Exposing the QCD Splitting Function with CMS Open Data

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1103/PhysRevLett.119.132003">http://dx.doi.org/10.1103/PhysRevLett.119.132003</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>American Physical Society</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Wed Mar 13 19:18:30 EDT 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/113593">http://hdl.handle.net/1721.1/113593</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,1,* Simone Marzani,2,† Jesse Thaler,3,‡ Aashish Tripathee,3,§ and Wei Xue3,∥
1Physics Department, Reed College, Portland, Oregon 97202, USA
2University at Buffalo, The State University of New York, Buffalo, New York 14260-1500, USA
3Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
(Received 9 May 2017; revised manuscript received 27 July 2017; published 26 September 2017)

The splitting function is a universal property of quantum chromodynamics (QCD) which describes how energy is shared between partons. Despite its ubiquitous appearance in many QCD calculations, the splitting function cannot be measured directly, since it always appears multiplied by a collinear singularity factor. Recently, however, a new jet substructure observable was introduced which asymptotes to the splitting function for sufficiently high jet energies. This provides a way to expose the splitting function through jet substructure measurements at the Large Hadron Collider. In this Letter, we use public data released by the CMS experiment to study the two-prong substructure of jets and test the $1 \to 2$ splitting function of QCD. To our knowledge, this is the first ever physics analysis based on the CMS Open Data.

DOI: 10.1103/PhysRevLett.119.132003

Quantum chromodynamics (QCD), like any weakly coupled gauge theory, exhibits universal behavior in the small angle limit. When two partons become collinear in QCD, the cross section for a $2 \to n$ scattering process factorizes into a $2 \to n - 1$ scattering cross section multiplied by a universal $1 \to 2$ splitting probability, with corrections suppressed by the degree of collinearity. Collinear universality is a fundamental property of QCD and appears in many applications, most famously in deriving the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi evolution equations [1–3] (see also [4–13]), and it is at the heart of the factorization theorem in hadron-hadron collisions [14,15]. In addition, parton shower generators are based on recursively applying $1 \to 2$ splittings [16–18], fixed-order subtraction schemes utilize the $1 \to 2$ splitting function [19–21], and the $k_t$ jet clustering metric is based on $2 \to 1$ recombination [22–24]. Collinear universality can be extended to multiparton splittings at tree level and beyond [25–41]; however, its all-orders validity [42,43] is spoiled in the presence of Glauber modes [44–47]. More recently, jet substructure techniques [48–52] have been introduced to distinguish $1 \to n$ decays of heavy particles from $1 \to n$ splittings in QCD in order to enhance the search for new physics at the Large Hadron Collider (LHC) [53–56].

Despite its ubiquity, however, the $1 \to 2$ splitting function cannot be directly measured at a collider, since collinear universality is inseparable from the existence of collinear singularities and closely related nonperturbative fragmentation functions. Specifically, when two partons are separated by an angle $\theta$, the $1 \to 2$ splitting probability takes the form

$$dP_{1\to jk} = \frac{d\theta}{\theta} dz P_{1\to jk}(z),$$

where the $P_{1\to jk}$ are the Altarelli-Parisi QCD splitting functions [3] which depend on the momentum fraction $z$ and the parton flavors $i$, $j$, and $k$. Crucially, this expression has a real emission singularity in the $\theta \to 0$ limit, as required to cancel corresponding virtual singularities from loop diagrams. In this sense, there is no way to directly measure the splitting function $P_{1\to jk}(z)$ in data, though there is of course overwhelming indirect evidence that $P_{1\to jk}(z)$ is a universal function from the many successes of QCD in describing high-energy scattering (see, e.g., [57–67]).

In this Letter, we present a semidirect method to test the $1 \to 2$ splitting function in QCD by studying the two-prong substructure of jets. Our method is based on soft drop declustering [68] (see also [52,69,70]), which recursively removes soft radiation from a jet until hard two-prong substructure is found. When applied to ordinary quark- and gluon-initiated jets with no intrinsic substructure, soft drop exposes the collinear core of the jet. As shown in Ref. [71], the momentum sharing between the two prongs (denoted $z_g$) is closely related to the momentum fraction $z$ appearing in Eq. (1), and the cross section for $z_g$ asymptotes to the QCD splitting function in the high-energy limit. While variants of $z_g$ have appeared in many jet substructure studies (notably the $\sqrt{y}$ parameter in Refs. [52,72]), to the best of our knowledge, no published $z_g$ distribution has ever been presented using actual collider data, though there are preliminary $z_g$ results from CMS [73], STAR [74], and ALICE [75] Collaborations. Here, we present the first analysis of $z_g$ using LHC data, taking advantage for the first time of public data released by the CMS experiment [76].

