Jet Substructure at the Large Hadron Collider

by

Aashish Tripathee

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Bachelor of Science in Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Signature redacted

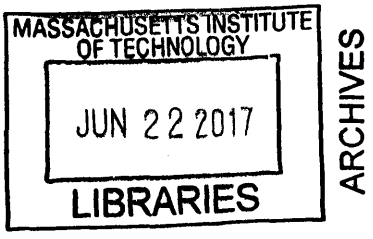
Author ... .
Department of Physics
May 19, 2017

Signature redacted

Certified by..... .....
Jesse D. Thaler
Associate Professor
Thesis Supervisor

Signature redacted

Accepted by ....
Nergis Mavalvala
Senior Thesis Coordinator, Department of Physics



Jet Substructure at the Large Hadron Collider

by

Aashish Tripathee

Submitted to the Department of Physics
on May 19, 2017, in partial fulfillment of the
requirements for the degree of
Bachelor of Science in Physics

Abstract

In this thesis, we use the CMS Open Data to study the 2-prong substructure of jets. We use CMS's particle flow reconstruction algorithm to obtain jet constituents, which we then use to perform various jet substructure studies. After validating our basic kinematics and substructure results through a comparison to results from parton shower generators, we extract the 2-prong substructure of the leading jet using the soft drop algorithm. We find good agreement between the results from the Open Data and those obtained from parton shower generators. For the 2-prong substructure, we also compare to analytic calculations performed to modified leading-logarithmic accuracy. To our best knowledge, this is the first ever physics analysis based on the CMS Open Data.

Thesis Supervisor: Jesse D. Thaler
Title: Associate Professor

Acknowledgments

First of all, I am grateful to my advisor Jesse Thaler not only for his indispensable guidance, and support throughout the research process but also for his invaluable mentorship regarding learning physics and pursuing an academic career in physics.

This thesis is based on two papers: "Exposing the QCD Splitting Function with CMS Open Data" [1] and "Jet Substructure Studies with CMS Open Data" [2]. So I am thankful to my collaborators: Andrew Larkoski, Simone Marzani, and Wei Xue in addition to Jesse. Moreover, I would like to express sincere gratitude to Salvatore Rappoccio and Kati Lassila-Perini for their guidance in navigating the CMS analysis framework.

I would also like to extend my gratitude to my academic advisor Tracy Slatyer for her unmatched level of support, mentorship, and patience with me; and for an amazing semester of 8.033 (and 8.323), which was my first real physics class at MIT and instrumental in my realization of how much I liked physics. I also want to thank Janet Conrad for believing in me and giving me the opportunity to work in her group and for a fantastic semester of Junior Lab. I am also thankful to Daniel Winklehner for teaching me so much about experimental particle physics, and coding.

I am also thankful to my family for their constant, unconditional support. I would also like to thank Ioana for continuously inspiring me, for pushing me so hard and for forcing me to get out of my comfort zone so many times. I am also indebted to the incredible support of Bishwa, Chandan, CJ, Ishwar, Kalyan, Kishore, Krupa, Mahesh, Pramod, Prasant, Sapna, Sarah, Shambhu, Simanta, Subekshya, Uddhav, Vijay, Zeo, and all other friends from Student House.

Thanks is also due to the extremely helpful and understanding staff in the Physics Academic Office: Nancy Boyce, Catherine Modica, Nancy Savioli, and Denise Wahkor. I am also grateful to the MIT UROP program without which, none of this would have been possible. I am also indebted to the kindness of Dean Justin Kasarsky at S^3.

Last but not the least, I would like to thank everyone in *The Melatonin/Breakfast Club* for always being there for me and for caring so much about me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Quantum Chromodynamics (QCD) is a highly successful theory of the strong interactions and is a non-abelian gauge theory with the symmetry group SU(3). The strong interaction governs the strong nuclear force responsible for confining quarks and gluons into hadrons like protons and neutrons. Gluons are the force carriers of the strong force.

The conserved quantity in QCD is the color charge, conveniently labeled by Red, Green, and Blue because three color charges combine to give a chargeless (white) hadron. Color confinement is a phenomenon which dictates that color charges cannot be isolated. This means, quarks and gluons cannot exist in isolation and instead always combine with each other into chargeless clumps called hadrons. Because of this, quarks (and gluons) cannot be studied in isolation and we need an alternate, indirect way to study them. One such way is through jets.

During a hard QCD process, when partons (quarks and gluons) fly apart, instead of appearing in isolation, they produce quark-antiquark pairs from the vacuum and produce other hadrons in a process called hadronization. A jet is a narrow collimated spray of these hadrons coming from the hadronization of quarks and gluons. Experimentally, a "jet" is identified using a jet algorithm, which is an important phenomenological tool to study the strongly interacting particles. The final states in hard QCD interactions can be calculated through the study of the structure of these jets. After a hard QCD process, the partons are sprayed around into fragments with

some color charge of the underlying particle. However, because of color confinement, these colored fragments create other colored objects to balance out the color charge, thereby making colorless jets. We need to algorithmically reconstruct such jets to study them, since a jet is not an actual physical object and so detectors detect individual particles and not jets.

Because jets arise from strong interactions, they provide fertile ground for attempts to understand the strong force. Of particular value is the study of jet substructure, which exposes the underlying structure of jets, thereby giving a window into the behavior of quarks and gluons and their interactions. Jet substructure methods are in wide use at the Large Hadron Collider (LHC), especially to search for physics beyond the standard model.

In this thesis, we study various jet substructure properties using the 2010 CMS Open Data. This thesis is based on two manuscripts currently in review at Physical Review Letters [1] and Physical Review D [2].

Chapter 2

CMS Open Data

For the first time in the history of particle physics, the CMS experiment at the Large Hadron Collider (LHC) released data it collected during Run B in 2010, corresponding to 7 TeV proton-proton collisions. This corresponds to a low-pileup, high luminosity dataset, making it a particularly attractive for physics analyses. To our best knowledge, our analysis [1, 2] is the first ever physics analysis done using the CMS Open Data.

The CMS Open Data can be accessed from the CERN Open Data Portal [5], which incidentally now also contains data from the ATLAS experiment, although ATLAS's Open Data is education-grade and not research-grade. The primary datasets are in the form of Analysis Object Data (AOD) files. AOD is a ROOT-based file format used internally by CMS. To process this data, one first needs to install a CERNVM virtual machine. CERNVM is a CMS-provided virtual machine based on SCIENTIFIC LINUX CERN 5, that comes pre-bundled with all the tools— including the CMS Software Framework (CMSSW)— required for analyzing the provided data.

Our analysis uses the Jet Primary Dataset [6]. This, fully described in Sec. 2.2, is a subset of all the recorded events that pass a certain set of triggers. There are 1664 AOD files in the Jet Primary Dataset, holding 20,022,826 events and occupying 2.0 terabytes of disk space. While the recommended way of acquiring the AOD files is through the XROOTD interface [7], we chose, for our own convenience, to instead download all the AOD files and process them locally. We did make sure to preserve the same directory structure as on the Open Data server to prevent our method of

data-retrieval from affecting our analysis in any possible way. We processed these AOD files and converted them into a text-based MIT Open Data (MOD) format (see Sec. 2.3) before performing actual physics analyses.

2.1 The CMS Software Framework

CMSSW is a framework written in Python and C++ to analyze and process CMS's data. All analyses inside the CMS collaboration are performed using this framework. The version of CMSSW that needs to be used with the 2010 CMS Open Data is 4.2.8, the current version, as of writing this, being 9.2.0. While performing the entire analysis using CMSSW is certainly the prescribed method, we decided to take a different route and used CMSSW only to extract data relevant for our analysis out of the AOD format.

Multiple data-tiers exist within CMS, Analysis Object Data (AOD) being the one the Open Data provides. The foremost detector-level data is called DAQ-RAW. This comes directly from the detectors and the L1 triggers. When DAQ-RAW data has been formatted, along with performing HLT (Higher Level Trigger) based selections, the data is in the RAW tier. This data is then used to reconstruct objects like tracks, jets and vertices to produce the RECO tier. The next tier is the Analysis Object Data (AOD) which is a subset of RECO and contains limited refitting of tracks and clusters. AOD is sufficient for most analyses and so is the format that the Open Data is distributed on. While other tiers exist (TAG, FEVT, GEN, SIM, DIGI), they are not relevant to our analysis and we omit discussions about them.

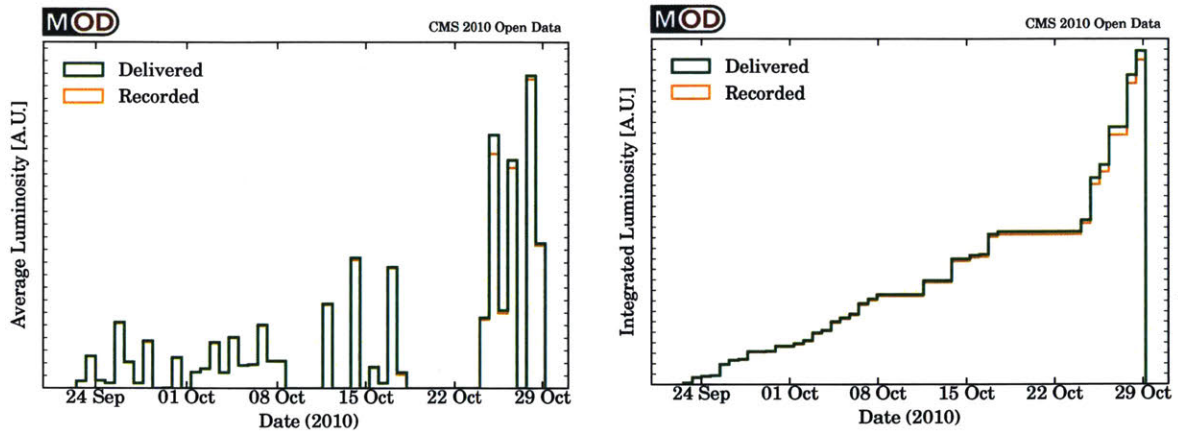
To extract data out of the retrieved AOD files, we wrote a chain of user-defined modules; a `Source` module that read events from the AOD files, and an `EDProducer` called `MODProducer` to write out the extracted information to a text-based format called MIT Open Data (MOD) format (See Sec. 2.3). The `ED` in `EDProducer` stands for "Event Data". The CMS-recommended method is to use an `EDProducer` to "produce" new data files, an `EDAnalyzer` for data analysis, and an `OutputModule` to write down the processed result. However, we chose to use an `EDProducer` instead of an

EDAnalyzer, even though we were not actually producing new processed data, because the name aligned better with our intended purpose of MODProducer: producing MOD data files. Moreover, to keep things simpler, we used standard C++ libraries for file output instead of the recommended OutputModule. MODProducer is available through our GitHub repository [8].

To enable data validation at a file-level and to keep file sizes manageable, we wanted to generate a separate MOD file for each of the AOD files, thereby maintaining a one-to-one relationship between the two file formats. The easiest way to do that would be to run MODProducer on each AOD file, one at a time. However, we encountered a hurdle on doing this when we realized that, for MODProducer to extract trigger prescales, it needs to load FrontierConditions_GlobalTag_cff and the appropriate global tag (GR_R_42_V25::All). Loading this information takes around 10 minutes and this needs to happen before MODProducer can extract any data. To circumvent this waiting period of 10 minutes per each of the 1664 files, we created another EDProducer called FilenameMapProducer, which creates a map of event and run numbers against the corresponding AOD filename. This allowed us to get away with running MODProducer just once, thereby negating the need of loading the tags multiple times.

MODProducer extracts PFCandidates (PFCs), jets clustered from these PFCs, corresponding jet calibration information, luminosity information[1], and basic identification tags like event and run numbers into the MOD format. Particle Flow is CMS's proprietary algorithm that uses information from the calorimeters and various other detector elements to provide a unique particle-like interpretation in terms of reconstructed photons, electrons, muons, charged hadrons, and neutral hadrons. CMS clusters these PFCs using the anti-k_T algorithm. While the dataset contains these jets for multiple values of the jet radius R, we output just the $R = 0.5$ jets as that is the

[1]The luminosity obtained directly from the AOD files, without the aid of a separate luminosity database, is not correct and the luminosity we calculate is roughly ten times the official value of 31.79 pb^{-1}. Despite this huge discrepancy, we realized that the luminosity distribution against time looks qualitatively very similar to results published by CMS [9] (see Fig. 2.1) and so we decided to keep the information. It gives us great pleasure to report that CMS is now providing a luminosity database for its Open Data releases and that the authors of this analysis [1, 2] were able to provide concrete feedback on the best way to provide this information.



(a) (b)

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

most commonly used value of R, particularly for substructure studies. For the AK5 jets, in addition to their four-momenta, we also output their calibration information the Jet Energy Correction (JEC) factors and jet quality parameters. Moreover, in our analysis pipeline, we validate these jets by performing our own clustering of the PFCandidates using FastJet 3.1.3 [10].

