Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,[1,*] Simone Marzani,[2,†] Jesse Thaler,[3,‡] Aashish Tripathee,[3,§] and Wei Xue[3,∥]

[1]Physics Department, Reed College, Portland, Oregon 97202, USA
[2]University at Buffalo, The State University of New York, Buffalo, New York 14260-1500, USA
[3]Center 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 → 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 θ, the $1 \to 2$ splitting probability takes the form

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

where the $P_{i\to jk}$ 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 $P_{i\to jk}(z)$ in data, though there is of course overwhelming indirect evidence that $P_{i\to jk}(z)$ 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 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.

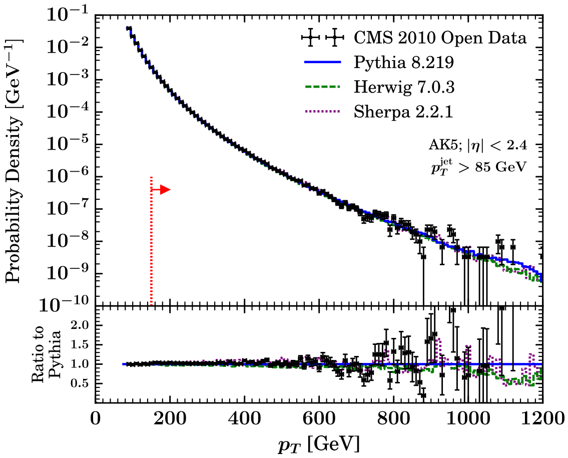

FIG. 1. Jet p_T spectrum from the CMS Open Data compared to three parton shower generators. Indicated is the $p_T > 150$ GeV cut used in later analyses.

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 $p_T^{\mathrm{min}} = 1$ 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

$$z > z_{\mathrm{cut}}\theta^{\beta}, \qquad z \equiv \frac{\min[p_{T1}, p_{T2}]}{p_{T1} + p_{T2}}, \qquad \theta = \frac{R_{12}}{R}. \tag{2}$$

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

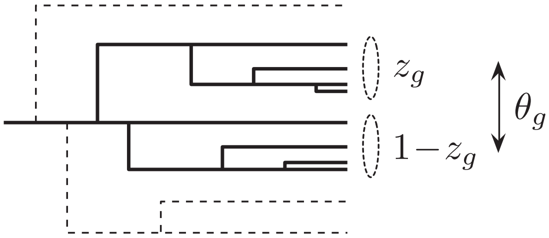

FIG. 2. Schematic of the soft drop algorithm, which removes angular-ordered branches whose momentum fraction z is below $z_{\mathrm{cut}}\theta^{\beta}$. The final groomed kinematics are indicated by the g subscript.

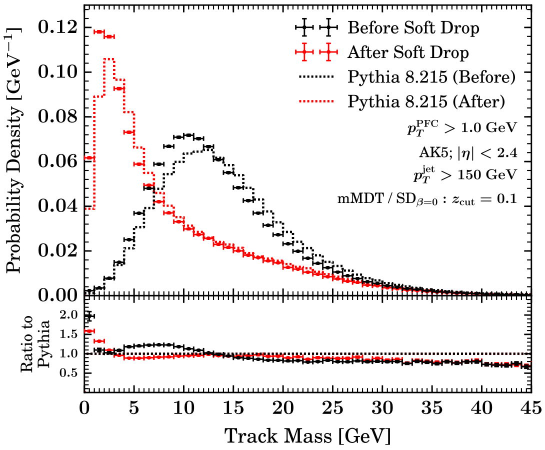

FIG. 3. Track jet mass spectra before and after the soft drop procedure with $\beta = 0$ (i.e., mMDT with $\mu = 1$), comparing the CMS Open Data to PYTHIA.

