PHYSICAL REVIEW D **81,** 094023 (2010)

Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches

Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh
University of Washington, Seattle, Washington 98195-1560, USA
(Received 15 February 2010; revised manuscript received 26 April 2010; published 20 May 2010)

We discuss jet substructure in recombination algorithms for QCD jets and single jets from heavy particle decays. We demonstrate that the jet algorithm can introduce significant systematic effects into the substructure. By characterizing these systematic effects and the substructure from QCD, splash-in, and heavy particle decays, we identify a technique, pruning, to better identify heavy particle decays into single jets and distinguish them from QCD jets. Pruning removes protojets typical of soft, wide-angle radiation, improves the mass resolution of jets reconstructing heavy particle decays, and decreases the QCD background to these decays. We show that pruning provides significant improvements over unpruned jets in identifying top quarks and W bosons and separating them from a QCD background, and may be useful in a search for heavy particles.

DOI: 10.1103/PhysRevD.81.094023 PACS numbers: 13.87.−a, 29.85.Fj

I. INTRODUCTION

The Large Hadron Collider (LHC) will present an exciting and challenging environment. Efforts to tease out hints of beyond the standard model (BSM) physics from complicated final states, typically dominated by standard model (SM) interactions, will almost surely require the use of new techniques applied to familiar quantities. Of particular interest is the question of how we think about hadronic jets at the LHC [1]. Historically jets have been employed as surrogates for *individual* short-distance energetic partons that evolve semi-independently into showers of energetic hadrons on their way from the interaction point through the detectors. An accurate reconstruction of the jets in an event then provides an approximate description of the underlying short-distance, hard-scattering kinematics. With this picture in mind, it is not surprising that the internal structure of jets, e.g., the fact that the experimentally detected jets exhibit nonzero masses, has rarely been used in analyses at the Tevatron. However, we can anticipate that large-mass objects, which yield multijet decays at the Tevatron, e.g., W/Z's (two jets) or top quarks (three jets), will often be produced with sufficient boosts to appear as single jets at the LHC. Thus the masses of jets and further details of the internal structure of jets will be useful in identifying single jets not only as familiar objects like the aforementioned vector bosons and top quarks, but also as less familiar cascade decays of supersymmetry particles or the decays of V particles [2]. In fact, the idea of studying the subjet structure of jets has been around for some time, but initially this study took the form of discussing the number of jets as a function of the jet resolution scale, typically at e^+e^- colliders, or the p_T distribution within the cone of (cone) jets at the Tevatron. (See, for example, the analyses in [3–5].) Recently a variety of studies have appeared suggesting a range of techniques for identifying jets with specific properties to find W's [6], top quarks [7–11], Higgs bosons [11,12], and cascades of supersymmetric particles [13,14]. It is to this discussion that we intend to contribute. Not surprisingly the current literature focuses on ''tagging'' the single jet decays of specific heavy particles. However, since we cannot be certain as to the full spectrum of new physics to be found at the LHC, it is important to keep in mind the underlying goal of separating QCD jets from *any* other type of jet. This will be challenging and the diversity of approaches currently being discussed in the literature is essential. Successful searches for new physics at the LHC will likely employ a variety of techniques. The analysis described below presents detailed properties of the ''pruning'' procedure outlined in [15].

In the following discussion we will focus on jets defined by k_T-type jet algorithms. The iterative recombination structure of these algorithms yields jets that, by definition, are assembled from a sequence of protojets, or subjets. It is natural to try to use this subjet structure (along with the p_T and mass of the jet) to distinguish different types of jets. A combination of cuts and likelihood methods applied to this subjet structure can be used to identify jets, and thus events, likely to be enriched with vector or Higgs bosons, top quarks, or BSM physics. Such jet-labeling techniques can then be used in conjunction with more familiar jet- and lepton-counting methods to isolate new physics at the LHC.

An essential aspect of high-p_T jets at the LHC is that the jet algorithm ensures nonzero masses not only for the individual jets, but also for the subjets. For recombination algorithms, we can analyze the $1 \rightarrow 2$ branching structure inherent in the substructure of the jet in terms of concepts familiar from usual two-body decays. In fact, it is exactly such decays (say from W/Z and top quark decays) that we want to compare in the current study to the structure of ''ordinary'' QCD (light quark and gluon) jets. As we analyze the internal structure of jets we will attempt to keep in mind the various limitations of jets. Jets are not intrinsically well defined, but exhibit (often broad) distributions that are shaped by the very algorithms that define

them. Further, true experimental QCD jets are not identical to the leading-logarithm parton showers produced by Monte Carlos, but also include subleading effects in the parton shower as well as (perturbative) contributions from hard emissions, which may be important for precisely the properties of jets we want to discuss here, including masses. Finally, the background particles from the underlying event, and from pileup at higher luminosities, will influence the properties of the jets observed in the detector.

In this paper, rather than developing a technique to find a particular signal, we study general features of jets found with recombination algorithms. Our goal is to use this knowledge to distinguish jets produced by heavy particle decays from jets produced purely from the showering of light quarks and gluons. We conclude that heavy particle jets can be effectively identified by first pruning away soft radiation from the edges of the jet, then looking for bumps in jet and subjet mass distributions. In the case of known particles, mass cuts can be applied directly.

The following discussion includes a review of jet algorithms (Sec. II) and a review of the expected properties of jets from QCD (Sec. III) and those from heavy particles (Sec. IV). For both QCD and heavy particle jets, we begin by considering the substructure predicted by parton-level (few-particle) models with appropriate (but approximate) dynamics. We will see that at this level the two types of signals have distinct substructure kinematics. We then study, with Monte Carlo simulation, the effects of the parton shower and the jet algorithm. The jet algorithm is intended to ''undo'' the shower, but can only do so on average and hence introduces its own biases. The systematic effects of recombination algorithms shape the kinematics of the observed substructure. We attempt to summarize the most relevant features. That jet algorithms introduce systematic effects is not a new idea (for example, the effect of the underlying event on k_T jet masses was noted in [16]), but we believe our focus on substructure kinematics due to the jet algorithm is novel. The expert reader may want to skim these sections and begin with Sec. V, where we summarize Secs. II, III, and IV and include a discussion of global effects such as the underlying event.

In Sec. VI, we show how the systematic effects seen so far can by reduced by a procedure we call pruning. Pruning is based on the same ideas as other jet substructure methods such as ''filtering'' [12] and ''top tagging'' [9], in that these techniques also modify the jet substructure to improve heavy particle identification. Pruning differs from these methods in that it is built as a broad jet substructure analysis tool, and one that can be used in a variety of searches. To this end, the mechanics of the pruning procedure differ from other methods, allowing it to be generalized more easily. Pruning can be performed using either the Cambridge-Aachen (CA) or k_T algorithms to generate substructure for a jet, and the procedure can be implemented on jets identified by any algorithm, since the procedure is independent of the initial jet finder.

We explore many aspects of pruning's performance to demonstrate its utility. Sections VII and VIII describe our Monte Carlo studies of pruning and their results. Additional computational details are provided in the Appendix. In Sec. IX we summarize these results and provide concluding remarks.

II. RECOMBINATION ALGORITHMS AND JET SUBSTRUCTURE

Jet algorithms can be broadly divided into two categories, recombination algorithms and cone algorithms [1]. Both types of algorithms form jets from protojets, which are initially generic objects such as calorimeter towers, topological clusters, or final-state particles. Cone algorithms fit protojets within a fixed geometric shape, the cone, and attempt to find stable configurations of those shapes to find jets. In the cone-jet language, ''stable'' means that the direction of the total four-momentum of the protojets in the cone matches the direction of the axis of the cone. Recombination algorithms, on the other hand, give a prescription to *pairwise* (re)combine protojets into new protojets, eventually yielding a jet. For the recombination algorithms studied in this work, this prescription is based on an understanding of how the QCD shower operates, so that the recombination algorithm attempts to undo the effects of showering and approximately trace back to objects coming from the hard scattering. The anti-k_T algorithm [17] functions more like the original cone algorithms, and its recombination scheme is not designed to backtrack through the QCD shower. Cone algorithms have been the standard in collider experiments, but recombination algorithms are finding more frequent use. Analyses at the Tevatron [18] have shown that the most common cone and recombination algorithms agree in measurements of jet cross sections.

A general recombination algorithm uses a distance measure ρ_{ij} between protojets to control how they are merged. A ''beam distance'' ρ_i determines when a protojet should be promoted to a jet. The algorithm proceeds as follows:

(0) Form a list L of all protojets to be merged.

(1) Calculate the distance between all pairs of protojets in L using the metric ρ_{ij}, and the beam distance for each protojet in L using ρ_i.

(2) Find the smallest overall distance in the set $\{\rho_i, \rho_{ij}\}$.

(3) If this smallest distance is a ρ_{ij}, merge protojets i and j by adding their four vectors. Replace the pair of protojets in L with this new merged protojet. If the smallest distance is a ρ_i, promote protojet i to a jet and remove it from L.

(4) Iterate this process until L is empty, i.e., all protojets have been promoted to jets.[1]

[1]This defines an *inclusive* algorithm. For an *exclusive* algorithm, there are no promotions, but instead of recombining until L is empty, mergings proceed until all ρ_{ij} exceed a fixed ρ_{cut}.

For the k_T [19–21] and CA [22] recombination algorithms the metrics are

$$\begin{aligned} &k_T: \rho_{ij} \equiv \min(p_{Ti}, p_{Tj})\Delta R_{ij}/D, \qquad \rho_i \equiv p_{Ti}; \\ &\text{CA}: \rho_{ij} \equiv \Delta R_{ij}/D, \qquad \rho_i \equiv 1. \end{aligned} \tag{1}$$

Here p_{Ti} is the transverse momentum of protojet i and $\Delta R_{ij} \equiv \sqrt{(\phi_i - \phi_j)^2 + (y_i - y_j)^2}$ is a measure of the angle between two protojets that is invariant under boosts along and rotations around the beam direction. ϕ is the azimuthal angle around the beam direction, $\phi = \tan^{-1} p_y/p_x$, and y is the rapidity, $y = \tanh^{-1} p_z/E$, with the beam along the z axis. The angular parameter D governs when protojets should be promoted to jets: it determines when a protojet's beam distance is less than the distance to other objects. D provides a rough measure of the typical angular size (in $y - \phi$) of the resulting jets.