2.2 The Jet Primary Dataset

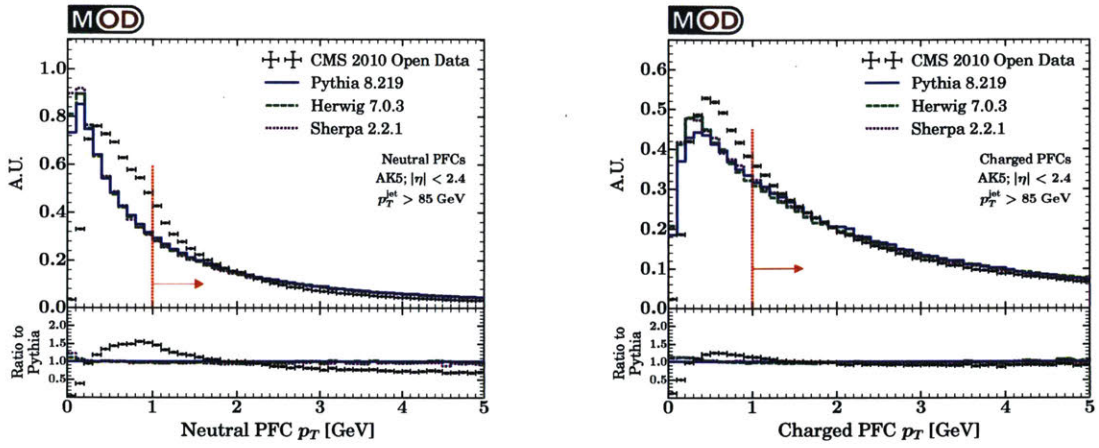
The CMS Open Data is grouped into various primary datasets based on which triggers were used for event selection. We restrict our analysis to the Jet Primary Dataset. The triggers present in this dataset are listed in Table 2.1. As evident from the table, the dataset has single-jet, di-jet, quad-jet, and H_T triggers. However, our analysis uses only the single-jet triggers marked by *. Each trigger has an associate prescale factor, corresponding to the ratio of frequency of trigger criteria being met to the number of events that are actually recorded for that specific trigger. When events corresponding to a certain trigger are encountered very frequently, only a fraction of them are recorded and a large prescale value is assigned to that trigger. This is done

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

Table 2.1: Jet triggers provided in the Jet Primary Dataset, including the number of events for which the trigger was present and/or fired. Entries marked by * are used in this analysis (see Table 2.2). HNF stands for HcalNoiseFiltered.

to avoid overwhelming data acquisition. The prescale factor we obtain for each trigger is the product of the prescale factors from the underlying Level 1 Trigger (based on low-level objects) and the final High Level Trigger (HLT).

The single-jet triggers (see Table 2.1) are designed to fire whenever any jet in the event is above a given p_T threshold. Because our analyses are based only on the hardest jet, we need to make sure that the trigger corresponding to the hardest jet ("trigger jet") is fired. Moreover, we also need to check that this trigger is nearly 100% efficient for jets of the given p_T. This necessitates the determination of p_T boundaries above which a given trigger is (nearly) 100% efficient. Fig. 2-3a shows the p_T spectrum of the hardest jet for the six triggers used in our analysis. We impose a "loose" jet quality cut after rescaling with appropriate Jet Energy Corrections (JEC) factors on these "trigger jets" (see Table A.1 and Fig. A-1a). We also require these jets to pass a



(a) (b)

Figure 2-2: Transverse momentum spectrum of raw PFCs, for (a) neutral candidates and (b) charged candidates. These histograms are populated only with PFCs from the hardest jet in the stated jet p_T range, comparing the CMS Open Data to parton shower generators. The cuts used in our jet substructure studies are $p_T^{\text{min}} = 1.0$ GeV, applied to both neutral and charged PFCs. For this and all remaining plots in this paper, one must keep in mind that the detector-level CMS Open Data and the particle-level parton showers are not directly comparable.

pseudorapidity cut of $|\eta| < 2.4$ to ensure that jets are reconstructed in the central part of the CMS detector where tracking information is available. We see good overlap of the p_T spectra as desired, except for the `Jet140U` trigger which is systematically low. The reason is that the `Jet140U` trigger was not present for the entirety of Run 2010B, so we revert to the `Jet100U` trigger when needed.

Using `HLT_Jet15U_HcalNoiseFiltered` as the baseline, the trigger efficiencies of the five remaining triggers are shown in Fig. 2-3b. Because we want to work with triggers that are nearly 100% efficient beyond the appropriate p_T values, we define trigger boundaries based on Fig. 2-3b. These boundaries are presented in Table 2.2, where the $p_T > 250$ GeV bin uses either `Jet100U` or `Jet140U` depending on whether the latter is present. Because each trigger selects a homogeneous event sample, we can use the average prescale value for the assigned trigger when filling histograms, which is statistically preferable to using the individual event prescale values. We show the distribution of all prescale values in Fig. A-2 in App. A.

Table 2.3 is a summary of our event selection summary. We start with 20 million

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

Table 2.2: Assigned triggers for the hardest jet in a given p_T range, along with the average prescale value that determines subsequent histogram weights. Since the Jet140U trigger was not present for all of Run 2010B, we use Jet100U when needed for the highest p_T bin.

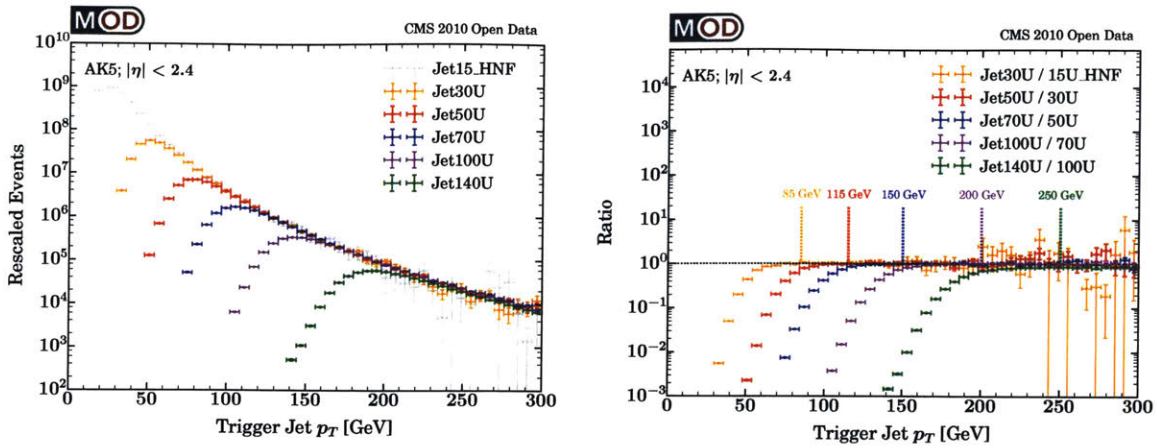
	Events	Fraction
Jet Primary Dataset	20,022,826	1.000
Validated Run	16,341,187	0.816
Assigned Trigger Fired (Table 2.2)	894,366	0.045
Loose Jet Quality (Table A.1)	843,129	0.042
AK5 Match	843,128	0.042
$\|\eta\| < 2.4$	768,687	0.038
Passes Soft Drop ($z_g > z_{\text{cut}}$)	760,055	0.038

Table 2.3: Overall workflow to go from the events in the Jet Primary Dataset to the events used in our jet substructure analysis.

events in the Jet Primary Dataset. We reduce this to about 82% by discarding events that are absent from the CMS-provided list of validated runs. Then, we restrict ourselves to events whose assigned trigger was fired. This drops the number of events to around 900 thousand. This defines the skimmed dataset. Next, imposing the loose jet quality criteria (see Sec. A.1) removes a small number of events, as does verifying that the AK5 jet provided by CMS matches those clustered by FastJet on the PFCs directly (see Secs. 2.3 and 2.4). An event is used for substructure analyses (see Sec. 2.4) only if its passes $|\eta| < 2.4$.

2.3 The MIT Open Data (MOD) Format

The MIT Open Data (MOD) format is a data-format we designed to hold a subset of the AOD data. We designed MOD format with the goal of having a light-weight, easy



(a) (b)

Figure 2-3: (a) Hardest jet p_T spectrum in the CMS Open Data from the six triggers used in this analysis (see Table 2.1). (b) Ratios of the jet p_T spectra from adjacent triggers used to determine when the triggers are nearly 100% efficient, which determine the jet trigger boundaries in Table 2.2.

to parse, human-readable (hence, text-based) data format for storing particle physics data. It uses space-separated entries with keyword labels. In addition to experimental data, we use MOD to record data generated from parton showers as well (see Sec. 3).

The representation of an event under the MOD format has the following keywords:

1. `BeginEvent`: This marks the beginning of an event. It also includes the data source (CMS Open Data or a certain parton shower generator) and the version number of the data format, currently at version 5.

2. `Cond`: This includes identification information viz. the run and event numbers. Additionally, it contains the timestamp of when the data was taken, number of primary vertices and information about the luminosity block. Luminosity block is later used in luminosity calculation.

3. `Trig`: This section contains a list of triggers present in the Jet Primary Dataset, along with their information like the associated prescale factors and whether or not the trigger fired.

4. `AK5`: List of CMS-clustered anti-k_T jets with $R = 0.5$. This includes the jets'

Code	Candidate	Total Count	$p_T > 1$ GeV
11	electron (e^-)	32,917	32,900
−11	positron (e^+)	32,984	32,968
13	muon (μ^-)	12,941	12,653
−13	antimuon (μ^+)	13,437	13,110
211	positive hadron (π^+)	6,908,914	5,183,048
−211	negative hadron (π^-)	6,729,328	5,027,146
22	photon (γ)	9,436,530	4,805,173
130	neutral hadron (K_L^0)	2,214,385	1,658,892

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

four-momenta, jet energy correction factor (calibration information), jet area, and information about jet quality factors.

5. `PFC`: List of PFCandidates. This includes their four-momenta and particle identification codes (based on pdgId [11]).

6. `EndEvent`: This marks the end of an event.

An example event in the MOD format has been included in Appendix B. For storing data generated by parton shower generators, we replace `Cond` and `Trig` by information about event weights, and rename `PFC` to `Part` as PFC is a CMS-specific construct.

Table 2.4 contains a list of all the particle identification codes we found in the Jet Primary Dataset. These correspond to the PFCandidates of the hardest jet that pass the criteria $p_T > 85$ GeV, $|\eta| < 2.4$). These identification codes are based on the Monte Carlo particle number scheme, pdgId [11]. CMS assigns ± 211 to all charged hadrons, which belongs to charged pions. CMS chose to assign all charged hadrons the identification code of charged pions instead of charged kaons as the former is more prevalent. Similarly, because neutral pions decay as $\pi^0 \to \gamma\gamma$, they are reconstructed as photons and so assigned code 22. This is particularly interesting because we had to make a conscious choice about whether or not to allow our neutral pions in parton

showers to decay into photons. We experimented with this with the hope that it would improve agreement between the Open Data and parton showers but it made little difference, if any.

As noted earlier, in addition to performing our own clustering, we also obtain CMS-clustered jets. This is necessary to extract JEC factors (jet calibration information) and impose jet quality cuts. The parameters for various level of jet quality cuts are shown in Table A.1 in Sec. A. Throughout our analysis, we impose the recommended "loose" jet quality cut.

The conversion from AOD to MOD files, after running `gzip` compression reduces the file size by roughly 10 times. Further, when we restrict ourselves to a skimmed dataset, corresponding to only those jets with hardest $p_T > 85$ GeV and whose assigned trigger fired, we reduce the gzip compressed 198.8 gigabytes MOD files to 11.6 gigabytes.

2.4 Analysis Tools

After getting the data in our own text-based format, we were no longer restricted to writing analysis software based on CMSSW. We decided to write our own analysis framework based on `FastJet` [10]. This allowed us the luxury of not having to implement many jet substructure tools from scratch. Our analysis framework is written C++ and is called `MODAnalyzer`. It is also available as a `GitHub` repository [12]. For the soft drop study in our analysis, we used the `RecursiveTools` package from FastJet contrib 1.019 [13].

We based the structure of MODAnalyzer on the structure of the MOD files. MODAnalyzer's core class `Event` holds all the event information by parsing the MOD files in addition to selecting the assigned trigger for the hardest jet. The AK5 and PFC information in the MOD events are stored as FastJet `PseudoJet` objects. Because both of these objects contain more information than what `PseudoJet` can store by default, we defined `InfoCalibratedJet` and `InfoPFC` classes that inherit from FastJet's `UserInfoBase` which extends the `PseudoJet` objects. The `Cond` and

`Trig` information are stored in `Condition` and `Trigger` classes respectively.

Because we store two kinds of jets: CMS-clustered AK5 jets, and jets clustered internally from the PFCs, it is important to be careful about which jet's information we use for what purpose. To define the hardest jet in the event, which defines the trigger jet, we use the CMS-clustered jets rescaled by the corresponding JEC factor. We then select the "assigned trigger" based on the p_T of this jet and keep the event only if that trigger fired. We also discard any event whose trigger jet fails to pass the loose jet quality cut. Because JEC factors, triggers, and jet quality factors are defined on CMS jets, it is important to make sure we are not using jets we cluster ourselves. However, to perform substructure studies, we need the underlying jet structure. So we find the internal PFC jet that is closest to the trigger jet in the rapidity-azimuth plane. If this jet matches the number of constituents in the CMS jet and if the four-momenta of the two jets agree up to a 1 MeV precision after rescaling the internal jet with the JEC of the corresponding CMS jet, we regard it as a "trigger-matched-jet" and use it for subsequent analyses. If this match fails, we discard the event although this only affects 1 event out of the 843,129 events in our analysis (see Table 2.3).

A large number of events in the dataset are not suitable for our analysis. This could be for multiple reasons: the hardest jet's p_T might be less than the 85 GeV minimum threshold set in Table 2.2, the assigned trigger might not have fired, or the trigger jet might not pass the loose jet quality cut we impose. We remove these events in a step we call event skimming where we read in each MOD file and write out another MOD file with only the events where the trigger jet has fired.[2] Additionally, because our substructure study is only based on the hardest jet, we further filter the events by outputting MOD files with a `Hardest_Jet_Selection` header, where we store only the PFC candidates of the hardest jet and minimal `Trig`, `Cond`, and `AK5` information which we consolidate under the `1JET` keyword. After this step, gzip compressed `Hardest_Jet_Selection` MOD files take only 725 megabytes of disk space.

[2]Exceptions to this are trigger and luminosity studies where we use the unskimmed, full dataset.

Chapter 3

Parton Shower Generators

A standard way to validate results from particle collider experiments is to compare the results to results obtained from parton shower generators. For our analysis, we compare the results from the CMS Open Data with results from three different parton shower generators: PYTHIA 8.219 [14], HERWIG 7.0.3 [15], and SHERPA 2.2.1 [16]. It is helpful to compare results to multiple generators because they work in different ways in terms of how they generate parton showers, primarily in terms of their choice of the evolution variable.

