec. II E, we consider the hardest and second-hardest jets for our analysis, after correcting the jet p_T using the JEC factors and imposing the "medium" JQC [116,117]. To access the constituents of jets, we first recluster the complete set of PFCs into AK5 jets and then compare against the CMS-provided preclustered AK5 objects. Due to rare numerical rounding issues, there are cases where the AK5 objects disagree, and we discard jets whose transverse momenta differ from the CMS-provided jets by more than one part in 10^6 or whose four-vectors are more than 10^{-6} apart in the rapidity-azimuth plane. When the AK5 objects agree, we associate them in the HDF5 files.

A number of substantial improvements have been made to `MODProducer` compared to Ref. [38]. We have added additional physics information in the MOD format, including metadata about files, LBs, and triggers. We have added primary vertex information to implement CHS for pileup mitigation (see Sec. III B below), made possible because the AOD files have a `VertexCollection` handle that can assign a charged-particle track to the closest collision vertex. We also added the ability to process MC files provided by CMS in the AODSIM format, including both generation-level particles and reconstructed PFCs.

After the jet selection stage in `MODAnalyzer`, the rest of our workflow is in PYTHON 3. We used NUMPY [118] for data manipulation, MATPLOTLIB [119] to produce figures, PYTHON OPTIMAL TRANSPORT [120] to calculate the EMD, and ENERGYFLOW 0.13 [84] for a variety of jet analysis tasks. In addition, we embedded our code in JUPYTER notebooks [121] for enhanced transparency and portability. To assist future jet studies on the CMS Open Data, our complete set of JUPYTER notebooks is available [85], and the corresponding reduced jet datasets are on the ZENODO platform [86–94].

C. Triggers, prescales, and luminosities

The `Jet` primary dataset contains 30 triggers [107]. We summarize these triggers in Table I, indicating the number of valid LBs and events for which the trigger is present, as well as the number of valid events for which the trigger fired. There are single jet and dijet triggers, where the trigger names include the nominal p_T requirement for the jet(s). For simplicity, we do not distinguish between trigger versions, denoted by suffixes like `_v2`, in our analysis. (The documentation for Ref. [66] lists 5 `L1FastJet` trigger variants in the `Jet` primary dataset, but as far as we can tell, these triggers were introduced after Run 2011A was complete.)

There are 7 triggers that were operational during the entire 2011A run, corresponding to 109,339 LBs. This can be compared to the luminosity information in Ref. [110], which lists 109,428 valid LBs in this run, leaving 89 LBs unaccounted for in the `Jet` primary dataset. These "missing" LBs only contribute 6 nb^{-1} to the recorded integrated luminosity, so their absence has a negligible impact on our studies. We investigate the missing LBs in more detail in Appendix A. There are also 643 LBs that are on the list of validated runs [111] but absent from the luminosity table [110]; we omit these from our analysis under the assumption that they are not in fact valid runs. Finally, we omit 143 valid LBs that contain events but have zero recorded luminosity, and we investigate these "zeroed" LBs further in Appendix A.

Because the total data-taking rate is limited, the lower p_T jet triggers are prescaled to only fire a fraction of the time they are active. The prescale factors satisfy $p^{\text{trig}} \geq 1$, with $p^{\text{trig}} = 1$ indicating an unprescaled trigger. (Strictly speaking, there are separate and independent prescale factors for the Level 1 (L1) trigger and the high-level trigger (HLT), but we always use p^{trig} to refer to the product of these factors.) The trigger prescale factors are fixed within a LB but can change between LBs. The effective luminosity for a given trigger is:

$$\mathcal{L}_{\text{eff}}^{\text{trig}} = \sum_{b \in \text{LBs}} \frac{\mathcal{L}_b}{p_b^{\text{trig}}}, \tag{1}$$

where b labels a LB, $\mathcal{L}_b$ is the recorded integrated luminosity in that block, and p_b^{trig} is the associated prescale factor. The effective luminosities for the `Jet` primary dataset triggers are reported in Table I, along with their average prescale factors and effective cross sections:

$$\langle p^{\text{trig}} \rangle = \frac{\mathcal{L}_{\text{total}}^{\text{trig}}}{\mathcal{L}_{\text{eff}}^{\text{trig}}}, \qquad \sigma_{\text{eff}}^{\text{trig}} = \frac{N^{\text{trig}}}{\mathcal{L}_{\text{eff}}^{\text{trig}}}, \tag{2}$$

where $\mathcal{L}_{\text{total}}^{\text{trig}} = \sum_b \mathcal{L}_b$ is the total luminosity of the run while the trigger was present, and N^{trig} is the total number of events for which the trigger fired.

Our analysis is based on the substructure of individual jets, so we focus our attention on the 9 single-jet triggers in Table I, omitting `HLT_Jet800` since it contains relatively few events. Their effective luminosities as a function of the number of cumulative time-ordered LBs are plotted in Fig. 1(a). We see that as the integrated luminosity increases, some of jet triggers have to be prescaled. We also see that the `HLT_Jet300` trigger only starts acquiring data partway through the 2011A run, coinciding with the `HLT_Jet240` trigger being prescaled.

TABLE I. Triggers in the CMS 2011A `Jet` primary dataset [66], restricted to LBs that have been identified as valid for physics analyses by CMS [111] and that have nonzero recorded luminosity [110]. Shown are the number of valid LBs and events for which the trigger is present and the number of valid events for which the trigger fired. Also provided are the effective luminosity $\mathcal{L}_{\text{eff}}^{\text{trig}}$ defined in Eq. (1), and the average prescale value $\langle p^{\text{trig}} \rangle$ and effective cross section $\sigma_{\text{eff}}^{\text{trig}}$ defined in Eq. (2). As discussed in Appendix A, there are 89 "missing" LBs in the CMS 2011A luminosity table [110] that are not represented in the `Jet` primary dataset, but they have a negligible impact on our analysis. We also omit 143 "zeroed" LBs during which events were detected but zero luminosity was recorded. The **HLT_Jet300** trigger (bolded) is the one used for the jet studies in Secs. III and IV.

Trigger name	LBs	Events	Fired	$\mathcal{L}_{\text{eff}}^{\text{trig}}$ [nb^{-1}]	$\langle p^{\text{trig}} \rangle$	$\sigma_{\text{eff}}^{\text{trig}}$ [nb]
HLT_Jet30	109,196	26,254,892	1,884,768	12.567	185,672.632	149,981.925
HLT_Jet60	109,196	26,254,892	1,829,490	293.986	7,936.716	6,223.060
HLT_Jet80	102,304	24,742,482	1,512,638	901.352	2,293.846	1,678.188
HLT_Jet110	109,196	26,254,892	2,212,878	6,172.430	378.016	358.510
HLT_Jet150	102,304	24,742,482	2,616,716	33,521.114	61.679	78.062
HLT_Jet190	109,196	26,254,892	2,715,282	114,843.687	20.317	23.643
HLT_Jet240	109,196	26,254,892	2,806,220	392,659.479	5.942	7.147
HLT_Jet300	**98,462**	**22,788,815**	**4,616,184**	**2,284,792.618**	**1.000**	**2.020**
HLT_Jet370	109,196	26,254,892	1,514,305	2,333,280.071	1.000	0.649
HLT_Jet800	47,156	10,578,173	23,332	1,414,462.687	1.000	0.016
HLT_DiJetAve30	98,462	22,788,815	1,394,369	20.585	110,990.490	67,735.556
HLT_DiJetAve60	98,462	22,788,815	1,440,740	539.491	4,235.090	2,670.555
HLT_DiJetAve80	91,570	21,276,405	1,059,885	1,474.722	1,369.123	718.702
HLT_DiJetAve110	98,462	22,788,815	1,714,381	10,583.561	215.881	161.985
HLT_DiJetAve150	91,570	21,276,405	2,162,760	59,292.115	34.053	36.476
HLT_DiJetAve190	98,462	22,788,815	2,343,401	208,109.103	10.979	11.260
HLT_DiJetAve240	98,462	22,788,815	2,697,899	800,844.351	2.853	3.369
HLT_DiJetAve300	98,462	22,788,815	2,356,128	2,284,792.618	1.000	1.031
HLT_DiJetAve370	98,462	22,788,815	741,410	2,284,792.618	1.000	0.324
HLT_DiJetAve15U	10,734	3,466,077	225,367	1.841	26,335.253	122,404.801
HLT_DiJetAve30U	10,734	3,466,077	353,409	45.628	1,062.680	7,745.523
HLT_DiJetAve50U	10,734	3,466,077	339,051	298.084	162.664	1,137.434
HLT_DiJetAve70U	10,734	3,466,077	624,758	2,061.075	23.525	303.122
HLT_DiJetAve100U	10,734	3,466,077	301,727	4,314.114	11.239	69.940
HLT_DiJetAve140U	10,734	3,466,077	415,806	25,144.074	1.928	16.537
HLT_DiJetAve180U	10,734	3,466,077	255,163	48,487.453	1.000	5.262
HLT_DiJetAve300U	10,734	3,466,077	21,347	48,487.453	1.000	0.440
HLT_Jet240_CentralJet30_BTagIP	47,156	10,578,173	2,216,488	1,414,462.687	1.000	1.567
HLT_Jet270_CentralJet30_BTagIP	47,156	10,578,173	1,280,355	1,414,462.687	1.000	0.905
HLT_Jet370_NoJetID	109,196	26,254,892	1,711,067	2,333,280.071	1.000	0.733
Missing	89			6.066		
Zeroed	143	20,876				
Total	109,428	26,275,768	26,275,768	2,333,286.137		

In Fig. 1(b), we plot the effective cross section in each time-ordered LB for the 9 single-jet triggers. The trigger behaviors are relatively stable over the course of the 2011A run, though there is a noticeable shift in the `HLT_Jet80` trigger when its selection criteria changed. One can also see when the `HLT_Jet300` trigger turned on and when the `HLT_Jet80` and `HLT_Jet150` triggers were turned off.

Since `HLT_Jet300` is the lowest p_T single-jet trigger that is unprescaled, it will be the sole trigger used in our substructure and EMD studies (see further discussion in Sec. II E). For reference, the recorded luminosity for `HLT_Jet300` as a function of time is plotted in Fig. 18 of Appendix A.

D. Monte Carlo event samples

A key feature of the CMS 2011 data release compared to the initial one from 2010 is the inclusion of MC event samples. (Some MC samples corresponding to the 2010 dataset have been subsequently released.) For our analysis, we use samples of hard QCD scattering generated by PYTHIA 6.4.25 [82] with tune Z2 [122]. As summarized in Table II, there are 15 samples with nonoverlapping hard-scattering parton $\hat{p}_T$ ranges [67–81], totaling 13.4 TB. They are labeled by CMS as `QCD_Pt-MINtoMAX_TuneZ2_7TeV_pythia6`, where $\hat{p}_T \in [\text{MIN}, \text{MAX}]$ GeV. These events are then simulated and

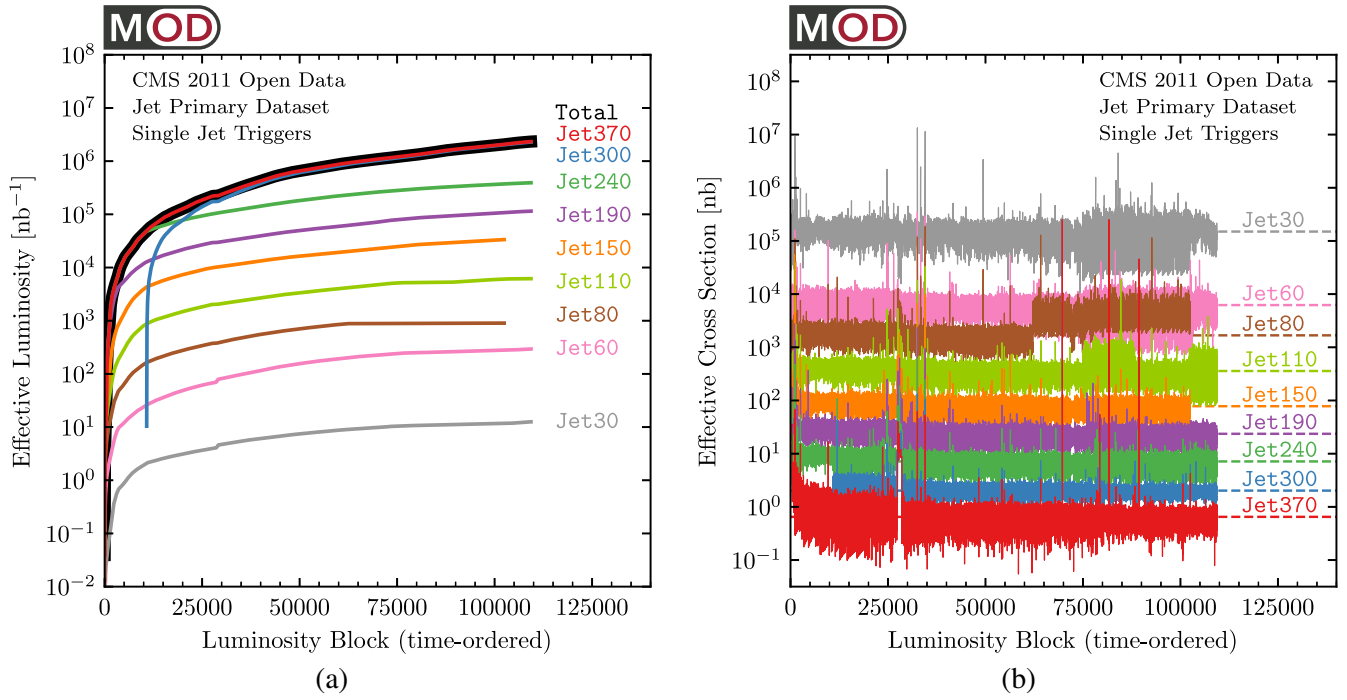

FIG. 1. (a) Effective luminosity for the single-jet triggers as a function of the cumulative number of LBs, ordered in time. Note that the `Jet300` trigger used for our jet studies turns on after around 50 pb^{-1} has already been collected, but this is a relatively small fraction of the total 2.3 fb^{-1} collected over the course of Run 2011A. The luminosity profile as a function of date is shown in Fig. 18 of Appendix A. (b) Effective cross section for the single-jet triggers in each LB where the trigger fired. The flatness of these curves indicates that the trigger behavior is roughly constant across the entire run, apart from moments where the trigger criteria or prescale factors changed. The horizontal dashed lines correspond to the total effective cross section for that trigger from Table I.

reconstructed using the CMS detector simulation based on GEANT 4 [83]. Throughout this paper, we use "generation" to refer to the output of the parton shower generator, and "simulation" to refer to the output of the detector simulation.

Both the generation-level and simulation-level objects are stored in AODSIM format by CMS, and we convert them to our MOD format using `MODProducer`. Apart from the generation-level event record from PYTHIA, the AODSIM format is very similar to AOD. In particular, AODSIM includes reconstructed AK5 jets, simulated trigger information, as well as the addition of pileup. We store the simulated PFCs, the final-state particles in the PYTHIA event record, and the $2 \to 2$ hard-scattering process for anticipated future studies related to parton flavor. If an association between simulation-level and generation-level jets is needed, jets are matched if their jet axes are within $\Delta R = 0.5$ of each other. To enable future jet flavor studies, generation-level jets are also matched to hard-process partons if they are less than $\Delta R = 1.0$ apart.

Because of the steep dependence of the QCD dijet cross section on $\hat{p}_T$, the MC events have different weights, though the weights for all events in a single MC sample are the same. Therefore, when filling histograms, we have to

TABLE II. Information about the MC event samples provided by CMS [67–81] from the PYTHIA 6 hard QCD scattering process. Shown are the generator-level $\hat{p}_T$ ranges, the number of files and events in each sample, and the effective cross section $\sigma_{\text{eff}}^{\text{MC}}$. Only the 8 samples with $\hat{p}_T > 170$ GeV are required for the jet studies in Secs. III and IV.

$\hat{p}_T^{\text{min}}$–$\hat{p}_T^{\text{max}}$ [GeV]	Files	Events	$\sigma_{\text{eff}}^{\text{MC}}$ [nb]	DOI
0–5	55	1,000,025	4.84×10^{7}	[67]
5–15	83	1,495,884	3.68×10^{7}	[68]
15–30	5,519	9,978,850	8.16×10^{5}	[69]
30–50	277	5,837,856	5.31×10^{4}	[70]
50–80	299	5,766,430	6.36×10^{3}	[71]
80–120	317	5,867,864	7.84×10^{2}	[72]
120–170	334	5,963,264	1.15×10^{2}	[73]
170–300	387	5,975,592	2.43×10^{1}	[74]
300–470	382	5,975,016	1.17×10^{0}	[75]
470–600	274	3,967,154	7.02×10^{-2}	[76]
600–800	271	3,988,701	1.56×10^{-2}	[77]
800–1000	295	3,945,269	1.84×10^{-3}	[78]
1000–1400	131	1,956,893	3.32×10^{-4}	[79]
1400–1800	182	1,991,792	1.09×10^{-5}	[80]
1800–∞	75	996,500	3.58×10^{-7}	[81]

weight each MC event by the generated cross section $\sigma_{\text{eff}}^{\text{MC}}$ divided by the number of events in the MC sample, as given in Table II. As discussed in Appendix B, we also weight the MC events according to the number of primary vertices in order to match the distribution of pileup seen in the data.

One subtlety in using the generation-level PYTHIA information is that there is a cutoff on the hadron lifetime above which they are considered stable. This cutoff is set to $c\tau_{\text{stable}} = 10$ mm, which means that various hadrons with nonzero strangeness are considered stable, notably the K_S^0 meson. Typically, these strange hadrons decay within the CMS detector volume and are often reconstructed as if the decay products came from the primary vertex. For example, $K_S^0 \to \pi^+\pi^-$ will typically be reconstructed as two pion-labeled PFCs. This leads to a mismatch in observables like track multiplicity unless we manually decay these strange hadrons. As a work-around, we load the generation-level event record into PYTHIA 8.235 [123] and adjust the hadron lifetime threshold to $c\tau_{\text{stable}} = 1000$ mm. Because the kinematics and flavors of the hadron decay will not be the same as in the CMS detector simulation, there is a slight mismatch when comparing a generation-level event to its simulation-level counterpart, though this issue does not arise when comparing histograms.

E. Jet and trigger selection

The jet studies in Secs. III and IV are based on the two hardest p_T jets in an event. This is motivated by the fact that $2 \to 2$ QCD dijet production at leading order yields two jets of equal p_T. Therefore, considering the substructure of just the hardest p_T jet (as in the studies of Refs. [37,38]) is IRC unsafe, since an infinitesimally soft emission can change the relative jet ordering. On the other hand, considering more than two jets requires information beyond leading order, so we only consider the two hardest p_T jets in our analysis. (See Ref. [124] for further discussions of single-jet inclusive cross section definitions.)

TABLE III. The jet quality criteria based on CMS recommendations for $|\eta^{\text{jet}}|<2.4$. For $|\eta^{\text{jet}}|>2.4$, where tracking information is not available, the charged particle criteria are not applied and all particles are treated as neutral. For our analysis, we impose the "medium" criteria.

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

The CMS single-jet triggers are designed to fire any time an event has a jet whose p_T is above a given threshold. We independently analyze the two hardest jets in an event, correcting their p_T values by the appropriate JEC factors. When we perform our substructure analysis, we require that

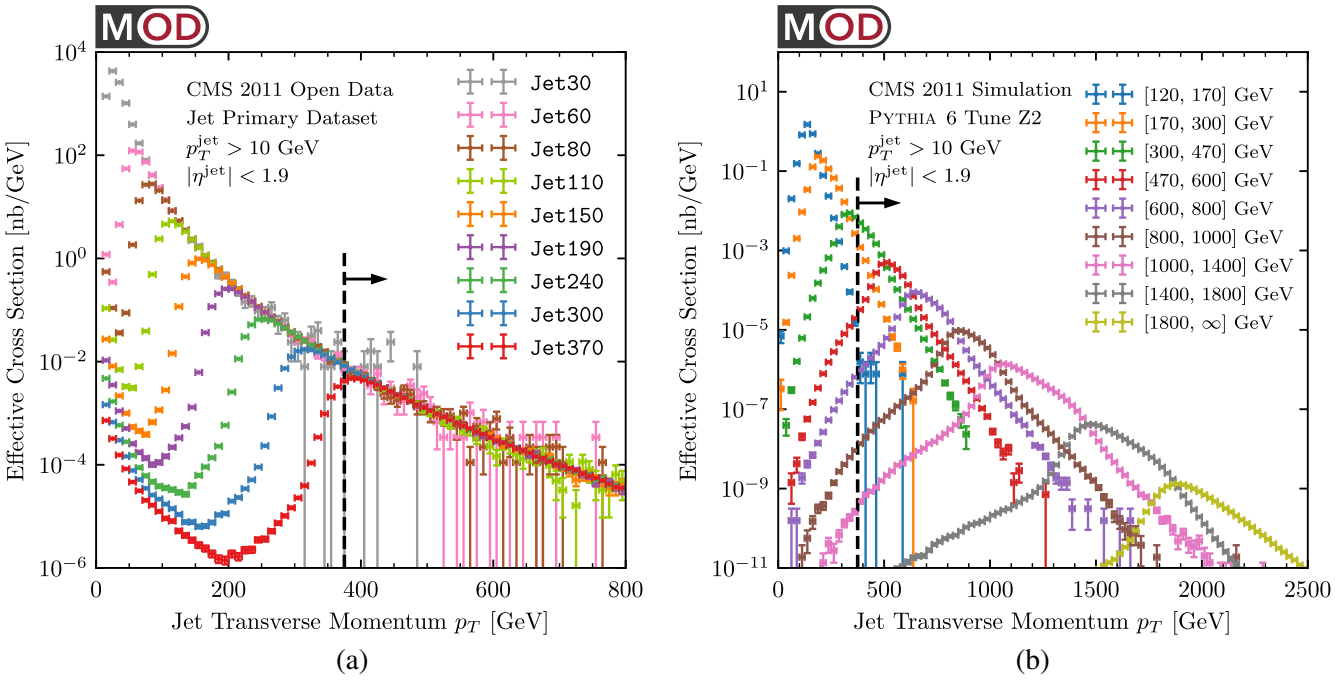

FIG. 2. The p_T spectrum for the hardest jet in (a) the 9 single-jet triggers and (b) the 9 relevant simulated MC samples, restricted to $|\eta^{\text{jet}}| < 1.9$. These jet spectra have JEC factors included and medium JQC imposed. The vertical dashed lines at 375 GeV indicate the jet p_T threshold used in this analysis.