The recombination metric ρ_{ij} determines the *order* in which protojets are merged in the jet, with recombinations that minimize the metric performed first. From the definitions of the recombination metrics in Eq. (1), it is clear that the k_T algorithm tends to merge low-p_T protojets earlier, while the CA algorithm merges pairs in strict angular order. This distinction will be very important in our subsequent discussion.

A. Jet substructure

A recombination algorithm naturally defines substructure for the jet. The sequence of recombinations tells us how to construct the jet in step-by-step $2 \to 1$ mergings, and we can unfold the jet into two, three, or more subjets by undoing the last recombinations. Because the jet algorithm begins and ends with physically meaningful information (starting at calorimeter cells, for example, and ending at jets), the intermediate (subjet) information generated by the k_T and CA (but not the anti-k_T[2]) recombination algorithms is expected to have physical significance as well. In particular, we expect the earliest recombinations to approximately reconstruct the QCD shower, while the last recombinations in the algorithm, those involving the largest-p_T degrees of freedom, may indicate whether the jet was produced by QCD alone or a heavy particle decay plus QCD showering. To discuss the details of jet substructure, we begin by defining relevant variables.

B. Variables describing branchings and their kinematics

In studying the substructure produced by jet algorithms, it will be useful to describe branchings using a set of kinematic variables. Since we will consider the substructure of (massive) jets reconstructing kinematic decays and of QCD jets, there are two natural choices of variables. Jet rest frame variables are useful to understand decays because the decay cross section takes a simple form. Lab frame variables are useful because jet algorithms are formulated in the lab frame, so algorithm systematics are most easily understood there. The QCD soft/collinear singularity structure is also easy to express in lab frame variables.

Naively, there are 12 variables completely describing a $1 \to 2$ splitting. Here we will focus on the top branching (the last merging) of the jet splitting into two daughter subjets, which we will label $J \to 1, 2$. Imposing the four constraints from momentum conservation to the branching leaves eight independent variables. The invariance of the algorithm metrics under longitudinal boosts and azimuthal rotations removes two of these (they are irrelevant). For simplicity we will use this invariance to set the jet's direction to be along the x axis, defining the z axis to be along the beam direction. Therefore there are six relevant variables needed to describe a $1 \to 2$ branching. Three of these variables are related to the three-momenta of the jet and subjets, and the other three are related to their masses.

Of the six variables, only one needs to be dimensionful, and we can describe all other scales in terms of this one. We choose the mass m_J of the jet. In addition, we use the masses of the two daughter subjets scaled by the jet mass:

$$a_1 \equiv \frac{m_1}{m_J} \quad \text{and} \quad a_2 \equiv \frac{m_2}{m_J}. \tag{2}$$

We choose the particle labeled by 1 to be the heavier particle, $a_1 > a_2$. The three masses, m_J, a_1, and a_2, will be common to both sets of variables. Additionally, we will typically want to fix the p_T of the jet and determine how the kinematics of a system change as p_{T_J} is varied. For QCD, a useful dimensionless quantity is the ratio of the mass and p_T of the jet, whose square we call x_J:

$$x_J \equiv \frac{m_J^2}{p_{T_J}^2}. \tag{3}$$

For decays, we will opt instead to use the familiar magnitude γ of the boost of the heavy particle from its rest frame to the lab frame, which is related to x_J by

$$\gamma = \sqrt{\frac{1}{x_J} + 1}, \qquad x_J = \frac{1}{\gamma^2 - 1}. \tag{4}$$

The remaining two variables, which are related to the momenta of the subjets, will differ between the rest frame and lab frame descriptions of the splitting.

Unpolarized $1 \to 2$ decays are naturally described in their rest frame by two angles. These angles are the polar and azimuthal angles of one particle (the heavier one, say) with respect to the direction of the boost to the lab frame, and we label them θ_0 and ϕ_0 respectively. Since we are

[2]The anti-k_T algorithm has the metrics $\rho_{ij} \equiv \min(p_{Ti}^{-1}, p_{Tj}^{-1})\Delta R_{ij}/D$, $\rho_i \equiv p_{Ti}^{-1}$, so it tends to cluster protojets with the hardest protojet, resulting in conelike jets with uninteresting substructure.

choosing that the final jet be in the $\hat{x}$ direction, θ_0 is measured from the $\hat{x}$ direction while ϕ_0 is the angle in the $y-z$ plane, which we choose to be measured from the $\hat{y}$ direction. Putting these variables together, the set that most intuitively describes a heavy particle decay is the ''rest frame'' set

$$\{m_J, a_1, a_2, \gamma, \cos\theta_0, \phi_0\}. \tag{5}$$

In the lab frame, we want to choose variables that are invariant under longitudinal boosts and azimuthal rotations. The angle ΔR_{12} between the daughter particles is a natural choice, as is the ratio of the minimum daughter p_T to the parent p_T, which is commonly called z:

$$z \equiv \frac{\min(p_{T_1}, p_{T_2})}{p_{T_J}}. \tag{6}$$

These variables make the recombination metrics for the k_T and CA algorithms simple:

$$\rho_{12}(k_T) = p_{T_J} z \Delta R_{12} \quad \text{and} \quad \rho_{12}(\mathrm{CA}) = \Delta R_{12}. \tag{7}$$

Note that for a generic recombination, the momentum factors in the denominator of Eq. (6) and in the k_T metric in Eq. (7) should be p_{Tp}, the momentum of the parent or combined subjet of the $2 \longrightarrow 1$ recombination.

From these considerations we choose to describe recombinations in the lab frame with the set of variables

$$\{m_J, a_1, a_2, x_J, z, \Delta R_{12}\}. \tag{8}$$

In using these variables it is essential to understand the structure of the corresponding phase space, especially for the last two variables in both sets. If we require that the decay ''fits'' in a jet, constraints and correlations appear. These are clearest in terms of the lab frame variables ΔR_{12} and z. As a first step in understanding these correlations, we plot in Fig. 1 the contour $\Delta R_{12} = D(= 1.0)$ in the $(\cos\theta_0, \phi_0)$ phase space for different values of γ and over different choices for a_1 and a_2. These specific values of a_1 and a_2 correspond to a variety of interesting processes: $a_1 = a_2 = 0$ gives the simplest kinematics and is therefore a useful starting point; $a_1 = 0.46$, $a_2 = 0$ gives the kinematics of the top quark decay; $a_1 = 0.9$, $a_2 = 0$ and $a_1 = 0.3$, $a_2 = 0.1$ are reasonable values for subjet masses from the CA and k_T algorithms, respectively. The contour $\Delta R_{12} = D$ defines the boundary in phase space where a $1 \longrightarrow 2$ process will no longer fit in a jet, with the interior region corresponding to splittings with $\Delta R_{12} < D$. Note that the contour is nearly vertical, increasingly so for larger γ. This is a reflection of the fact that ΔR_{12} is nearly independent of ϕ_0, up to terms suppressed by γ^{-2}.

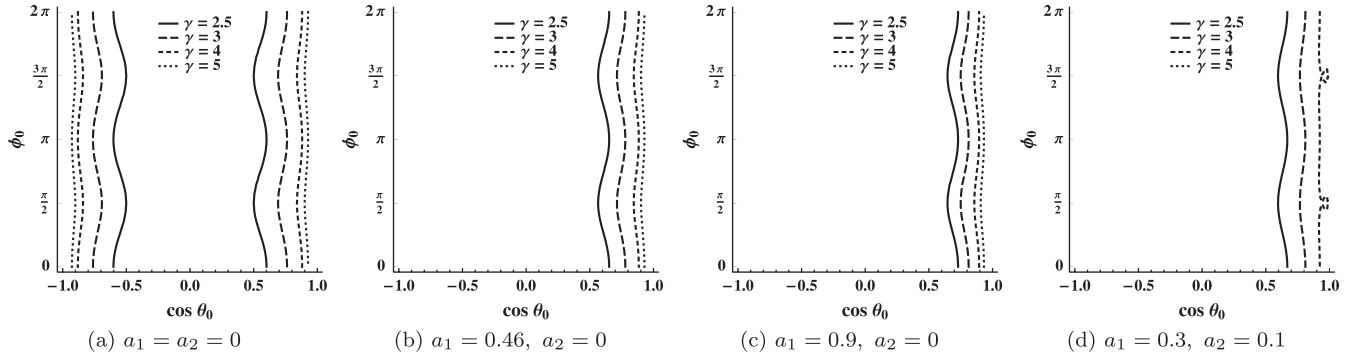
While the constraint $\Delta R_{12} < D$ becomes simpler in the $(z, \Delta R_{12})$ phase space, the boundaries of the phase space become more complex. In Fig. 2, we plot the available phase space in $(z, \Delta R_{12})$ for the same values of x_J, a_1, and a_2 as in Fig. 1, translating the value of γ into x_J. The most striking feature is that for fixed x_J, a_1, and a_2, the phase space in $(z, \Delta R_{12})$ is nearly one dimensional; this is again due to the fact that ΔR_{12} and also z are nearly independent of ϕ_0. In particular, for $a_1 = a_2 = 0$ [as in Fig. 2(a)], the phase space approximates the contour describing fixed x_J for small ΔR_{12}, which takes the simple form

$$x_J \equiv \frac{m_J^2}{p_{T_J}^2} \approx z(1-z)\Delta R_{12}^2. \tag{9}$$

This approximation is accurate even for larger angles, $\Delta R_{12} \approx 1$, at the 10% level. Note also that the width of the band about the contour described by Eq. (9) is itself of order x_J. As we decrease x_J the band moves down and becomes narrower as indicated in Fig. 2(a).