All parton shower samples were generated with the default di-jet production settings as the single-jet trigger processes are predominantly di-jet productions. We use a p_T based weighing scheme to maximize the phase space utilization. Similar to the Open Data analyses, we restrict basic jet observable studies to $p_T > 85$ GeV and jet substructure analyses to p_T GeV.

These generators typically produce their output in the `HEPMC` data format [17]. However, to ensure consistent analysis pipeline across all data sources, we convert these `HEPMC` files to MOD files, with certain modifications outlined in Sec. 2.3. While the outputted MOD data files certainly look different for data coming from CMS Open Data versus for data coming from these parton shower generators, after skimming and applying the Hardest Jet Selection (as discussed in Sec. 2.4), the end-point MOD files are identical. This enables us to use the same analysis code, which helps minimize differences in the results coming outside of the underlying physics and the differences

in the inner workings of the generators.

Because the results obtained directly from these parton shower generators are truth-level, comparisons to experimental data need to be done carefully. One way to enable a more direct comparison is to use a fast detector simulation software like DELPHES. We did attempt to use DELPHES 3.3.2 [18] but upon using the default CMS-like detector configuration, we noticed that the distributions were over-smeared. This can be attributed to the fact that the default CMS-like configuration is meant for basic jet studies and not jet substructure studies. And because no any simulated parton shower datasets were provided with the 2010 CMS Open Data release[1], we had to limit our study to truth-level comparisons.

Even in the absence of the use of a detector simulation, in an attempt to make the comparisons fairer, we tried to account for the finite energy resolution of the CMS detector by imposing a constraint of $p_T > 1.0$ GeV on the PFCandidates and truth-level particles for Open Data and Monte Carlo respectively. This cut is placed only for substructure studies and not for basic jet observables. This cut is motivated by Fig. 2-2. As is evident from the disagreement between data and Monte Carlo for $p_T < 1.0$ GeV, the PFCs below this cutoff are affected by detector inefficiencies. This strategy is similar in spirit to the SOFTKILLER approach to pileup mitigation [20].

A notable aspect of Fig. 2-2 is that the agreement for charged PFCs is better than the agreement for neutral PFCs. This can be explained by the fact that, for charged particles the PFCandidates utilize information from the tracker as well as the calorimeter thereby giving a higher angular resolution than for neutral particles. This is also evident in the jet substructure plots in Sec. 4.2 where there is a significant improvement in agreement for track-based variants of the observables.

[1]The 2011 CMS Open Data [19] release does include simulated parton shower datasets, allowing for a detector-level comparison.

Chapter 4

Hardest Jet Properties

In this section, we present various basic kinematic and substructure observables of the hardest jets. For all observables, we compare our results from the CMS Open Data with results from the parton shower generators. The jets have been rescaled with corresponding Jet Energy Correction (JEC) factors before calculating the appropriate observable. Except for the pseudorapidity distribution in Fig. 4-3b, we impose a $|\eta| < 2.4$ cut. We impose the restriction $p_T > 85$ GeV for basic jet kinematics and a higher cut of $p_T > 150$ GeV for substructure observables. This is done to avoid skewing the results because of the large prescales coming from the 15U and 30U triggers (see Table 2.2).

4.1 Jet Kinematics

Fig. 4-2a shows the p_T spectrum of the hardest jet. The lower threshold of 85 GeV, as mentioned earlier, is set by the lower bound of the lowest trigger `HLT_Jet_30U` as can be seen in Table 2.2. There is excellent agreement between Open Data and the parton shower generators. Fig. 4-2b shows the effect of jet energy corrections. It is clear that the effect of JEC becomes increasingly important for harder jets and without this calibration information, the agreement with parton shower generators would not have been as strong.

The azimuthal distribution can be seen in Fig. 4-3a and as expected, it is flat.

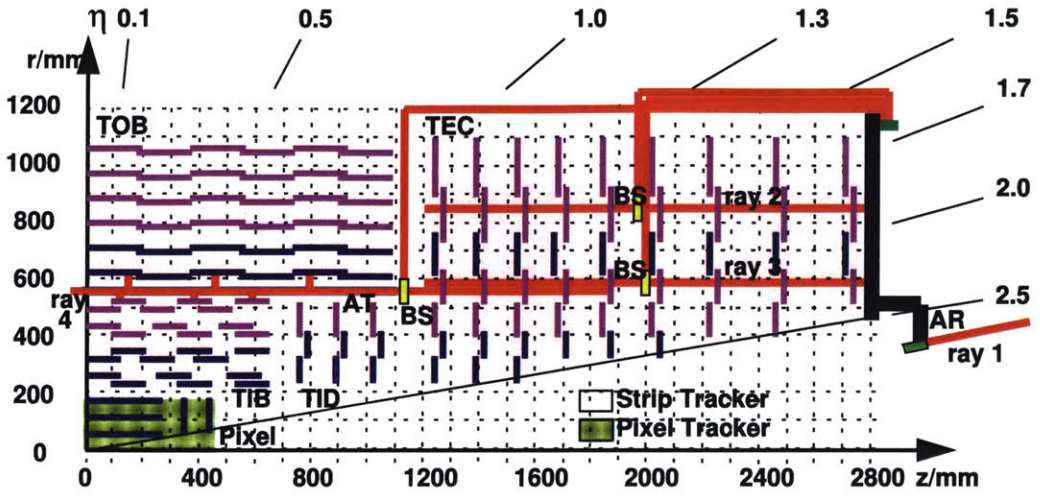
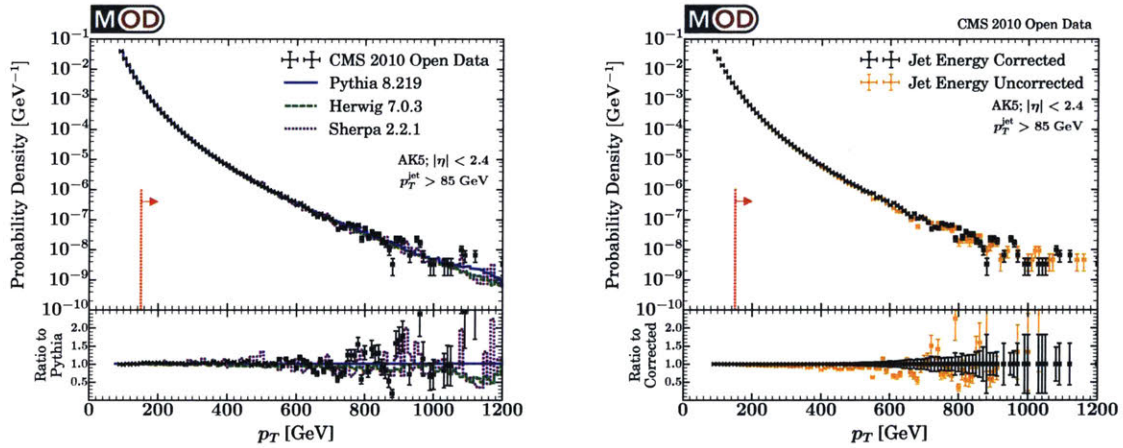
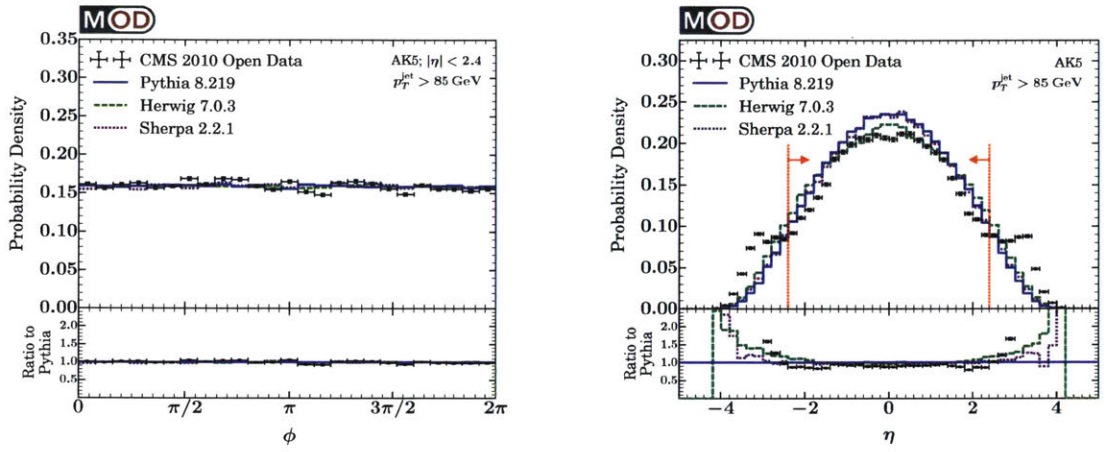


Figure 4-1: An illustration of the CMS silicon tracker. Notice that the tracker ends at $\eta = 2.5$ and so, beyond that region, no tracking information is available [3].



(a) (b)

Figure 4-2: (a) Hardest jet p_T spectrum, comparing the CMS Open Data with PYTHIA 8.219, HERWIG 7.0.3, and SHERPA 2.2.1. The maximum jet p_T in the Jet Primary Dataset is 1277 GeV. (b) Hardest jet p_T before and after applying the appropriate JEC factors. Because these are normalized histograms with the same $p_T > 85$ GeV cut, the mismatch in JEC values is only apparent at high p_T.



(a) (b)

Figure 4-3: (a) Azimuthal angle of the hardest jet, which is flat as desired. (b) Pseudorapidity spectrum for the hardest jet. Note the population of anomalous jets at $|\eta| > 2.4$, coming from the edge of tracking acceptance, which is why we enforce $|\eta| < 2.4$ in our analysis.

Again, the agreement with parton shower generators is excellent. Fig. 4-3b shows the jet pseudorapidity distribution. The regions $|\eta| < 2.4$ have been clearly marked because we impose this constraint for all of our analyses. It is notable that for the Open Data, the bins in the $|\eta| > 2.4$ region have an excess compared to the parton shower generators. As can be seen in Fig. 4-1, the CMS tracker extends only to $\eta = \pm 2.5$ and so, beyond this region, tracking information is not available so the excess jets most likely represent jets that did not pass the jet quality criteria.

4.2 Jet Substructure Observables

The most ubiquitous jet substructure observables are jet multiplicity and jet mass. Jet multiplicity, however, is very sensitive to CMS's particle flow reconstruction and because our parton shower results are truth-level particles, we must be careful while comparing the two. Because it is difficult to reconstruct very soft particles, we avoid counting them by imposing a $p_T > 1.0$ GeV cut on the PFCandidates. Moreover, note that real-world detectors cannot resolve arbitrary angles and so particles separated by very small angles are likely to be merged together by CMS's particle flow algorithm.

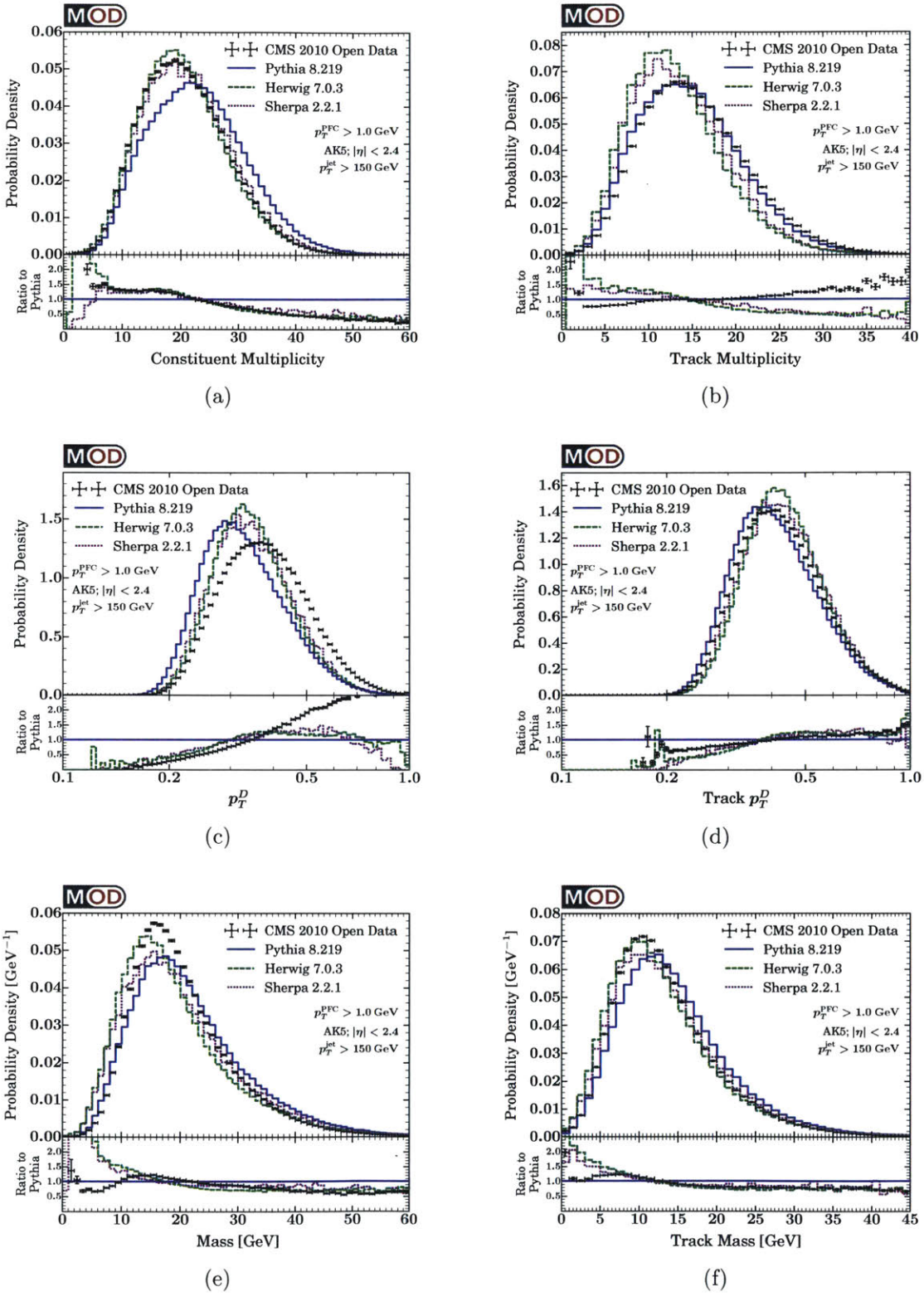


Figure 4-4: Basic substructure observables for the hardest jet. We emphasize that in this and all subsequent figures, the distributions are not directly comparable, since the CMS Open Data has not been unfolded to account for detector effects and the parton shower generators have not been folded with detector effects.