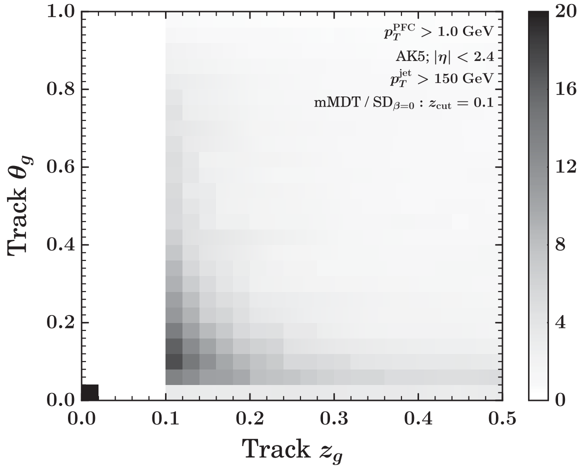

FIG. 4. Double-differential distribution of track z_g versus θ_g in the CMS Open Data, i.e., the dimensionless probability density $p(z_g, \theta_g)$ whose integral is 1.

analysis, we take the momentum fraction cut and angular exponent to be, respectively,

$$z_{\mathrm{cut}} = 0.1, \qquad \beta = 0, \tag{3}$$

such that soft drop acts like the modified mass drop tagger (mMDT) [69] with $\mu = 1$. The values of z and θ obtained after the soft drop are referred to as z_g and θ_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 $p_T^{\mathrm{min}} = 1$ 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 $m > 10$ GeV; below 10 GeV, one expects deviations from the finite detector resolution of CMS and the fact that the PFCs do not include full hadron mass information. We emphasize that no additional corrections have been applied to the CMS Open Data, apart from the JEC factor needed to impose the $p_T > 150$ GeV criteria and the $p_T^{\mathrm{min}} = 1$ GeV PFC restriction needed to account for finite energy resolution and efficiency. Similarly, we are showing particle-level predictions from PYTHIA using the default tune with no detector simulation (but the same restriction to charged hadrons with $p_T^{\mathrm{min}} = 1$ GeV). Because we do not have access to detector-simulated Monte Carlo samples, and because there is insufficient information in the AOD format to estimate systematic uncertainties, the error bars shown include only statistical uncertainties.

To see the two-prong structure revealed by the soft drop, Fig. 4 shows the double-differential track (z_g, θ_g) spectrum seen in the CMS Open Data. The peak towards small values of z_g and θ_g reflects the double-logarithmic structure in Eq. (1), since soft gluon emission from a hard quark or gluon is approximated by

$$dP_{i\to ig} \simeq \frac{2\alpha_s C_i}{\pi}\frac{d\theta}{\theta}\frac{dz}{z}, \tag{4}$$

where α_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_{cut}, which regulates the soft singularity of QCD. In principle, the θ_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 $\theta_g \simeq \Lambda_{\mathrm{QCD}}/(z_{\mathrm{cut}} p_T R) \simeq 10^{-1}$. 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 α_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

$$p(z_g) = \int d\theta_g p(\theta_g) p(z_g|\theta_g), \tag{5}$$

where $p(\theta_g)$ is the probability distribution for θ_g and $p(z_g|\theta_g)$ is the conditional probability distribution for z_g given a fixed value of θ_g. While z_g is collinear unsafe, the conditional probability distribution $p(z_g|\theta_g)$ is calculable as a perturbative expansion, since any finite value of θ_g will remove the one-parton region of phase space. By resumming the $p(\theta_g)$ distribution to all orders in α_s, the $\theta_g \to 0$ 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]

$$p(z_g) = \sum_i f_i p_i(z_g), \tag{6}$$

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

$$p_i(z) = \frac{\bar{P}_i(z)}{\int_{z_{\mathrm{cut}}}^{1/2} dz' \bar{P}_i(z')} \Theta(z > z_{\mathrm{cut}}) + O(\alpha_s), \tag{7}$$

where

$$\bar{P}_i(z) = \sum_{j,k} [P_{i\to jk}(z) + P_{i\to jk}(1-z)]. \tag{8}$$

The z_g distribution is a flavor-averaged, z-symmetrized, z_{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

$$p(z_g) \simeq \frac{2\frac{z_g}{1-z_g} + 2\frac{1-z_g}{z_g} + 1}{\frac{3}{2}(2z_{\mathrm{cut}} - 1) + 2\log\frac{1-z_{\mathrm{cut}}}{z_{\mathrm{cut}}}}, \tag{9}$$

and the probability distribution for z_g is independent of α_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 $p(\theta_g)$ to modified leading-logarithmic (MLL) accuracy, which includes running coupling effects and subleading terms in the splitting functions. We also calculate $p(z_g|\theta_g)$ 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].