The CMS Open Data are derived from 7 TeV center-of-mass proton-proton collisions recorded in 2010 and released to the public on the CERN Open Data Portal in November 2014 [77]. The data are provided in analysis object data (AOD) format, which is a CMS-specific data scheme based
on the ROOT framework [78]. Crucially for the purposes of studying jet substructure, the AOD format contains all of the particle flow candidates (PFCs) [79,80] used for jet finding within CMS [81], and we can apply jet substructure techniques directly on the PFCs themselves. The AOD files have an associated conditions database which include jet energy correction (JEC) factors and recommended jet quality cuts, though no specific calibration tools for jet substructure studies. The main limitation of the 2010 CMS Open Data release is that it does not come accompanied by detector-simulated Monte Carlo samples, though this issue has been partially addressed in the 2011 CMS Open Data release [82]. Even without a detector simulation, we can improve the robustness of our analysis by using a charged-particle subset of PFCs with better angular resolution. Overall, this study highlights the fantastic performance of CMS’s particle flow algorithm and the exciting physics opportunities made possible by this public data release.

Our analysis is based on $31.8 \, \text{pb}^{-1}$ [83,84] of data collected using the Jet Primary Dataset [76], which contains events selected by single-jet triggers and dijet triggers, as well as some quadjet and $H_T$ triggers. We use the HLT_Jet30U/50U/70U/100U/140U triggers for this analysis, which gives us near 100% efficiency to select single jets with transverse momentum $p_T > 85$ GeV. All jets in our analysis are clustered using the anti-$k_t$ jet clustering algorithm [85] with radius parameter $R = 0.5$; we validated that the anti-$k_t$ jets reported by CMS in the AOD format agree with those found by directly clustering the PFCs with FASTJET 3.1.3 [86]. To gain a more transparent understanding of the CMS data, we converted the AOD file format into our own text-based MIT Open Data (MOD) file format. Information about the MOD format as well as a broader suite of jet substructure analyses will be presented in a companion paper [87]. The substructure results shown here use the RECURSIVETOOLS 1.0.0 package from FASTJET CONTRIB 1.019 [88].

To validate initial jet reconstruction, Fig. 1 shows the $p_T$ spectrum of the hardest jet in the event, with a pseudorapidity cut of $|\eta| < 2.4$ and transverse momentum cut of $p_T > 85$ GeV. This spectrum is obtained after applying the “loose” jet quality criteria provided by CMS as well as rescaling the jet $p_T$ by the provided JEC factors. For comparison, we show the same spectrum obtained from three parton shower generators with their default settings: PYTHIA 8.219 [89], HERWIG 7.0.3 [90], and SHERPA 2.2.1 [91]. The qualitative agreement between all four samples is excellent. Note that this spectrum is obtained after combining five different CMS triggers with prescale factors that changed over the course of the 2010 run. No kinks are observed at the transitions between the various triggers, giving us confidence that we can derive jet spectra using the trigger and prescale values provided in the AOD files.

We now turn to an analysis of the two-prong substructure of the hardest jet, imposing a further cut of $p_T > 150$ GeV in order to avoid the large prescale factors present in the HLT_Jet30U/50U triggers. To partially account for the finite energy resolution and efficiency of the CMS detector, we consider only PFCs within the hardest jet above $p_T^{\text{min}} = 1$ GeV. Moreover, because charged particles have better angular resolution than neutral ones, our analysis will be based only on charged particles with associated tracks; we refer the reader to Ref. [87] for substructure analyses with both charged and neutral PFCs. The charged PFCs are reclustered with the Cambridge-Aachen (CA) algorithm [92,93] to form an angular-ordered clustering tree. We then apply the soft drop declustering procedure [68] in Fig. 2, which recursively declusters the CA tree, removing the softer $p_T$ branch until two-prong substructure is found which satisfies

$$z > z_{\text{cut}} \theta^g, \quad z \equiv \frac{\min[p_{T1}, p_{T2}]}{p_{T1} + p_{T2}}, \quad \theta = \frac{R_{12}}{R}. \quad (2)$$