In Fig. 4-4a, we see that there is a good agreement between data, HERWIG 7.0.3, and SHERPA 2.2.1but PYTHIA 8.219differs considerably. However, this agreement is flipped for track multiplicity where there is close agreement with PYTHIA 8.219but not with HERWIG 7.0.3and SHERPA 2.2.1. It is very difficult to comment on the possible reasons for this, especially without detector simulations, particularly because constituent multiplicity is both infrared and collinear (IRC) unsafe and so very sensitive to angular and energy resolution.

There is a much better agreement in Fig. 4-4e between data and Monte Carlo for the jet mass spectrum. This improvement in agreement can be attributed to the jet mass being IRC-safe. Again, as expected, the agreement in the track-variant seems to be slightly better.

Next, we consider the observable p_T^D, defined as:

$$p_T^D = \frac{\sqrt{\sum_{i \in \text{jet}} p_{T_i}^2}}{\sum_{i \in \text{jet}} p_{T_i}} \tag{4.1}$$

p_T^D is infrared safe but collinear unsafe. This is a particularly useful observable for quark / gluon discrimination studies [21]. It can be seen in Fig. 4-4c that, p_T^D is systematically higher for the CMS Open Data than for the parton showers. This noticeably large difference, similar to for constituent multiplicity and jet mass, is reduced for the track version.

4.3 Jet Angularities

Jet angularities are a broad class of observables that are both infrared and collinear safe. This makes them excellent observables, especially to study the radiation pattern of quark and gluons as they are not very sensitive to detector resolution [22, 23, 24, 25, 26]. We define jet angularities as:

$$e^{(\alpha)} = \sum_{i \in \text{ jet}} z_i \theta_i^\alpha \ , \tag{4.2}$$

with

$$z_i = \frac{p_{T_i}}{\sum_{j \in \text{jet}} p_{T_j}}, \quad \theta_i = \frac{R_i}{R} \,. \tag{4.3}$$

R_i is the distance in the rapdity-azimuth plane to a recoil-free axis [27, 28, 29, 30, 25]. For getting a recoil-free axis, we use the winner-take-all-axis [25, 31, 32] defined from Cambridge/Aachen clustering [33, 34]. Note that, $e^{(\alpha)}$ is IRC-safe only for $\alpha > 0$.

Setting $\alpha < 1$ allows us to test the radiation patterns in the core while setting $\alpha > 1$ allows use to test it in the periphery of the jet. In our analysis, we focus on the following most-commonly used values: $\alpha = 1/2$ (Les Houches Angularity (LHA)) [35, 36], $\alpha = 1$ (Jet Width) [27, 37, 38], and $\alpha = 2$ (Jet Thrust) [39]. The corresponding distributions are shown in Fig. 4-5. Even though these are IRC-safe observables and therefore not very sensitive to detector resolution, we still place the $p_T > 1.0$ GeV cut on the PFCandidates for consistency.

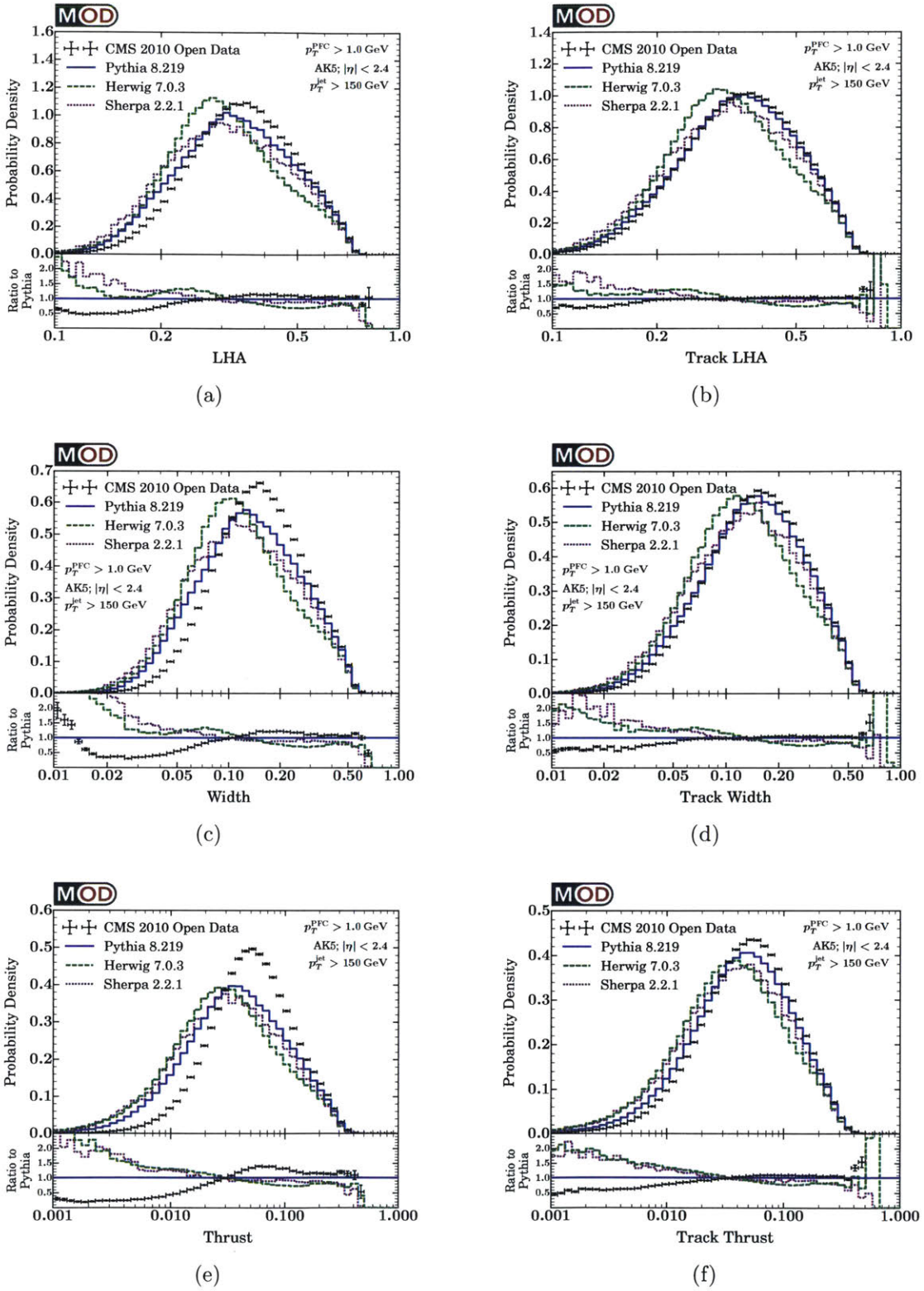


Figure 4-5: Same as Fig. 4-4 but for the IRC-safe recoil-free jet angularities: (top row) LHA with $\alpha = 1/2$, (middle row) jet width with $\alpha = 1$, and (bottom row) jet thrust with $\alpha = 2$. Once again we compare (left column) all particle distributions to (right column) track-only variants.

Chapter 5

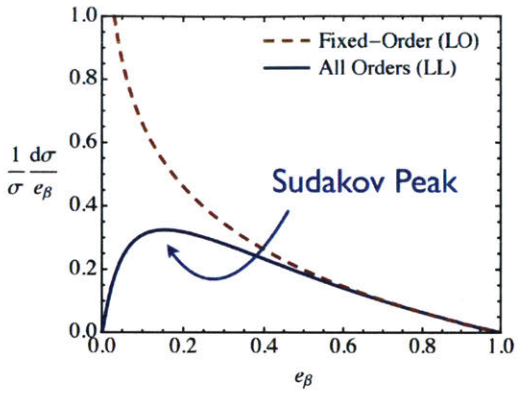
Two-Prong Substructure of Jets

We measured the jet kinematics and substructure observables as a way to validate our workflow. Having done that, we now test the 2-prong substructure of jets using the soft drop algorithm [40]. Although the full power of soft drop goes much beyond just this, soft drop removes soft contaminations from a jet, making it a jet grooming algorithm. This makes the observables coming out of soft drop robust to pileup and detector effects. Additionally, soft drop has a corresponding first-principle QCD calculation, making it possible to directly compare results from calculations with measurements from data and parton showers.

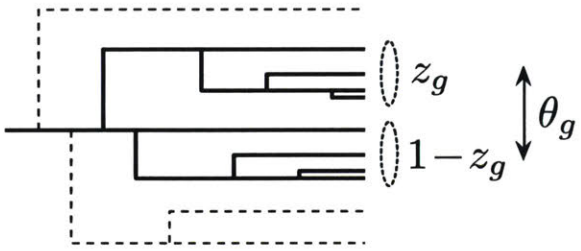
5.1 Introduction to Soft Drop

The soft drop algorithm goes as follows:

1. Take the hardest anti-k_T jet (for our analysis, with $R = 0.5$).
2. Decluster the hardest jet and recluster the constituents with the Cambridge-Aachen (C/A) algorithm. [33, 34]. This produces an angular-ordered clustering tree.
3. Systematically decluster the tree from the top of the tree. At each step, remove the softer branch until we find a $1 \rightarrow 2$ branch that passes the softdrop condition, $z > z_{\text{cut}}\theta^{\beta}$.



(a)



(b)

Figure 5-1: (a) Sudakov peak suppressing the isolated singularity when calculation is carried out to all orders of the coupling. (b) Schematic of the soft drop algorithm, which recursively removes branches from the C/A clustering tree if the momentum fraction z fails to satisfy $z > z_{\text{cut}}\theta^\beta$.

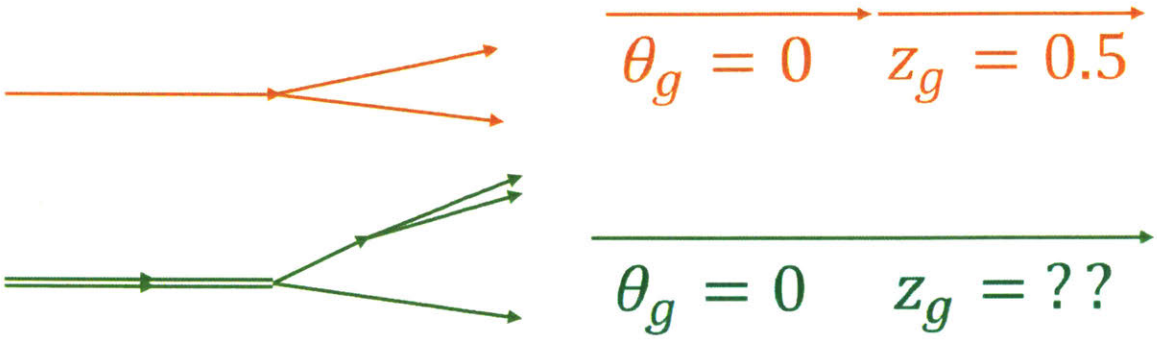
In the soft drop condition defined above, z_{cut} is an energy fraction cut, β is an adjustable angular exponent, and

$$z = \frac{\min\left[p_{T_1}, p_{T_2}\right]}{p_{T_1} + p_{T_2}}, \quad \theta = \frac{R_{12}}{R}. \tag{5.1}$$

When we have found a branch that satisfies the soft drop condition, we denote the momentum fraction by z_g and the opening angle by θ_g. These two kinematic observables characterize the hard 2-prong substructure of the jet. The g-subscript is to indicate that these are groomed observables.

Soft drop performs three different tasks simultaneously: it removes soft contaminations from the jets, which like mentioned earlier, helps avoid jet contamination from pileup, Initial State Radiation (ISR), and the Underlying Event (UE); it dynamically reduces the effective jet radius to match the radius of the hardest jet core; and, it provides the 2-prong kinematic observables z_g and θ_g, which can be used not only for fundamental QCD tests [41, 42, 43, 44, 45], but also to discriminate between quark/gluon jets and boosted W/Z/Higgs jets [46, 47], thereby making soft drop a great tool for new physics searches.

For our analyses, we restrict ourselves to $z_{\text{cut}} = 0.1$, and $\beta = 0$. Notice that when $\beta = 0$, the soft drop condition reduces to $z > z_{\text{cut}}$. In addition to z_g and θ_g, we present



(a) (b)

Figure 5-2: (a) Example jet where z_g is collinear safe. (b) Demonstration of a jet structure for which z_g is not defined. We need to make sure z_g is always defined, necessitating some restriction on θ_g to stop us from ever getting into such a territory where z_g is not well-defined.

results for the observable $e_g^{(\alpha)}$ with $\alpha \in \{1/2, 1, 2\}$. So the following five variables are of particular interest to us:

$$z_g, \quad \theta_g, \quad e_g^{(1/2)}, \quad e_g^{(1)}, \quad e_g^{(2)}, \tag{5.2}$$

$e_g^{(\alpha)}$ is a single-emission groomed variant of the angularities mentioned in Sec. 4.3, with the corresponding values for α.