As illustrated in Figs. 2(b) and 2(d), we can also see a double-band structure to the $(z, \Delta R_{12})$ phase space. The upper band corresponds to the case where the lighter daughter is softer (smaller p_T) than the heavier daughter (and determines z), while the lower band corresponds to the case where the heavier daughter is softer. This does not occur in Fig. 2(a) because $a_1 = a_2$ (the single band is double covered), or in Fig. 2(c) because the heavier particle is never the softer one for the chosen values of x_J.

We have said nothing about the density of points in phase space for either pair of variables. This is because the weighting of phase space is set by the dynamics of a



(a) $a_1 = a_2 = 0$ (b) $a_1 = 0.46,\ a_2 = 0$ (c) $a_1 = 0.9,\ a_2 = 0$ (d) $a_1 = 0.3,\ a_2 = 0.1$

FIG. 1. Boundaries in the $\cos\theta_0 - \phi_0$ plane for a recombination step to fit in a jet of size $D = 1.0$, for several values of the boost γ and the subjet masses $\{a_1, a_2\}$. The ''interior'' region has $\Delta R_{12} < D$.

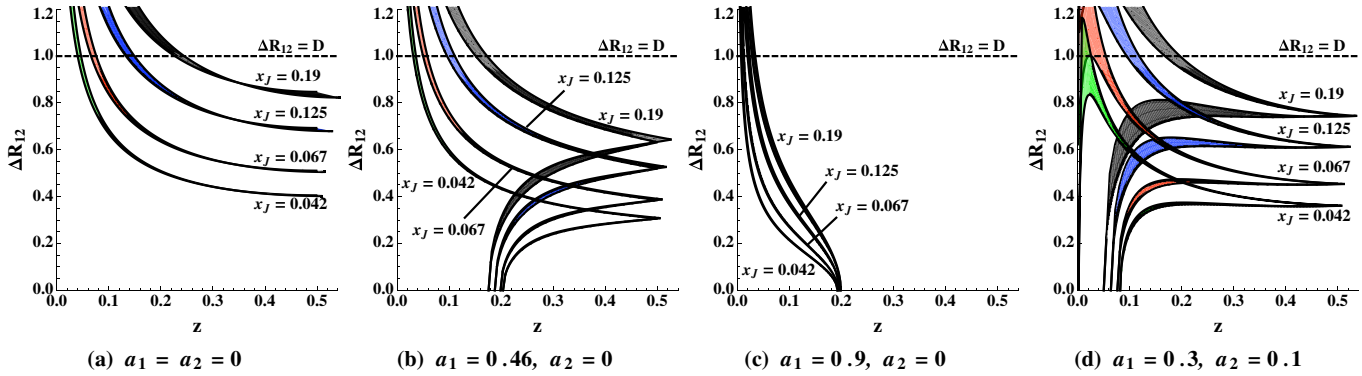


FIG. 2 (color online). Boundaries in the $z - \Delta R_{12}$ plane for a recombination step of fixed $\{a_1, a_2, x_J\}$, for various values of x_J and the subjet masses $\{a_1, a_2\}$. Configurations with $\Delta R_{12} < D$ fit in a jet; $D = 1.0$ is shown, for example.

process, while the boundaries are set by the kinematics. Decays and QCD splittings weight the phase space differently, as we will see in Secs. III and IV.

C. Ordering in recombination algorithms

Having laid out variables useful to describe $1 \to 2$ processes, we can discuss how the jet algorithm orders recombinations in these variables. Recombination algorithms merge objects according to the pairwise metric ρ_{ij}. The sequence of recombinations is almost always monotonic in this metric: as the algorithm proceeds, the value increases. Only certain kinematic configurations will decrease the metric from one recombination to the next, and the monotonicity violation is small and rare in practice.

This means it is straightforward to understand the typical recombinations that occur at different stages of the algorithm. We can think in terms of a phase space boundary: the algorithm enforces a boundary in phase space at a constant value of the recombination metric that evolves to larger values as the recombination process proceeds. If a recombination occurs at a certain value of the metric, ρ_0, then subsequent recombinations are very unlikely to have $\rho_{ij} < \rho_0$, meaning that this region of phase space is unavailable for further recombinations.

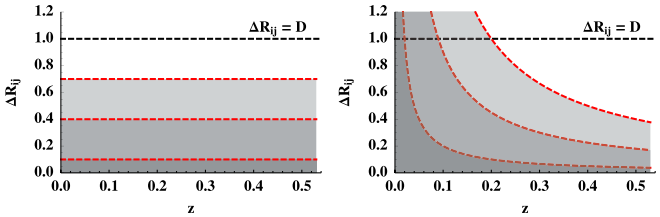


FIG. 3 (color online). Typical boundaries (red, dashed lines) on phase space due to ordering in the CA and k_T algorithms. The shaded region below the boundaries is cut out, and the more heavily shaded regions correspond to earlier in the recombination sequence. The cutoff $\Delta R_{ij} = D = 1.0$ is shown for reference (black, dashed lines).

In Fig. 3, we plot typical boundaries for the CA and k_T algorithms in the $(z, \Delta R_{12})$ phase space. For CA, these boundaries are simply lines of constant ΔR_{12}, since the recombination metric is $\rho_{ij}(\mathrm{CA}) = \Delta R_{ij}$. For k_T, these boundaries are contours in $z\Delta R_{12}$, and implicitly depend on the p_T of the parent particle in the splitting. Because the k_T recombination metric for $i, j \to p$ is $\rho_{ij}(k_T) = z\Delta R_{ij} p_{Tp}$, increasing the value of p_{Tp} will shift the boundary in to smaller $z\Delta R_{ij}$. These algorithm-dependent ordering effects will be important in understanding the restrictions on the kinematics of the last recombinations in a jet. For instance, we expect to observe no small-angle late recombinations in a jet defined by the CA algorithm.

Having considered some generic features of jet substructure and the systematic effects of recombination algorithms, we now explore how these effects combine with the underlying dynamics of QCD and heavy particle decays to produce the jets we observe.

III. SUBSTRUCTURE OF QCD JETS

The LHC will be the first collider where jet masses play a serious role in analyses. The proton-proton center of mass energy at the LHC is sufficiently large that the mass spectrum of QCD jets will extend far into the regime of heavy particle production (m_W and above). Because masses are such an important variable in jet substructure, masses of QCD jets will play an essential role in determining the effectiveness of jet substructure techniques at separating QCD jets from jets with new physics. We expect that the jet mass distribution in QCD is smoothly falling due to the lack of any intrinsic mass scale above Λ_{QCD}, while jets containing heavy particles are expected to exhibit enhancements in a relatively narrow jet mass range (given by the particle's width, detector effects, and the systematics of the algorithm).

Understanding the more detailed substructure of QCD jets (beyond the mass of the jet) presents an interesting challenge. QCD jets are typically characterized by the soft and collinear kinematic regimes that dominate their evo-

lution, but QCD populates the entire phase space of allowed kinematics. Because of its immense cross section relative to other processes, small effects in QCD can produce event rates that still dominate other signals, even after cuts. Furthermore, the full kinematic distributions in QCD jet substructure currently can only be approximately calculated, so we focus on understanding the key features of QCD jets and the systematic effects that arise from the algorithms that define them. Note that even when an on-shell heavy particle is present in a jet, the corresponding kinematic decay(s) will contribute to only a few of the branchings within the jet. QCD will still be responsible for the bulk of the complexity in the jet substructure, which is produced as the colored partons shower and hadronize, leading to the high multiplicity of color singlet particles observed in the detector.

It is a complex question to ask whether the jet substructure is accurately reconstructing the parton shower, and somewhat misguided, as the parton shower represents colored particles while the experimental algorithm only deals with color singlets. A more sensible question, and an answerable one, is to ask whether the algorithm is faithful to the dynamics of the parton shower. This is the basis of the metrics of the k_T and CA recombination algorithms—the ordering of recombinations captures the dominant kinematic features of branchings within the shower. In particular, the cross section for an extra real emission in the parton shower contains both a soft (z) and a collinear (ΔR) singularity:

$$d\sigma_{n+1} \sim d\sigma_n \frac{dz}{z} \frac{d\Delta R}{\Delta R}. \tag{10}$$

While these singularities are regulated (in perturbation theory) by virtual corrections, the enhancement remains, and we expect emissions in the QCD parton shower to be dominantly soft and/or collinear. Because of their different metrics, the k_T and CA algorithms will recombine these emissions differently, producing distinct substructure. In the next two subsections, we will discuss the interplay between the dynamics of QCD and the recombination algorithms, first using a toy analytic model, then with more realistic simulated events.

A. Jets in a toy QCD

To establish an intuitive level of understanding of jet substructure in QCD we consider a toy model description of jets in terms of a single branching and the variables x_J, z, and ΔR_{12}. We take the jet to have a fixed p_{T_J}. We combine the leading-logarithmic dynamics of Eq. (10) with the approximate expression for the jet mass in Eq. (9), and we label this combined approximation as the LL approximation. Recall that this approximation for the jet mass is useful for small subjet masses and small opening angles. From Sec. II B, recall that fixing x_J provides lower bounds on both z and ΔR_{12} and ensures finite results for the LL approximation. This approach leads to the following simple form for the x_J distribution:

$$\begin{aligned}\frac{1}{\sigma}\frac{d\sigma_{\mathrm{LL}}}{d(m_J^2/p_{T_J}^2)} &\equiv \frac{1}{\sigma}\frac{d\sigma_{\mathrm{LL}}}{dx_J} \\ &\sim \int_0^{1/2}\int_0^D \frac{dz}{z}\frac{d\Delta R_{12}}{\Delta R_{12}} \\ &\quad \times \delta(x_J - z(1-z)\Delta R_{12}^2) \\ &= \frac{-\ln(1-\sqrt{1-4x_J/D^2})}{2x_J}\Theta[D^2/4 - x_J].\end{aligned} \tag{11}$$

Note we are integrating over the phase space of Fig. 2(a), treating it as one dimensional. The resulting distribution is exhibited in Fig. 4 for $D = 1.0$ where we have multiplied by a factor of x_J to remove the explicit pole. We observe both the cutoff at $x_J = D^2/4$ arising from the kinematics discussed in Sec. II B and the $-\ln(x_J)/x_J$ small-x_J behavior arising from the singular soft/collinear dynamics. Even if the infrared singularity is regulated by virtual emissions and the distribution is resummed, we still expect QCD jet mass distributions (with fixed p_{T_J}) to be peaked at small mass values and be rapidly cutoff for $m_J > p_{T_J}D/2$.