Here, $p_{T1}$ and $p_{T2}$ are the transverse momenta of the two branches of the CA tree, and $R_{12} = \sqrt{(y_{12})^2 + (\phi_{12})^2}$ is their relative rapidity-azimuth distance. Throughout our

![FIG. 1. Jet $p_T$ spectrum from the CMS Open Data compared to three parton shower generators. Indicated is the $p_T > 150$ GeV cut used in later analyses.](image)

![FIG. 2. Schematic of the soft drop algorithm, which removes angular-ordered branches whose momentum fraction $z$ is below $z_{\text{cut}} \theta^g$. The final groomed kinematics are indicated by the $g$ subscript.](image)
For the analysis, we take the momentum fraction cut and angular exponent to be, respectively,

\[ z_{\text{cut}} = 0.1, \quad \beta = 0, \]

such that soft drop acts like the modified mass drop tagger (mMDT) [69] with \( \mu = 1 \). The values of \( z \) and \( \theta \) obtained after the soft drop are referred to as \( z_g \) and \( \theta_g \), where the \( g \) subscript is a reminder that these values were obtained after jet grooming. These two observables encode information about the two nontrivial kinematic variables in the unpolarized \( 1 \to 2 \) QCD splitting function from Eq. (1). Note that \( z_g \) is a ratio of \( p_T \) scales, so not affected by the JEC factor applied to the jet \( p_T \) as a whole. Similarly, as a dimensionless quantity, \( z_g \) is relatively insensitive to the absolute energy scale of the PFCs and is only mildly affected by the \( p_T^{\text{min}} = 1 \) GeV restriction.

The key observable used in jet substructure analyses at ATLAS and CMS is the jet invariant mass [94–96]. The track-only jet mass spectrum before and after the soft drop is shown in Fig. 3 and compared to predictions from PYTHIA. There is reasonable qualitative agreement between the CMS Open Data and PYTHIA for \( \mu = 1 \). The peak towards small values of \( z_g \) is ambiguous for a jet containing a single parton, and therefore real emission singularities associated with two partons (where \( z_g \) is well defined) cannot cancel against virtual emission singularities associated with one parton (where \( z_g \) is ill defined). That said, we can follow the strategy outlined in Refs. [71,97] and express the normalized \( z_g \) probability distribution \( p(z_g) \) as

\[ p(z_g) = \int d\theta_g p(\theta_g) p(z_g|\theta_g), \]

where \( \alpha_s \) is the strong coupling constant and \( C_i \) is the Casimir factor (\( 4/3 \) for quarks, 3 for gluons). The \( z_g \) distribution is cut off by \( z_{\text{cut}} \), which regulates the soft singularity of QCD. In principle, the \( \theta_g \) distribution could extend all the way to zero, but it is cut off both by the angular resolution of the CMS detector and by nonperturbative QCD effects which are relevant for \( \theta_g = \lambda_{\text{QCD}}/(z_{\text{cut}} p_T) \approx 10^{-1} \). In addition, the perturbative \( \theta_g \to 0 \) singularity in Eq. (1) is regulated by a single-logarithmic form factor [68], which we now exploit to perform analytic calculations of the \( z_g \) distribution.

In perturbative QCD, \( z_g \) with \( \beta = 0 \) is a collinear-unsafe observable and therefore not calculable order by order in an expansion in the strong coupling constant \( \alpha_s \). In particular, \( z_g \) is ambiguous for a jet containing a single parton, and therefore real emission singularities associated with two partons (where \( z_g \) is well defined) cannot cancel against virtual emission singularities associated with one parton (where \( z_g \) is ill defined). That said, we can follow the strategy outlined in Refs. [71,97] and express the normalized \( z_g \) probability distribution \( p(z_g) \) as

\[ p(z_g) = \int d\theta_g p(\theta_g) p(z_g|\theta_g), \]
where \( p(\theta_y) \) is the probability distribution for \( \theta_y \) and \( p(z_g|\theta_y) \) is the conditional probability distribution for \( z_g \) given a fixed value of \( \theta_y \). While \( z_g \) is collinear unsafe, the conditional probability distribution \( p(z_g|\theta_y) \) is calculable as a perturbative expansion, since any finite value of \( \theta_y \) will remove the one-parton region of phase space. By resumming the \( p(\theta_y) \) distribution to all orders in \( \alpha_s \), the \( \theta_y \to 0 \) limit is regulated, and the integral in Eq. (5) yields a finite distribution for \( p(z_g) \). In this way, \( z_g \) is a collinear unsafe but “Sudakov safe” observable [97].