5.2 Theory Calculations

z_g is an infrared safe but collinear unsafe observable. Consider the schematic in Fig. 5-2a. Soft drop gives the same result for both jets and so, is not dependent on collinear splitting. For this particular example jet, z_g is collinear safe.

Consider now the schematic in Fig. 5-2b. z_g is clearly not defined for the "green" jet. So z_g is not collinear safe for this jet, for example. One way of getting around this is to measure some collinear safe variable and relate it back to z_g. The most natural

choice for this is the opening angle between the sub-jets, θ_g, which is in fact collinear safe. However, we need to be careful about using θ_g here.

While we can certainly use θ_g, z_g is not defined when $\theta_g = 0$ so we need to make sure that θ_g never reaches 0. But this is not a problem, as θ_g never reaches zero anyway. This allows use to indirectly calculate $p(z_g)$ using $p(z_g \mid \theta_g > \theta_{\text{cut}})$,

$$p(z_g) = \int p(z_g, \theta_g)\, \mathrm{d}\theta_g = \int p(\theta_g)\, p(z_g \mid \theta_g)\, \mathrm{d}\theta_g \tag{5.3}$$

$p(\theta_g)$ is safe and $p(z_g \mid \theta_g)$ is calculable except for $\theta_g \neq 0$. This makes $p(z_g)$ Sudakov-safe, where the infinity is regulated by the Sudakov-peak when $p(z_g)$ is expanded to all orders of α_s. Fig. 5-1a shows the Sudakov-peak in effect, where, to Leading Order (LO), the cross-section blows up as $e_\beta \to 0$ but when we calculate to Leading Log (LL), the Sudakov-peak suppresses the isolated singularity, producing a well-defined cross-section that decays with e_β.

We can use the same process to calculate $e_g^{(\alpha)}$. Just like Eq. (5.3), we can write:

$$p(e_g^{(\alpha)}) = \int p(e_g^{(\alpha)}, \theta_g)\, \mathrm{d}\theta_g = \int p(\theta_g)\, p(e_g^{(\alpha)} \mid \theta_g)\, \mathrm{d}\theta_g \quad , \tag{5.4}$$

which allows use to write the probability distribution for $e_g^{(\alpha)}$,

$$p(e_g^{(\alpha)}, \theta_g) \equiv \frac{1}{\sigma} \frac{\mathrm{d}^2\sigma}{\mathrm{d}e_g^{(\alpha)}\, \mathrm{d}\theta_g} \,. \tag{5.5}$$

Notice that $z_g = e_g^{(\alpha=0)}$, so Eq. (5.3) is same as Eq. (5.4) with $\alpha = 0$.

We obtain the uncertainties in the probability distribution by varying the scale, and the quark/gluon composition and from uncertainties in the running coupling [40, 48].

There are two known effects which have not been accounted for in our estimation of uncertainties. The first one, is non-perturbative corrections. Because our calculations are carried out perturbatively, when non-perturbative physics dominates z_g and θ_g, those effects need to be taken into account. For double-differential distributions, this

occurs when

$$z_g \theta_g \lesssim \frac{\Lambda}{p_T R}, \tag{5.6}$$

where $\Lambda \sim \mathcal{O}(\text{GeV})$ and p_T is the lowest value in the plotted range. For our analyses, we use $\Lambda = 2$ GeV. Projecting to the single observables, non-perturbative dynamics becomes relevant when

$$\theta_g \lesssim \frac{\Lambda}{z_{\text{cut}} p_T R}, \quad e_g^{(\alpha)} \lesssim \max\{1, z_{\text{cut}}^{1-\alpha}\} \left(\frac{\Lambda}{p_T R}\right)^{\alpha}. \tag{5.7}$$

We indicate the region where non-perturbative effects dominate, we change the theory curves to dashed style. In Fig. 5-3 and Fig. 5-4, we mark the region with white hashing.

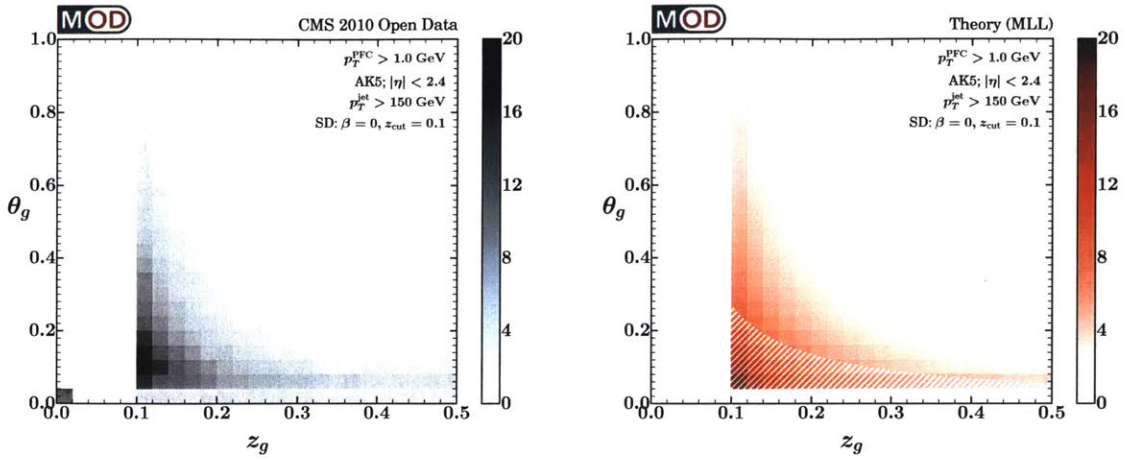
The second effect that we have ignored is matching to fixed-order matrix elements. This will, however, have only a small effect for a reasonably small jet radius and for the $e_g^{(\alpha)} \ll 1$ limit, both of which holds for our studies.

5.3 Measurements

Fig. 5-3 shows the two-dimensional distribution for $p(z_g, \theta_g)$ for the Open Data, analytic calculations, and the three parton shower generators. The soft and collinear singularities of QCD can be clearly seen as the peak at small values of z_g and θ_g. Also notice that, for all but analytical calculations, there is a non-zero bin for $z_g = \theta_g = 0$. This corresponds to jets with only constituent left after soft drop.

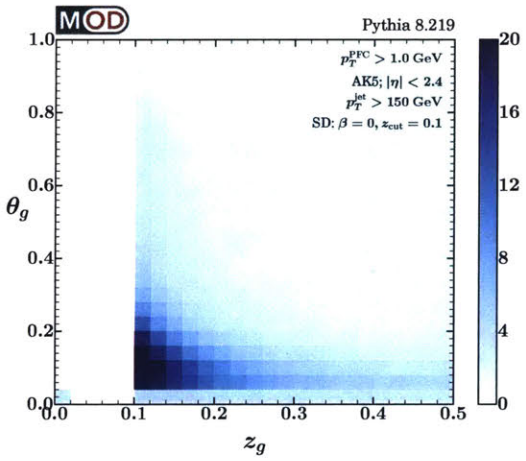
We also plot the same two-dimensional distribution on a log scale to better highlight the logarithmic nature of the soft / collinear singularity structure of QCD. The peak at around $\theta_g = 0.1$ is particularly interesting as it seems to be suppressed in the Open Data. This is very difficult to explain though as it lies in the non-perturbative regime as suggested by Eq. (5.6).

We next plot the single-variable distributions from Eq. (5.2). For each variable, we include a track-only variant and as has been seen throughout our analyses, the agreement is clearly better for track-based observable than for the regular variants.

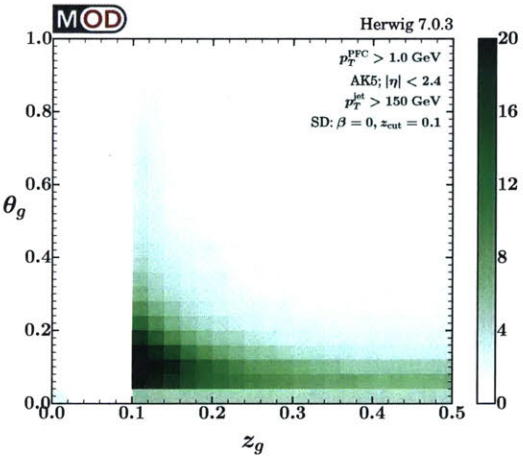


(a)

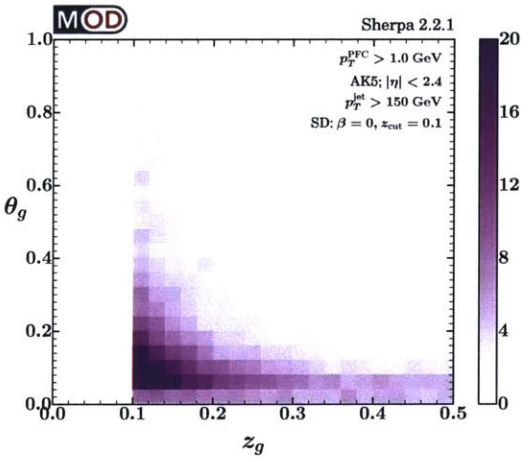
(b)



(c)

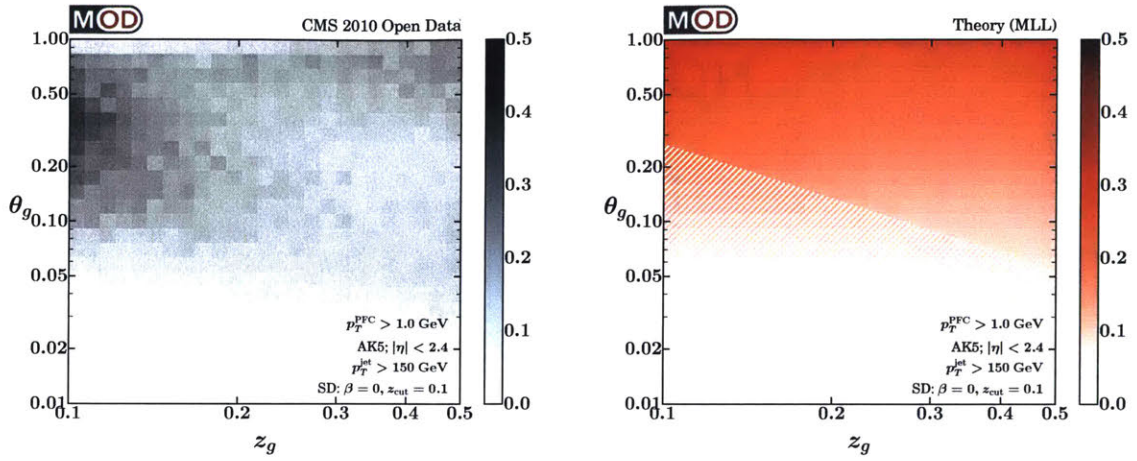


(d)



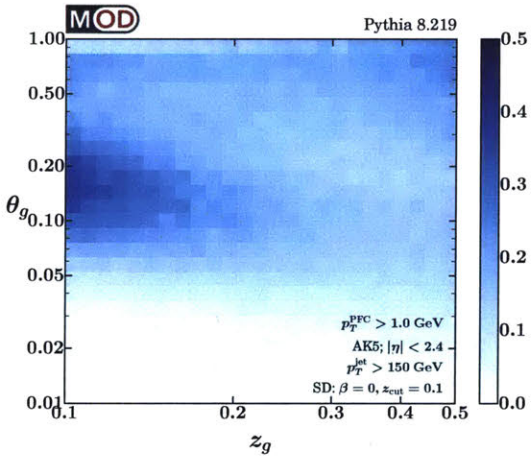
(e)

Figure 5-3: Two dimensional distributions of z_g versus θ_g. The hard vertical cut corresponds to $z_g = z_{\mathrm{cut}}$. The white hashing in the MLL distribution corresponds to where nonperturbative physics dominates.

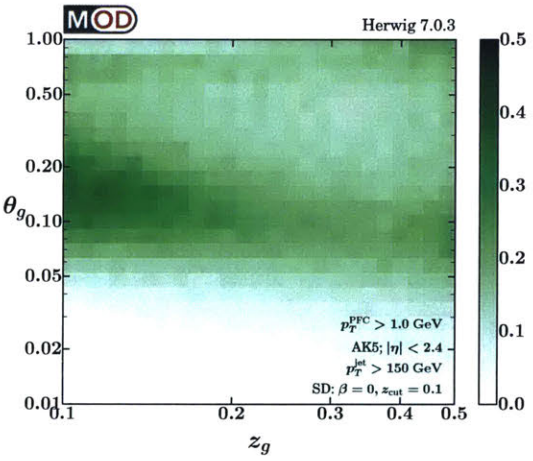


(a)

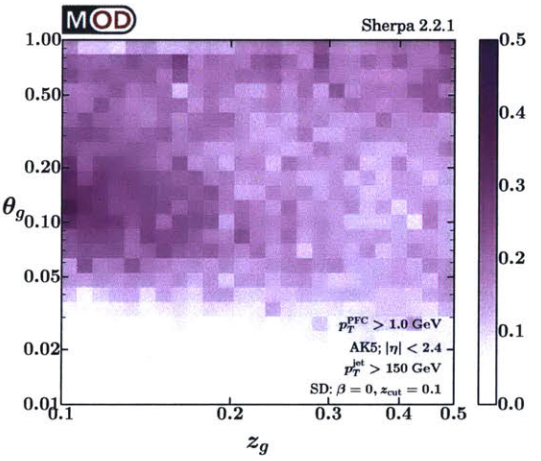
(b)



(c)



(d)



(e)

Figure 5-4: Same as Fig. 5-3 but on a logarithmic scale to highlight the soft/collinear limit.