We can improve this approximation somewhat by using the more quantitative perturbative analysis described in [1]. In perturbation theory jet masses appear at next-to-leading order (NLO) in the overall jet process where two (massless) partons can be present in a single jet. Strictly, the jet mass is then being evaluated at leading order (i.e., the jet mass vanishes with only one parton in a jet) and one would prefer a NNLO result to understand scale dependence (we take $\mu = p_{T_J}/2$). Here we will simply use the available NLO tools [23]. This approach leads to the very similar x_J distribution displayed in Fig. 5, plotted for two values of p_{T_J} (at the LHC, with $\sqrt{s} = 14$ TeV). We are correctly including the full NLO matrix element (not simply the singular parts), the full kinematics of the jet mass (not just the small-angle approximation), and the effects of

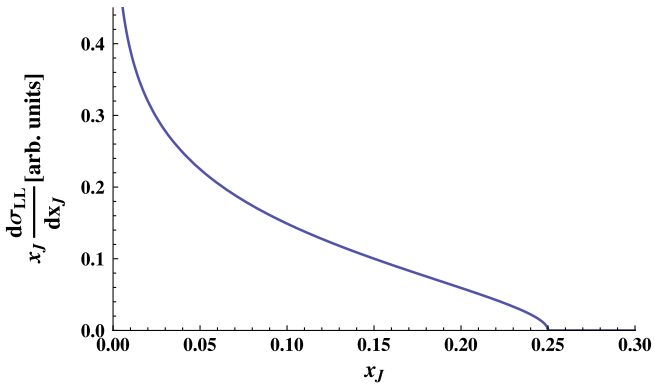


FIG. 4 (color online). Distribution in x_J for a simple LL toy model with $D = 1.0$.

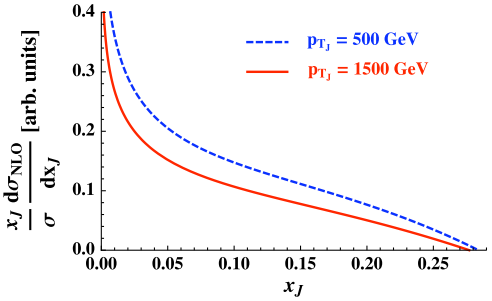


FIG. 5 (color online). NLO distribution in x_J for k_T-style QCD jets with $D = 1.0$, $\sqrt{s} = 14$ TeV, and two values of p_{T_J}.

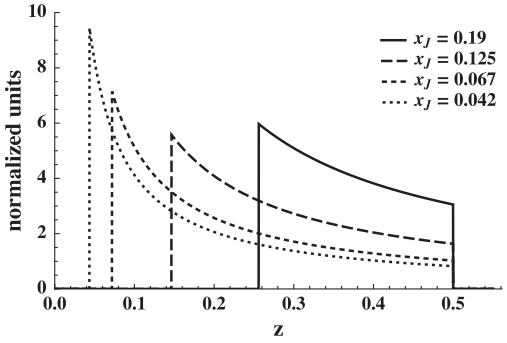


FIG. 6. Distribution in z for LL QCD jets for $D = 1.0$ and various values of x_J. The curves are normalized to have unit area.

the parton distribution functions. In this case the distribution is normalized by dividing by the Born jet cross section. Again we see the dominant impact of the soft/collinear singularities for small jet masses. Note also that there is little residual dependence on the value of the jet momentum and that again the distribution essentially vanishes for $x_J \gtrsim 0.25$, $m_J/p_{T_J} \gtrsim 0.5 = D/2$. The average jet mass suggested by these results is $\langle m_J/p_{T_J}\rangle \approx 0.2D$. Because the jet only contains two partons at NLO, we are still ignoring the effects of the nonzero subjet masses and the effects of the ordering of mergings imposed by the algorithm itself. For example, at this order there is no difference between the CA and k_T algorithms.

Next we consider the z and ΔR_{12} distributions for the LL approximation where a single recombination of two (massless) partons is required to reconstruct as a jet of definite p_{T_J} and mass (fixed x_J). To that end we can undo one of the integrals in Eq. (11) and consider the distributions for z and ΔR_{12}. We find for the z distribution the form

$$\frac{1}{\sigma}\frac{d\sigma_{\rm LL}}{dx_J dz} \sim \frac{1}{2zx_J}\Theta\left[z - \frac{1-\sqrt{1-4x_J/D^2}}{2}\right]\Theta\left[\frac{1}{2} - z\right]. \tag{12}$$

As expected, we see the poles in z and x_J from the soft/collinear dynamics, but, as in Sec. II B, the constraint of fixed x_J yields a lower limit for z. Recall that the upper limit for z arises from its definition, again applied in the small-angle limit. Thus the LL QCD distribution in z is peaked at the lower limit but the characteristic turn-on point is fixed by the kinematics, requiring the branching at fixed x_J to be in a jet of size D. This behavior is illustrated in Fig. 6 for various values of $x_J = 1/(\gamma^2 - 1)$ corresponding to those used in Sec. II B.

The expression for the ΔR_{12} dependence in the LL approximation is

$$\frac{1}{\sigma}\frac{d\sigma_{\rm LL}}{dx_J d\Delta R_{12}} \sim \frac{2}{\Delta R_{12}^2}\frac{\Theta[\Delta R_{12} - 2\sqrt{x_J}]\Theta[D - \Delta R_{12}]}{\sqrt{\Delta R_{12}^2 - 4x_J}(1-\sqrt{1-4x_J/\Delta R_{12}^2})}. \tag{13}$$

This distribution is illustrated in Fig. 7 for the same values of x_J as in Fig. 6. As with the z distribution the kinematic constraint of being a jet with a definite x_J yields a lower limit, $\Delta R_{12} \gtrsim 2\sqrt{x_J}$, along with the expected upper limit, $\Delta R_{12} \leq D$. However, for ΔR_{12} the change of variables also introduces an (integrable) square root singularity at the lower limit. This square root factor tends to be numerically more important than the $1/\Delta R_{12}^2$ factor.[3] Since this square root singularity arises from the choice of variable (a kinematic effect), we will see that it is also present for heavy particle decays, suggesting that the ΔR_{12} variable will not be as useful as z in distinguishing QCD jets from heavy particle decay jets.

Thus, in our toy QCD model with a single recombination, leading-logarithm dynamics and the small-angle jet mass definition, the constraints due to fixing x_J tend to dominate the behavior of the z and ΔR_{12} distributions, with limited dependence on the QCD dynamics and no distinction between the CA and k_T algorithms. However, this situation changes dramatically when we consider more realistic jets with full showering, a subject to which we now turn.

B. Jet substructure in simulated QCD events

To obtain a more realistic understanding of the properties of QCD jet masses we now consider jet substructure that arises in more fully simulated events. In particular, we focus on Monte Carlo QCD jets with transverse momenta in the range $p_{T_J} = 500$–700 GeV ($c = 1$ throughout this paper) found in matched QCD multijet samples, created as described in the Appendix. The matching process means that we are including, to a good approximation, the full NLO perturbative probability for energetic, large-angle emissions in the simulated showers, and not just the soft and collinear terms. As suggested earlier, we anticipate two important changes from the previous discussion. First, the showering ensures that the daughter subjets at the last

[3]One factor of ΔR_{12} arises from the collinear QCD dynamics while the other comes from a change of variables. The soft QCD singularity is contained in the denominator factor $(1-\sqrt{1-4x_J/\Delta R_{12}^2}) \to 2z$ for $x_J \ll \Delta R^2$ (equivalently, $z \ll 1$).

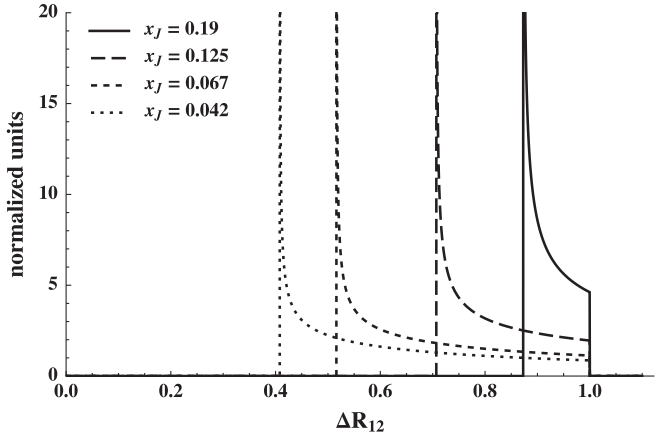


FIG. 7. Distribution in ΔR_{12} for LL QCD jets for $D = 1.0$ and various values of x_J. The curves are normalized to have unit area.

recombination have nonzero masses. More importantly and as noted in Sec. II C, the sequence of recombinations generated by the jet algorithm tends to force the final recombination into a particular region of phase space that depends on the recombination metric of the algorithm. For the CA algorithm this means that the final recombination will tend to have a value of ΔR_{12} near the limit D, while the k_T algorithm will have a large value of $z\Delta R_{12} p_{T_J}$. This issue will play an important role in explaining the observed z and ΔR_{12} distributions.