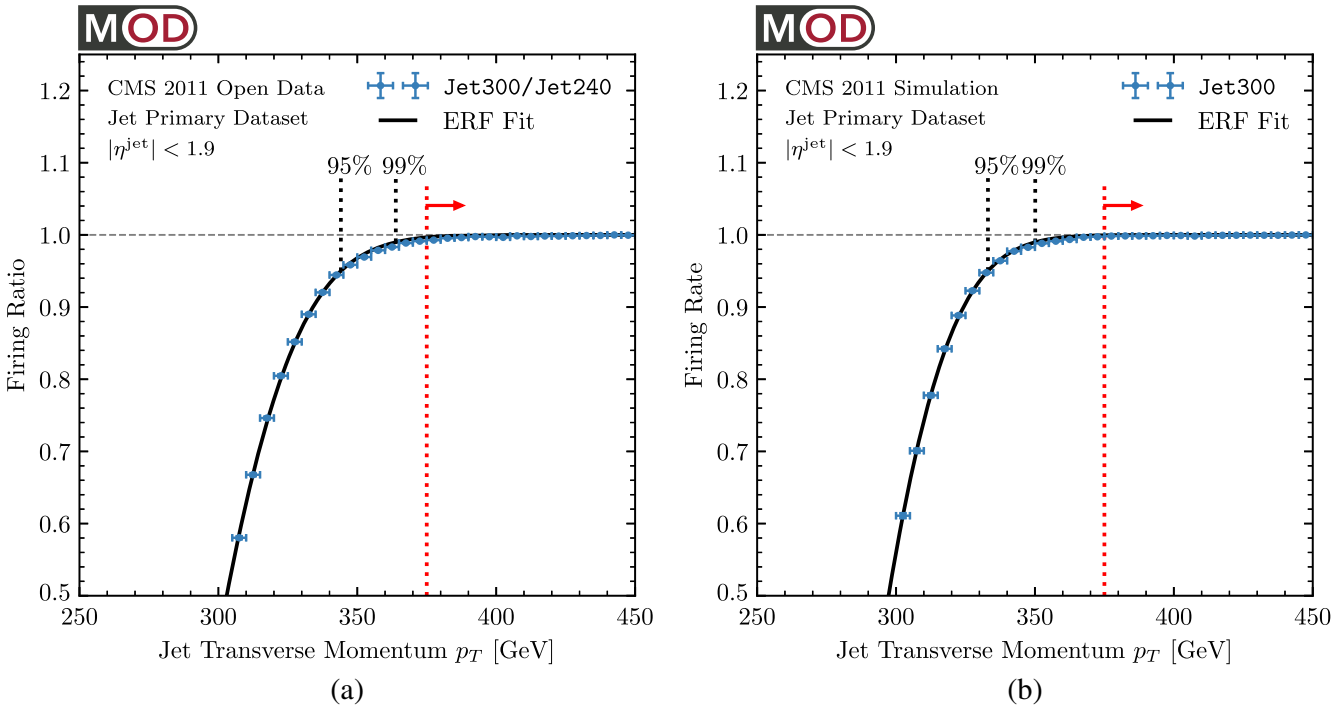

FIG. 3. Trigger turn on behavior as a function of reconstructed hardest jet p_T for the Jet300 trigger, including JEC factors. Shown are (a) the relative efficiency of the Jet300 trigger with respect to Jet240 in the CMS Open Data, and (b) the absolute efficiency of the Jet300 trigger in the MC simulation. Both of these curves are fit to an error function (ERF) to estimate the efficiency boundaries. From these, we conclude that the Jet300 trigger is fully efficient above $p_T^{\text{jet}} > 375$ GeV. This analysis is repeated for the other triggers in Fig. 21 of Appendix C.

the jets satisfy $|\eta^{\text{jet}}| < 1.9$ to make sure that the $R = 0.5$ jets are reconstructed fully within the tracking volume that covers $|\eta^{\text{tracker}}| < 2.4$. We impose "medium" JQC (see Table III) [116,117] throughout this study.

In Fig. 2(a), we show the p_T spectrum of just the hardest jet in the CMS 2011 Open Data, separated into the 9 single-jet triggers. [The spectrum for the two hardest jets will be shown in Fig. 5(a).] We see that the triggers start to collect an appreciable number of jets when the jet p_T matches the trigger name, asymptoting to a common smooth p_T spectrum. The small population of jets at low p_T values below the turn on is due primarily to trigger misfirings, for example from fake jets that do not satisfy the jet quality criteria. In Fig. 2(b), we show the same p_T spectrum in the CMS simulation, separated into the 9 most relevant MC samples for our analysis (out of 15 total). We see that the MC files have support mainly in their designated $\hat{p}_T$ ranges, albeit with a spread due to phenomena like initial state radiation (ISR) that change the overall event kinematics.

To simplify our physics studies, we use just one of the single-jet triggers. As mentioned above, we select HLT_Jet300 since this has the lowest p_T threshold among the unprescaled single-jet triggers. Looking at Fig. 2(a), we can estimate that Jet300 is fully efficient above $p_T > 375$ GeV. Looking at Fig. 2(b), we see that all of the MC samples with $\hat{p}_T > 170$ GeV contribute appreciably to the $p_T > 375$ GeV region, corresponding to 8 required MC event samples.

To determine where the Jet300 trigger is fully efficient, we compare its behavior to the Jet240 trigger; see related trigger efficiency studies in Refs. [107,125]. In Fig. 3(a), we consider events where the Jet300 trigger is present and the Jet240 trigger fired. We then plot the fraction of events where Jet300 fired as a function of jet p_T. Fitting the resulting fraction to an error function, we estimate that the Jet300 trigger is 99% efficient (relative to Jet240) at 367 GeV, justifying our choice of $p_T > 375$ GeV. We can cross check our trigger efficiency study using the simulated MC samples. In Fig. 3(b), we plot the fraction of events where the simulated Jet300 trigger fired as a function of jet p_T. Doing the same error function fit, we find that the simulated Jet300 trigger is 99% efficient (relative to an absolute scale) at 350 GeV, which is again consistent with our $p_T > 375$ GeV choice. For completeness, we provide efficiency plots for all of the triggers in Fig. 21 of Appendix C. Since we are performing an exploratory jet study, we do not correct for this small trigger inefficiency in our analysis.

Our initial workflow is summarized in Table IV. Because we consider the two hardest jets with $p_T^{\text{jet}} > 10$ GeV, there are about twice as many jets in the analysis as the number

TABLE IV. Initial workflow and event selection for the jet studies in Secs. III and IV. The selections in the first block ensure that the `Jet300` trigger fired in a valid LB, the requirements in the second block ensure that the `Jet300` trigger is fully efficient, and the cuts in the third block impose the JQC and the baseline analysis criteria. Because our analysis is based on the two hardest jets, there is an increase by a factor of about two between the first and second blocks.

	CMS 2011 open data	CMS 2011 simulation	PYTHIA 6 generation
Total events	30,726,331	28,796,917	21,802,470
Valid	26,254,892		
`Jet300` Trigger present	22,788,815		
`Jet300` Trigger fired	4,616,184	22,108,599	
Two hardest jets, $p_T^{\text{jet}} > 10$ GeV	9,106,775	44,217,198	43,604,940
$p_T^{\text{jet}} > 375$ GeV	1,785,625	35,155,818	35,267,080
AK5 match	1,785,625	35,155,790	
Medium JQC	1,731,255	35,145,175	
$\lvert\eta^{\text{jet}}\rvert < 1.9$	1,690,984	34,969,900	35,089,120
$p_T^{\text{jet}} \in [375, 425]$ GeV	879,046	2,379,525	2,203,305

of events. In order to have a more homogenous jet sample, we impose the narrower $p_T^{\text{jet}} \in [375, 425]$ GeV range for our substructure and EMD studies below. An example event from the CMS 2011 Open Data passing our kinematic jet selections is displayed in Fig. 4, including information about the charges and vertices of the PFCs.

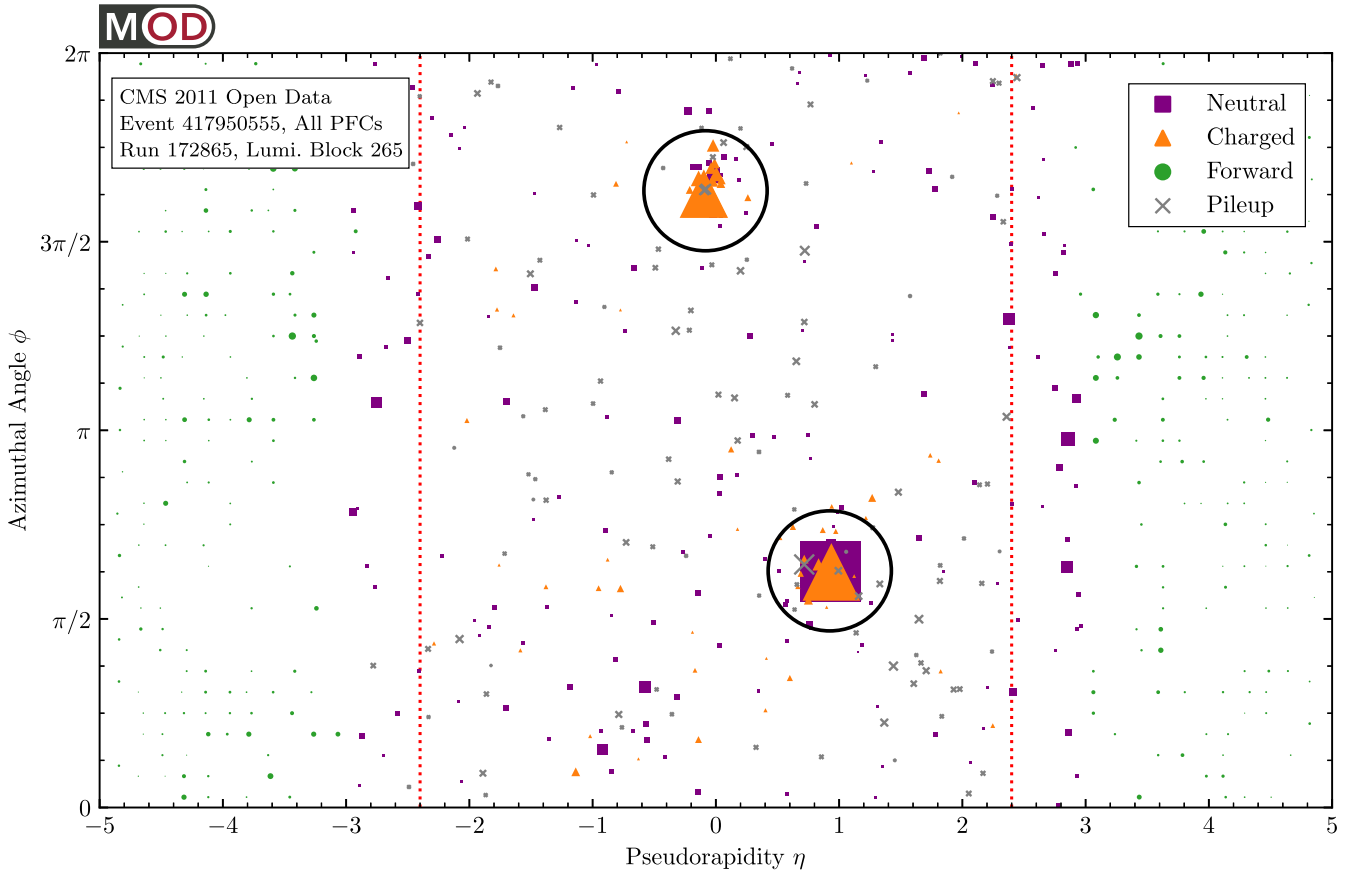

FIG. 4. Reconstructed PFCs in an example event from the CMS Open Data passing our jet selection criteria. The size of the symbol indicates the PFC transverse momentum and the style indicates its charge, with purple squares for neutral PFCs, orange triangles for charged PFCs, and green circles for PFCs in the forward region where no charge information is available. Charged pileup PFCs removed by CHS are indicated as gray crosses. The leading two jets are shown as circles of radius $R = 0.5$, and the tracking region $|\eta| < 2.4$ is within the dashed, vertical lines.

III. ANALYZING JET SUBSTRUCTURE

To validate the performance of the CMS detector for jet reconstruction, we present a variety of jet kinematics and jet substructure distributions derived from the CMS 2011 Open Data. There are two main differences compared to a similar analysis performed in Ref. [38]. First, we can now compare the open data distributions to detector-simulated MC samples to check for robustness. Second, we have proper luminosity information [110] such that we can plot (uncorrected) differential cross sections, instead of just normalized probability distributions.

A. Overall jet kinematics

In Fig. 5(a), we show the p_T spectrum of the two hardest jets (i.e., two histogram entries per event), restricted to the region $|\eta^{\text{jet}}| < 1.9$ and $p_T^{\text{jet}} > 375$ GeV. Here, we compare the CMS Open Data in black to the simulated MC samples in orange. We find very good agreement in the shape of the p_T spectrum after including appropriate K-factors described below, though there are small disagreements and discontinuities for $p_T^{\text{jet}} > 750$ GeV. We also show the generation-level PYTHIA distribution without detector simulation in blue, which matches very well to the orange simulation-level distribution with detector response, indicating that the overall JEC factors have been chosen appropriately. (Of course, the JEC factors also include data-driven corrections beyond just those captured by the detector simulation.) Note that these distributions only include statistical uncertainties, without any estimate of systematic uncertainties.

Because PYTHIA is a leading-order generator, we have rescaled the MC events by a next-to-leading-order (NLO) K-factor. This p_T-dependent K-factor is derived from Ref. [126] for $R = 0.5$ jets, with $K_{\text{NLO}} \simeq 1.135$ in the vicinity of 400 GeV. As discussed further in Appendix B, we reweight the MC in order that the pileup level in the simulation matches the data. Finally, we multiply by an additional factor of $K_{375} = 0.961$ to ensure that the lowest bin in the simulation has the same normalization as the actual data. This factor partially accounts for effects like the efficiency of the medium JQC, which is difficult to extract reliably from the CMS simulation, as well as QCD corrections beyond NLO and uncertainties on the recorded luminosity.

In Fig. 5(b), we show the jet pseudorapidity spectrum. After relaxing the $|\eta^{\text{jet}}| < 1.9$ requirement, we find a small number of jets at larger pseudorapidities. Compared to the simulated data, the open data has more jets in the vicinity of $|\eta^{\text{jet}}| \simeq 1.2$ and fewer in the vicinity of $|\eta^{\text{jet}}| \simeq 0.0$, indicating a possible issue with the PYTHIA prediction or with the pseudorapidity dependence of the JEC factors. That said,

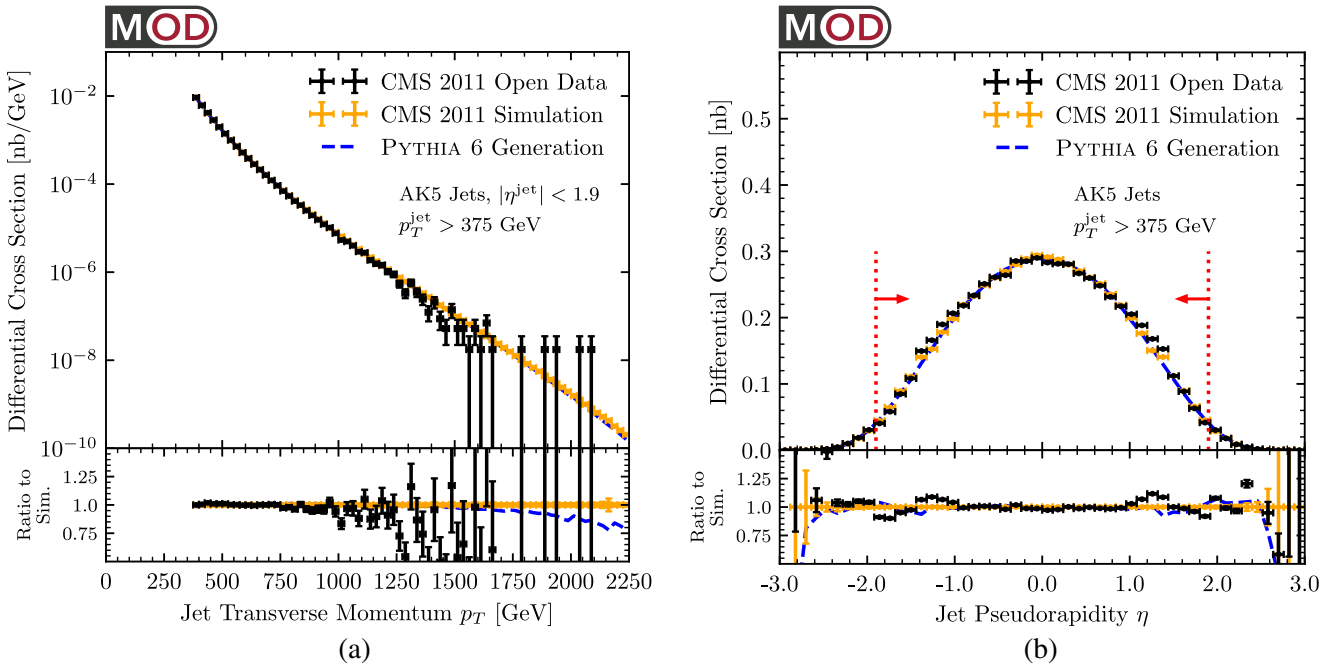

FIG. 5. (a) Jet transverse momentum spectrum, comparing the CMS Open Data to MC event samples at the simulation level and generation level. We consider up to two of the hardest p_T jets, restricted to $|\eta^{\text{jet}}| < 1.9$ and $p_T^{\text{jet}} > 375$ GeV. In addition to having a p_T-dependent NLO K-factor, the MC events have been normalized to match the lowest p_T bin. (b) Jet pseudorapidity spectrum, with the $|\eta^{\text{jet}}|$ requirement removed. For both jet spectra, we see very good agreement between data and simulation, indicating that we have properly processed the CMS Open Data, including appropriate JEC factors. In these and all subsequent plots, the error bars indicate statistical uncertainties only, with no attempt at estimating systematic uncertainties. The jet azimuth spectrum is shown in Fig. 22 of Appendix C.

the overall agreement is very good, giving us confidence that we can make basic kinematic jet selections. For completeness, the jet azimuth spectrum is shown in Fig. 22 of Appendix C, which exhibits the expected flat spectrum with small fluctuations due to detector inhomogeneities.

B. Jet constituents

In addition to the reconstructed AK5 jets, the CMS Open Data contains the complete list of PFCs, which allows us to calculate a wide range of jet substructure observables. Due to detector effects, one has to be careful when interpreting the PFC information. Ultimately, we will focus on track-based observables which have better reconstruction performance as well as better pileup stability.

In Table V, we list the PID codes of the PFCs and their absolute counts in the jet sample with $|\eta^{\text{jet}}| < 1.9$ and $p_T^{\text{jet}} \in [375, 425]$ GeV. Note that there are more events in the MC samples than in the open data, so there is a corresponding increase in the number of total PFCs. The PID codes indicate the most likely particle candidate, using the PDG MC numbering scheme [127]. In particular, code 211 includes π^+, K^+, and proton candidates, code 22 includes photon and merged $\pi^0 \to \gamma\gamma$ candidates, and code 130 includes K_L^0 and neutron candidates.

The counts in Table V include contamination from pileup. As shown in Fig. 19(a) of Appendix C, there are typically ~5 pileup events per beam crossing. While the CMS Open Data already includes a pileup correction for the jet p_T via the JEC factors, this is insufficient to correct substructure distributions. We have two ways to mitigate the effect of pileup. First, we apply the CHS procedure [65] to remove charged particles not associated with the primary vertex. This is possible since `MODProducer` now stores vertex information (see Sec. II B above), so we can remove charged jet constituents assigned to pileup vertices. Though CHS cannot remove neutral particles from pileup, it does reduce the overall pileup contamination by a factor of ~2/3. Second, inspired by the SoftKiller procedure [128], we impose a $p_T^{\text{PFC}} > 1$ GeV cut on all PFCs, where this value is motivated by Fig. 6 below. This helps control the level of neutral pileup, though we will still focus on track-based observables in our subsequent analyses.

The p_T spectrum of neutral PFCs is shown Fig. 6(a). The neutral PFCs do not benefit from CHS, so there is a significant excess of neutral PFCs from pileup below around 2 GeV, compared to generation-level expectations. That said, the CMS simulation appropriately captures this neutral pileup contamination. Because of finite calorimeter granularity, there is a depletion of moderate p_T neutral PFCs as a result of merging. This merging results in an excess of higher p_T neutral PFCs, which can be seen in Fig. 23(a) of Appendix C.

The p_T spectrum of charged PFCs is shown in Fig. 6(b). With CHS, the PFC p_T spectrum is rather similar between the CMS Open Data and the MC event samples, even at the generator level and even going out to higher p_T in Fig. 23(b) of Appendix C. The main difference is below 1 GeV, where one sees the impact of tracking inefficiencies and momentum misreconstruction. For this reason, we impose a cut of $p_T^{\text{PFC}} > 1$ GeV for all of our jet substructure studies, which results in better data/MC agreement for observables like track multiplicity that are sensitive to such effects. Note that this same p_T^{PFC} cut was advocated for in Ref. [38], though a looser cut of 500 MeV is used by CMS in its track multiplicity study [129].

C. Jet substructure observables

We now plot a representative sample of jet substructure observables, comparing the CMS Open Data to the MC samples, both before and after detector simulation. Based on the conclusions of Sec. III B, we always implement CHS and impose the $p_T^{\text{PFC}} > 1$ GeV cut. In order to analyze jets with similar total p_T, we focus on the relatively narrow range of $p_T^{\text{jet}} \in [375, 425]$ GeV.