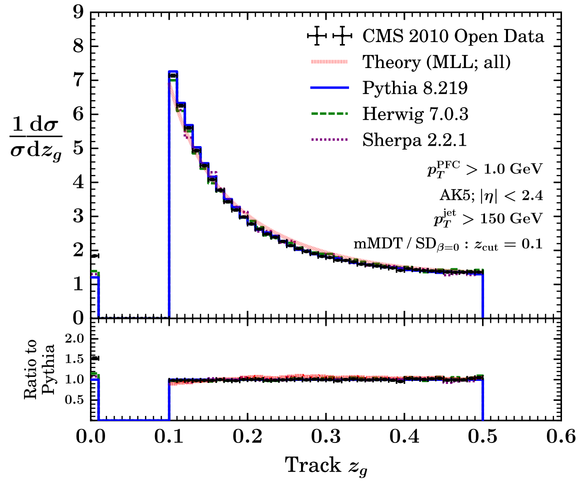

FIG. 5. Distribution of z_g from mMDT and soft drop. The theory distribution is from an all-particle prediction yet agrees very well with the track-based distributions.

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 $z_g \simeq z_{\mathrm{cut}}$ is due to growing logarithms of z_g that are not resummed in our MLL approach. We verified that these discrepancies are suppressed for $z_{\mathrm{cut}} = 0.2$ and enhanced for $z_{\mathrm{cut}} = 0.05$, 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 Lipeles, Peter Loch, Marcello Maggi, David Miller, Ben Nachman, Christoph Paus, Alexx Perloff, Andreas Pfeiffer, Maurizio Pierini, Ana Rodriguez, Gunther Roland, Ariel Schwartzman, Liz Sexton-Kennedy, Maria Spiropulu, Nhan Tran, Ana Trisovic, Chris Tully, Marta Verweij, Mikko Voutilainen, and Mike Williams. This work is supported by the MIT Charles E. Reed Faculty Initiatives Fund. The work of J. T., A. T., and W. X. is supported by the U.S. Department of Energy (DOE) under 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.

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

[1] V. N. Gribov and L. N. Lipatov, Deep inelastic e p scattering in perturbation theory, Sov. J. Nucl. Phys. **15**, 438 (1972).

[2] Y. L. Dokshitzer, Calculation of the structure functions for deep inelastic scattering and $e^+ e^-$ annihilation by perturbation theory in quantum chromodynamics, Sov. Phys. JETP **46**, 641 (1977).

[3] G. Altarelli and G. Parisi, Asymptotic freedom in parton language, Nucl. Phys. **B126**, 298 (1977).

[4] E. G. Floratos, D. A. Ross, and C. T. Sachrajda, Higher order effects in asymptotically free gauge theories: The anomalous dimensions of Wilson operators, Nucl. Phys. **B129**, 66 (1977); Erratum, Nucl. Phys. **B139**, 545(E) (1978).

[5] E. G. Floratos, D. A. Ross, and C. T. Sachrajda, Higher order effects in asymptotically free gauge theories. 2. Flavor singlet Wilson operators and coefficient functions, Nucl. Phys. **B152**, 493 (1979).

[6] A. Gonzalez-Arroyo, C. Lopez, and F. J. Yndurain, Second order contributions to the structure functions in deep inelastic scattering. 1. Theoretical calculations, Nucl. Phys. **B153**, 161 (1979).

[7] A. Gonzalez-Arroyo and C. Lopez, Second order contributions to the structure functions in deep inelastic scattering. 3. The singlet case, Nucl. Phys. **B166**, 429 (1980).

[8] G. Curci, W. Furmanski, and R. Petronzio, Evolution of parton densities beyond leading order: The nonsinglet case, Nucl. Phys. **B175**, 27 (1980).

[9] W. Furmanski and R. Petronzio, Singlet parton densities beyond leading order, Phys. Lett. B **97**, 437 (1980).

[10] E. G. Floratos, C. Kounnas, and R. Lacaze, Higher order QCD effects in inclusive annihilation and deep inelastic scattering, Nucl. Phys. **B192**, 417 (1981).

[11] R. Hamberg and W. L. van Neerven, The correct renormalization of the gluon operator in a covariant gauge, Nucl. Phys. **B379**, 143 (1992).

[12] A. Vogt, S. Moch, and J. A. M. Vermaseren, The three-loop splitting functions in QCD: The singlet case, Nucl. Phys. **B691**, 129 (2004).