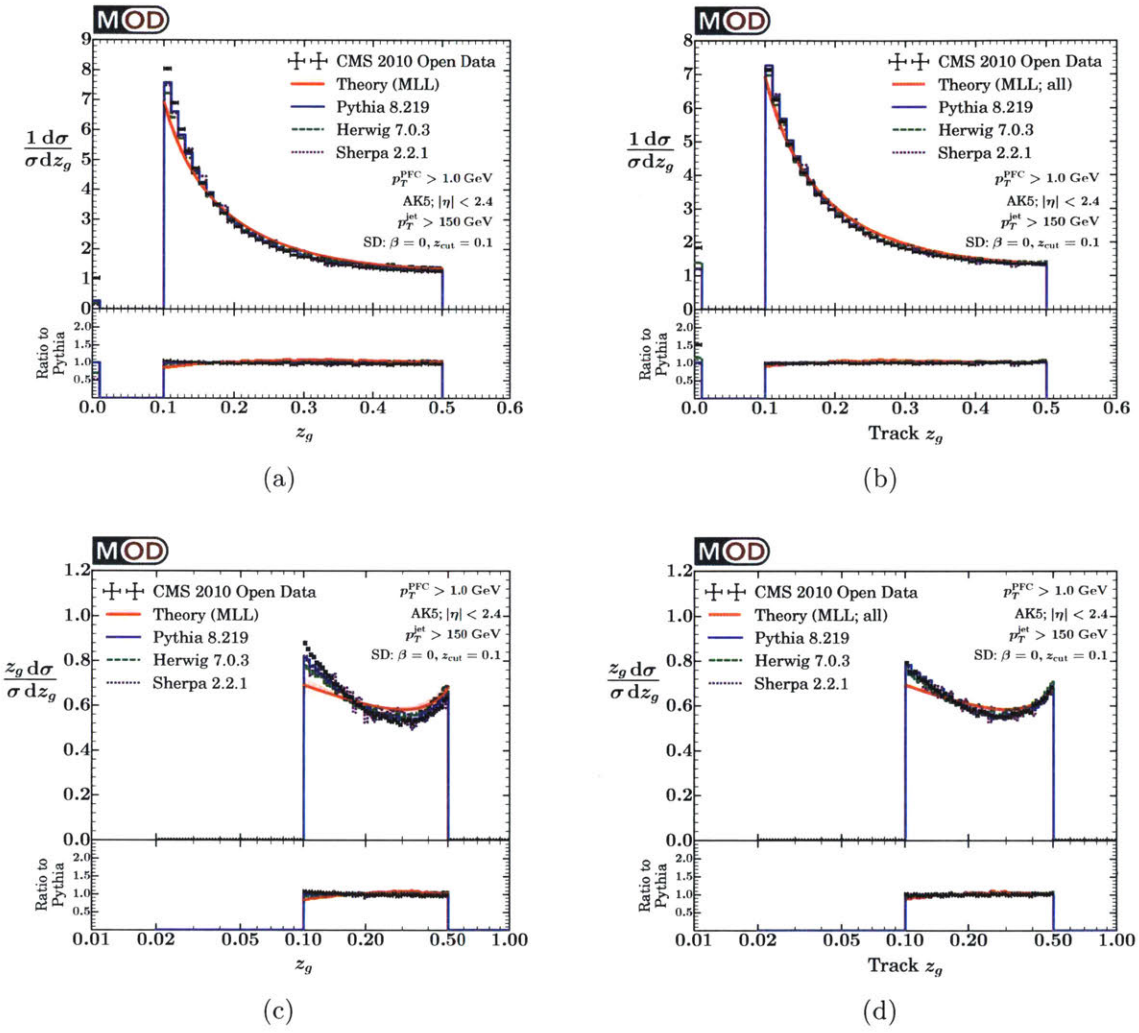


Figure 5-5: Soft-dropped distributions for z_g using (left column) all particles and (right column) only charged particles. In this and subsequent plots, the MLL distributions are the same in both columns and do not account for the $p_T^{\mathrm{min}} = 1$ GeV cut on PFCs or the switch to charged particles (hence the dashed version on the right). The top row shows the linear distributions while the bottom row shows the logarithmic distributions.

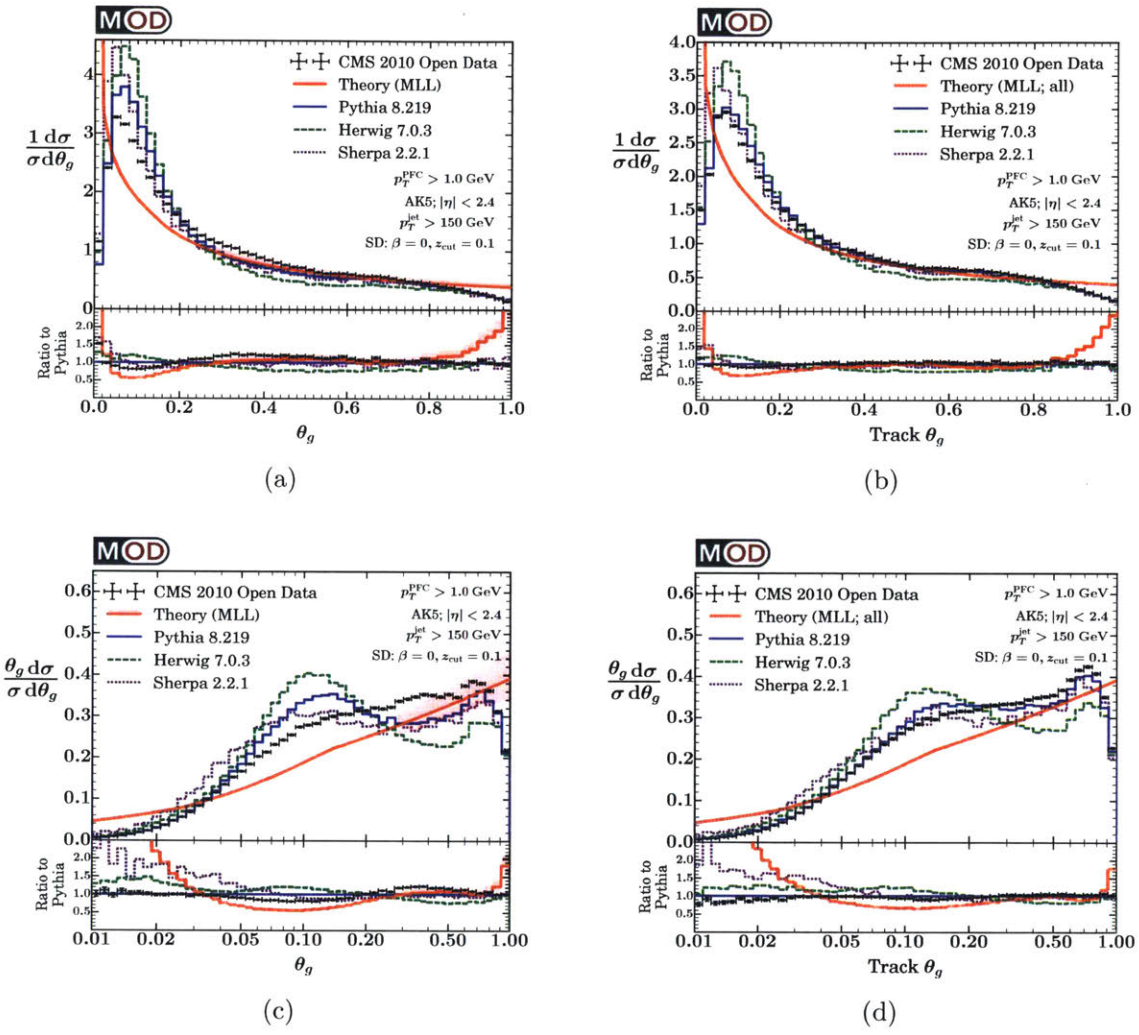


Figure 5-6: Same as Fig. 5-5 but for θ_g. For the MLL distributions, the region where nonperturbative dynamics matters is indicated by the use of dashing. We do not indicate the regime where fixed-order corrections matter, since we have no first-principles estimate for the transition point.

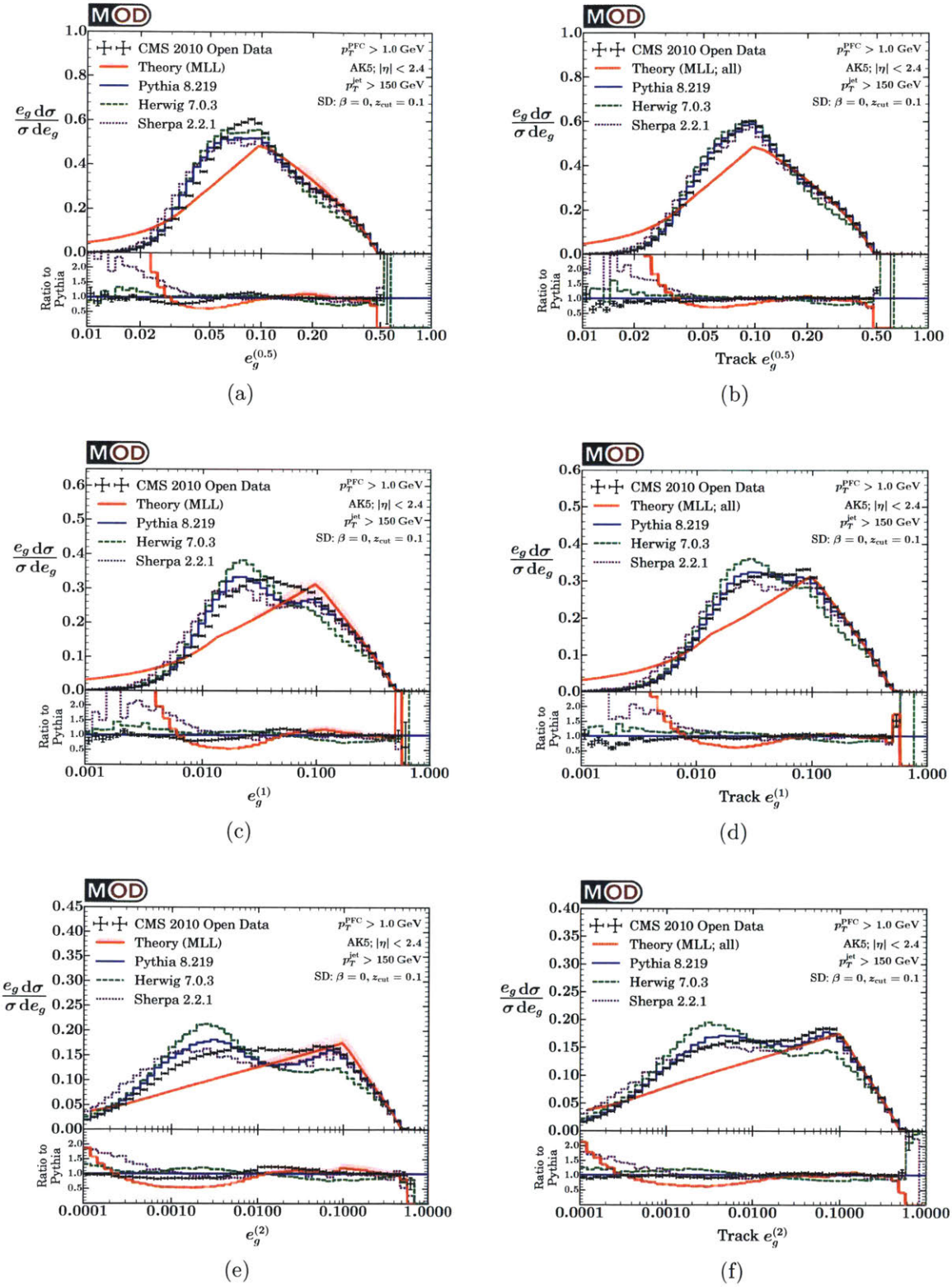


Figure 5-7: Logarithmic distributions for (top row) $e_g^{(1/2)} = z_g\sqrt{\theta_g}$, (middle row) $e_g^{(1)} = z_g\theta_g$, and (bottom row) $e_g^{(2)} = z_g\theta_g^2$, using (left column) all particles and (right column) only charged particles. Dashing indicates the region where non-peturbative physics dominates.

We begin with z_g. There are two notable features in Fig. 5-5a and Fig. 5-5c: one, notice that the log version is approximately flat, as predicted by the logarithmic structure of the singularity structure; two, the excess on the $z_g = 0$ bin in Open Data is noticeably larger than for parton showers. This can be attributed to CMS's limited angular resolution, particularly for neutral particles.

Next, the θ_g distributions in Fig. 5-6a and Fig. 5-6c are helpful to better understand the angular effects. θ_g represents the dynamically reduced opening angle of the 2-prongs and so, is representative of the angularity. As is evident from the plots, noticeable differences appear in the non-perturbative region. Also, notice the kink at $\theta_g \approx 0.1$ which also appeared as a clearly-visible feature for $p(z_g, \theta_g)$ in Fig. 5-4.

Moving on to the single-emission angularities $e_g^{(\alpha)}$, there is a decent agreement between CMS Open Data and the parton shower generators. Like in other distributions, this agreement improves when we switch to track-based observables. As expected, the kink in the analytical distribution lies at $e_g^{(\alpha)} = z_{\text{cut}} = 0.1$.

Chapter 6

Additional Soft-Dropped Observables

We now present observables we have already seen in Sec. 4.2, but obtained after soft drop declustering the jets. We compare these observables to before applying soft drop on them to highlight the effect the algorithm has on them.

First, let us consider the fractional p_T loss: the fraction of the original jet p_T discarded after soft drop. Fig. 6-1 shows excellent agreement between the CMS Open Data and parton showers.

Next, consider the basic jet substructure observables. Fig. 6-2 shows the same basic substructure observables as Fig. 4-4 (constituent multiplicity, p_T^D, and jet mass) but shows the effect soft drop has on the respective observables. We can see that there is a decent agreement between CMS Open Data and parton showers and the agreement, similar as for other distributions, gets better for track-based versions. Also notable is that soft drop does not seem to have much, if any, effect on the amount of agreement.