TABLE V. Counts of PFCs by PID code, considering the constituents of the two hardest jets with the restriction $|\eta^{\text{jet}}| < 1.9$ and $p_T^{\text{jet}} \in [375, 425]$ GeV. The MC simulation has a larger number of events than the CMS Open Data, and therefore more total PFCs. Note that the PID code is based on the PDG MC numbering scheme, but a code like ± 211 indicates any charged hadron candidate, not solely $\pi^{\pm}$.

		CMS 2011 open data			CMS 2011 simulation		
PID	Candidate	Total count	After CHS	$p_T > 1$ GeV	Total count	After CHS	$p_T > 1$ GeV
11	Electron (e^-)	31,297	30,304	30,284	76,819	73,937	73,906
−11	Positron (e^+)	31,444	30,470	30,448	75,651	72,920	72,868
13	Muon (μ^-)	16,779	14,957	14,912	47,871	42,604	42,511
−13	Antimuon (μ^+)	17,453	15,373	15,310	50,009	44,256	44,149
211	Positive hadron (e.g. π^+)	10,731,634	8,159,520	6,950,019	31,682,518	23,267,103	19,775,066
−211	Negative hadron (e.g. π^-)	10,414,733	7,987,681	6,780,597	30,718,965	22,837,987	19,361,736
22	Photon (γ)	14,102,402	14,102,402	7,157,772	39,487,711	39,487,711	19,805,470
130	Neutral hadron (e.g. K_L^0)	2,955,136	2,955,136	2,317,806	7,509,228	7,509,228	5,974,028

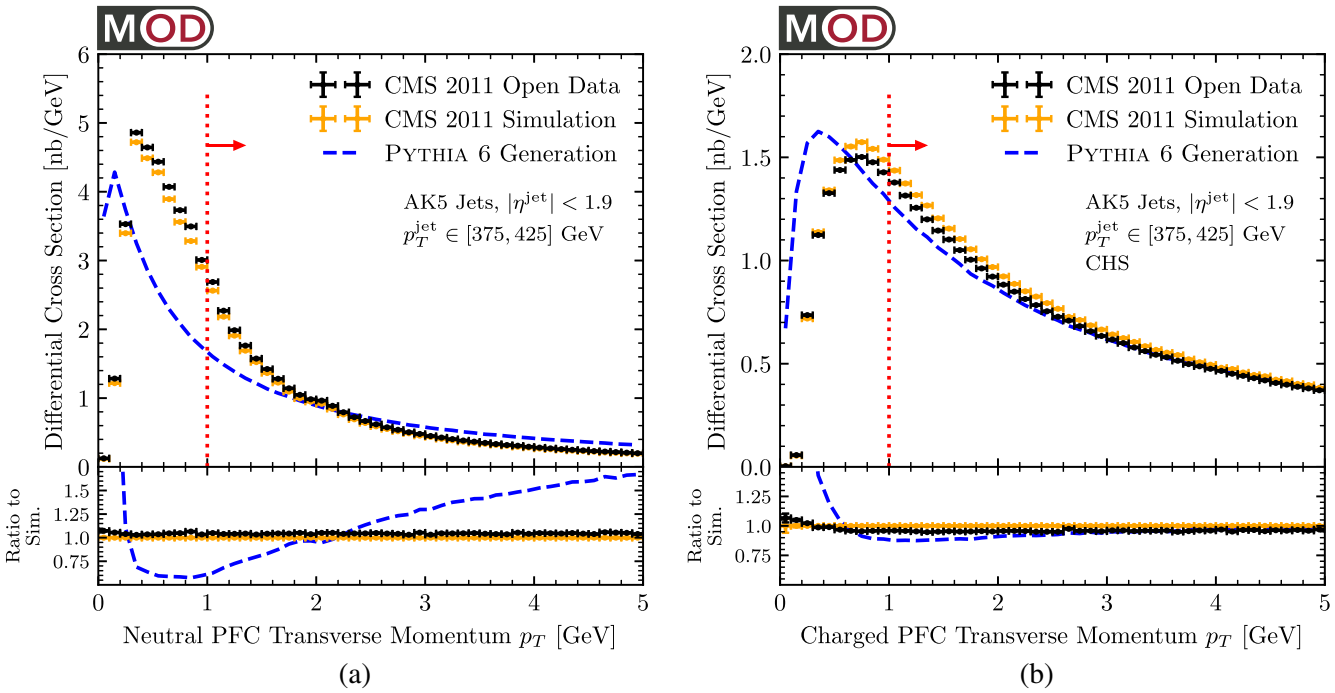

FIG. 6. Transverse momentum spectra for (a) neutral PFCs and (b) charged PFCs, including CHS to mitigate charged pileup, restricted to PFCs that are within the analyzed jets. The CMS simulation captures the key features of the CMS Open Data. Only for charged PFCs with $p_T^{\rm PFC} > 1$ GeV is there reasonable agreement with the generation-level expectations from PYTHIA. The complete PFC p_T spectrum is shown in Fig. 23 of Appendix C.

In Fig. 7, we show three classic substructure distributions: jet mass, constituent multiplicity, and p_T^D [130]. Using all PFCs, shown in the left column of Fig. 7, there is good agreement between the CMS Open Data and the simulation-level MC events. This suggests that PYTHIA 6 with tune Z2 has a reasonable model for jet fragmentation and that the CMS simulation provides a faithful characterization of the detector response; see related studies in Ref. [129], as well as Ref. [131] for alternative PYTHIA tunes.

That said, there are significant differences when comparing the generation-level and simulation-level MC distributions, even after applying CHS for pileup mitigation. Roughly speaking, the CMS detector reconstructs fewer PFCs than expected, which is consistent with merging of neutral PFCs due to finite calorimeter granularity. On the other hand, the CMS detector reconstructs a larger jet mass than expected, which is consistent with residual neutral pileup contamination.

We can improve the generation-level and simulation-level agreement by restricting our analysis to just charged PFCs, as shown in the right column of Fig. 7. The agreement improves most notably for the IRC-unsafe observables of multiplicity and p_T^D. While the CMS detector reconstructs fewer charged PFCs than expected from PYTHIA at the generation level, the difference is well within the theoretical uncertainties in MC generation (see further discussion in Ref. [132]). Since we will not attempt to unfold the data in this paper, it is important for us to use observables that are robust to detector effects. For this reason, the focus of our EMD studies will be on track-based observables.

It is worth remarking that the good agreement in the track multiplicity distribution in Fig. 7(d) is due in part to using the medium JQC. If we were to use the loose JQC, there would be an excess of events with very low track multiplicity in the CMS Open Data. Most likely, these are prompt photons which barely pass the loose JQC, and to describe these properly, we would need to include photon-plus-jet MC samples. This excess is removed by the medium JQC, with only a modest impact on other substructure distributions.

We investigate three additional jet substructure distributions in Fig. 8: N_{95} [133], z_g [40], and D_2 [134] with $\beta = 1$. These observables probe, respectively, the uniformity of jet activity, the momentum sharing between subjets, and the two-prong substructure of jets. We implement N_{95} as the minimum number of pixels in a 33×33 jet image from $-R$ to R required to account for at least 95% of the total p_T. The soft drop jet grooming [135,136] parameters used to define the groomed momentum fraction z_g are $z_{\rm cut} = 0.1$ and $\beta = 0$. Jets with $z_g = 0$ indicate that the grooming procedure results in just a single remaining particle. Again, we find good agreement between the CMS Open Data and the simulation-level MC samples when using all PFCs, but the detector-level and simulation-level distributions agree somewhat better when restricted to track-based observables. Using our released samples [86–94], it is straightforward to plot a wide range of jet substructure observables

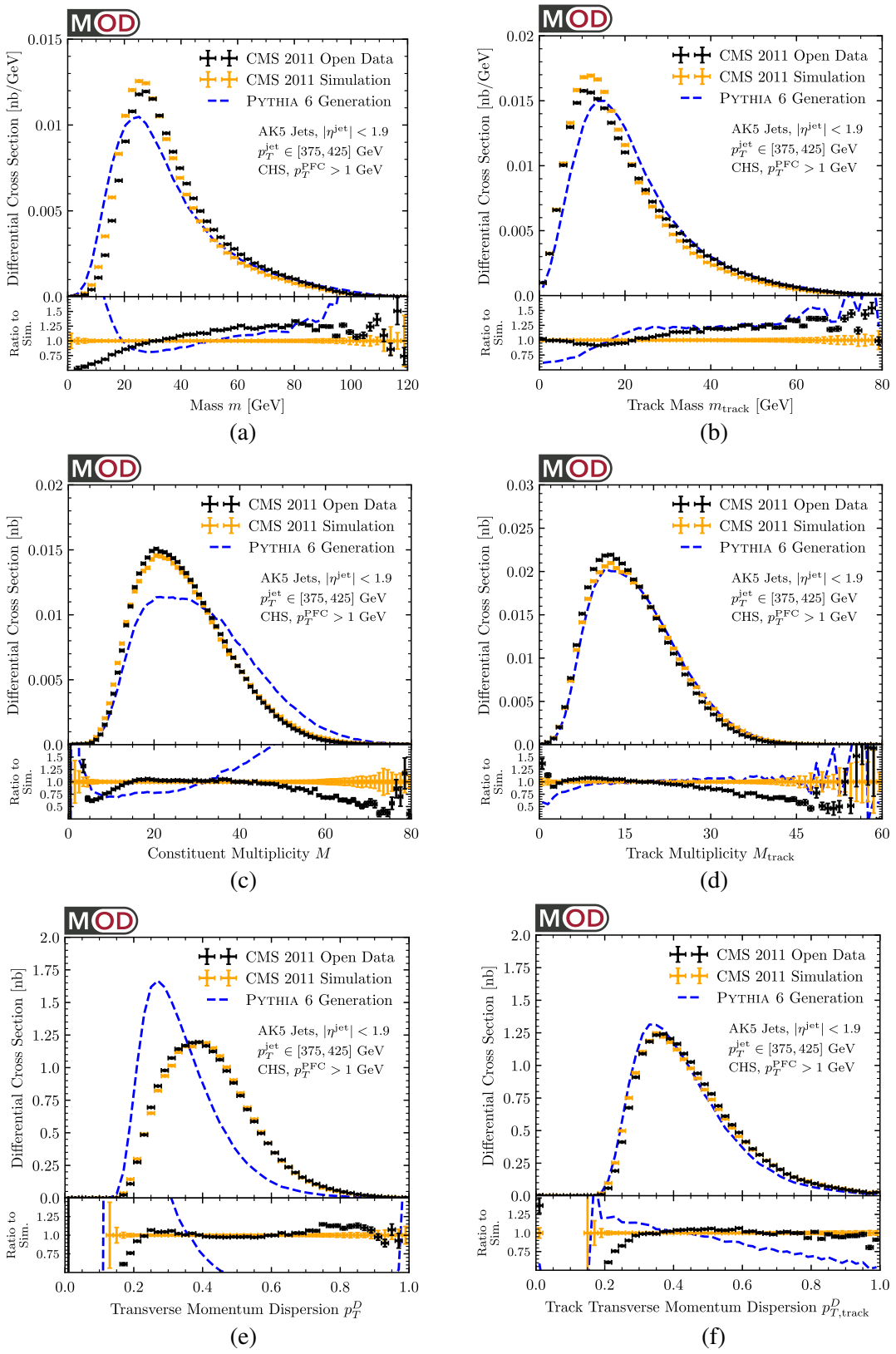

FIG. 7. Jet substructure observables using (left column) all PFCs and (right column) charged PFCs. In all cases, we apply CHS and enforce $p_T^{\text{PFC}} > 1$ GeV. The observables are (top row) jet mass, (middle row) constituent multiplicity, and (bottom row) transverse momentum dispersion (p_T^D).

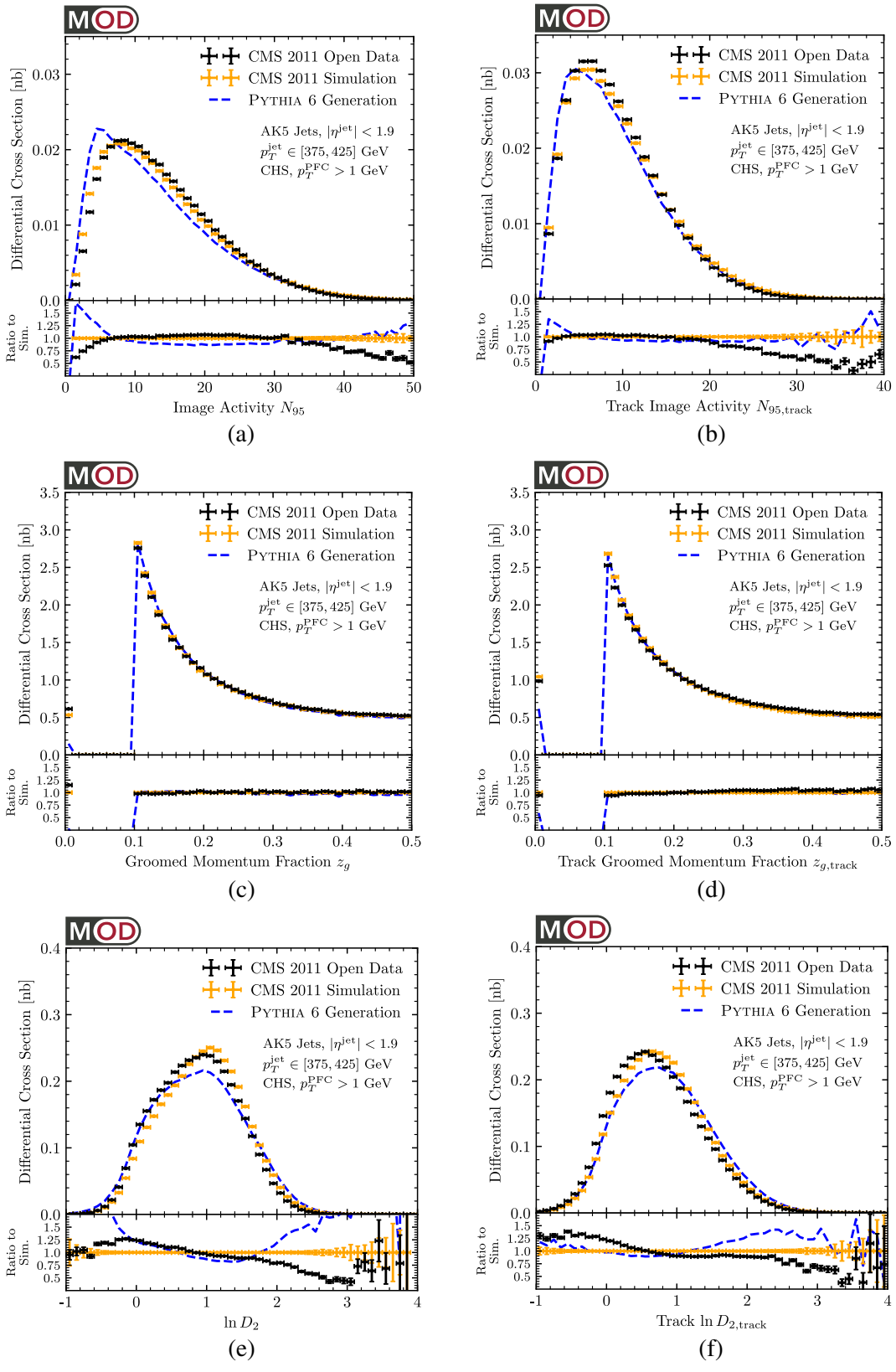

FIG. 8. Same as Fig. 7 for three more jet substructure observables: (top row) N_{95}, (middle row) z_g, and (bottom row) D_2.

[137], a number of which have already been implemented in the ENERGYFLOW package [84].

IV. EXPLORING THE SPACE OF JETS

We now turn from considering individual substructure observables at the histogram level to studying the radiation pattern in jets more broadly. In this section, we will use the energy mover's distance [56] as a metric to compare the energy flow of jets. We perform a range of exploratory EMD studies on the CMS Open Data to universally probe jet modifications, explore the space of jets, and visualize the most representative jets.

A. Review of the energy mover's distance

The jet energy flow can be characterized by an energy density on a two-dimensional surface, corresponding to an idealized detector at infinity [21–23]. For proton-proton collisions, we typically use transverse momentum p_T instead of energy and we indicate angular directions via rapidity y and azimuth ϕ. In these coordinates, the energy flow (more precisely, the transverse momentum flow) is

$$\rho(y,\phi) = \sum_{j\in\mathcal{J}} p_{Tj}\delta(y-y_j)\delta(\phi-\phi_j), \tag{3}$$

where j labels the constituents of the jet $\mathcal{J}$.

The expression in Eq. (3) is IRC safe by construction, since a particle with zero p_T does not contribute to the sum and a collinear splitting does not change the sum. The energy flow does not include any PID information, which is important to ensure IRC safety. To handle constituent masses, one could include velocity information [138], but that is beyond the scope of this paper.

Given two jets $\mathcal{I}$ and $\mathcal{J}$, the EMD is [56]

$$\mathrm{EMD}(\mathcal{I},\mathcal{J}) = \min_{\{f_{ij}\}} \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} f_{ij}\frac{R_{ij}}{R} + \left|\sum_{i\in\mathcal{I}} p_{Ti} - \sum_{j\in\mathcal{J}} p_{Tj}\right|, \tag{4}$$

where $R_{ij}^2 = (y_i-y_j)^2 + (\phi_i-\phi_j)^2$ is the rapidity-azimuth distance, R is the jet radius, and f_{ij} is the amount of transverse momentum "transported" from particle i in jet $\mathcal{I}$ to particle j in jet $\mathcal{J}$, subject to the constraints:

$$f_{ij} \geq 0, \qquad \sum_{j\in\mathcal{J}} f_{ij} \leq p_{Ti}, \qquad \sum_{i\in\mathcal{I}} f_{ij} \leq p_{Tj}, \tag{5}$$

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} f_{ij} = \min\left(\sum_{i\in\mathcal{I}} p_{Ti}, \sum_{j\in\mathcal{J}} p_{Tj}\right). \tag{6}$$

Finding the minimum over $\{f_{ij}\}$ in Eq. (4) is an optimal transport problem which can be solved efficiently using the network simplex algorithm [139–141].

The expression in Eq. (4) is non-negative, symmetric, and satisfies the triangle inequality:

$$\begin{aligned} &\mathrm{EMD}(\mathcal{I},\mathcal{J}) \geq 0, \\ &\mathrm{EMD}(\mathcal{I},\mathcal{J}) = \mathrm{EMD}(\mathcal{J},\mathcal{I}), \\ &\mathrm{EMD}(\mathcal{I},\mathcal{J}) \leq \mathrm{EMD}(\mathcal{I},\mathcal{K}) + \mathrm{EMD}(\mathcal{K},\mathcal{J}). \end{aligned} \tag{7}$$

Therefore, EMD is a proper metric on the space of energy flows, with units of energy (i.e., GeV). If the EMD between two jets is zero, then they are treated as identical. For this reason, it is often convenient to perform symmetry transformations on the jets prior to calculating the EMD. (This transformation procedure is closely related to the tangent earth mover's distance [142].) For all of the EMD studies in this paper, we longitudinally boost and azimuthally rotate each jet such that its four-vector is at the (y,ϕ) origin.

The second term in Eq. (4) is a cost term when two jets have different values of their scalar sum p_T. Because we are primarily interested in relative jet energy flows and not absolute jet energy scales, it is convenient to rescale the jets to make this cost term vanish. For jets with $p_T^{\mathrm{jet}} \in [375, 425]$ GeV, we rescale the jet constituents uniformly such that

$$\sum_{j\in\mathcal{J}} p_{Tj} \Rightarrow 400 \text{ GeV}. \tag{8}$$

Since we are working in relatively narrow p_T range and since QCD is a quasi-scale-invariant theory, this rescaling has only a mild impact on our results. Experimentally, this rescaling has the nice feature of reducing the dependence of our results on the JEC factors and on any PFC selection criteria. Theoretically, this rescaling has the nice feature of making the EMD identical (up to an overall energy scale) to the 1-Wasserstein metric between probability densities [143,144]. Changing the baseline from 400 GeV to some other scale would just proportionally rescale all the results below.

As motivated by Sec. III (and further motivated by Sec. IV B below), we often restrict our attention to charged particles with $p_T^{\mathrm{PFC}} > 1$ GeV. Strictly speaking, such a PFC restriction breaks the collinear safety (though not the soft safety) of the EMD, though there are calculational strategies to account for this using track functions [145–148]. Note that we always apply the rescaling in Eq. (8) *after* applying any PFC-level restrictions, such that our track-only results are similar in spirit to track-assisted observables [149,150]. Crucially, the PFC restriction and overall rescaling still preserve the metric properties of the EMD in Eq. (7).