First, consider the jet mass distributions from the simulated event samples. In Fig. 8, we plot the jet mass distributions for the k_T and CA algorithms for all jets in the stated p_T bin (500–700 GeV). As expected, for both algorithms the QCD jet mass distribution smoothly falls from a peak only slightly displaced from zero [the remnant of the perturbative $-\ln(m^2)/m^2$ behavior]. There is a more rapid cutoff for $m_J > p_{T_J} D/2$, which corresponds to the expected kinematic cutoff from the LL approximation, but smeared by the nonzero width of the p_T bin, the nonzero subjet masses and the other small corrections to the LL

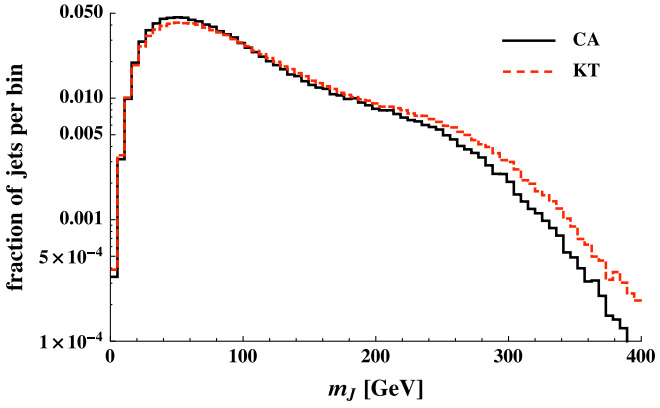


FIG. 8 (color online). Distribution in m_J for QCD jets with p_T between 500 and 700 GeV with $D = 1.0$.

approximation. The average jet mass, $\langle m_J \rangle \approx 100$ GeV, is in crude agreement with the perturbative expectation $\langle m_J / p_{T_J} \rangle \approx 0.2$. Note that the two algorithms now differ somewhat in that the k_T algorithm displays a slightly larger tail at high masses. As we will see in more detail below, this distinction arises from the difference in the metrics leading to recombining protojets over a slightly larger angular range in the k_T algorithm. On the other hand, the two curves are remarkably similar. Note that we have used a logarithmic scale to ensure that the difference is apparent. Without the enhanced number of energetic, large-angle emissions characteristic of this *matched* sample, the distinction between the two algorithms is much smaller, i.e., a typical dijet, LO Monte Carlo sample yields more similar distributions for the two algorithms.

Other details of the QCD jet substructure are substantially more sensitive to the specific algorithm than the jet mass distribution. To illustrate this point we will discuss the distributions of z, ΔR_{12}, and the subjet masses for the last recombination in the jet. We can understand the observed behavior by combining a simple picture of the geometry of the jet with the constraints induced on the phase space for a recombination from the jet algorithm. In particular, recall that the ordering of recombinations defined by the jet algorithm imposes relevant boundaries on the phase space available to the late recombinations (see Fig. 3).

While the details of how the k_T and CA algorithms recombine protojets within a jet are different, the overall structure of a large-p_T jet is set by the shower dynamics of QCD, i.e., the dominance of soft/collinear emissions. Typically the jet has one (or a few) hard core(s), where a hard core is a localized region in $y - \phi$ with large energy deposition. The core is surrounded by regions with substantially smaller energy depositions arising from the radiation emitted by the energetic particles in the core (i.e., the shower), which tend to dominate the area of the jet. In particular, the periphery of the jet is occupied primarily by the particles from soft radiation, since even a wide-angle hard parton will radiate soft gluons in its vicinity. This simple picture leads to very different recombinations with the k_T and CA algorithms, especially the last recombinations.

The CA algorithm orders recombinations only by angle and ignores the p_T of the protojets. This implies that the protojets still available for the last recombination steps are those at large angle with respect to the core of the jet. Because the core of the jet carries large p_T, as the recombinations proceed the directions of the protojets in the core do not change significantly. Until the final steps, the recombinations involving the soft, peripheral protojets tend to occur only locally in $y - \phi$ and do not involve the large-p_T protojets in the core of the jet. Therefore, the last recombinations defined by the CA algorithm are expected to involve two very different protojets. Typically

one has large p_T, carrying most of the four-momentum of the jet, while the other has small p_T and is located at the periphery of the jet. The last recombination will tend to exhibit large ΔR_{12}, small z, large a_1 (near 1), and small a_2, where the last two points follow from the small z and correspond to the $(z, \Delta R_{12})$ phase space of Fig. 2(c).

In contrast, the k_T algorithm orders recombinations according to both p_T and angle. Thus the k_T algorithm tends to recombine the soft protojets on the periphery of the jet earlier than with the CA algorithm. At the same time, the reduced dependence on the angle in the recombination metric implies the angle between protojets for the final recombinations will be lower for k_T than CA. While there is still a tendency for the last recombination in the k_T algorithm to involve a soft protojet with the core protojet, the soft protojet tends to be not as soft as with the CA algorithm (i.e., the z value is larger), while the angular separation is smaller. Since this final soft protojet in the k_T algorithm has participated in more previous recombinations than in the CA case, we expect the average a_2 value to be farther from zero and the a_1 value to be farther from 1. Generally the $(z, \Delta R_{12})$ phase space for the final k_T recombination is expected to be more like that illustrated in Figs. 2(b) and 2(d) [coupled with the boundary in Fig. 3(b)].

To illustrate this discussion we have plotted distributions of z, ΔR_{12}, and a_1 for the last recombination in a jet for the k_T and CA algorithms in Fig. 9 for the matched QCD sample described previously. We plot distributions with and without a cut on the jet mass, where the cut is a narrow window ($\approx$ 15 GeV) around the top quark mass. This cut selects heavy QCD jets, and for the p_T window of 500–700 GeV it corresponds to a cut on x_J of 0.06–0.12. These distributions reflect the combined influence of the QCD shower dynamics, the restricted kinematics from being in a jet, and the algorithm-dependent ordering effects discussed above. Most importantly, note the very strong enhancement at the smallest values of z for the CA algorithm in Fig. 9(a), which persists even after the heavy jet mass cut. Note the log scale in Fig. 9(a). While the k_T result in Fig. 9(b) is still

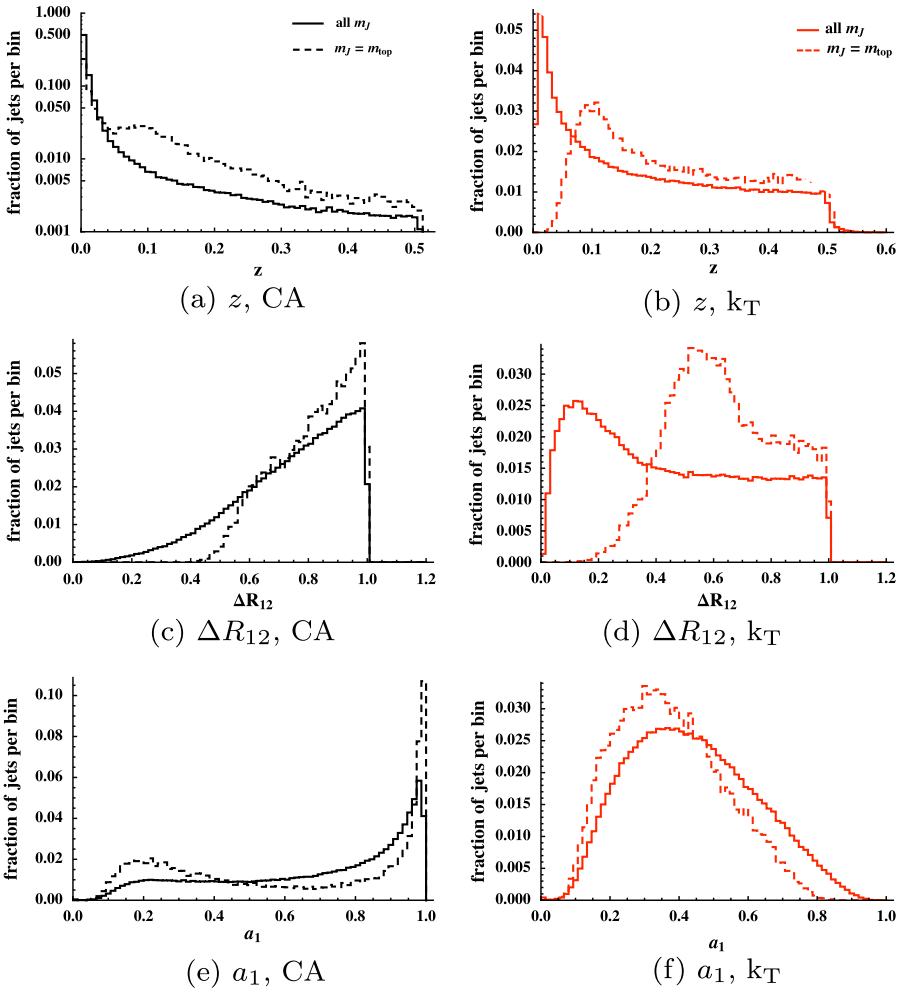


FIG. 9 (color online). Distribution in z, ΔR_{12}, and the scaled (heavier) daughter mass a_1 for QCD jets, using the CA and k_T algorithms, with (dashed lines) and without (solid lines) a cut around the top quark mass. The jets have p_T between 500 and 700 GeV with $D = 1.0$. Note the log scale for the z distribution of CA jets.

peaked near zero when summed over all jet masses, the enhancement is not nearly as strong. After the heavy jet mass cut is applied, the distribution shifts to larger values of z, with an enhancement remaining at small values. Only in this last plot is there evidence of the lower limit on z of order 0.1 expected from the earlier LL approximation results.

Figure 9(c) illustrates the expected enhancement near $\Delta R_{12} = D = 1.0$ for CA. Figure 9(d) shows that k_T exhibits a much broader distribution than CA with an enhancement for small ΔR_{12} values. Once the heavy jet mass cut is applied, both algorithms exhibit the lower kinematic cutoff on ΔR_{12} suggested in the LL approximation results, as both distributions shift to larger values of the angle. This shift serves to enhance the CA peak at the upper limit and moves the lower end enhancement in k_T to substantially larger values of ΔR_{12}.