Fig. 6-3 shows a similar comparison for the jet angularities from Fig. 4-5. One particularly interesting aspect here is that the soft-dropped versions of the observables are approximately flat. This is illustrative to the fact that soft drop transforms the double logarithmic structure of the angularities to a single logarithmic structure (and since the distributions are log-scaled, it appears flat). This happens because, soft

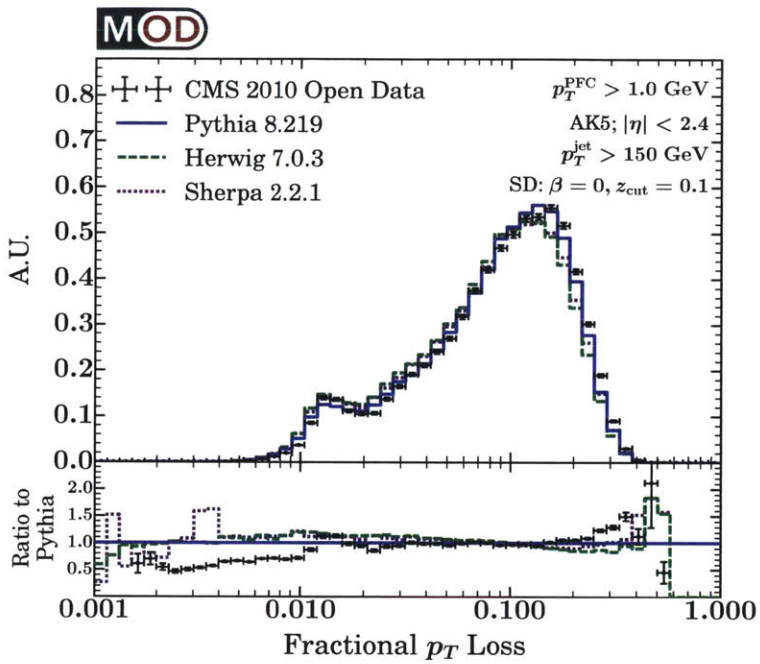


Figure 6-1: Fraction of the original jet p_T lost after performing soft drop declustering. Because this is a fraction, no JEC factors are applied.

drop, as outlined earlier, removes all soft-collinear radiation with $z < z_{\mathrm{cut}}$, thereby making the angularities exhibit single logarithmic structure [49, 50, 40].

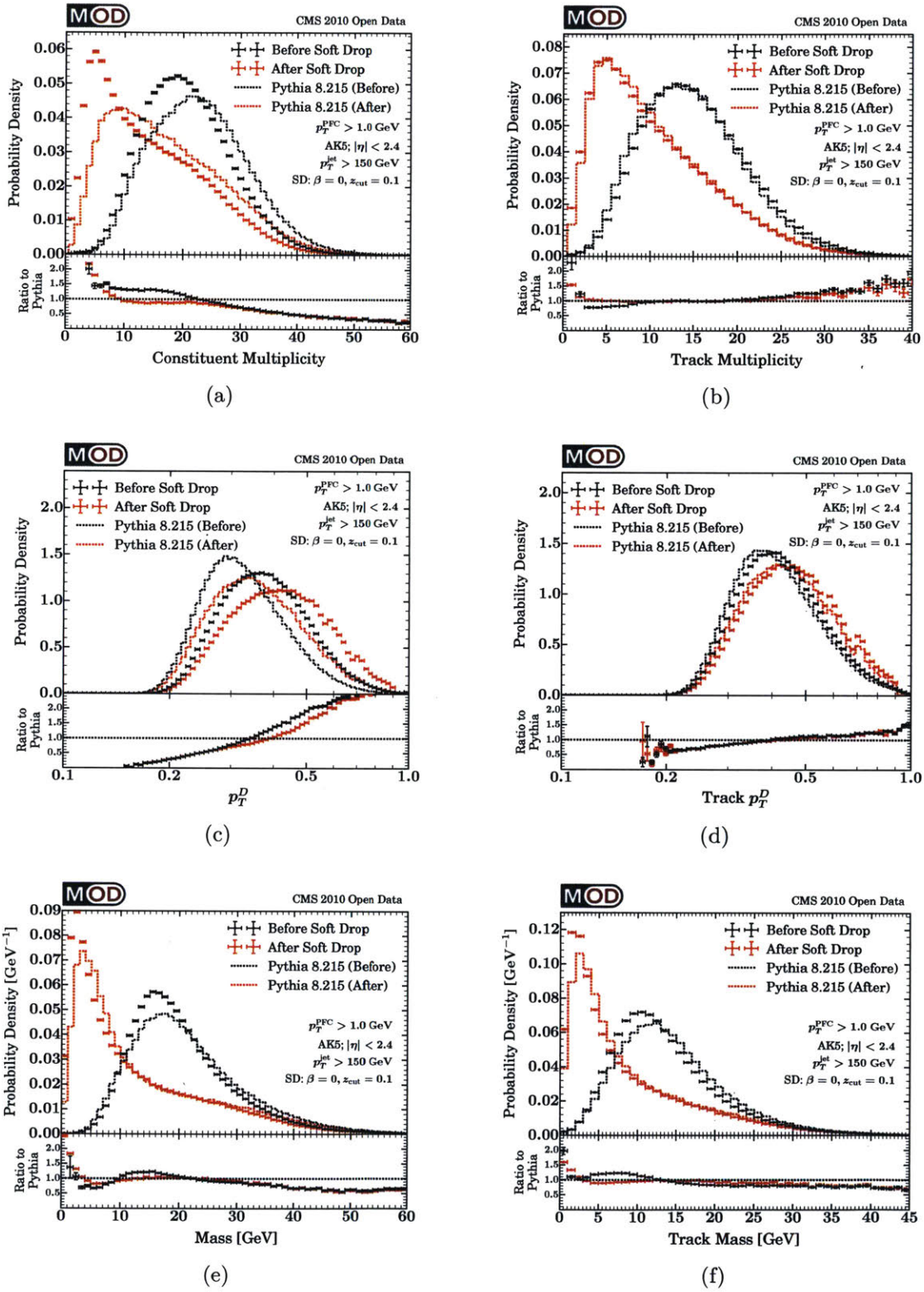


Figure 6-2: Same observables as in Fig. 4-4, but now showing the original distributions (black) compared to those obtained after soft drop declustering (red).

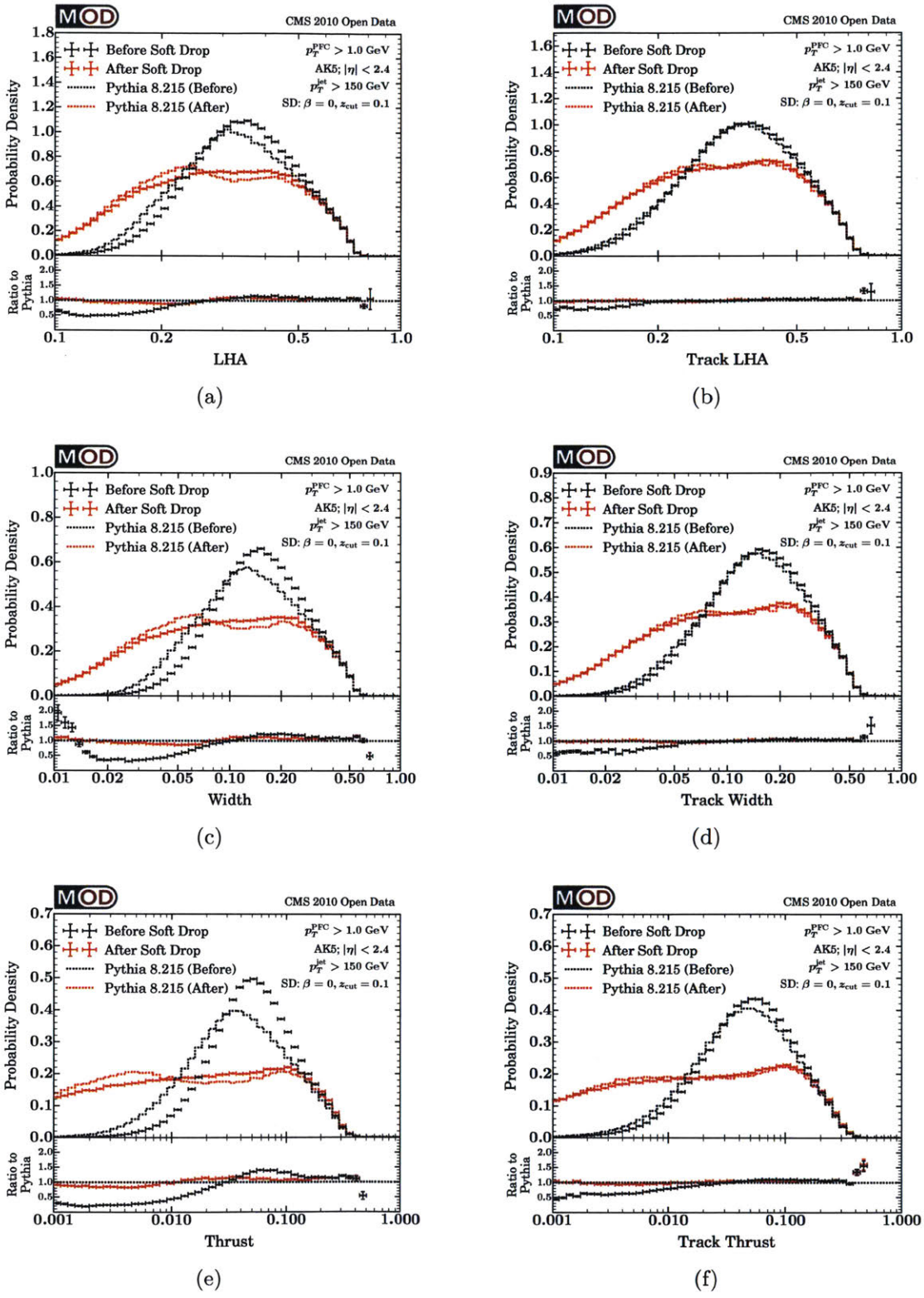


Figure 6-3: Same observables as in Fig. 6-3, but now showing the original distributions (black) compared to those obtained after soft drop declustering (red).

Chapter 7

Conclusion

The LHC is at the forefront of high-energy physics, pushing the boundaries of the entire field of physics. While this endeavor of moving physics forward through precision measurements and through new discoveries is certainly led by the collaborations within the LHC (CMS, ATLAS, LHCb, MoEDAL, TOTEM, LHC-forward, and ALICE), releasing its datasets has begun a new chapter for an entire generation of physicists. Even though there are definite challenges to successfully carrying out physics analyses from outside a collaboration, Open Data provides an unrivaled opportunity for physicists outside the collaborations to experiment with new ways of looking at the collected data. In this thesis, based on `arXiv:1704.05066` [1] and `arXiv:1704.05842` [2], we presented the first such analysis made with 2010 CMS Open Data of 7 TeV collisions.

We showed how to extract information out of the provided AOD files, and then we validated our basic kinematics and jet substructure observables by comparing with results obtained from parton shower generators. We then exposed the 2-prong substructure of QCD, comparing our results to those obtained from parton showers and first-principle QCD calculations. We believe that our analysis is a small but important step in the direction of the full potential that open data releases like this has in store, and we hope our experience motivates the LHC collaborations to expand their investment in public data release and encourages the wider particle physics community to explore these datasets to fully exploit the huge opportunities they provide.

Appendix A

Additional Open Data Information

In this appendix, we provide additional information about the overall CMS Open Data extraction. Fig. A-2, shows the distribution of prescales for the triggers from in Table 2.2. As expected, higher trigger thresholds have lower prescale values, but there is substantial variation in the prescale values which changed over the duration of the run.

The jet quality criteria are shown in Table A.1. We use the "loose" selection throughout our entire analysis.

Fig. A-1a shows a distribution of the Jet Energy Corrections for the hardest jet. The JEC factors not only account for detector effects but also takes pileup into consideration through area subtraction [4]. We show the distribution of jet areas for the hardest jet in Fig. A-1b, which peaks at πR^2 for $R = 0.5$ as expected.

Table A.2 shows the number of primary interactions per bunch crossing. Notice that the number of primary vertices is less than 5 for over 90% of the events, suggesting effectively no pileup for 90% of the events and modest pileup (NPV < 15) for the remaining 10%.

	Loose	Medium	Tight
Neutral Hadron Fraction	< 0.99	< 0.95	< 0.90
Neutral EM Fraction	< 0.99	< 0.95	< 0.90
Number of Constituents	> 1	> 1	> 1
Charged Hadron Fraction	> 0.00	> 0.00	> 0.00
Charged EM Fraction	< 0.99	< 0.99	< 0.99
Charged Multiplicity	> 0	> 0	> 0

Table A.1: Recommended jet quality criteria provided by CMS for $|\eta| < 2.4$. For $|\eta| > 2.4$, where no tracking is available, the last three requirements are not applied, and all jet constituents are treated as neutral. For our analysis, we always impose the "loose" criteria.