An example EMD computation for two jets in the CMS Open Data is shown in Fig. 9. In the top row, we show two jets plotted in the style of Fig. 4. Here, the size of the dots indicates the transverse momenta of the PFCs, the colors indicate whether the PFCs are neutral or charged, and the crosses indicate charged PFCs that have been removed by CHS. In the bottom row, we drop the PID information and

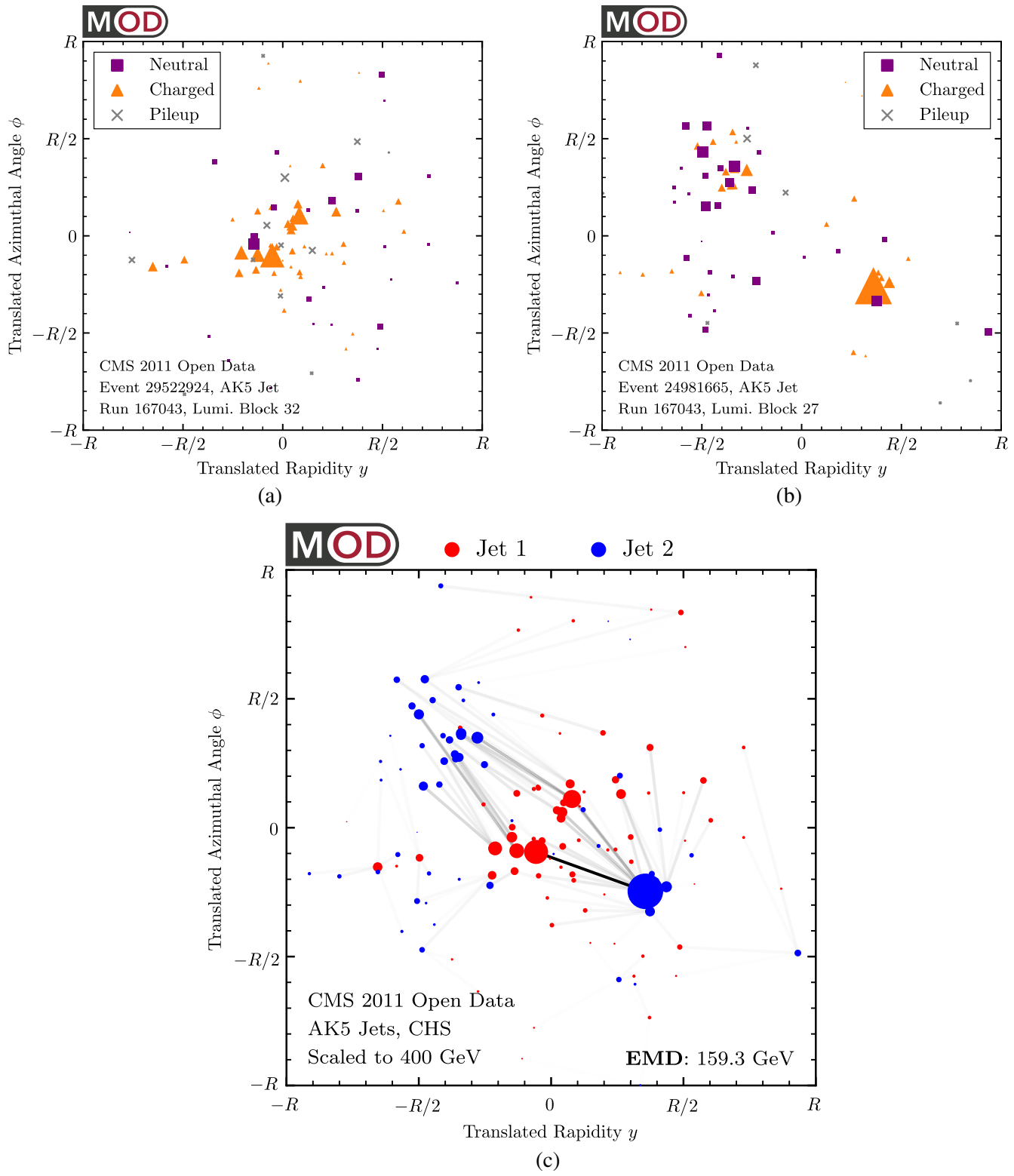

FIG. 9. Example EMD computation. (top row) Two jets from the CMS Open Data shown in the style of Fig. 4, with the size of each symbol indicating the particle transverse momentum and the style indicating the charge. Pileup particles removed by CHS are indicated by gray crosses. (bottom) Both jets represented as energy flow distributions via Eq. (3), along with the optimal transportation plan to rearrange one jet into the other, with the intensity of each line corresponding to $\{f_{ij}\}$ of Eq. (4).

switch to the energy flow representation in Eq. (3). We overlay the two jets, with the red dots corresponding to the first jet, the blue dots corresponding to the second jet, and the gray lines indicating the optimal transport $\{f_{ij}\}$. Because we have rescaled the jets by Eq. (8), all p_T from the first jet can be transported to the second jet.

B. Quantifying detector effects

As a first application of the EMD, we investigate a novel way to quantify the impact of detector effects and pileup. An example MC jet is shown in Fig. 10, where the EMD is computed between the same jet before and after detector simulation. See Sec. II D for how we associate simulation-level and generation-level jets. Pileup is removed with CHS and a variety of PFC cuts are applied to improve the agreement between the particle-level and detector-level jets. This is explicitly shown by the decreasing EMD as the cuts are applied, quantifying the fact that the radiation patterns within the jets are becoming more similar.

To see the impact of these cuts on the jet ensemble as a whole, in Fig. 11 we histogram the EMDs between the same MC jet evaluated at generation level and simulation level. Here, we impose $p_T^{\text{jet}} \in [375, 425]$ GeV on the simulation-level jet, while the generation-level jet could fall outside of this range. We emphasize that these EMD calculations are performed *after* the rescaling in Eq. (8), so this only quantifies the change in the radiation pattern, not the change in radiation intensity. As emphasized in Ref. [56], jets that are close in EMD are close in any (Lipschitz-bounded) IRC-safe measure, so small values of the generation-to-simulation EMD correspond to small differences between, for example, the generation- and simulation-level jet mass. In this way, the EMD provides a universal bound on the impact detector effects can have on IRC-safe observables, which is a convenient alternative to studying the impact on specific observables individually.

Considering all PFCs in Fig. 11(a), the generation-to-simulation EMD peaks at around 17 GeV. We can decrease the generation-to-simulation difference by sequentially applying CHS and the $p_T^{\text{PFC}} > 1$ GeV cut, though the impact is relatively modest. In evaluating the EMD, the $p_T^{\text{PFC}} > 1$ GeV restriction is applied at both the generation and simulation levels. Imposing the track-only restriction in Fig. 11(b), the generation-to-simulation EMD peak is shifted downward by a factor of about 2. Now, CHS has a much more pronounced impact, since it decreases substantially the relative pileup contamination. The $p_T^{\text{PFC}} > 1$ GeV cut has a modest, but non-negligible, impact. As expected, the impact of detector effects and pileup is minimized for track-based observables after CHS. In Fig. 20 in Appendix B, we further investigate the performance of CHS for pileup mitigation. In Fig. 24 in Appendix C, we investigate the impact of the p_T^{PFC} cut in more detail.

From these studies, we conclude that our default selection (charged PFCs with $p_T^{\text{PFC}} > 1$ GeV) is a reasonable compromise between reconstruction performance and substructure sensitivity. More generally, we see that the EMD is an effective and intuitive way to quantify the impact of detector effects and pileup contamination.

C. Visualizing the space

It is interesting to directly visualize the metric space of jets defined by EMD. There are a variety of techniques to visualize high-dimensional data in low dimensions, which provide a fascinating way to see the broad features of a dataset. Here, we apply t-distributed stochastic neighbor embedding (t-SNE) [151–154], which finds a low-dimensional embedding of the data in a way that respects the distance between data points. We run t-SNE with a two-dimensional embedding space, in which the procedure defines two axes and attempts to place data points in this two-dimensional plane in such a way that jets close in EMD are nearby and jets far in EMD are distant.

Though there are techniques to implement t-SNE on N data points in $\mathcal{O}(N \log N)$ runtime [154], due to current limitations in the `scikit-learn` [155] implementation that we use, we have to perform $\mathcal{O}(N^2)$ operations. To make this computationally tractable, we restrict our attention to the $p_T^{\text{jet}} \in [399, 401]$ GeV range, which yields approximately 40,000 jets in the CMS Open Data. We also subsample and unweight the MC events to obtain around 40,000 generation-level and 40,000 simulation-level jets as well. (Because there are insufficient events in the $\hat{p}_T \in [170, 300]$ GeV MC sample [74], we have to downweight them by a factor of around 10 to achieve an approximately unweighted sample.) We apply CHS, the $p_T^{\text{PFC}} > 1$ GeV cut, and the track-only restriction on all jets. To reduce the effective dimensionality of the dataset and remove a trivial isometry, we rotate the jets around the jet axis such that the principle component of the transverse radiation pattern is aligned vertically in the rapidity-azimuth plane, breaking the two-fold degeneracy by enforcing that the jet has more scalar sum p_T at positive azimuth. We also keep only the particles within a jet radius of the jet axis.

The results of t-SNE embedding into a two-dimensional space are shown in Fig. 12, for the CMS Open Data and for the simulation-level and generation-level MC samples. For visual clarity, we rotate the t-SNE manifold such that the three embeddings exhibit roughly the same large-scale structure. The gray contours represent the density of the embedded jets. Example jets are sprinkled throughout the space and color coded by their jet mass fractile (i.e., fraction of events with smaller jet mass than the color coded value).

For the CMS Open Data in Fig. 12(a), the t-SNE embedding exhibits a dominant cluster of jets with typically low jet mass, with a long slope extending out to typically higher jet masses. The most exotic jets are furthest away from the dominant cluster. The t-SNE embeddings of the

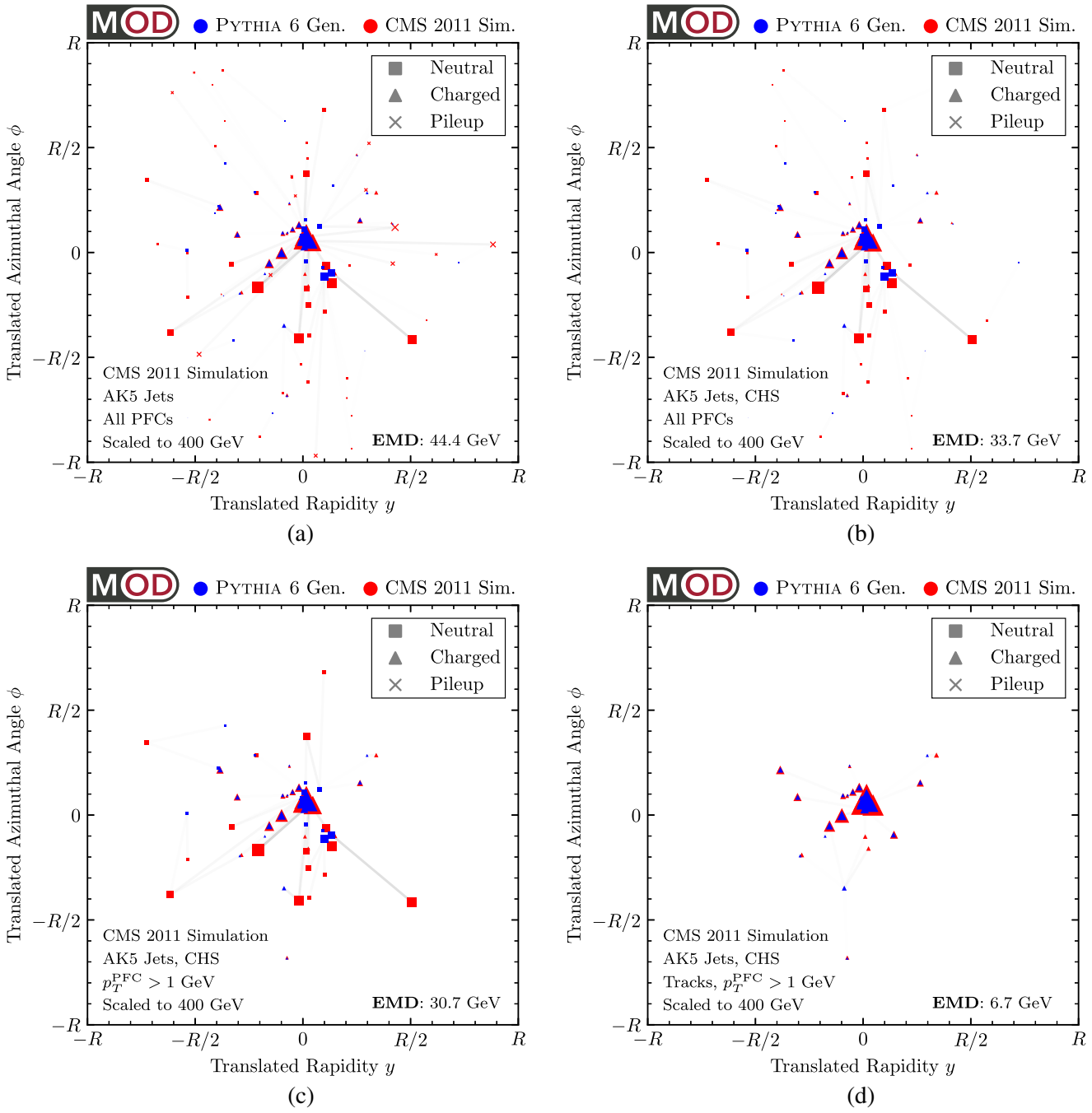

FIG. 10. A jet from the PYTHIA hard QCD MC sample shown (blue) before and (red) after the GEANT-based CMS detector simulation, with the size of each symbol indicating the particle transverse momentum and the shapes indicating the charge. To improve visibility and clarity, the sizes of the symbols in the generator-level jet have been uniformly decreased. Pileup particles removed by CHS are indicated by crosses, and the optimal transportation plans between the jets are shown as gray lines. The jets are shown (a) with all PFCs, (b) after applying CHS to remove charged pileup, (c) after an additional $p_T^{\mathrm{PFC}} > 1$ GeV cut, and (d) after further restricting only to tracks. The EMD between the jet before and after the detector simulation decreases as these cuts are applied, highlighting that these PFC cuts minimize the impact of detector effects.

MC samples in Figs. 12(b) and 12(c) are qualitatively similar, though the specific density distributions differ. Using smaller jets samples, we find that the variability between the data and MC t-SNE embeddings is comparable to the variability when running t-SNE multiple times on the same sample. No obvious anomalies in the CMS Open Data appear visually, though we return to anomalous jet configurations in Sec. IV F.

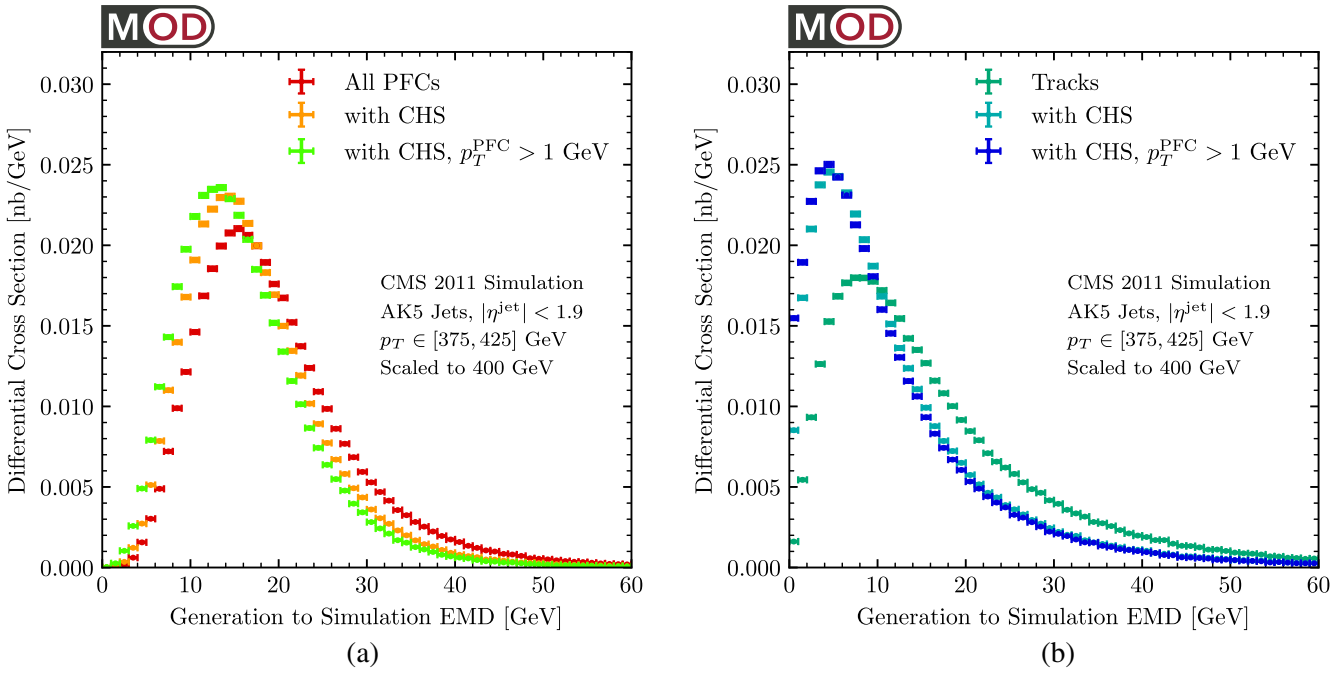

FIG. 11. Quantifying detector effects through the distribution of generation-to-simulation EMDs. Starting from the same jet generated by PYTHIA, we compute the EMD between the jet before and after the GEANT-based CMS detector simulation. These are shown for (a) all PFCs and (b) tracks only, with the subsequent application of CHS and the $p_T^{\mathrm{PFC}} > 1$ GeV restriction. The agreement between the generation-level and simulation-level radiation patterns (as quantified by EMD) can indeed be seen to improve as the selections tighten. See Fig. 20 in Appendix B for a study of the impact of CHS for different levels of pileup contamination. See Fig. 24 in Appendix C for a study of the impact of the p_T^{PFC} cut.

D. Correlation dimension

To gain more quantitative insight into the space of jets, we can use the EMD to compute its dimensionality. While a variety of definitions exist for intrinsic dimension, we use the correlation dimension [156,157], which is a type of fractal dimension and was the measure used in Ref. [56]. From a matrix of pairwise EMDs between jets, the correlation dimension is defined as:

$$\dim(Q) = Q\frac{\partial}{\partial Q}\ln \sum_{1\le k<\ell\le N} \Theta[\mathrm{EMD}(\mathcal{J}_k, \mathcal{J}_\ell) < Q]. \qquad (9)$$

Here, N is the total number of jets in the sample and the Heaviside theta function indicates whether the jet k is within an EMD Q of jet ℓ. To gain an intuition for this formula, note that for a uniform data sample in d dimensions, the expected number of neighbors B within a ball of radius Q scales like Q^d, such that $d \simeq \partial \ln B/\partial \ln Q$. The expression in Eq. (9) has this same relation, where the number of neighbors B is averaged over balls of radius Q centered around each data point.

The computational cost of implementing Eq. (9) is $\mathcal{O}(N^2)$, so we restrict our attention to the same $p_T^{\mathrm{jet}} \in [399, 401]$ GeV subsample as in Sec. IV C. (Because it is straightforward to compute $\dim(Q)$ using MC weights, this time we do not need to downweight the $\hat{p}_T \in [170, 300]$ GeV MC sample [74].) We also perform the same jet rotation in Sec. IV C.

After the rescaling in Eq. (8), the maximum possible value of the EMD is 400 GeV, so $\dim(Q)$ always equals zero for $Q > 400$ GeV. Because we cluster jets with the anti-k_T algorithm, though, the jet configurations that could in principle lead to this maximum EMD value are not present in our samples. For example, consider two jets of equal scalar sum p_T: one consists of a single particle; the second consists of two particles, each with transverse momentum $p_T/2$, separated by ΔR. The EMD between these configurations is $\frac{1}{2} p_T \Delta R$. Within a jet region of size R, ΔR could in principle be as large as $2R$ (i.e., EMD as large as p_T), but anti-k_T would split the second jet in two unless $\Delta R < R$ (i.e., EMD of $p_T/2$). In practice, we find that $\dim(Q)$ indeed goes to zero around $Q \simeq 200$ GeV.

In Fig. 13(a), we compare the correlation dimension between the CMS Open Data and the MC samples, again with CHS and tracks only with $p_T^{\mathrm{PFC}} > 1$ GeV. The agreement between the open data and the MC sample at simulation-level is very good, though the correlation dimension is roughly 0.5 above the generation-level curve for much of the plotted Q range. Naively, one might think that detector effects would decrease the correlation dimension, since finite granularity effects decrease the relative complexity of jet configurations. Instead, the added half dimension suggests that the detector has more of a smearing

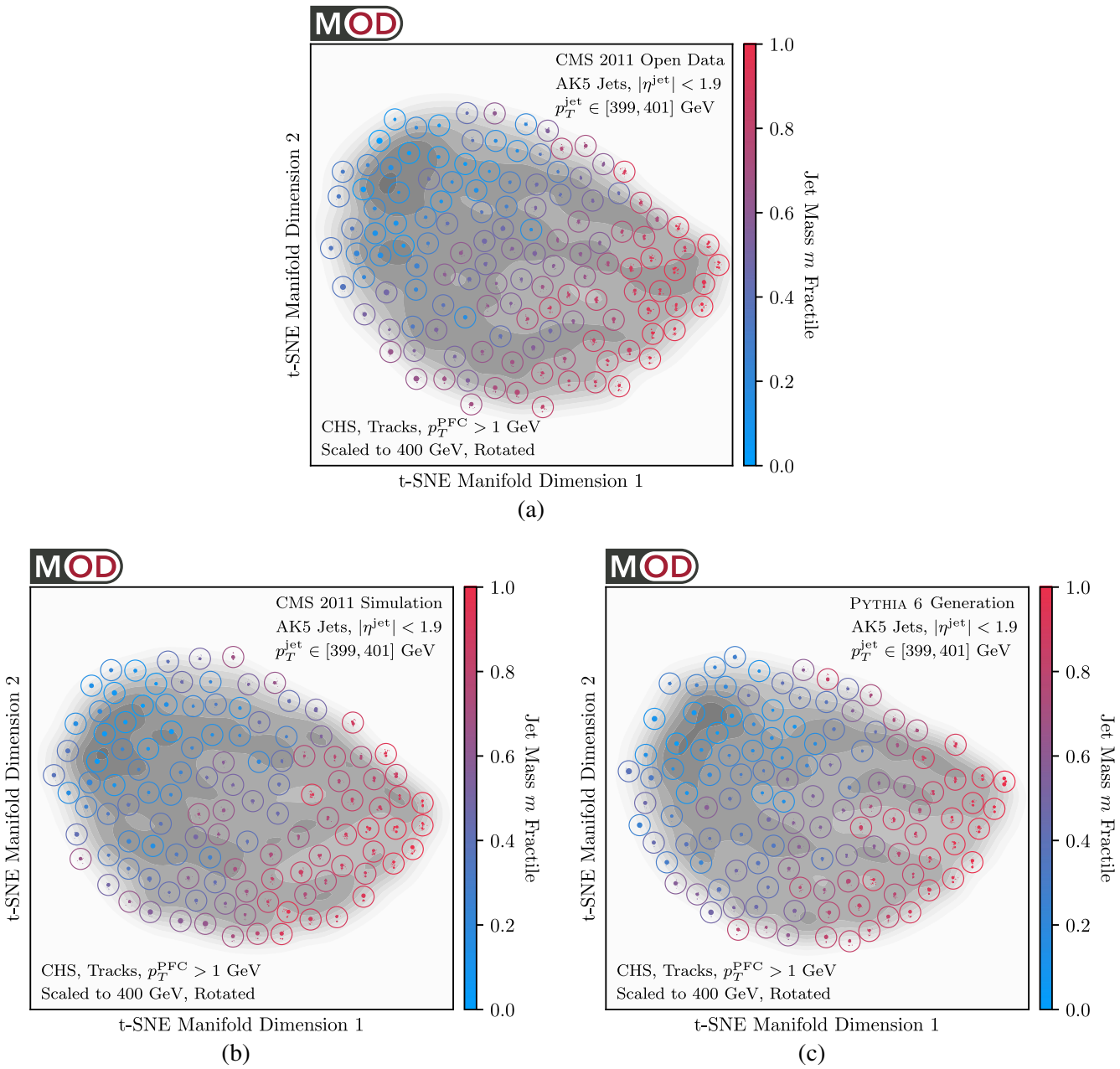

FIG. 12. Two-dimensional t-SNE embedding of jets in the $p_T^{\text{jet}} \in [399, 401]$ GeV range from the (a) CMS Open Data, (b) simulation-level MC, and (c) generation-level MC. The gray contours indicate the density of embedded jets, and the example jets are color coded by the jet mass fractile in the corresponding dataset.

effect, analogous to the way that smearing a zero-dimensional point generates a higher-dimensional manifold.