The CA algorithm bias toward large a_1 is demonstrated in Fig. 9(e). We can see that requiring a heavy jet enhances the large-a_1 peak. The k_T distribution in a_1, shown in Fig. 9(f), exhibits a broad enhancement around $a_1 \approx 0.4$. This distribution is relatively unchanged after the jet mass cut. To give some insight into the correlations between z and ΔR_{12}, in Fig. 10 we plot the distribution of both variables simultaneously for both algorithms, with no jet mass cut applied. The very strong enhancement at small z and large ΔR_{12} for CA is evident in this plot. For k_T, there is still an enhancement at small z and large ΔR_{12}, but there is support over the whole range in z and ΔR_{12} with the impact of the shaping due to the $z \times \Delta R_{12}$ dependence in the metric clearly evident. Note that the k_T distribution is closer to what one would expect from QCD alone, with enhancements at *both* small z and small ΔR_{12}, while the CA distribution is asymmetrically shaped away from the QCD-like result. Finally we should recall, as indicated by Fig. 8, that the jets found by the two algorithms tend to be slightly different, with the k_T algorithm recombining slightly more of the original (typically soft) protojets at the periphery and leading to slightly larger jet masses.

Because the QCD shower is present in all jets, and is responsible for the complexity in the jet substructure, the systematic effects discussed above will be present in all jets. While the kinematics of a heavy particle decay is distinct from QCD in certain respects, we will find that these effects still present themselves in jets containing the decay of a heavy particle. This reduces our ability to identify jets containing a heavy particle, and will lead us to propose a technique to reduce them. In the following section, we study the kinematics of heavy particle decays and discuss where these systematic effects arise.

IV. SUBSTRUCTURE OF HEAVY PARTICLE JETS

Recombination algorithms have the potential to reconstruct the decay of a heavy particle. Ideally, the substructure of a jet may be used to identify jets coming from a decay and reject the QCD background to those jets. In this section, we investigate a pair of unpolarized parton-level decays, a heavy particle decaying into two massless quarks (a $1 \rightarrow 2$ decay) and a top quark decay into three massless quarks (a two-step decay). For each decay, we study the available phase space in terms of the lab frame variables ΔR_{12} and z and the shaping of kinematic distributions imposed by the requirement that the decay be reconstructed in a single jet. We will determine the kinematic regime where decays are reconstructed, and contrast this with the kinematics for a $1 \rightarrow 2$ splitting in QCD.

A. $1 \rightarrow 2$ decays

We begin by considering a $1 \rightarrow 2$ decay with massless daughters. An unpolarized decay has a simple phase space in terms of the rest frame variables $\cos\theta_0$ and ϕ_0:

$$\frac{d^2N_0}{d\cos\theta_0 d\phi_0} = \frac{1}{4\pi}. \tag{14}$$

Recall from Sec. II B that $\cos\theta_0$ and ϕ_0 are the polar and azimuthal angles of the heavier daughter particle in the parent particle rest frame relative to the direction of the

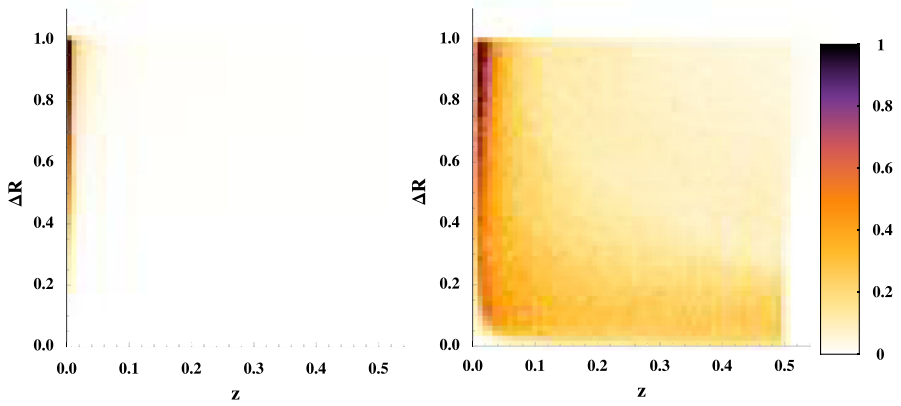


FIG. 10 (color online). Combined distribution in z and ΔR_{12} for QCD jets, using the CA (left panel) and k_T (right panel) algorithms, for jets with p_T between 500 and 700 GeV with $D = 1.0$. Each bin represents a *relative* density, normalized to 1 for the largest bin.

boost to the lab frame. In general, we will use N_0 to label the distribution of *all* decays, while N will label the distribution of decays *reconstructed* inside a single jet. N_0 is normalized to unity, so that for any variable set Φ,

$$\int d\Phi \frac{dN_0}{d\Phi} = 1. \tag{15}$$

The distribution N is defined from N_0 by selecting those decays that fit in a single jet, so that generically

$$\frac{dN}{d\Phi} \equiv \int d\Phi' \frac{dN_0}{d\Phi'} \delta(\Phi' - \Phi)\Theta(\text{single jet reconstruction}). \tag{16}$$

N is naturally normalized to the total fraction of reconstructed decays. The constraints of single jet reconstruction will depend on the decay and on the jet algorithm used, and abstractly take the form of a set of Θ functions. For a $1 \to 2$ decay and a recombination-type algorithm, the only constraint is that the daughters must be separated by an angle less than D:

$$\Delta R_{12} < D. \tag{17}$$

Since the kinematic limits imposed by reconstruction are sensitive to the boost γ of the parent particle, we will want to consider the quantities of interest at a variety of γ values. To illustrate this γ dependence, we first find the total fraction of all decays that are reconstructed in a single jet for a given value of the boost. We call this fraction $f_R(\gamma)$:

$$f_R(\gamma) \equiv \int d\cos\theta_0 d\phi_0 \frac{d^2N_0}{d\cos\theta_0 d\phi_0} \Theta(D - \Delta R_{12}). \tag{18}$$

In Fig. 11, we plot $f_R(\gamma)$ vs γ for several values of D. The reconstruction fraction rises rapidly from no reconstruction to nearly complete reconstruction in a narrow range in γ. This indicates that ΔR_{12} is strongly dependent on γ for fixed $\cos\theta_0$ and ϕ_0, which we will see below. Conversely, the minimum boost necessary for a decay to fit in a jet depends strongly on D. The turn-on for increasing γ is the same effect as the $(z, \Delta R_{12})$ phase space moving into the allowed region below $\Delta R_{12} = D$ in Fig. 2(a) as x_J is reduced.

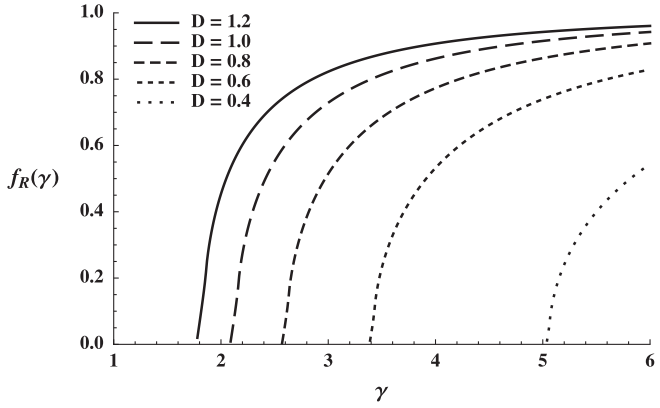


FIG. 11. Reconstruction fractions $f_R(\gamma)$ as a function of γ for various D.

To better understand the effect that reconstruction has on the phase space for decays, we would like to find the distribution of $1 \to 2$ decays in terms of lab frame variables,

$$\frac{d^2N_0}{dz d\Delta R_{12}}. \tag{19}$$

With two massless daughters, ΔR_{12} is given in terms of rest frame variables by

$$\Delta R_{12}^2 = \left[\tanh^{-1}\left(\frac{2\gamma \sin\theta_0 \sin\phi_0}{\sin^2\theta_0(\beta^2\gamma^2 + \sin^2\phi_0) + 1}\right)\right]^2 + \left[\tan^{-1}\left(\frac{2\beta\gamma \sin\theta_0 \cos\phi_0}{\sin^2\theta_0(\beta^2\gamma^2 + \sin^2\phi_0) - 1}\right)\right]^2 \tag{20}$$

with $\beta \equiv \sqrt{1 - \gamma^{-2}}$. This relation is analytically noninvertible, meaning we cannot write the Jacobian for the transformation

$$\frac{d^2N_0}{d\cos\theta_0 d\phi_0} \to \frac{d^2N_0}{dz d\Delta R_{12}} \tag{21}$$

in closed form. However, ΔR_{12} has some simple limits. In particular, when the boost γ is large, to leading order in γ^{-1},

$$\Delta R_{12} = \frac{2}{\gamma \sin\theta_0} + \mathcal{O}(\gamma^{-3}). \tag{22}$$

This limit is only valid for $\sin\theta_0 \gtrsim \gamma^{-1}$, but as we will see this is the region of phase space where the decay will be reconstructed in a single jet. The large-boost approximation describes the key features of the kinematics and is useful for a simple picture of kinematic distributions when particles are reconstructed in a single jet.

Since $\gamma = \sqrt{1 + 1/x_J}$, this limit is equivalent to the small-angle limit we took in Sec. III A. [For $\Delta R^2 \ll 1$, $x_J \approx z(1 - z)\Delta R^2 \ll 1$.] We can see this in Eq. (20), where $\Delta R \approx 1/\gamma$.