[13] S. Moch, J. A. M. Vermaseren, and A. Vogt, The three loop splitting functions in QCD: The nonsinglet case, Nucl. Phys. **B688**, 101 (2004).

[14] J. C. Collins, D. E. Soper, and G. F. Sterman, All order factorization for Drell-Yan cross-sections, Phys. Lett. B **134**, 263 (1984).

[15] J. C. Collins, D. E. Soper, and G. F. Sterman, Factorization for short distance hadron-hadron scattering, Nucl. Phys. **B261**, 104 (1985).

[16] P. Mazzanti and R. Odorico, A Monte Carlo program for QCD event simulation in e^+e^- annihilation at LEP energies, Z. Phys. C **7**, 61 (1980).

[17] T. Sjostrand, The Lund Monte Carlo for jet fragmentation, Comput. Phys. Commun. **27**, 243 (1982).

[18] G. Marchesini and B. R. Webber, Simulation of QCD jets including soft gluon interference, Nucl. Phys. **B238**, 1 (1984).

[19] R. Keith Ellis, D. A. Ross, and A. E. Terrano, The perturbative calculation of jet structure in e^+e^- annihilation, Nucl. Phys. **B178**, 421 (1981).

[20] K. Fabricius, I. Schmitt, G. Kramer, and G. Schierholz, Higher order perturbative QCD calculation of jet cross-sections in e^+e^- annihilation, Z. Phys. C **11**, 315 (1982).

[21] S. Catani and M. H. Seymour, A general algorithm for calculating jet cross sections in NLO QCD, Nucl. Phys. **B485**, 291 (1997).

[22] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, New clustering algorithm for multi-jet cross-sections in e^+e^- annihilation, Phys. Lett. B **269**, 432 (1991).

[23] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, Longitudinally invariant K_t clustering algorithms for hadron hadron collisions, Nucl. Phys. **B406**, 187 (1993).

[24] S. D. Ellis and D. E. Soper, Successive combination jet algorithm for hadron collisions, Phys. Rev. D **48**, 3160 (1993).

[25] S. Catani and M. Grazzini, Collinear factorization and splitting functions for next-to-next-to-leading order QCD calculations, Phys. Lett. B **446**, 143 (1999).

[26] S. Catani and M. Grazzini, Infrared factorization of tree level QCD amplitudes at the next-to-next-to-leading order and beyond, Nucl. Phys. **B570**, 287 (2000).

[27] Z. Bern, V. D. Duca, and C. R. Schmidt, The infrared behavior of one loop gluon amplitudes at next-to-next-to-leading order, Phys. Lett. B **445**, 168 (1998).

[28] Z. Bern, V. D. Duca, W. B. Kilgore, and C. R. Schmidt, The infrared behavior of one loop QCD amplitudes at next-to-next-to leading order, Phys. Rev. D **60**, 116001 (1999).

[29] S. D. Badger and E. W. Nigel Glover, Two loop splitting functions in QCD, J. High Energy Phys. 07 (2004) 040.

[30] F. A. Berends and W. T. Giele, Recursive calculations for processes with n gluons, Nucl. Phys. **B306**, 759 (1988).

[31] M. L. Mangano and S. J. Parke, Multiparton amplitudes in gauge theories, Phys. Rep. **200**, 301 (1991).

[32] J. M. Campbell and E. W. Nigel Glover, Double unresolved approximations to multiparton scattering amplitudes, Nucl. Phys. **B527**, 264 (1998).

[33] V. D. Duca, A. Frizzo, and F. Maltoni, Factorization of tree QCD amplitudes in the high-energy limit and in the collinear limit, Nucl. Phys. **B568**, 211 (2000).

[34] T. G. Birthwright, E. W. Nigel Glover, V. V. Khoze, and P. Marquard, Multi-gluon collinear limits from MHV diagrams, J. High Energy Phys. 05 (2005) 013.

[35] T. G. Birthwright, E. W. Nigel Glover, V. V. Khoze, and P. Marquard, Collinear limits in QCD from MHV rules, J. High Energy Phys. 07 (2005) 068.

[36] Z. Bern, G. Chalmers, L. J. Dixon, and D. A. Kosower, One Loop N Gluon Amplitudes with Maximal Helicity Violation via Collinear Limits, Phys. Rev. Lett. **72**, 2134 (1994).