The fact that the correlation dimension in Fig. 13 increases logarithmically with decreasing Q is expected from first principles QCD. The number of jet constituents scales up logarithmically with decreasing energy scale (see e.g., [158,159]), as does the entropy of a jet [160], and both of these quantities are related to the effective dimensionality of the space of QCD jets. We leave a QCD calculation of $\dim(Q)$ to future work, noting that the result will depend on the strong coupling constant α_s as well as on the relative fraction of quark and gluon jets in the sample.

The correlation dimension gives us an interesting handle to understand the impact of applying cuts on the PFCs, complementary to the studies in Sec. IV B. In the bottom row of Fig. 13, we show $\dim(Q)$ for all PFCs and just tracks, as well as the effect of the $p_T^{\text{PFC}} > 1$ GeV cut, always with CHS applied. For the CMS Open Data in Fig. 13(b) and for the simulation-level MC in Fig. 13(c), there is relatively little impact on the correlation dimension

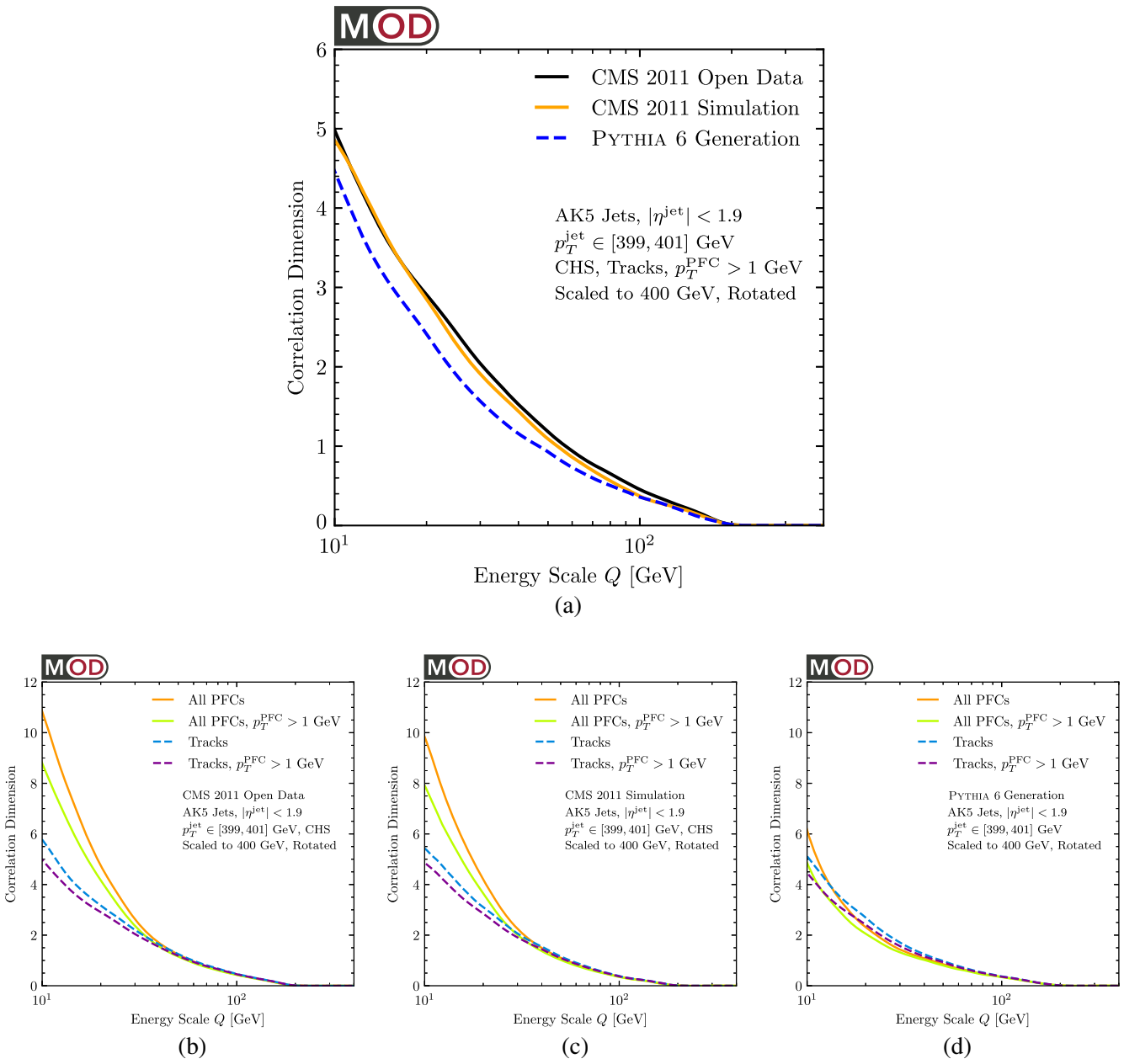

FIG. 13. The correlation dimension of the space of jets as a function of energy scale Q, (a) comparing the CMS Open Data to the generation-level and simulation-level MC samples. There is good agreement between the MC simulation-level and the open data, while the MC generation-level jets have a systematically smaller correlation dimension over much of the energy range. Also shown are different PFC selections in the (b) CMS Open Data, (c) simulation-level MC, and (d) generation-level MC which either impose the $p_T^{\text{PFC}} > 1$ GeV cut or restrict to only tracks or both. In all cases, the high-energy limit of the correlation dimension is robust to the PFC selection, with significant differences only appearing for $Q \lesssim 40$ GeV.

for $Q \gtrsim 40$ GeV. Below this scale, though, the correlation dimension is significantly smaller when restricting to just tracks and/or when imposing $p_T^{\text{PFC}} > 1$ GeV. Interestingly, for the generation-level curves in Fig. 13(d), there is a much more modest impact from these restrictions. In fact, restricting to charged PFCs can sometimes *increase* the correlation dimension, since after applying the rescaling in Eq. (8), the charged PFC restriction acts like a kind of smearing. From this we conclude that $\dim(Q)$ is a robust measure of dimensionality at high Q, and very sensitive to QCD fragmentation and detector effects at small Q.

E. The most representative jets

Computing the EMD also allows us to visualize the space of jets in such a way that observable values can be correlated with jet topologies. Specifically, given a set of

jets, we can find the k jets $\{\mathcal{K}_1, \ldots, \mathcal{K}_k\}$ (called medoids) that minimize the sum of the distances of each jet to its closest medoid:

$$\mathcal{V}_k = \frac{1}{N}\sum_{i=1}^{N} \min\{\mathrm{EMD}(\mathcal{J}_i, \mathcal{K}_1), \ldots, \mathrm{EMD}(\mathcal{J}_i, \mathcal{K}_k)\}. \quad (10)$$

The value of Eq. (10) provides a quantitative notion of how well approximated the dataset is by the k jets. Inspired by the N-subjettiness observables of Ref. [161,162], this quantity can be thought of as the "k-eventiness" of the dataset.

While naively optimizing the choice of the medoids takes $\mathcal{O}(N^{K+1})$ runtime, we use a fast iterative approximation techniques from the `pyclustering` PYTHON package [163]. This k-medoids procedure provides a significantly more representative selection of jets than a random subsample, as quantified by the $\mathcal{V}_k$ distribution in Fig. 14 for the case of $k = 25$. Along these lines, one might also consider clustering the full dataset of jets, for instance using iterative reclustering similar to techniques used to cluster particles into jets [112,164–167], though we leave further explorations in this direction to future work.

In Fig. 15, we show the 25 most representative jets in the $p_T^{\mathrm{jet}} \in [399, 401]$ GeV subsample from Sec. IV C, arranged according to t-SNE and sized according the number of closest neighbors. Because these medoids are representative (and not just randomly selected) in that they try to minimize $\mathcal{V}_k$, there is a rigorous sense in which understanding the structure of these 25 jets captures the structure of the CMS Open Data jet ensemble as a whole.

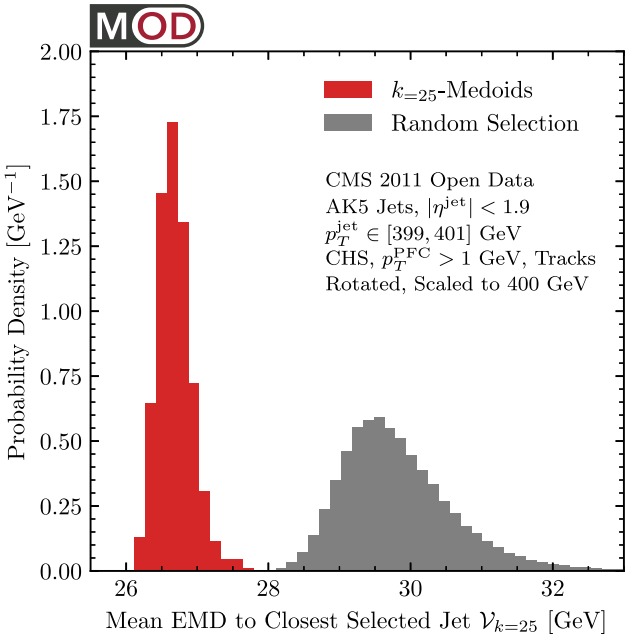

FIG. 14. The distance of our jet dataset to a selection of 25 representative jets, shown for (red) jets selected with the k-medoids algorithm as well as (gray) randomly selected jets. The k-medoids are systematically closer to the dataset, demonstrating that jets chosen in this way are significantly more representative than a random selection of jets.

If we apply the k-medoid procedure to jets occupying the same histogram bins of a specific observable, we can then visualize how the jet topology changes as observable values change. In Fig. 16, we show histograms for the six substructure observables from Sec. III C, using the CMS Open Data with CHS and only tracks with $p_T^{\mathrm{PFC}} > 1$ GeV. In each histogram bin, we show the four most representative jets, as determined by the 4-medoids procedure. For jet mass in Fig. 16(a), we see a steady evolution from one-prong topologies to two-prong topologies. The reverse behavior is shown for D_2 in Fig. 16(b), with two-prong topologies evolving into one-prong ones. One low-D_2 medoid jet consists of two highly overlapping prongs, distinct from the one-prong high-D_2 configurations, highlighting the Sudakov safety of D_2 [40,134]. For the IRC-unsafe observables of track multiplicity in Fig. 16(c) and p_T^D in Fig. 16(e), we see evolutions between simple topologies and jets with more complex substructure. For N_{95} in Fig. 16(d), there is a progression from narrow jets to diffuse jets. Finally, for z_g in Fig. 16(f), there is an evolution from unbalanced subjets to balanced subjets, with its Sudakov safety apparent from the one-prong configurations throughout. While all of these behaviors can be understood from the definition of these observables, the k-medoids procedure offer an intuitive visualization of the jet configurations that contribute to each observable value.

F. Toward anomaly detection

As the last application of the EMD in this paper, we present a first step toward using it for anomaly detection. Instead of finding the most representative jets as in Sec. IV E, we can find the least representative jets. As one way to quantify this, we can find the nth moment of the EMD distribution of one jet to the rest of the dataset,

$$\bar{Q}_n(\mathcal{I}) = \sqrt[n]{\frac{1}{N}\sum_{k=1}^{N}(\mathrm{EMD}(\mathcal{I}, \mathcal{J}_k))^n}, \quad (11)$$

where we applied the nth root such that $\bar{Q}_n$ has units of GeV. Small values of $\bar{Q}_n$ indicate a common jet configuration. Large values of $\bar{Q}_n$ indicate a jet which is far from the rest of the dataset, and therefore anomalous.

In Fig. 17, we show the distribution of $\bar{Q}_n$ for $n = 1$ (i.e., mean EMD) along with the four medoids in each histogram bin. As expected from the t-SNE visualization in Fig. 12, the most typical jet configurations have a single hard prong, while the least typical configurations have multiprong or diffuse topologies. In Fig. 25 of Appendix C, we show a simular plot for $n = \frac{1}{2}$ and $n = 2$. The most anomalous jets isolated by $\bar{Q}_n$ for $n = \frac{1}{2}$, 1, and 2 agree for the six most

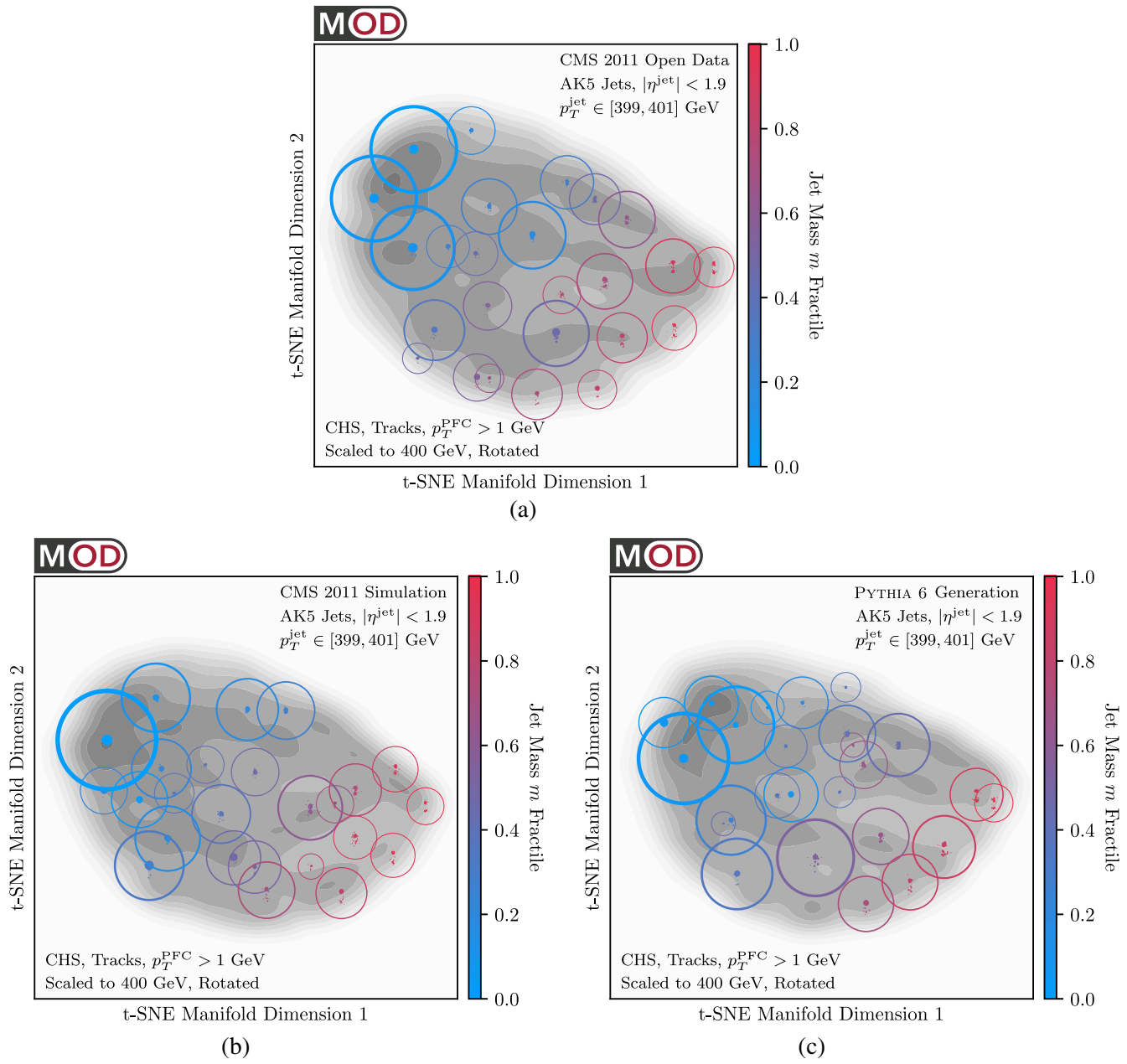

FIG. 15. The 25 most representative jets (medoids) in the (a) CMS Open Data, (b) simulation-level MC, and (c) generation-level for $p_T^{\text{jet}} \in [399, 401]$ GeV. The jets are arranged according the t-SNE algorithm as in Fig. 12 and their area is proportional to the number of jets nearest to them. The medoid jets try to "tile" the space in a rigorous sense.

anomalous jets, with the top three such jets shown in the bottom row of Fig. 17. The most anomalous jets are all highly complex three-prong topologies, hinting at a close relationship between this measure of anomalousness and observables such as N-subjettiness [161,162].

The anomalousness of a jet, quantified by $\bar{Q}_n$, is non-trivially correlated with the jet mass, which is easily confirmed by observing the medoids in each bin in Fig. 17. While this is expected and understandable from QCD, this correlation can complicate searches for resonant new physics by sculpting the background. To circumvent this correlation in the case of these searches, the EMD-based approach can be combined with mass decorrelation techniques [168–170] or with ideas such as CWoLa hunting [171] to look for anomalies within mass bins compared to sidebands.

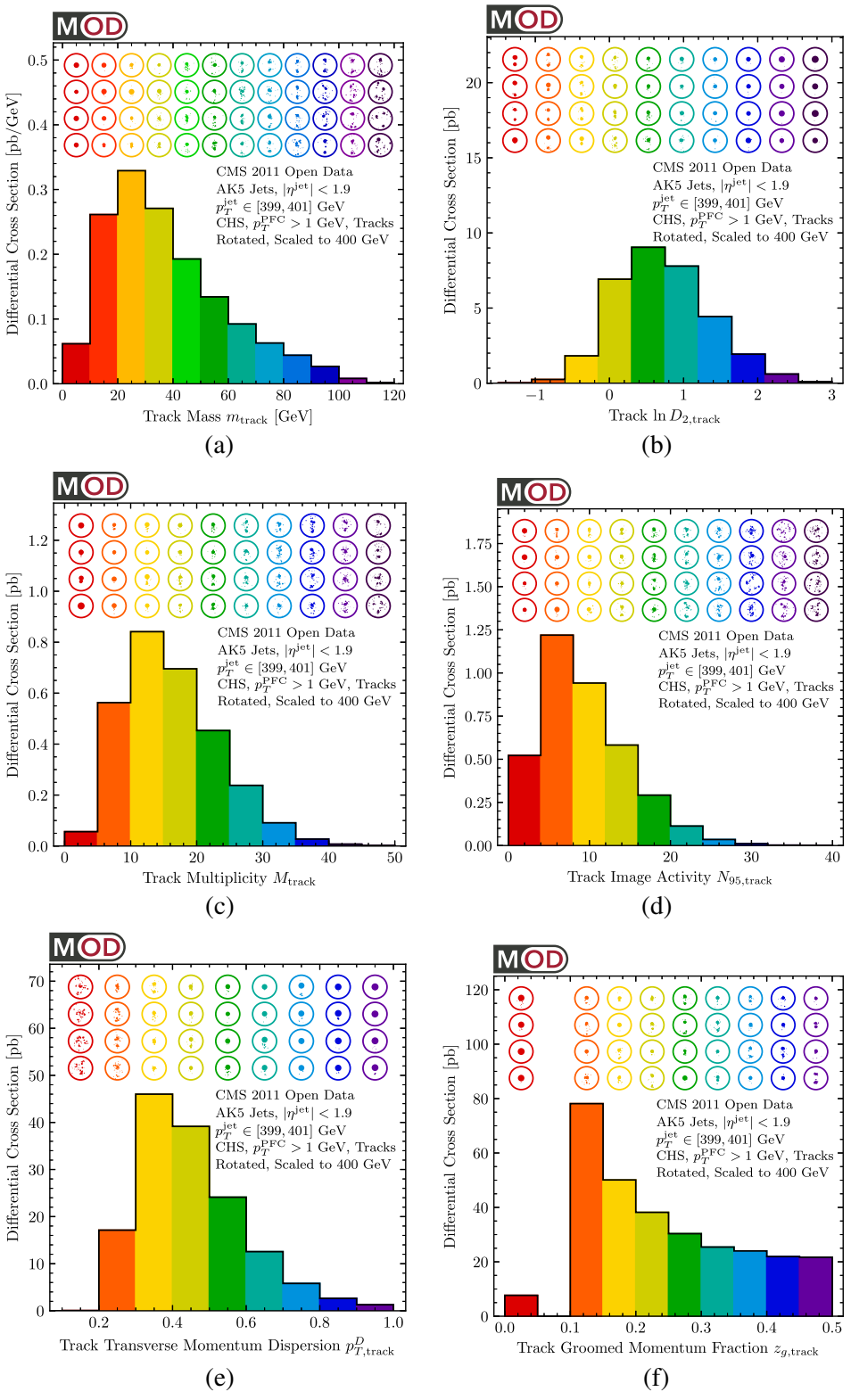

FIG. 16. The same jet substructure observables from Sec. III C, but now showing the four most representative jets (medoids) in each histogram bin. These distributions are obtained from the CMS Open Data after applying CHS, the $p_T^{\text{PFC}} > 1$ GeV cut, the track-only restriction, as well as the rotation and rescaling in Eq. (8). As in Fig. 7, we show (a) jet mass, (c) track multiplicity, and (e) p_T^D. As in Fig. 8, we show (b) D_2, (d) N_{95}, and (f) z_g. Track multiplicity and p_T^D are IRC-unsafe observables, and hence are not fully described by the energy flow in the jet.