The value of z is also simple in the large-boost approximation. In this limit,

$$z = \frac{1 - |\cos\theta_0|}{2} + \mathcal{O}(\gamma^{-2}). \tag{23}$$

With the large-boost approximation, z and ΔR_{12} are both independent of ϕ_0. As noted earlier both ΔR_{12} and z depend on ϕ_0 only through terms that are suppressed by inverse powers of γ (cf. Figs. 1 and 2). In this limit we can integrate out ϕ_0 and find the distributions in z and ΔR_{12} for all decays. For z the distribution is simply flat:

$$\frac{dN_0}{dz} \approx 2\Theta\left(\frac{1}{2} - z\right)\Theta(z). \tag{24}$$

We have included the limits for clarity. For ΔR_{12}, the

distribution is

$$\frac{dN_0}{d\Delta R_{12}} \approx \frac{4}{\gamma^2 \Delta R_{12}^2} \frac{\Theta(\Delta R_{12} - 2\gamma^{-1})}{\sqrt{\Delta R_{12}^2 - 4\gamma^{-2}}}. \quad (25)$$

This distribution has a lower cutoff requiring $\Delta R_{12} \geq 2\gamma^{-1}$. This is close to the true lower limit on ΔR_{12}, $\Delta R_{12} \geq 2\csc^{-1}\gamma$. Note that in Eq. (25), there is an enhancement at the lower cutoff in ΔR_{12} due to the square root singularity arising from the change of variables, just as there was in the QCD result in Eq. (14).

In Fig. 12, we plot the exact distribution dN_0/dz, found numerically, for several values of γ. The true distribution is qualitatively similar to the approximate one in Eq. (24), which is flat. The peak in the distribution at small z values comes from the reduced phase space as $z \to 0$, and the peak is lower for larger boosts. In Fig. 13, we plot the exact distribution $dN_0/d\Delta R_{12}$, which is again qualitatively similar to the large-boost result. The distribution in ΔR_{12} is localized at the lower limit, especially for larger boosts. This provides a useful rule: the opening angle of a decay is strongly correlated with the transverse boost of the parent particle. Note that the relevant boost is the transverse one because the angular measure ΔR is invariant under longitudinal boosts (recall that in the example here, we have set the parent particle to be transverse).

The constraint imposed by reconstruction is simple in the large-boost approximation. In terms of $\sin\theta_0$, the constraint $\Delta R_{12} < D$ requires $\sin\theta_0 > 2/\gamma D$, which excludes the region where the approximation breaks down. Therefore the large-boost approximation is apt for describing the kinematics of a reconstructed decay. In Fig. 14, we plot the distribution, $dN/d\cos\theta_0$, where the implied sharp cutoff is apparent [and should be compared to what we observed in Fig. 1(a)]. This distribution is easy to understand in the rest frame of the decay. When $|\cos\theta_0|$ is close to 1, one of the daughters is nearly collinear with the direction of the boost to the lab frame, and the other is nearly anticollinear. The anticollinear daughter is not sufficiently boosted to have $\Delta R_{12} < D$ with the collinear daughter, and the parent particle is not reconstructed. As $|\cos\theta_0|$ decreases, the two daughters can be recombined in the same jet; this transition is rapid because the ϕ_0 dependence of the kinematics is small. We now look at the distributions of z and ΔR_{12} when we require reconstruction.

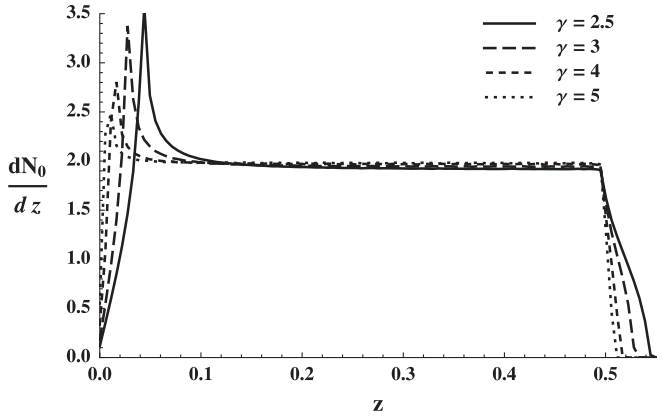


FIG. 12. The distribution of all decays in z for several values of γ.

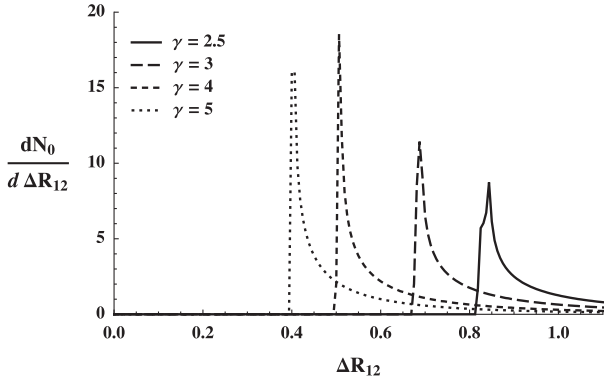


FIG. 13. The distribution of all decays in ΔR_{12} for several values of γ.

Because z is linearly related to $\cos\theta_0$ at large boosts, the distribution in z has a simple form:

$$\frac{dN}{dz} \approx 2\Theta\left(z - \frac{1 - \sqrt{1 - 4/(\gamma^2 D^2)}}{2}\right)\Theta\left(\frac{1}{2} - z\right). \quad (26)$$

Comparing to Eq. (24), we see that requiring reconstruction simply cuts out the region of phase space at small z. This is confirmed in the exact distribution dN/dz, shown in Fig. 15. The small-z decays that are not reconstructed come from the regions of phase space with $|\cos\theta_0|$ near 1, just as in the previous discussion. In these decays, the backward-going (anticollinear) daughter is boosted to have small p_T in the lab frame. Comparing to Fig. 6, the distribution in z for QCD splittings, we see first that the cutoffs on the distributions are similar (they are not identical because of the LL approximation used in Fig. 6). However, the QCD

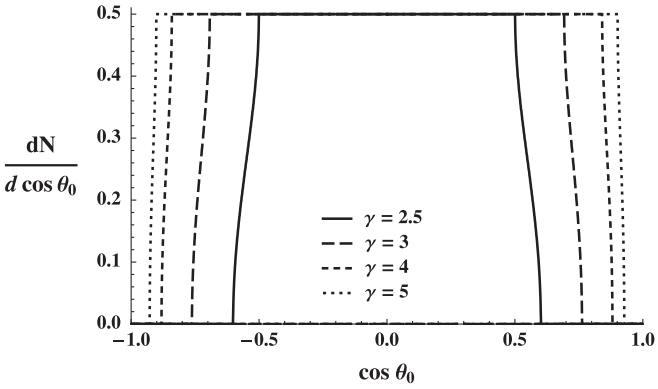


FIG. 14. The reconstructed distribution $dN/d\cos\theta_0$ with $D = 1.0$ for various values of γ.

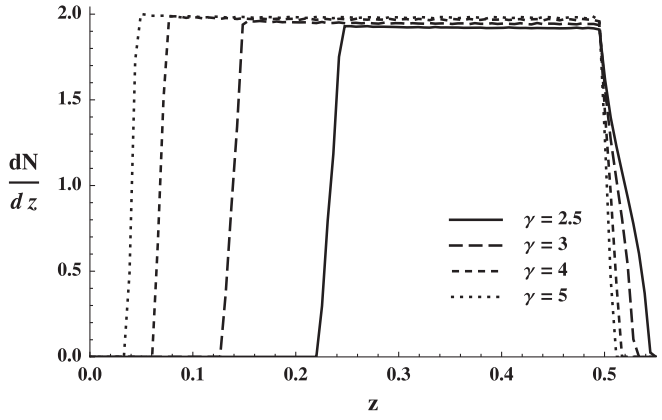


FIG. 15. The distribution of reconstructed decays in z for several values of γ.

distribution has an enhancement at small-z values, due to the QCD soft singularity, that the distribution for reconstructed decays does not exhibit.

The distribution of reconstructed particles in the variable ΔR_{12} is related simply to the distribution of all decays in the same variable:

$$\frac{dN}{d\Delta R_{12}} = \frac{dN_0}{d\Delta R_{12}}\Theta(D - \Delta R_{12}), \tag{27}$$

which means that the distribution $dN/d\Delta R_{12}$ is given by Fig. 13 with a cutoff at $\Delta R_{12} = D$. Note that this distribution is very close in shape to the distribution of QCD branchings versus ΔR_{12} displayed in Eq. (14) and Fig. 7. This similarity arises from the fact that the most important factor in the shape is the square root singularity, which arises from the change of variables in both cases and hides the underlying differences in dynamics.

B. Two-step decays

We now turn our attention to two-step decays, which exhibit a more complex substructure. Two-step decays offer new insights into the ordering effects of the k_T and CA algorithms, highlight the shaping effects from the algorithm on the jet substructure, and offer a surrogate for the cascade decays that are often featured in new physics scenarios. Even at the parton level the choice of jet algorithm matters in reconstructing a multistep decay; different algorithms can give different substructure. In studying this substructure we take the same approach as for the $1 \to 2$ decay, translating the simple kinematics of a parton-level decay into the lab frame variables ΔR_{12} and z.

The top quark is a good example of a two-step decay, and we focus on it in this section. We will label the top quark decay $t \to Wb$, with $W \to qq'$. In this discussion requiring that the top quark be reconstructed means that the W must be recombined from q and q' first, then merged with the b. The observed (3-parton) ''jet'' will then have the W as one of its daughter subjets.

For the k_T algorithm, reconstructing the top quark in a single jet imposes the following constraints on the partons:

$$\begin{aligned} \min(p_{Tq}, p_{Tq'})\Delta R_{qq'} &< \min(p_{Tq}, p_{Tb})\Delta R_{bq}, \\ \min(p_{Tq}, p_{Tq'})\Delta R_{qq'} &< \min(p_{Tq'}, p_{Tb})\Delta R_{bq'}, \\ \Delta R_{qq'} &< D, \\ \text{and} \quad \Delta R_{bW} &< D. \end{aligned} \tag{28}$$

For the CA algorithm the relations are strictly in terms of the angle:

$$\begin{aligned} \Delta R_{qq'} &< \Delta R_{bq}, \\ \Delta R_{qq'} &< \Delta R_{bq'}, \\ \Delta R_{qq'} &< D, \\ \text{and} \quad \Delta R_{bW} &< D. \end{aligned} \tag{29}$$

The kinematic limits requiring the decay to be reconstructed in a single jet are the same for the two algorithms, but fixing the ordering of the two recombinations requires a different restriction for each algorithm, which in turn biases the distributions of kinematic variables.