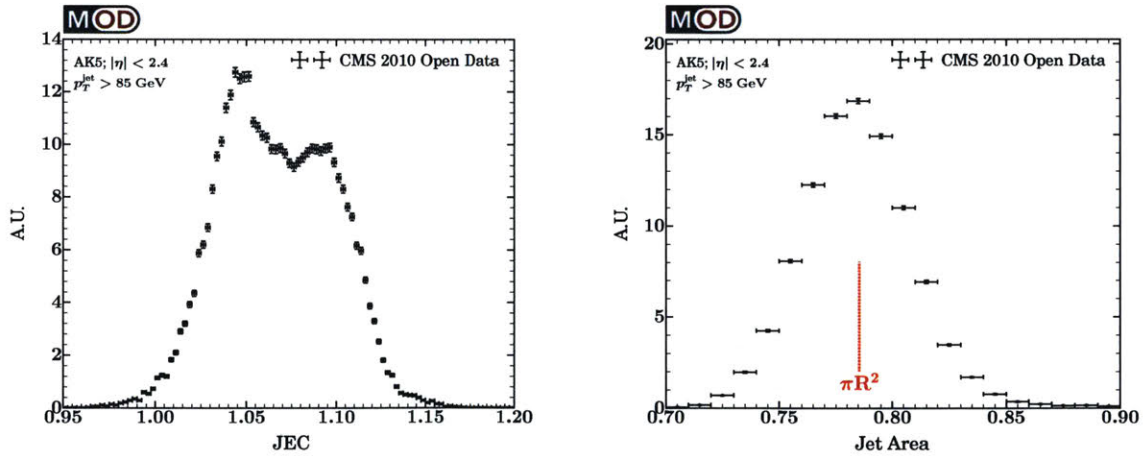


Figure A-1: Range of (a) JEC factors and (b) active jet areas [4] encountered for the hardest jet.

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

Table A.2: Number of primary interactions per bunch crossing. Since Run 2010B was a relatively low luminosity run, a large fraction of the event sample has $N_{PV} = 1$, corresponding to no pileup contamination.

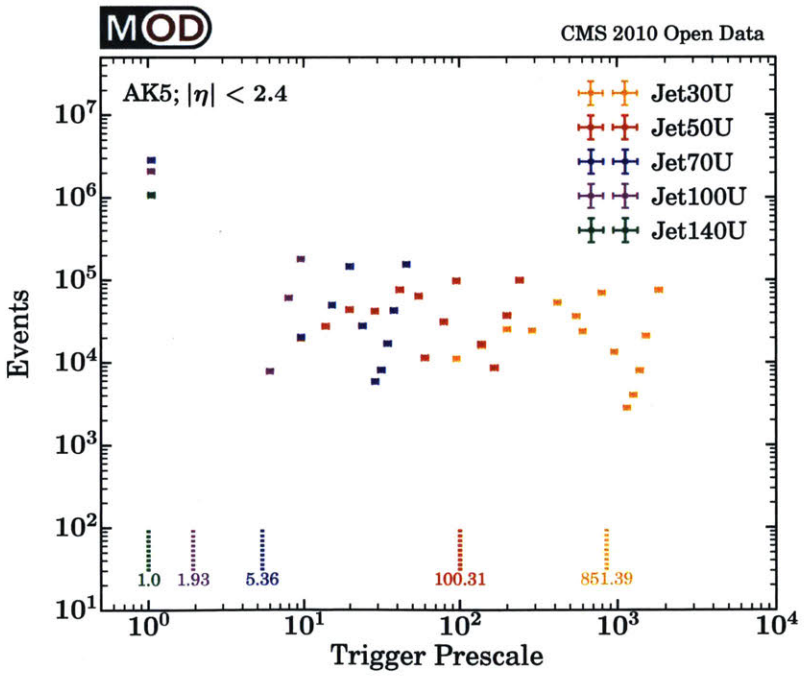


Figure A-2: Trigger prescale values for jets that pass the criteria in Table 2.2. When filling histograms in this paper, we always use the average prescale values, not the individual ones.

Appendix B

MIT Open Data (MOD) Sample Event

We provide a sample event in the MOD data format in this appendix. Please note that the list of PFCandidates in the sample event selected has been truncated to shorten the output. So, clustering the given PFCs will not produce the corresponding AK5 jets. For an unaltered sample event, please refer to the ancillary file in `arXiv:1704.05842` [2].

BeginEvent Version 5 CMS_2010 Jet_Primary_Dataset

#	Cond	RunNum	EventNum	LumiBlock	validLumi	intgDelLumi	intgRecLumi	AvgInstLumi
		NPV	timestamp	msOffset				
	Cond	147114	152963276	259	1	11513.9	10445.3	49.2951
		1	1286106631	410947				

#	Trig	Name	Prescale_1	Prescale_2	Fired?
	Trig	HLT_DiJetAve15U	280	10	0
	Trig	HLT_DiJetAve30U	1	280	0
	Trig	HLT_DiJetAve50U	1	28	0
	Trig	HLT_DiJetAve70U	1	1	1
	Trig	HLT_EcalOnly_SumEt160	1	1	0
	Trig	HLT_HT100U	1	1	1
	Trig	HLT_HT120U	1	1	1
	Trig	HLT_HT140U	1	1	1
	Trig	HLT_Jet100U	1	1	1
	Trig	HLT_Jet15U	280	20	0
	Trig	HLT_Jet15U_HcalNoiseFiltered	280	20	0
	Trig	HLT_Jet30U	1	560	0
	Trig	HLT_Jet50U	1	56	0

```
   Trig                HLT_Jet70U            1             1             1
   Trig              HLT_QuadJet20U          1             1             0
   Trig              HLT_QuadJet25U          1             1             0
# AK5            px            py            pz         energy           jec           area    no_of_const
   chrg_multip   neu_had_frac    neu_em_frac  chrg_had_frac   chrg_em_frac
  AK5    28.13078665 -157.73707632  228.62372934  279.54120919    1.12216985    0.79786479           31
                21     0.01451194    0.28866803    0.53939581    0.15742429
  AK5   -49.72790461  144.13004209 -102.17612654  185.02054234    1.06568718    0.78789151           43
                31     0.03131541    0.11848661    0.85019807    0.00000000
  AK5    13.91300205   14.46203663  -32.36586274   38.47387898    1.24642992    0.84773135           18
                 9     0.00000000    0.56448937    0.43551072    0.00000000
  AK5     6.76953145    0.42620028   -2.06816746    7.21711671    1.42414188    0.81781143            6
                 1     0.13133221    0.75329578    0.11537200    0.00000000
# PFC            px            py            pz         energy  pdgId
  PFC     0.62619645   -3.40553349    6.06523567    6.98543799    211
  PFC    -0.03002445   -0.67901639    1.20540196    1.38382030     22
  PFC     0.34554418   -1.76035851    3.07534181    3.56307030    211
  PFC    -0.92202425    2.86300494   -1.13828921    3.21902285    211
  PFC    -4.96452123   13.62500583   -9.58583145   17.38375389    211
  PFC     0.12325423    0.44456433   -0.32053494    0.56175768     22
  PFC    -0.10387208   -0.14109921    0.26991964    0.32179964     22
```

PFC	0.18553281	0.30619199	-0.61890901	0.71499952	22
PFC	0.30781106	-0.00395793	-0.02677169	0.30899844	22
PFC	-0.18802100	-0.19204160	0.13056783	0.29879730	22
PFC	0.24072418	0.03465248	-0.15824331	0.29015490	22
PFC	-0.00658784	-0.13697054	-0.25310287	0.28786349	22
PFC	-0.05069927	-0.22571411	-0.16914340	0.28657767	22

EndEvent

Bibliography

[1] A. Larkoski, S. Marzani, J. Thaler, A. Tripathee, and W. Xue, "Exposing the QCD Splitting Function with CMS Open Data," 2017.

[2] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, and J. Thaler, "Jet Substructure Studies with CMS Open Data," 2017.

[3] W. Adam *et al.*, "Alignment of the CMS Silicon Strip Tracker during stand-alone Commissioning," *JINST*, vol. 4, p. T07001, 2009.

[4] M. Cacciari, G. P. Salam, and G. Soyez, "The Catchment Area of Jets," *JHEP*, vol. 0804, p. 005, 2008.

[5] "CERN Open Data Portal." `http://opendata.cern.ch`.

[6] "Jet primary dataset in AOD format from RunB of 2010 (/Jet/Run2010B-Apr21ReReco-v1/AOD)," *CMS Collaboration, CERN Open Data Portal.*

[7] "XRootD Project." `http://xrootd.org`.

[8] "MIT Open Data Producer." `https://github.com/tripatheea/MODProducer`.

[9] "Public CMS Luminosity Information." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`.

[10] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual," *Eur.Phys.J.*, vol. C72, p. 1896, 2012.

[11] C. Patrignani *et al.*, "Review of Particle Physics," *Chin. Phys.*, vol. C40, no. 10, p. 100001, 2016.

[12] "MIT Open Data Analyzer." `https://github.com/tripatheea/MODAnalyzer`.

[13] "Fastjet contrib." `http://fastjet.hepforge.org/contrib/`.

[14] T. Sjostrand, S. Mrenna, and P. Skands, "A Brief Introduction to PYTHIA 8.1," *Comput.Phys.Commun.178:852-867,2008*, Oct. 2007.

[15] J. Bellm *et al.*, "Herwig 7.0/Herwig++ 3.0 release note," *Eur. Phys. J.*, vol. C76, no. 4, p. 196, 2016.

[16] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter, "Event generation with SHERPA 1.1," *JHEP*, vol. 02, p. 007, 2009.

[17] M. Dobbs and J. B. Hansen, "The HepMC C++ Monte Carlo event record for High Energy Physics," *Comput. Phys. Commun.*, vol. 134, pp. 41–46, 2001.

[18] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment," *JHEP*, vol. 02, p. 057, 2014.

[19] "Jet primary dataset in AOD format from RunA of 2011 (/Jet/Run2011A-12Oct2013-v1/AOD)," *CMS Collaboration, CERN Open Data Portal.*

[20] M. Cacciari, G. P. Salam, and G. Soyez, "SoftKiller, a particle-level pileup removal method," *Eur. Phys. J.*, vol. C75, no. 2, p. 59, 2015.

[21] "Performance of quark/gluon discrimination in 8 TeV pp data," Tech. Rep. CMS-PAS-JME-13-002, 2013.

[22] C. F. Berger, T. Kúcs, and G. Sterman, "Event shape / energy flow correlations," *Phys. Rev. D*, vol. 68, p. 014012, 2003.

[23] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung, *et al.*, "Substructure of high-p_T Jets at the LHC," *Phys.Rev.*, vol. D79, p. 074017, 2009.

[24] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, and C. Lee, "Jet Shapes and Jet Algorithms in SCET," *JHEP*, vol. 1011, p. 101, 2010.

[25] A. J. Larkoski, D. Neill, and J. Thaler, "Jet Shapes with the Broadening Axis," *JHEP*, vol. 1404, p. 017, 2014.

[26] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, "Gaining (Mutual) Information about Quark/Gluon Discrimination," *JHEP*, vol. 11, p. 129, 2014.

[27] S. Catani, G. Turnock, and B. Webber, "Jet broadening measures in e^+e^- annihilation," *Phys.Lett.*, vol. B295, pp. 269–276, 1992.

[28] Y. L. Dokshitzer, A. Lucenti, G. Marchesini, and G. Salam, "On the QCD analysis of jet broadening," *JHEP*, vol. 9801, p. 011, 1998.

[29] A. Banfi, G. P. Salam, and G. Zanderighi, "Principles of general final-state resummation and automated implementation," *JHEP*, vol. 0503, p. 073, 2005.

[30] A. J. Larkoski, G. P. Salam, and J. Thaler, "Energy Correlation Functions for Jet Substructure," *JHEP*, vol. 1306, p. 108, 2013.

[31] D. Bertolini, T. Chan, and J. Thaler, "Jet Observables Without Jet Algorithms," *JHEP*, vol. 1404, p. 013, 2014.

[32] G. Salam, "E_t^∞ Scheme," *Unpublished.*

[33] M. Wobisch and T. Wengler, "Hadronization corrections to jet cross-sections in deep inelastic scattering," 1998.

[34] Y. L. Dokshitzer, G. Leder, S. Moretti, and B. Webber, "Better jet clustering algorithms," *JHEP*, vol. 9708, p. 001, 1997.

[35] J. R. Andersen *et al.*, "Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report," in *9th Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2015) Les Houches, France, June 1-19, 2015*, 2016.

[36] P. Gras, S. Hoeche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer, A. Siódmok, P. Skands, G. Soyez, and J. Thaler, "Systematics of quark/gluon tagging," 2017.

[37] P. E. Rakow and B. Webber, "Transverse Momentum Moments of Hadron Distributions in QCD Jets," *Nucl.Phys.*, vol. B191, p. 63, 1981.

[38] R. K. Ellis and B. Webber, "QCD Jet Broadening in Hadron Hadron Collisions," *Conf.Proc.*, vol. C860623, p. 74, 1986.

[39] E. Farhi, "A QCD Test for Jets," *Phys.Rev.Lett.*, vol. 39, pp. 1587–1588, 1977.

[40] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft Drop," *JHEP*, vol. 1405, p. 146, 2014.

[41] A. J. Larkoski, S. Marzani, and J. Thaler, "Sudakov Safety in Perturbative QCD," *Phys. Rev.*, vol. D91, no. 11, p. 111501, 2015.

[42] "Splitting function in pp and PbPb collisions at 5.02 TeV," Tech. Rep. CMS-PAS-HIN-16-006, 2016.

[43] K. Kauder, "Measurement of the shared momentum fraction zg using jet reconstruction in p+p and au+au collisions with star." Conference talk at Hard Probes 2016.

[44] K. Lapidus, "Hard substructure of jets probed in p-pb collisions." Poster at Quark Matter 2017.

[45] P. Ilten, N. L. Rodd, J. Thaler, and M. Williams, "Disentangling Heavy Flavor at Colliders," 2017.

[46] 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.*, vol. 100, p. 242001, 2008.

[47] G. Aad *et al.*, "Search for high-mass diboson resonances with boson-tagged jets in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector," *JHEP*, vol. 12, p. 055, 2015.

[48] A. J. Larkoski and J. Thaler, "Unsafe but Calculable: Ratios of Angularities in Perturbative QCD," *JHEP*, vol. 1309, p. 137, 2013.

[49] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, "Towards an understanding of jet substructure," *JHEP*, vol. 09, p. 029, 2013.

[50] M. Dasgupta, A. Fregoso, S. Marzani, and A. Powling, "Jet substructure with analytical methods," *Eur. Phys. J.*, vol. C73, no. 11, p. 2623, 2013.