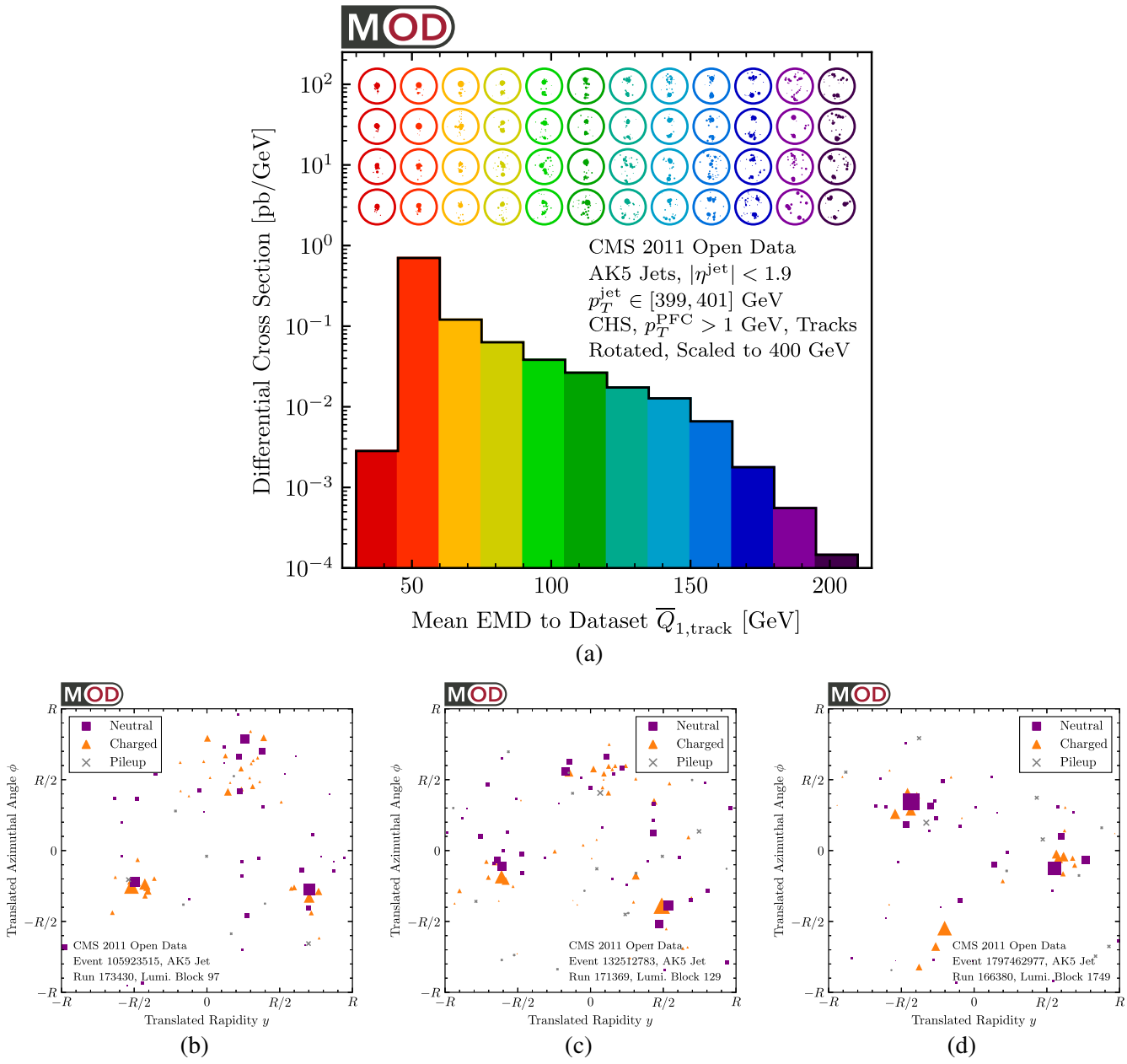

FIG. 17. (a) Distribution on the CMS Open Data of $\bar{Q}_1$ from Eq. (11) along with the 4-medoids in each histogram bin. The most typical (atypical) jets in the dataset have small (large) values of $\bar{Q}_1$. Event displays are shown for the (b) most, (c) second most, and (d) third most anomalous jets in our CMS Open Data sample.

V. CONCLUSIONS

The CMS Open Data is an exciting resource for performing exploratory studies in collider physics. In this paper, we performed the first ever exploration of the metric space of QCD jets on real collider data, using the EMD [56] as our measure of jet similarity. The EMD provides complementary information to traditional histogram-based analyses, and it also provides new strategies for data visualization in particle physics. In terms of quantitative measures, we showed how to use the EMD to characterize the impact of detector effects and to calculate the intrinsic dimension of a jet ensemble. For qualitative studies, we showed how to use the EMD to identify the most representative jets in a histogram bin and the least representative jets in the ensemble as a whole, where the latter analysis is particularly interesting in the context of anomaly detection for new physics searches [171–177].

Beyond the specific EMD studies here, a key outcome of this research is a processed and validated jet sample for use in future jet studies consisting of jets in the CMS 2011 Open Data with a p_T above 375 GeV. This processed single-jet dataset is available on the ZENODO platform [86–94] along with the analysis tools needed to make the bulk of plots in this

paper [84,85]. This sample is ready to use out-of-the-box by future users, since JEC factors and JQC are available and easy to apply, and baseline event selection criteria have been chosen to ensure that the `Jet300` trigger is fully efficient. Because we apply the same processing pipeline to corresponding simulated MC events, one can assess the impact of detector effects on new jet analysis strategies. While we have not performed detector unfolding or estimation of systematic uncertainties in this exploratory study, our dataset contains sufficient information to implement these important elements, which we leave to future work. As an important stress test of this archival strategy, we plan to perform our next jet physics analysis directly on the released datasets without ever accessing the underlying CMS AOD files.

There are a number of future directions to pursue using the EMD. We focused on a narrow p_T range of [375,425] GeV in this paper in order to have a more uniform jet sample, but it would be interesting to perform EMD studies on higher p_T jets. This is particularly relevant in the context of the intrinsic dimension; in a preliminary QCD calculation of the correlation dimension as a function of jet p_T, we find nontrivial dependence both on Q and on the quark/gluon composition of the sample. One application suggested in Ref. [56] is using EMD for jet classification, and it would be interesting to do a data/simulation classification study in the spirit of Refs. [178,179] to identify regions of phase space that are not well modeled by the current generation/simulation tools. In this study, we focused on applying the EMD to individual jets, but it could also be applied to events as a whole, which would be a novel strategy to explore the `MinimumBias` primary dataset. It would also be interesting to explore alternative EMD definitions that incorporate PID information.

Finally, we applaud the commitment shown by the CMS experiment to releasing research-grade public data. The inclusion of simulated datasets in the 2011 release was essential for us to gain confidence in the robustness of track-based observables for jet substructure studies. Even without the actual data files, the simulated datasets are a valuable resource for phenomenological studies, since they cover a wide range of final states with fully realistic detector information. As CMS continues to release research-grade data, we hope that more researchers take advantage of this unique resource for particle physics.

ACKNOWLEDGMENTS

We thank CERN, the CMS collaboration, and the CMS Data Preservation and Open Access (DPOA) team for making research-grade collider data available to the public. We specifically thank Edgar Carrera, Kati Lassila-Perini, and Tibor Simko for help processing the CMS Open Data, and Salvatore Rappoccio for help implementing CHS. We thank Maximilian Henderson, Edward Hirst, and Ziqi Zhou for collaboration in the early stages of this work. We benefitted from additional feedback from Cari Cesarotti, Kyle Cranmer, Achim Geiser, Matthew LeBlanc, David Miller, Benjamin Nachman, Jennifer Roloff, and Yotam Soreq. This work was supported by the Office of Nuclear Physics of the U.S. Department of Energy (DOE) under Grant No. DE-SC0011090 and the DOE Office of High Energy Physics under Grants No. DE-SC0012567 and No. DE-SC0019128. R. M. is additionally supported by a fellowship from the Heising-Simons Foundation. J. T. is additionally supported by the Simons Foundation through a Simons Fellowship in Theoretical Physics. We benefited from the hospitality of the Harvard Center for the Fundamental Laws of Nature and the Fermilab Distinguished Scholars program. Cloud computing resources were provided through a Google Cloud allotment from the MIT Quest for Intelligence.

APPENDIX A: MISSING AND ZEROED LUMINOSITY BLOCKS

As mentioned in Sec. II C, there are 89 valid LBs tabulated in Ref. [110] that do not appear anywhere in the `Jet` primary dataset [66]. There is of course the possibility that we made a mistake in processing the data, though we verified that `MODProducer` recovers the total number of events (both valid and not) quoted in Ref. [66]. Also, the missing LBs do not appear to represent a missing AOD file, which was an issue that had to be resolved for Ref. [38]. In particular, the missing LBs do not appear to be linked in time, whereas a given AOD file typically has consecutive sequences of LBs. Moreover, there are strange characteristics of the missing LBs that suggest that there might be more systematic issues at play.

We can classify the missing LBs into two main categories:

(1) *Near zero luminosity.* For 17 missing LBs, the recorded luminosity was less than 0.03 $\mu\,\mathrm{b}^{-1}$. It is plausible that none of the jet triggers fired during these LBs, in which case they should count (negligibly) toward the integrated luminosity of the run.

(2) *Large delivered/recorded discrepancy.* For 71 missing LBs, the recorded luminosity was at least an order of magnitude smaller than the delivered luminosity. It is plausible that these LBs should not have been classified as valid, in which case it is consistent to ignore them.

Curiously, there was one missing LB where the discrepancy between the delivered and recorded luminosities was only 2.3%. This is consistent with the typical delivered/recorded mismatch for the valid LBs in the `Jet` primary dataset, which is around 3%.

Another issue raised in Sec. II C is that there are 201 valid LBs present in Ref. [110] which have zero recorded luminosity. The 164 such LBs in Run A can be categorized as follows:

(1) *Exactly zero delivered luminosity.* For 3 zeroed LBs, the delivered luminosity was also zero. Of these, 1 LB contained 0 events; the other 2 contained a total of 3 events that were triggered in the `Jet` primary dataset.

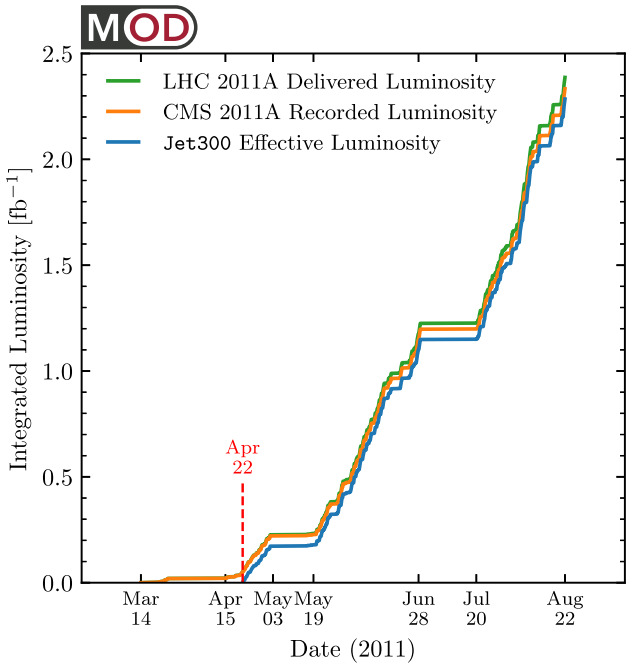

FIG. 18. The delivered and integrated luminosity for the Run 2011A dataset over time. Also shown is the effective luminosity of the Jet300 trigger, which was activated on April 22, 2011.

(2) *Near zero delivered luminosity.* For 20 zeroed LBs, the delivered luminosity was less than 0.05 μ b^{-1}, so it is expected that the recorded luminosity could be zero. Of these, 11 LBs contained 0 events; the other 9 contained a total of 23 events that were triggered on in the Jet primary dataset, so we can safely ignore these as well.

(3) *Sizable delivered luminosity.* For 141 zeroed LBs, the delivered luminosity was greater than 2.7 nb^{-1}, so one expects at least one of the Jet triggers to have fired. Of these, 9 LBs contained 0 events; the other 132 contained a total of 20,850 events, even though the recorded luminosity was zero. Most likely, these were misclassified as valid LBs.

Tallying these together, there are 21 zeroed LBs that have zero events, which are already counted as missing LBs above. The remaining 143 zerored LBs have a total of 20,876 events, which is the number listed in Table I. Following the recommendation of CMS, we omit all of the zeroed LBs from our analysis.

While these missing and zeroed LBs do not affect the conclusions of our physics studies, they do highlight the importance of stress-testing archival data strategies to make sure that there is validated information available to future generations of collider enthusiasts [180].

For completeness, in Fig. 18, we plot the total delivered and recorded luminosities for Run 2011A as a function of date, along with the effective luminosity for the Jet300 trigger. Note that the loss of luminosity due to the late turn-on of the Jet300 trigger has a negligible effect on our analyses.

APPENDIX B: ASPECTS OF PILEUP

The CMS simulated MC samples include the effect of pileup, but the number of overlapping events differs from

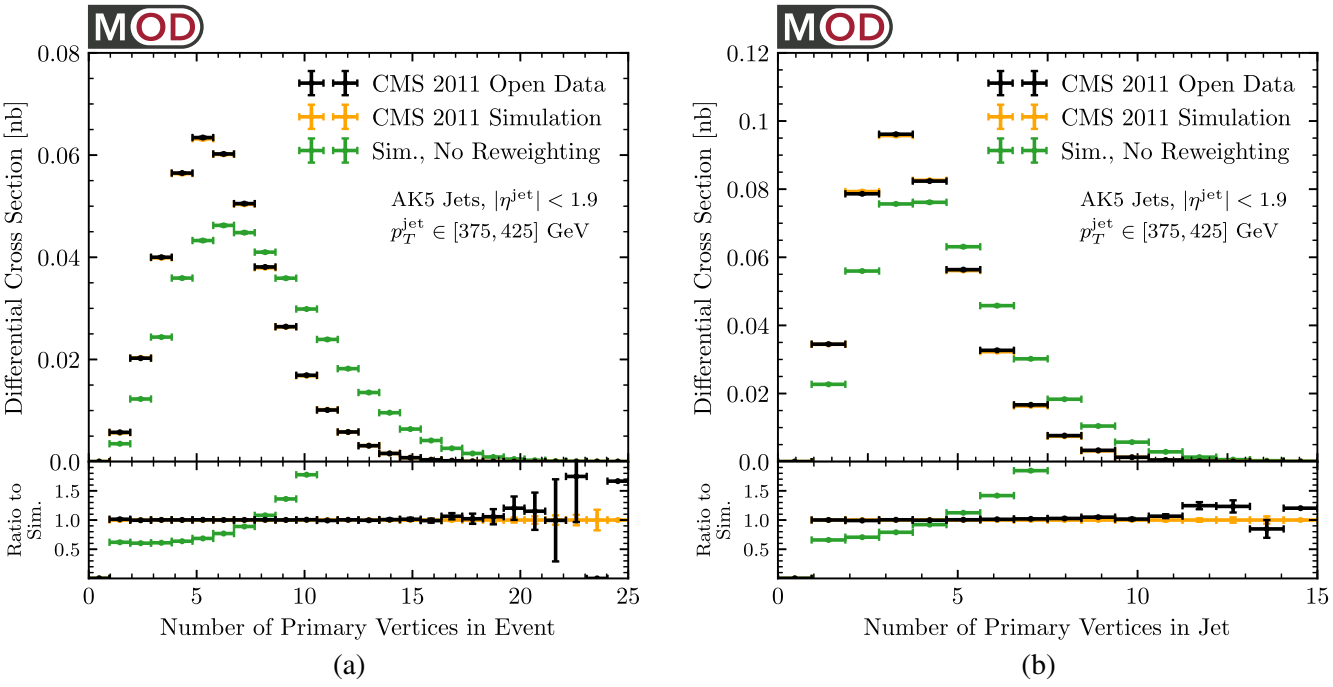

FIG. 19. Level of pileup contamination in the actual and simulated CMS datasets, with and without the pileup reweighting. Shown are the number of primary vertices (a) in the event as a whole and (b) associated with the reconstructed jets of interest. Larger values of N_{PV} correspond to more pileup.

what is observed in the CMS 2011 Open Data. To correct for this, we reweight the MC events to match the observed number of primary vertices (N_{PV}). Note that a larger number of primary vertices is associated with a larger amount of pileup contamination.

The effect of this reweighting is shown in Fig. 19(a), where we plot the number of primary vertices associated with each event in the CMS Open Data compared to the MC simulation, both before and after reweighting. The reweighting factor is derived from all "medium" quality jets with $p_T^{\mathrm{jet}} > 375$ GeV and $|\eta^{\mathrm{jet}}| < 1.9$, though the plot only shows the $p_T^{\mathrm{jet}} \in [375, 425]$ GeV range. As a cross check, in Fig. 19(b), we plot the number of primary vertices with at least one track associated with the reconstructed jet of

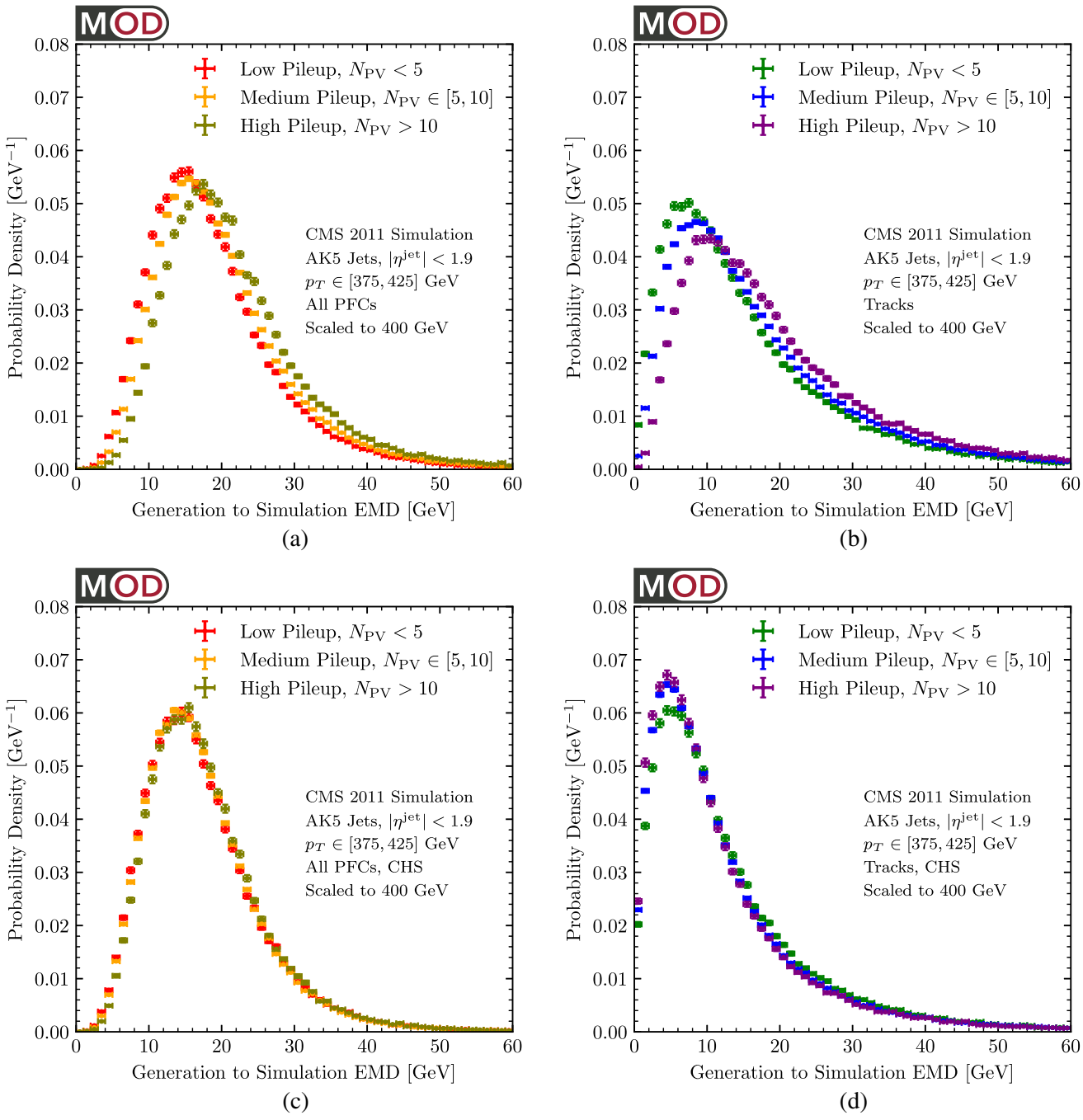

FIG. 20. The generation-to-simulation EMD in the style of Fig. 11 for different levels of pileup contamination, as quantified by the number of primary vertices (N_{PV}) in the event. Distributions are for (left column) all PFCs and (right column) just tracks, shown (top row) before and (bottom row) after CHS is applied.

interest. From this, we conclude that the event-wide reweighting does indeed correct the in-jet pileup contamination level.

We can quantify the performance of CHS for pileup mitigation by performing an EMD analysis analogous to Sec. IV B. In Fig. 20, we show the generation-to-simulation EMD before and after CHS is applied, split into low ($N_{\mathrm{PV}} < 5$), medium ($N_{\mathrm{PV}} \in [5, 10]$), and high ($N_{\mathrm{PV}} > 10$) levels of pileup contamination. First, we see that the EMD grows (i.e., reconstruction degrades) as the pileup levels increase, though for these modest levels of pileup, the distortions are not so large. As already shown in Fig. 11, CHS does mitigate the impact of pileup, with better performance when considering just tracks.