[37] Z. Bern, L. J. Dixon, D. C. Dunbar, and D. A. Kosower, One loop n point gauge theory amplitudes, unitarity and collinear limits, Nucl. Phys. **B425**, 217 (1994).

[38] Z. Bern and G. Chalmers, Factorization in one loop gauge theory, Nucl. Phys. **B447**, 465 (1995).

[39] D. A. Kosower and P. Uwer, One loop splitting amplitudes in gauge theory, Nucl. Phys. **B563**, 477 (1999).

[40] S. Catani, D. de Florian, and G. Rodrigo, The triple collinear limit of one loop QCD amplitudes, Phys. Lett. B **586**, 323 (2004).

[41] Z. Bern, L. J. Dixon, and D. A. Kosower, Two-loop $g \to gg$ splitting amplitudes in QCD, J. High Energy Phys. 08 (2004) 012.

[42] D. A. Kosower, All order collinear behavior in gauge theories, Nucl. Phys. **B552**, 319 (1999).

[43] I. Feige and M. D. Schwartz, Hard-soft-collinear factorization to all orders, Phys. Rev. D **90**, 105020 (2014).

[44] S. Catani, D. de Florian, and G. Rodrigo, Space-like (versus time-like) collinear limits in QCD: Is factorization violated?, J. High Energy Phys. 07 (2012) 026.

[45] J. R. Forshaw, M. H. Seymour, and A. Siodmok, On the breaking of collinear factorization in QCD, J. High Energy Phys. 11 (2012) 066.

[46] I. Z. Rothstein and I. W. Stewart, An effective field theory or forward scattering and factorization violation, J. High Energy Phys. 08 (2016) 025.

[47] M. D. Schwartz, K. Yan, and H. X. Zhu, Collinear factorization violation and effective field theory, arXiv:1703.08572 [Phys. Rev. D (to be published)].

[48] M. H. Seymour, in *Proceedings of the ECFA Large Hadron Collider (LHC) Workshop: Physics and Instrumentation Aachen, Germany, 1990* (CERN, Geneva, 1991).

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

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

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

[52] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Jet Substructure as a New Higgs Search Channel at the LHC, Phys. Rev. Lett. **100**, 242001 (2008).

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

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

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

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

[57] R. Brandelik *et al.* (TASSO Collaboration), Evidence for planar events in e^+e^- annihilation at high-energies, Phys. Lett. B **86**, 243 (1979).

[58] D. P. Barber *et al.*, Discovery of Three Jet Events and a Test of Quantum Chromodynamics at PETRA Energies, Phys. Rev. Lett. **43**, 830 (1979).

[59] C. Berger *et al.* (PLUTO Collaboration), Evidence for gluon bremsstrahlung in e^+e^- annihilations at high-energies, Phys. Lett. B **86**, 418 (1979).

[60] W. Bartel *et al.* (JADE Collaboration), Observation of planar three jet events in e^+e^- annihilation and evidence for gluon bremsstrahlung, Phys. Lett. B **91**, 142 (1980).

[61] P. Abreu *et al.* (DELPHI Collaboration), Measurement of the gluon fragmentation function and a comparison of the scaling violation in gluon and quark jets, Eur. Phys. J. C **13**, 573 (2000).

[62] M. Z. Akrawy *et al.* (OPAL Collaboration), A study of coherence of soft gluons in hadron jets, Phys. Lett. B **247**, 617 (1990).

[63] G. Abbiendi *et al.* (OPAL Collaboration), A simultaneous measurement of the QCD color factors and the strong coupling, Eur. Phys. J. C **20**, 601 (2001).

[64] A. Heister *et al.* (ALEPH Collaboration), Studies of QCD at e^+e^- centre-of-mass energies between 91-GeV and 209-GeV, Eur. Phys. J. C **35**, 457 (2004).

[65] J. Abdallah *et al.* (DELPHI Collaboration), A study of the energy evolution of event shape distributions and their means with the DELPHI detector at LEP, Eur. Phys. J. C **29**, 285 (2003).

[66] G. Abbiendi *et al.* (OPAL Collaboration), Measurement of event shape distributions and moments in $e^+e^- \to$ hadrons at 91-GeV–209-GeV and a determination of alpha(s), Eur. Phys. J. C **40**, 287 (2005).