Remarkably, to the lowest nontrivial order, the probability distribution for \( p(z_g) \) can be directly expressed in terms of the QCD splitting function as [71]

\[
p(z_g) = \sum_i f_i p_i(z_g),
\]

where \( f_i \) is the fraction of the event sample composed of jets initiated by partons of flavor \( i \) (i.e., quarks or gluons), and

\[
p_i(z) = \frac{\bar{P}_i(z)}{\int_{z_{\text{cut}}}^{1} dz' P_i(z')} \Theta(z > z_{\text{cut}}) + O(\alpha_s),
\]

where

\[
\bar{P}_i(z) = \sum_{j,k} [P_{i\to j k}(z) + P_{i\to j k}(1-z)].
\]

The \( z_g \) distribution is a flavor-averaged, \( z \)-symmetrized, \( z_{\text{cut}} \)-truncated, and normalized version of the QCD splitting function. Because of a supersymmetric relationship between the quark and gluon splitting functions [98,99], \( \bar{P}_i \) is the same for quarks and gluons to an excellent approximation, such that

\[
p(z_g) \approx \frac{2 z_g}{1 - z_g} + 2 \log \frac{z_g}{z_{\text{cut}}},
\]

and the probability distribution for \( z_g \) is independent of \( \alpha_s \) at leading order. In this way, measuring \( z_g \) exposes the QCD splitting function. The predicted \( z_g \) distribution can be refined by performing higher-order calculations. As in Ref. [71], we calculate \( p(\theta_y) \) to modified leading-logarithmic (MLL) accuracy, which includes running coupling effects and subleading terms in the splitting functions. We also calculate \( p(z_g|\theta_y) \) to leading fixed order in the collinear approximation and obtain an analytic prediction for \( p(z_g) \) using Eq. (5). While not shown below, the theoretical uncertainties on \( p(z_g) \) can be estimated by varying the different renormalization scales that enter the calculation [87].

In Fig. 5, we show the \( z_g \) distribution for our jet selection, comparing the analytic expression in Eq. (5) [which extends Eq. (9) to MLL accuracy], three parton shower generators, and the CMS Open Data. Strictly speaking, the theoretical calculation described above should be modified [100,101] to account for the fact that the current analysis is based only on charged particles; for this reason, we show \( p(z_g) \) without its uncertainty band to emphasize its qualitative nature. Notwithstanding the above, the CMS Open Data agree very well with the theory calculation as well as with the Monte Carlo parton showers, and the characteristic \( 1/z \) behavior expected from the QCD splitting function is seen in all distributions. The one point where there is a noticeable (but expected) difference between the open data and the parton showers is at \( z_g = 0 \), which corresponds to jets that have only one constituent after a soft drop. Because close-by particles can be reconstructed as a single PFC due to finite angular resolution, the CMS Open Data are expected to have more “one-particle” jets than the parton shower generators. We have evidence that the small difference between the parton showers and the theory distribution at \( z_g = z_{\text{cut}} \) is due to growing logarithms of \( z_g \) that are not resummed in our MLL approach. We verified that these discrepancies are suppressed for \( z_{\text{cut}} = 0.2 \) and enhanced for \( z_{\text{cut}} = 0.05 \), consistent with this expectation.

The CMS Open Data represent a new chapter in particle physics, since, for the first time, high-quality collider data has been released to scientists not affiliated with an experimental collaboration. In this Letter, we applied state-of-the-art jet substructure techniques on the CMS Open Data and exposed the QCD splitting function, which encodes the universal behavior of gauge theories in the collinear limit. This was possible only because of theoretical advances on Sudakov safe observables, which allowed us to predict the \( z_g \) distribution from first principles, and the fantastic experimental
performance of the CMS detector, which allowed us to perform a detailed study of the substructure of jets. We hope this Letter inspires scientists outside of the LHC Collaborations to incorporate CMS Open Data into their research and motivates the LHC Collaborations to continue their support of open data initiatives.

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

1 larkoski@reed.edu
2 larsm1@buffalo.edu
3 jthaler@mit.edu
4 ashishl@mit.edu
5 weixue@mit.edu