One surprise in Fig. 20(d) is that the track-only EMD gets *smaller* as the pileup contamination increases. We are not sure of the origin of this behavior. It might be related to the use of the rescaling factors in Eq. (8), or it might indicate a bias where low N_{PV} events often have unreconstructed primary vertices, so CHS does not removes tracks that it should. Regardless, we see that the EMD is a useful way to quantify the performance of pileup mitigation strategies.

APPENDIX C: ADDITIONAL PLOTS

In this appendix, we provide additional plots to complement the ones in the text.

In Fig. 21, we plot the turn-on behavior for all of the relevant single-jet triggers, to compare to the `Jet300` study in Fig. 3. In making this plot, we have to address the fact that some of the triggers share the same L1 trigger seed and their firing rates are therefore correlated. For uncorrelated triggers, if trigger A has prescale factor p_A^{trig} and trigger B has prescale factor p_B^{trig} and both triggers are fully efficient, then the probability of B firing given that A fired is

$$\mathcal{P}_{\mathrm{uncorr}}(B_{\mathrm{fired}}|A_{\mathrm{fired}}) = \frac{1}{p_B^{\mathrm{trig}}}, \tag{C1}$$

which is independent of p_A^{trig} since the triggers are uncorrelated. On the other hand, if two triggers have the same L1 seed, then the probability of B firing given that A fired is

$$\mathcal{P}_{\mathrm{corr}}(B_{\mathrm{fired}}|A_{\mathrm{fired}}) = \frac{\mathrm{gcd}(p_A^{\mathrm{trig}}, p_B^{\mathrm{trig}})}{p_B^{\mathrm{trig}}}, \tag{C2}$$

where gcd is the greater common divisor. This can be understood since if trigger A (B) fires deterministically every p_A^{trig} (p_B^{trig}) events, then they will overlap a factor of $\mathrm{gcd}(p_A^{\mathrm{trig}}, p_B^{\mathrm{trig}})$ more often than if the triggers fired randomly and independently. For example, if $\mathrm{gcd}(p_A^{\mathrm{trig}}, p_B^{\mathrm{trig}}) = p_B^{\mathrm{trig}}$,

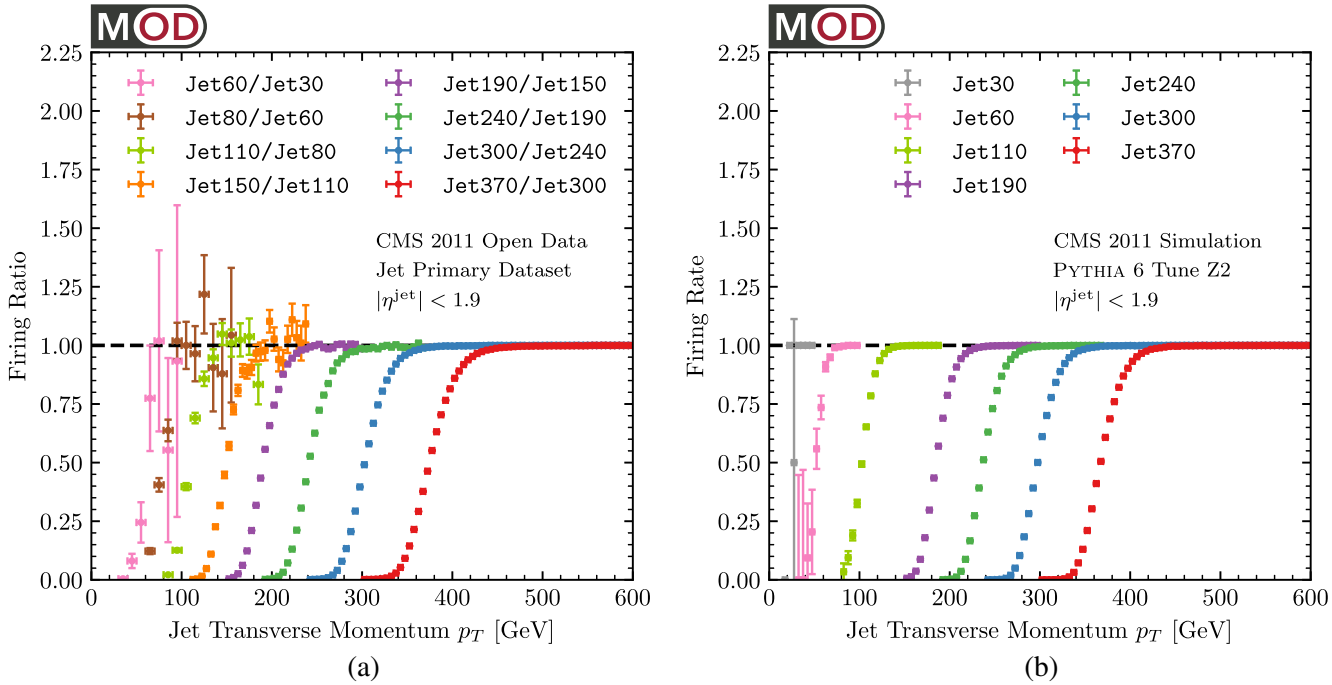

FIG. 21. (a) Relative trigger efficiency in the CMS Open Data, for 8 single-jet triggers compared to the adjacent trigger with lower p_T threshold. Up to statistical fluctuations, the firing ratio approaches 1 in all cases, after correcting for the L1 trigger correlation subtlety in Eq. (C2). (b) Absolute trigger efficiency in the MC simulation for seven single-jet triggers. The `Jet80` and `Jet150` triggers are not present in the simulated datasets, which are the two triggers that were turned off prior to the end of Run 2011A, as can be seen in Fig. 1(b). Efficiency information for the `Jet300` trigger is highlighted in Fig. 3.

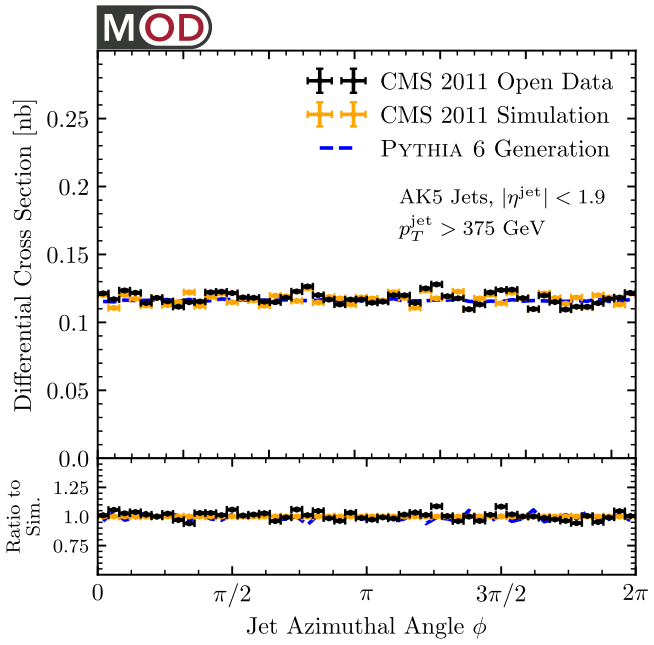

FIG. 22. Jet azimuthal angle (ϕ) distribution for the two hardest jets, comparing the CMS Open Data to MC event samples at the simulation level and generation level. See Fig. 5(b) for the pseudorapidity spectrum.

then the only time trigger B can fire is if trigger A has also fired, so $\mathcal{P}_{\text{corr}}(B_{\text{fired}}|A_{\text{fired}}) = 1$. In our case, this affects the `HLT_Jet150`, `HLT_Jet190`, `HLT_Jet240`, `HLT_Jet300`, and `HLT_Jet370` triggers, which are all seeded by the same `L1_SingleJet92` trigger [66].

In Fig. 22, we plot the azimuthal angle (ϕ) distribution for the two hardest jets. As expected, we observe a flat spectrum in both the CMS Open Data and the MC simulation, though the bin-to-bin fluctuations in the open data are larger than one would expect from statistics alone, possibly indicating an issue with the lack of ϕ-dependence of the JECs.

In Fig. 23, we plot the complete PFC p_T spectra for both neutral and charged constituents, going beyond the limited range shown in Fig. 6. This highlights the tighter relationship between generation-level and simulation-level information when using charged particles alone. Though not shown, we used this plot when deciding to impose the medium JQC, since with only the loose JQC, there was an excess of events with high-p_T neutral PFCs, most likely from photon-plus-jet events.

We now use EMD to study the impact of the p_T^{PFC} cut in our analysis. In the top row of Fig. 24, we do an apples-to-apples comparison with the same particle selection at generation level and simulation level. As the p_T cut on the PFCs gets more aggressive, the generation-to-simulation EMD decreases, indicating better agreement. Of course, this p_T^{PFC} cut removes information about jet substructure, so there is a balance between minimizing detector effects and maximizing sensitivity to the underlying radiation pattern. In the bottom row of Fig. 24, the baseline generation-level jet contains all particles ("raw"), regardless of what selections are made at simulation level. When using all PFCs in Fig. 24(c), the EMD decreases (i.e., reconstruction improves) as the p_T^{PFC} cut gets more

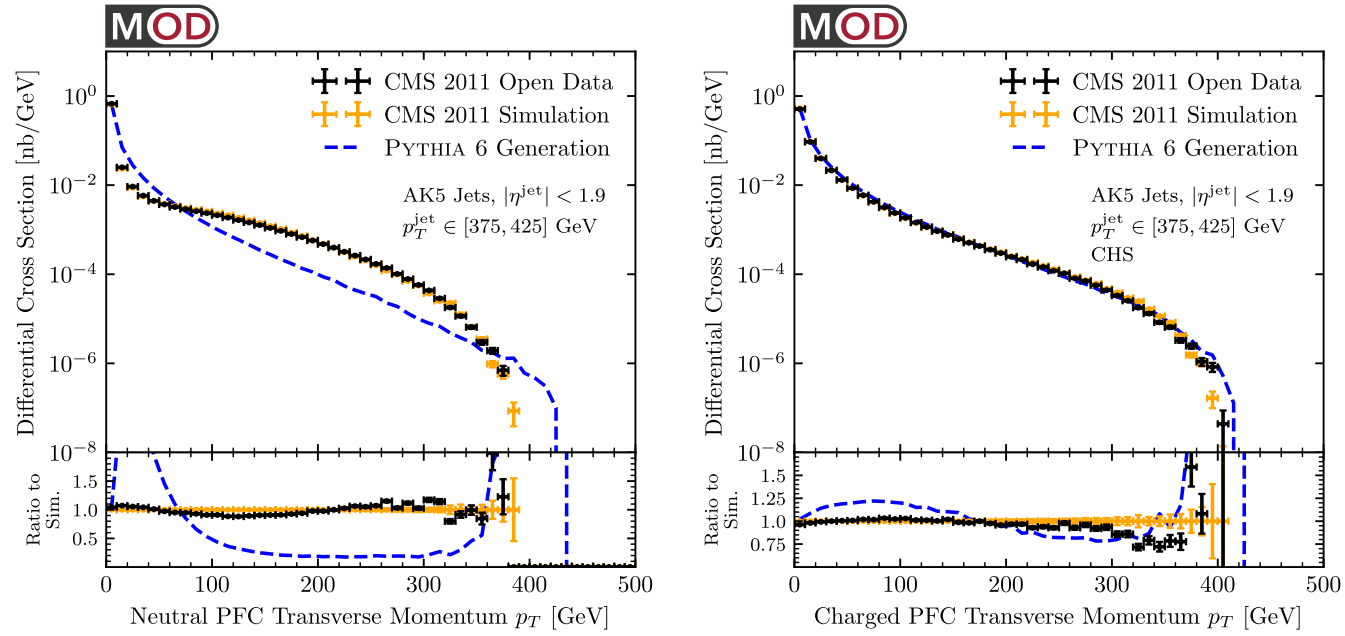

FIG. 23. Transverse momentum spectra for (a) neutral PFCs and (b) charged PFCs after CHS. A zoomed version of these plots highlighting the region below 5 GeV is shown in Fig. 6.

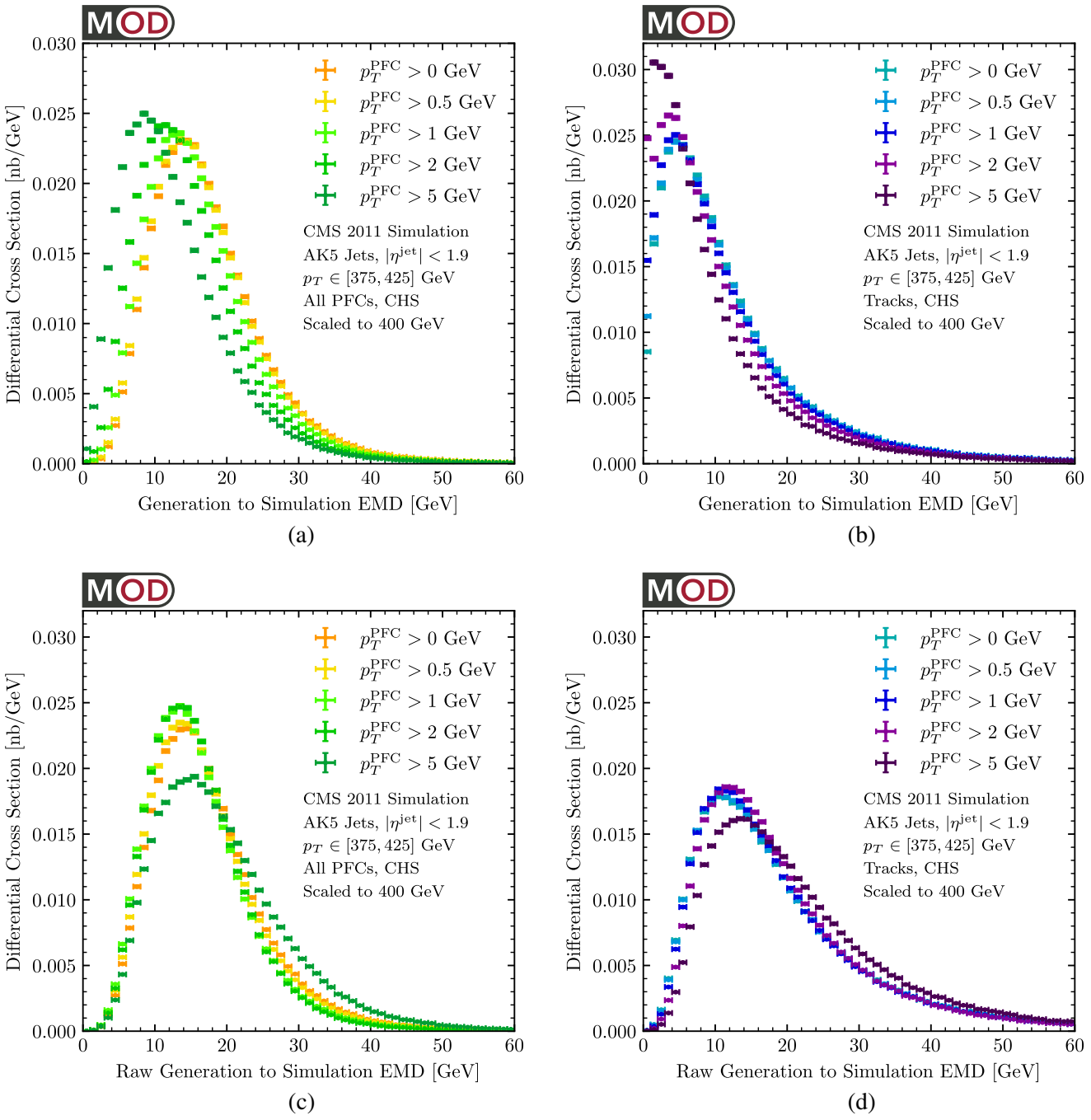

FIG. 24. Generation-to-simulation EMD as different PFC p_T selections are applied to the jet constituents, for (left column) all PFCs and (right column) just tracks. (top row) The baseline generation-level jet has the same p_T^{PFC} cut and track selection requirements as the simulation-level jets. (bottom row) The baseline generation-level jet uses all particles ("raw"), with no p_T cuts or track restrictions. In all cases, we apply the rescaling factor in Eq. (8).

stringent, up until the 2 GeV point where we start to see degradation. When using just charged PFCs in Fig. 24(d), the peak of the EMD distribution shifts to lower values but there is a long tail, and the reconstruction always degrades with increasing p_T^{PFC} cut.

In Fig. 25, we study the most anomalous jets according to $\bar{Q}_n$ from Eq. (11) for the additional choices of n of $n = \frac{1}{2}$ and $n = 2$. The results are comparable to the $n = 1$ case shown in Fig. 17, with all three choices of n agreeing on the three most anomalous jets.

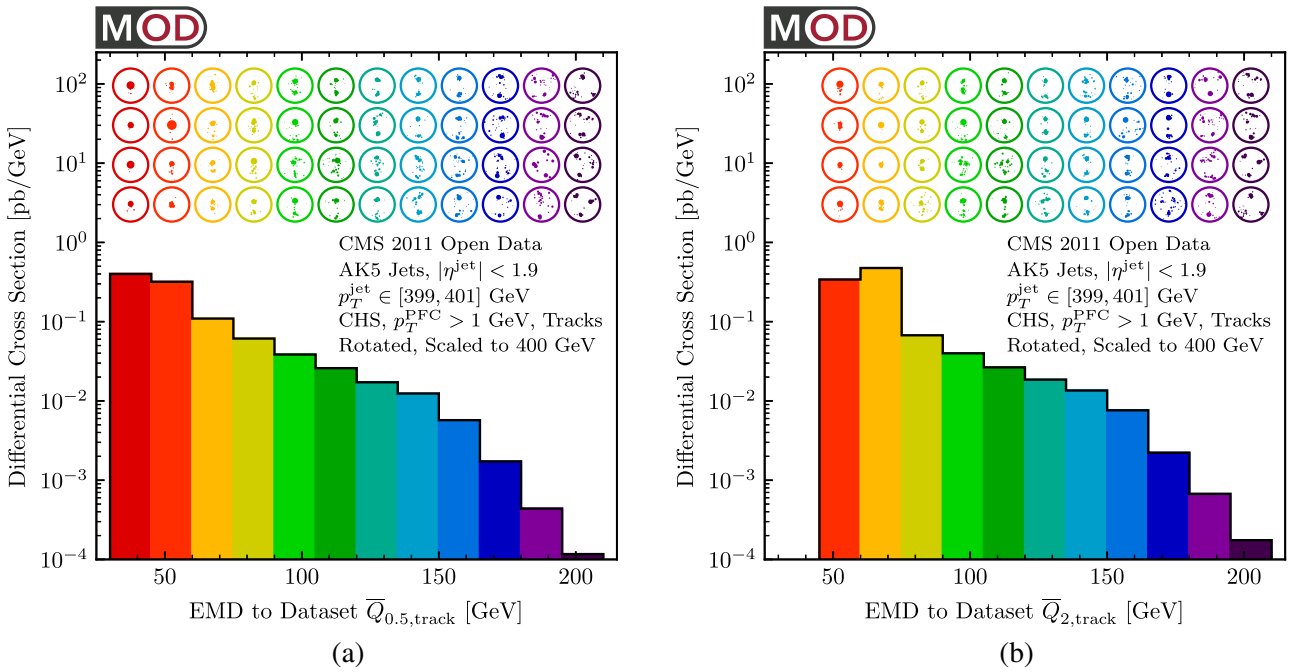

FIG. 25. Distribution on the CMS Open Data of $\bar{Q}_n$ from Eq. (11) for (a) $n = \frac{1}{2}$ and (b) $n = 2$ along with the 4-medoids in each histogram bin. See Fig. 17 for the analogous distribution for $n = 1$.

[1] G. Hanson *et al.*, Evidence for Jet Structure in Hadron Production by e+ e- Annihilation, Phys. Rev. Lett. **35**, 1609 (1975).

[2] J. D. Bjorken and S. J. Brodsky, Statistical model for electron-positron annihilation into hadrons, Phys. Rev. D **1**, 1416 (1970).

[3] J. R. Ellis, M. K. Gaillard, and G. G. Ross, Search for gluons in e+ e- annihilation, Nucl. Phys. **B111**, 253 (1976); Erratum, Nucl. Phys. **B130**, 516 (1977).

[4] H. Georgi and M. Machacek, A Simple QCD Prediction of Jet Structure in e+ e- Annihilation, Phys. Rev. Lett. **39**, 1237 (1977).

[5] E. Farhi, A QCD Test for Jets, Phys. Rev. Lett. **39**, 1587 (1977).

[6] G. Parisi, Super inclusive cross-sections, Phys. Lett. **74B**, 65 (1978).

[7] J. F. Donoghue, F. E. Low, and S.-Y. Pi, Tensor analysis of hadronic jets in quantum chromodynamics, Phys. Rev. D **20**, 2759 (1979).

[8] P. E. L. Rakow and B. R. Webber, Transverse momentum moments of hadron distributions in QCD jets, Nucl. Phys. **B191**, 63 (1981).

[9] M. H. Seymour, Tagging a heavy Higgs boson, in *ECFA Large Hadron Collider Workshop, Aachen, Germany, 1990: Proceedings.2.* (CERN, Geneva, 1991), pp. 557–569.

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

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

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

[13] 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).

[14] A. Abdesselam *et al.*, Boosted objects: A probe of beyond the Standard Model physics, Eur. Phys. J. C **71**, 1661 (2011).

[15] A. Altheimer *et al.*, Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks, J. Phys. G **39**, 063001 (2012).

[16] A. Altheimer *et al.*, Boosted objects and jet substructure at the LHC. Report of BOOST2012, held at IFIC Valencia, 23rd-27th of July 2012, Eur. Phys. J. C **74**, 2792 (2014).

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

[18] A. J. Larkoski, I. Moult, and B. Nachman, Jet substructure at the large hadron collider: A review of recent advances in theory and machine learning, arXiv:1709.04464.

[19] L. Asquith *et al.*, Jet substructure at the Large Hadron Collider: Experimental review, Rev. Mod. Phys. **91**, 045003 (2019).

[20] S. Marzani, G. Soyez, and M. Spannowsky, Looking inside jets: An introduction to jet substructure and boosted-object phenomenology, Lect. Notes Phys. **958** (2019).

[21] F. V. Tkachov, Measuring multi-jet structure of hadronic energy flow or what is a jet?, Int. J. Mod. Phys. A **12**, 5411 (1997).