[67] P. Achard *et al.* (L3 Collaboration), Studies of hadronic event structure in e^+e^- annihilation from 30-GeV to 209-GeV with the L3 detector, Phys. Rep. **399**, 71 (2004).

[68] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft drop, J. High Energy Phys. 05 (2014) 146.

[69] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, Towards an understanding of jet substructure, J. High Energy Phys. 09 (2013) 029.

[70] M. Dasgupta, A. Fregoso, S. Marzani, and A. Powling, Jet substructure with analytical methods, Eur. Phys. J. C **73**, 2623 (2013).

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

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

[73] CMS Collaboration, Technical Report No. CMS-PAS-HIN-16-006, 2016.

[74] K. Kauder (STAR Collaboration) (unpublished).

[75] K. Lapidus (ALICE Collaboration) (unpublished).

[76] CMS Collaboration, Jet primary dataset in AOD format from RunB of 2010 (/Jet/Run2010B-Apr21ReReco-v1/AOD), CERN Open Data Portal (2014), DOI: 10.7483/OPENDATA.CMS.3S7F.2E9W.

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

[78] R. Brun and F. Rademakers, ROOT: An object oriented data analysis framework, Nucl. Instrum. Methods Phys. Res., Sect. A **389**, 81 (1997).

[79] CMS Collaboration, Technical Report No. CMS-PAS-PFT-09-001, CERN, 2009.

[80] CMS Collaboration, Technical Report No. CMS-PAS-PFT-10-001, 2010.

[81] V. Khachatryan *et al.* (CMS Collaboration), Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV, J. Instrum. **12**, P02014 (2017).

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

[83] M. De Gruttola, Ph.D. thesis, Naples University, 2010.

[84] CMS Collaboration, Technical Report No. CMS-PAS-EWK-10-004, CERN, 1900.

[85] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k(t) jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.

[86] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, Eur. Phys. J. C **72**, 1896 (2012).

[87] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, and J. Thaler, Jet substructure studies with CMS open data, arXiv:1704.05842 [Phys. Rev. D. (to be published)].

[88] Fastjet contrib, http://fastjet.hepforge.org/contrib/.

[89] T. Sjostrand, S. Mrenna, and P. Skands, A brief introduction to PYTHIA 8.1, Comput. Phys. Commun. **178**, 852 (2008).

[90] J. Bellm *et al.*, Herwig 7.0/Herwig++ 3.0 release note, Eur. Phys. J. C **76**, 196 (2016).

[91] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter, Event generation with SHERPA 1.1, J. High Energy Phys. 02 (2009) 007.

[92] M. Wobisch and T. Wengler, Hadronization corrections to jet cross-sections in deep inelastic scattering, arXiv:hep-ph/9907280.

[93] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, Better jet clustering algorithms, J. High Energy Phys. 08 (1997) 001.

[94] G. Aad *et al.* (ATLAS Collaboration), Jet mass and substructure of inclusive jets in $\sqrt{s} = 7$ TeV pp collisions with the ATLAS experiment, J. High Energy Phys. 05 (2012) 128.

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

[96] T. Aaltonen *et al.* (CDF Collaboration), Study of substructure of high transverse momentum jets produced in proton-antiproton collisions at $\sqrt{s} = 1.96$ TeV, Phys. Rev. D **85**, 091101 (2012).

[97] A. J. Larkoski and J. Thaler, Unsafe but calculable: Ratios of angularities in perturbative QCD, J. High Energy Phys. 09 (2013) 137.

[98] Y. L. Dokshitzer, V. A. Khoze, A. H. Mueller, and S. I. Troian, *Basics of Perturbative QCD* (Frontieres, Gif-sur-Yvette, France, 1991).

[99] M. H. Seymour, Jet shapes in hadron collisions: Higher orders, resummation and hadronization, Nucl. Phys. **B513**, 269 (1998).

[100] H.-M. Chang, M. Procura, J. Thaler, and W. J. Waalewijn, Calculating Track-Based Observables for the LHC, Phys. Rev. Lett. **111**, 102002 (2013).

[101] H.-M. Chang, M. Procura, J. Thaler, and W. J. Waalewijn, Calculating track thrust with track functions, Phys. Rev. D **88**, 034030 (2013).