[22] N. A. Sveshnikov and F. V. Tkachov, Jets and quantum field theory, Phys. Lett. B **382**, 403 (1996).

[23] P. S. Cherzor and N. A. Sveshnikov, Jet observables and energy momentum tensor, in *Quantum Field Theory and High-Energy Physics. Proceedings, Workshop, QFTHEP'97, Samara, Russia, 1997* (MSU, Moscow, 1997), pp. 402–407.

[24] T. Kinoshita, Mass singularities of Feynman amplitudes, J. Math. Phys. (N.Y.) **3**, 650 (1962).

[25] T. D. Lee and M. Nauenberg, Degenerate systems and mass singularities, Phys. Rev. **133**, B1549 (1964).

[26] G. F. Sterman and S. Weinberg, Jets from Quantum Chromodynamics, Phys. Rev. Lett. **39**, 1436 (1977).

[27] C. F. Berger *et al.*, Snowmass 2001: Jet energy flow project, eConf **C010630**, P512 (2001).

[28] C. F. Berger, T. Kucs, and G. F. Sterman, Event shape / energy flow correlations, Phys. Rev. D **68**, 014012 (2003).

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

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

[31] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow polynomials: A complete linear basis for jet substructure, J. High Energy Phys. 04 (2018) 013.

[32] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, J. High Energy Phys. 01 (2019) 121.

[33] S. H. Lim and M. M. Nojiri, Spectral analysis of jet substructure with neural networks: Boosted Higgs case, J. High Energy Phys. 10 (2018) 181.

[34] A. Chakraborty, S. H. Lim, and M. M. Nojiri, Interpretable deep learning for two-prong jet classification with jet spectra, J. High Energy Phys. 07 (2019) 135.

[35] S. Chatrchyan *et al.* (CMS Collaboration), The CMS experiment at the CERN LHC, J. Instrum. **3**, S08004 (2008).

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

[37] 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).

[38] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, and J. Thaler, Jet substructure studies with cms open data, Phys. Rev. D **96**, 074003 (2017).

[39] CMS Collaboration, Jet primary dataset in AOD format from RunB of 2010, https://jet.com/Run2010B-Apr21R-eReco-v1/AOD, CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.3S7F.2E9W (2014).

[40] A. J. Larkoski, S. Marzani, and J. Thaler, Sudakov safety in perturbative QCD, Phys. Rev. D **91**, 111501 (2015).

[41] A. M. Sirunyan *et al.* (CMS Collaboration), Measurement of the Splitting Function in pp and Pb-Pb Collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV, Phys. Rev. Lett. **120**, 142302 (2018).

[42] S. Acharya *et al.* (ALICE Collaboration), Exploration of jet substructure using iterative declustering in pp and Pb-Pb collisions at LHC energies, arXiv:1905.02512.

[43] K. Kauder (STAR Collaboration), Measurement of the Shared Momentum Fraction z_g using Jet Reconstruction in $\mathrm{p}+\mathrm{p}$ and $\mathrm{Au}+\mathrm{Au}$ Collisions with STAR, Nucl. Phys. **A967**, 516 (2017).

[44] C. F. Madrazo, I. H. Cacha, L. L. Iglesias, and J. M. de Lucas, Application of a Convolutional Neural Network for image classification to the analysis of collisions in High Energy Physics, EPJ Web Conf. **214**, 06017 (2019).

[45] M. Andrews, M. Paulini, S. Gleyzer, and B. Poczos, End-to-End Physics Event Classification with the CMS Open Data: Applying Image-based Deep Learning on Detector Data to Directly Classify Collision Events at the LHC, arXiv:1807.11916.

[46] M. Andrews, J. Alison, S. An, P. Bryant, B. Burkle, S. Gleyzer, M. Narain, M. Paulini, B. Poczos, and E. Usai, End-to-end jet classification of quarks and gluons with the CMS open data, arXiv:1902.08276.

[47] S. P. Mehdiabadi and A. Fahim, Explicit jet veto as a tool to purify the underlying event in the drell-yan process using CMS open data, J. Phys. G **46**, 095003 (2019).

[48] C. Cesarotti, Y. Soreq, M. J. Strassler, J. Thaler, and W. Xue, Searching in CMS open data for dimuon resonances with substantial transverse momentum, Phys. Rev. D **100**, 015021 (2019).

[49] C. G. Lester and M. Schott, Search for non-standard sources of parity violation in jets at $\sqrt{s} = 8$ TeV with CMS open data, J. High Energy Phys. 12 (2019) 120.

[50] A. Apyan, W. Cuozzo, M. Klute, Y. Saito, M. Schott, and B. Sintayehu, Opportunities and challenges of standard model production cross section measurements at $\sqrt{s} = 8$ TeV using CMS open data, J. Instrum. **15**, P01009 (2020).

[51] ALEPH Collaboration, ALEPH Preservation Policy, https://hep-project-dphep-portal.web.cern.ch/content/aleph-preservation-policy (2003).

[52] A. Heister, Observation of an excess at 30 GeV in the opposite sign di-muon spectra of $\mathrm{Z} \to b\bar{b} + \mathrm{X}$ events recorded by the ALEPH experiment at LEP, arXiv:1610.06536.

[53] J. Kile and J. von Wimmersperg-Toeller, Monte Carlo tuning for $e^+e^- \to$ hadrons and comparison with unfolded LEP data, arXiv:1706.02242.

[54] J. Kile and J. von Wimmersperg-Toeller, Localized 4σ and 5σ dijet mass excesses in ALEPH LEP2 Four-Jet Events, J. High Energy Phys. 10 (2018) 116.

[55] J. Kile and J. von Wimmersperg-Toeller, Simulation of $e^+e^- \to$ hadrons and comparison to ALEPH data at full detector simulation with an emphasis on four-jet states, arXiv:1706.02269.

[56] P. T. Komiske, E. M. Metodiev, and J. Thaler, The Metric Space of Collider Events, Phys. Rev. Lett. **123**, 041801 (2019).

[57] S. Peleg, M. Werman, and H. Rom, A unified approach to the change of resolution: Space and gray-level, IEEE Trans. Pattern Anal. Mach. Intell. **11**, 739 (1989).

[58] Y. Rubner, C. Tomasi, and L. J. Guibas, A metric for distributions with applications to image databases, in *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98 (IEEE Computer Society, Washington, DC, USA, 1998), pp. 59–66.

[59] Y. Rubner, C. Tomasi, and L. J. Guibas, The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vis. **40**, 99 (2000).

[60] O. Pele and M. Werman, A linear time histogram metric for improved SIFT matching, in *Computer Vision—ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, 2008, Proceedings, Part III* (Springer-Verlag, Berlin, Heidelberg, 2008), pp. 495–508.

[61] O. Pele and B. Taskar, The tangent earth mover's distance, in *Geometric Science of Information—First International Conference, GSI 2013, Paris, France, 2013. Proceedings* (Kluwer Academic Publishers, Norwell, MA, 2013), pp. 397–404.

[62] CMS Collaboration, Particle-flow event reconstruction in CMS and performance for jets, taus, and MET, Tech. Rep. CMS-PAS-PFT-09-001, CERN, 2009. Geneva, 2009.

[63] CMS Collaboration, Commissioning of the particle-flow event reconstruction with the first LHC collisions recorded in the CMS detector, Tech. Rep. CMS-PAS-PFT-10-001, 2010.

[64] A. M. Sirunyan *et al.* (CMS Collaboration), Particle-flow reconstruction and global event description with the CMS detector, J. Instrum. **12**, P10003 (2017).

[65] CMS Collaboration, Pileup removal algorithms, Tech. Rep. CMS-PAS-JME-14-001, 2014.

[66] CMS Collaboration, Jet primary dataset in AOD format from RunA of 2011, (/Jet/Run2011A-12Oct2013-v1/AOD). CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.UP77.P6PQ (2016).

[67] CMS Collaboration, Simulated dataset QCD_Pt-0to5_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.2QR5.9P6G (2016).

[68] CMS Collaboration, Simulated dataset QCD_Pt-5to15_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.972P.PY4C (2016).

[69] CMS Collaboration, Simulated dataset QCD_Pt-15to30_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.RSKY.VC8C (2016).

[70] CMS Collaboration, Simulated dataset QCD_Pt-30to50_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.Q3BX.69VQ (2016).

[71] CMS Collaboration, Simulated dataset QCD_Pt-50to80_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.84VC.RU8W (2016).

[72] CMS Collaboration, Simulated dataset QCD_Pt-80to120_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.PUTE.7H2H (2016).

[73] CMS Collaboration, Simulated dataset QCD_Pt-120to170_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.QJND.HA88 (2016).

[74] CMS Collaboration, Simulated dataset QCD_Pt-170to300_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.WKRR.DCJP (2016).

[75] CMS Collaboration, Simulated dataset QCD_Pt-300to470_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.X3XQ.USQR (2016).

[76] CMS Collaboration, Simulated dataset QCD_Pt-470to600_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.BKTD.SGJX (2016).

[77] CMS Collaboration, Simulated dataset QCD_Pt-600to800_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.EJT7.KSAY (2016).

[78] CMS Collaboration, Simulated dataset QCD_Pt-800to1000_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.S3D5.KF2C (2016).

[79] CMS Collaboration, Simulated dataset QCD_Pt-1000to1400_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.96U2.3YAH (2016).

[80] CMS Collaboration, Simulated dataset QCD_Pt-1400to1800_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.RC9V.B5KX (2016).

[81] CMS Collaboration, Simulated dataset QCD_Pt-1800_TuneZ2_7TeV_pythia6 in AODSIM format for 2011 collision data (SM Exclusive), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.CX2X.J3KW (2016).

[82] T. Sjostrand, S. Mrenna, and P. Z. Skands, PYTHIA 6.4 Physics and Manual, J. High Energy Phys. 05 (2006) 026.

[83] S. Agostinelli *et al.* (GEANT4 Collaboration), GEANT4: A Simulation toolkit, Nucl. Instrum. Methods A **506**, 250 (2003).

[84] P. T. Komiske and E. M. Metodiev, EnergyFlow, https://energyflow.network/ (2019).

[85] P. T. Komiske and E. M. Metodiev, MOD GitHub Repository, https://github.com/pkomiske/MOD (2019).

[86] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A open data | jet primary dataset |

pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3340205 (2019).
[87] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A simulation | Pythia 6 QCD 170-300 | pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3341500 (2019).
[88] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A simulation | Pythia 6 QCD 300-470 | pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3341498 (2019).
[89] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A simulation | Pythia 6 QCD 470-600 | pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3341419 (2019).
[90] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A simulation | Pythia 6 QCD 600-800 | pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3364139 (2019).
[91] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A simulation | Pythia 6 QCD 800-1000 | pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3341413 (2019).
[92] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A simulation | Pythia 6 QCD 1000-1400 | pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3341502 (2019).
[93] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A simulation | Pythia 6 QCD 1400-1800 | pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3341770 (2019).
[94] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, CMS 2011A simulation | Pythia 6 QCD 1800-inf | pT > 375 GeV | MOD HDF5 format, Zenodo https://doi.org/10.5281/zenodo.3341772 (2019).
[95] CMS Collaboration, CMS releases first batch of high-level LHC open data, http://opendata.cern.ch/docs/cms-releases-first-batch-of-high-level-lhc-open-data (2014).
[96] CMS Collaboration, CMS releases new batch of research data from LHC, http://opendata.cern.ch/docs/cms-releases-new-batch-of-research-data-from-lhc (2016).
[97] CMS Collaboration, Observing the Higgs with over one petabyte of new CMS Open Data, http://opendata.cern.ch/docs/observing-higgs-over-one-petabyte-new-cms-open-data (2017).
[98] CMS Collaboration, CMS releases open data for Machine Learning, http://opendata.cern.ch/docs/cms-releases-open-data-for-machine-learning (2019).
[99] ALICE Collaboration, ALICE releases educational datasets, http://opendata.cern.ch/docs/alice-releases-educational-datasets (2014).
[100] ATLAS Collaboration, Explore ATLAS Open Data resources, http://opendata.cern.ch/docs/explore-atlas-open-data-resources (2016).
[101] LHCb Masterclasses, http://opendata.cern.ch/record/41.
[102] Release of the first set of data samples by the OPERA Collaboration, http://opendata.cern.ch/docs/opera-news-first-release-2018 (2018).
[103] HEPData, https://www.hepdata.net/.
[104] Rivet, https://rivet.hepforge.org/.
[105] Reana, http://reana.io/.
[106] X. Chen *et al.*, Open is not enough, Nat. Phys. **15**, 113 (2019).
[107] V. Khachatryan *et al.* (CMS Collaboration), The CMS trigger system, J. Instrum. **12**, P01020 (2017).
[108] CMS Collaboration, Jet primary dataset sample in RAW format from RunA of 2011 (from /Jet/Run2011A-v1/RAW), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.DRTR.53Q8 (2019).
[109] CMS Collaboration, MinimumBias primary dataset sample in RAW format from RunA of 2011 (from /MinimumBias/Run2011A-v1/RAW), CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.I8HN.DF32 (2017).
[110] CMS Collaboration, CMS luminosity information, for 2011 CMS open data, CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.FPPH.Q7S2 (2017).
[111] CMS Collaboration, CMS list of validated runs for primary datasets of 2011 data taking, CERN Open Data Portal, https://doi.org/10.7483/OPENDATA.CMS.3Q75.7835 (2016).
[112] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[113] M. Cacciari, G. P. Salam, and G. Soyez, The catchment area of jets, J. High Energy Phys. 04 (2008) 005.
[114] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, Eur. Phys. J. C **72**, 1896 (2012).
[115] The HDF Group, Hierarchical Data Format, version 5, (1997-NNNN), http://www.hdfgroup.org/HDF5/.
[116] CMS Collaboration, Jet Performance in pp Collisions at 7 TeV, Tech. Rep. CMS-PAS-JME-10-003, 2010.
[117] CMS Collaboration, Determination of jet energy calibration and transverse momentum resolution in CMS, J. Instrum. **6**, 11002 (2011).
[118] NumPy, https://www.numpy.org/.
[119] Matplotlib, https://matplotlib.org/.
[120] R. Flamary and N. Courty, *Pot Python Optimal Transport Library* (2017), https://github.com/rflamary/POT.
[121] Project Jupyter, https://jupyter.org/.
[122] R. Field, Min-Bias and the Underlying Event at the LHC, Acta Phys. Pol. B **42**, 2631 (2011).
[123] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015).
[124] M. Cacciari, S. Forte, D. Napoletano, G. Soyez, and G. Stagnitto, The single-jet inclusive cross-section and its definition, Phys. Rev. D **100**, 114015 (2019).
[125] G. Aad *et al.* (ATLAS Collaboration), The performance of the jet trigger for the ATLAS detector during 2011 data taking, Eur. Phys. J. C **76**, 526 (2016).
[126] M. C. Kumar and S.-O. Moch, Phenomenology of threshold corrections for inclusive jet production at hadron colliders, Phys. Lett. B **730**, 122 (2014).
[127] M. Tanabashi *et al.* (Particle Data Group), Review of particle physics, Phys. Rev. D **98**, 030001 (2018).
[128] M. Cacciari, G. P. Salam, and G. Soyez, SoftKiller, a particle-level pileup removal method, Eur. Phys. J. C **75**, 59 (2015).
[129] 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.
[130] CMS Collaboration, Performance of quark/gluon discrimination in 8 TeV pp data, Tech. Rep. CMS-PAS-JME-13-002, 2013.
[131] G. Aad *et al.* (ATLAS Collaboration), Study of jet shapes in inclusive jet production in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector, Phys. Rev. D **83**, 052003 (2011).
[132] P. Gras, S. Höche, 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.
[133] J. Pumplin, How to tell quark jets from gluon jets, Phys. Rev. D **44**, 2025 (1991).
[134] A. J. Larkoski, I. Moult, and D. Neill, Power counting to better jet observables, J. High Energy Phys. 12 (2014) 009.
[135] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, Towards an understanding of jet substructure, J. High Energy Phys. 09 (2013) 029.
[136] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft drop, J. High Energy Phys. 05 (2014) 146.
[137] P. T. Komiske and E. M. Metodiev, MOD Jet Demonstration, https://mybinder.org/v2/gh/pkomiske/EnergyFlow/master?filepath=demos/MOD%20Jet%20Demo.ipynb (2019).
[138] V. Mateu, I. W. Stewart, and J. Thaler, Power corrections to event shapes with mass-dependent operators, Phys. Rev. D **87**, 014025 (2013).
[139] J. B. Orlin, A polynomial time primal network simplex algorithm for minimum cost flows, Math. Program. **77**, 109 (1997).
[140] R. E. Tarjan, Dynamic trees as search trees via euler tours, applied to the network simplex algorithm, Math. Program. **77**, 169 (1997).
[141] J. B. Orlin, S. A. Plotkin, and É. Tardos, Polynomial dual network simplex algorithms, Math. Program. **60**, 255 (1993).
[142] O. Pele and B. Taskar, The tangent earth mover's distance, in *Geometric Science of Information*, edited by F. Nielsen and F. Barbaresco (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013), pp. 397–404.
[143] L. N. Wasserstein, Markov processes over denumerable products of spaces describing large systems of automata, Probl. Inf. Transm. **5**, 47 (1969).
[144] R. L. Dobrushin, Prescribing a system of random variables by conditional distributions, Theory Probab. Appl. **15**, 458 (1970).
[145] W. J. Waalewijn, Calculating the charge of a jet, Phys. Rev. D **86**, 094030 (2012).
[146] D. Krohn, M. D. Schwartz, T. Lin, and W. J. Waalewijn, Jet Charge at the LHC, Phys. Rev. Lett. **110**, 212001 (2013).
[147] H.-M. Chang, M. Procura, J. Thaler, and W. J. Waalewijn, Calculating Track-Based Observables for the LHC, Phys. Rev. Lett. **111**, 102002 (2013).
[148] H.-M. Chang, M. Procura, J. Thaler, and W. J. Waalewijn, Calculating track thrust with track functions, Phys. Rev. D **88**, 034030 (2013).
[149] The ATLAS collaboration, Jet mass reconstruction with the ATLAS Detector in early Run 2 data, Tech. Rep. ATLAS-CONF-2016-035, 2016.
[150] B. T. Elder and J. Thaler, Aspects of track-assisted mass, J. High Energy Phys. 03 (2019) 104.
[151] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. **9**, 2579 (2008).
[152] L. van der Maaten, Learning a parametric embedding by preserving local structure, in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, 2009* (PMLR, Clearwater Beach, Florida, 2009), pp. 384–391.
[153] L. van der Maaten and G. E. Hinton, Visualizing non-metric similarities in multiple maps, Mach. Learn. **87**, 33 (2012).
[154] L. van der Maaten, Accelerating t-sne using tree-based algorithms, J. Mach. Learn. Res. **15**, 3221 (2014).
[155] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).
[156] P. Grassberger and I. Procaccia, Characterization of Strange Attractors, Phys. Rev. Lett. **50**, 346 (1983).
[157] B. Kégl, Intrinsic dimension estimation using packing numbers, in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, Vancouver, British Columbia, Canada]* (MIT Press, Cambridge, MA, 2002), pp. 681–688.
[158] P. Bolzoni, B. A. Kniehl, and A. V. Kotikov, Gluon and Quark Jet Multiplicities at $N^3LO + NNLL$, Phys. Rev. Lett. **109**, 242002 (2012).
[159] P. Bolzoni, B. A. Kniehl, and A. V. Kotikov, Average gluon and quark jet multiplicities at higher orders, Nucl. Phys. **B875**, 18 (2013).
[160] D. Neill and W. J. Waalewijn, The entropy of a jet, Phys. Rev. Lett. **123**, 142001 (2019).
[161] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.
[162] J. Thaler and K. Van Tilburg, Maximizing Boosted Top Identification by Minimizing N-subjettiness, J. High Energy Phys. 02 (2012) 093.
[163] A. Novikov, PyClustering: Data mining library, J. Open Source Software **4**, 1230 (2019).
[164] S. D. Ellis and D. E. Soper, Successive combination jet algorithm for hadron collisions, Phys. Rev. D **48**, 3160 (1993).
[165] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, Longitudinally invariant K_t clustering algorithms for hadron hadron collisions, Nucl. Phys. **B406**, 187 (1993).
[166] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, Better jet clustering algorithms, J. High Energy Phys. 08 (1997) 001.
[167] M. Wobisch and T. Wengler, Hadronization corrections to jet cross-sections in deep inelastic scattering, in *Monte Carlo generators for HERA physics. Proceedings, Workshop, Hamburg, Germany, 1998–1999* (DESY, Hamburg, 1998), pp. 270–279.

[168] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure, J. High Energy Phys. 156 (2016) 05.
[169] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Søgaard, Decorrelated jet substructure tagging using adversarial neural networks, Phys. Rev. D **96**, 074034 (2017).
[170] I. Moult, B. Nachman, and D. Neill, Convolved substructure: Analytically decorrelating jet substructure observables, J. High Energy Phys. 05 (2018) 002.
[171] J. H. Collins, K. Howe, and B. Nachman, Anomaly Detection for Resonant New Physics with Machine Learning, Phys. Rev. Lett. **121**, 241803 (2018).
[172] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, A generic anti-QCD jet tagger, J. High Energy Phys. 11 (2017) 163.
[173] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, Novelty detection meets collider physics,arXiv:1807.10261.
[174] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or what?, SciPost Phys. **6**, 030 (2019).
[175] M. Farina, Y. Nakai, and D. Shih, Searching for new physics with deep autoencoders, arXiv:1808.08992.
[176] J. H. Collins, K. Howe, and B. Nachman, Extending the search for new resonances with machine learning, Phys. Rev. D **99**, 014038 (2019).
[177] T. S. Roy and A. H. Vijay, A robust anomaly finder based on autoencoder, arXiv:1903.02032.
[178] R. T. D'Agnolo and A. Wulzer, Learning new physics from a machine, Phys. Rev. D **99**, 015014 (2019).
[179] A. De Simone and T. Jacques, Guiding new physics searches with unsupervised learning, Eur. Phys. J. C **79**, 289 (2019).
[180] M. Strassler and J. Thaler, Slow and steady, Nat. Phys. **15**, 725 (2019).