The common requirements such that the top quark be reconstructed in a single jet, $\Delta R_{qq'} < D$ and $\Delta R_{Wb} < D$, are straightforward to understand in terms of the rest frame variable $\cos\theta_0$, which here is the polar angle in the top quark rest frame between the W and the boost direction to the lab frame. For $\cos\theta_0 \approx 1$, the W has a large transverse boost in the lab frame, so $\Delta R_{qq'} < D$, but the angle between the W and b will be large (as was the case for the corresponding $1 \to 2$ decay in the previous section). For $\cos\theta_0 \approx -1$, the W transverse boost is small, and $\Delta R_{qq'}$ will be large. Therefore, we only expect to reconstruct top quarks in a single jet when $|\cos\theta_0|$ is not near 1.

If the CA algorithm correctly reconstructs the top quark, the two quarks from the W decay must be the closest pair (in ΔR) of the three final-state particles. This requirement strongly selects for decays where the W opening angle, $\Delta R_{qq'}$, is smaller than the top quark opening angle, ΔR_{Wb}. Therefore, only decays with a large (transverse) W boost will be reconstructed by the CA algorithm. In terms of $\cos\theta_0$, the fraction of decays that are reconstructed will increase as we increase $\cos\theta_0$ toward the upper limit where $\Delta R_{Wb} \geq D$, and the reconstruction fraction will be small for lower values of $\cos\theta_0$.

The k_T algorithm orders recombinations by p_T as well as angle, and the set of reconstructed decays is understood most easily by contrasting with CA. As the transverse boost of the W decreases, on average the p_T of the q and q' decrease while the p_T of the b increases. Therefore, while $\Delta R_{qq'}$ is increasing, $\min(p_{Tq}, p_{Tq'})$ is decreasing, and these competing effects suggest that k_T reconstructs decays with smaller values of $\cos\theta_0$ than CA, and that the dependence on $\cos\theta_0$ is not as strong.

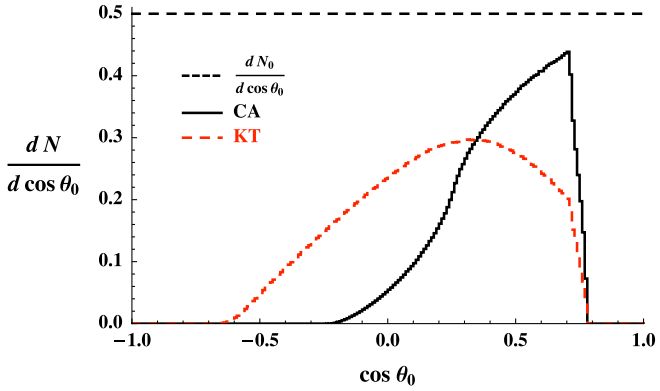


FIG. 16 (color online). $dN/d\cos\theta_0$ vs $\cos\theta_0$, with $\gamma = 3$, for both the k_T and CA algorithms. The underlying distribution $dN_0/d\cos\theta_0 = 1/2$ is plotted as the dotted line for reference.

The effect of the CA and k_T algorithms on the observed distribution in $\cos\theta_0$ is shown in Fig. 16, where we plot the distribution of $\cos\theta_0$ for reconstructed top quarks for both algorithms. The top boost is fixed to $\gamma = 3$. We observe the kinematic limit near $\cos\theta_0 \approx 0.8$ is common between algorithms, and that $\cos\theta_0 \approx -1$ is not accessed by either algorithm. As expected, the distribution for the CA algorithm falls off more sharply than for k_T at lower values of $\cos\theta_0$.

Next, we look at distributions in z and ΔR_{Wb}. Just as in the $1 \rightarrow 2$ decay, we expect decays with small z not to be correctly reconstructed. Small values of z will come when the W or b is soft, and therefore produced very backward going in the top rest frame. This corresponds to $\cos\theta_0 \approx \pm 1$, and from Fig. 16 these decays are not reconstructed. In Fig. 17, we plot the distribution in z for all decays, dN_0/dz, and the distribution for reconstructed decays, dN/dz, for a boost of $\gamma = 3$.

In dN_0/dz, the discontinuity at $z \approx 0.2$ arises from the fact that the W is sometimes softer than the b, but has a minimum p_T. The extra weight in dN_0/dz for z above this value comes from the decays where the W is softer than the b. Note that these decays are rarely reconstructed, especially for CA: the distribution dN/dz is smooth, and has little additional support in the region where the W is softer. This correlates with the fact that decays with negative $\cos\theta_0$ values are rarely reconstructed with CA, but more frequently with k_T. The distribution dN/dz has a lower cutoff that corresponds to the upper cutoff in Fig. 16. As the boost γ of the top increases, the cutoff at small z decreases, since the limit in $\cos\theta_0$ for which $\Delta R_{Wb} > D$ will increase toward 1.

The opening angle ΔR_{Wb} of the top quark decay also illustrates how strongly the kinematics are shaped by the jet algorithm. When $\cos\theta_0 \approx -1$, for sufficient boosts ΔR_{Wb} is small because the W is boosted forward in the lab frame, but these decays are not reconstructed because the ordering of recombinations will typically be incorrect and the W decay may not have $\Delta R_{qq'} < D$. For $\cos\theta_0 \approx 1$, ΔR_{Wb} will exceed D and the top will not be reconstructed. In Fig. 18, we plot the distribution $dN_0/d\Delta R_{Wb}$ of the angle between the W and b in all top decays for a top boost of $\gamma = 3$, as well as the distribution $dN/d\Delta R_{12}$ of the angle of the last recombination for reconstructed top quarks with the k_T and CA algorithms. Note that when the top quark is reconstructed at the parton level, $\Delta R_{12} = \Delta R_{Wb}$. The difference in $dN/d\Delta R_{12}$ between the k_T and CA algorithms reflects their different recombination orderings. Because CA orders strictly by angle, the angle ΔR_{12} tends to be larger than for k_T because CA requires $\Delta R_{12} = \Delta R_{Wb} > \Delta R_{qq'}$.

C. Hadron-level top quark jets

To this point, we have looked at parton-level kinematics of the top decay. However, we cannot expect the jet algorithm to faithfully represent the kinematics of the parton-level top decay in jets which include the physics of showering and hadronization. That is, the systematic effects of the jet algorithm, similar to those seen in QCD jets in Sec. III B, can be expected to appear in top quark jets as well. The substructure of a jet that reconstructs the top

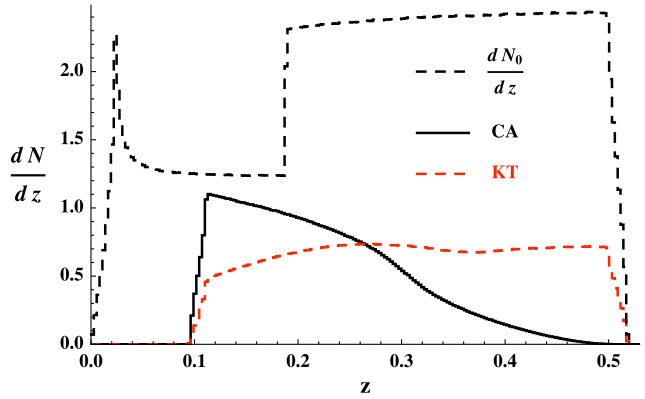


FIG. 17 (color online). dN_0/dz (all decays) and dN/dz (reconstructed decays), with $\gamma = 3$.

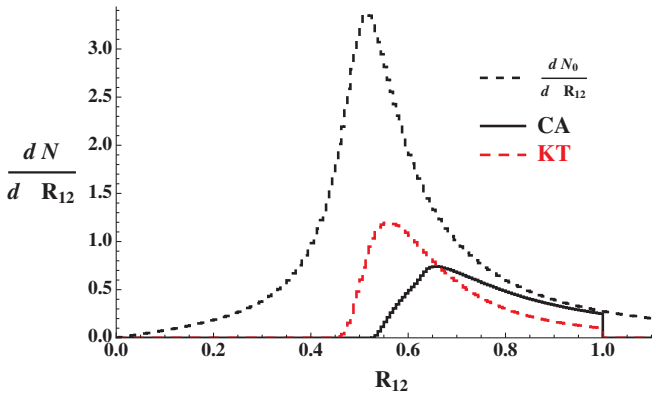


FIG. 18 (color online). $dN_0/d\Delta R_{Wb}$ (all decays) and $dN/d\Delta R_{12}$ (reconstructed decays), with $\gamma = 3$.

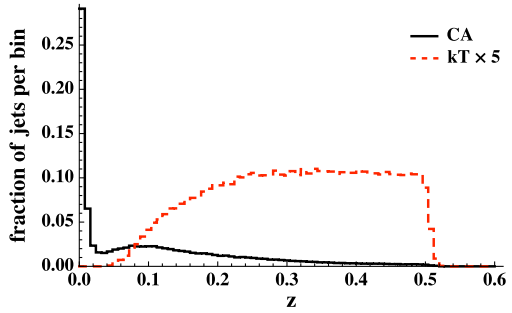


FIG. 19 (color online). Distribution in z for jets with the top mass in the $t\bar{t}$ sample. The jets have p_T between 500 and 700 GeV, and $D = 1.0$. Note the k_T distribution is scaled up by a factor of 5 to make the scales comparable.

quark mass may not match onto the kinematics of that decay. For instance, with the CA algorithm we expect that soft recombinations will occur at the last recombination step, even for jets that contain the decay products of a top quark. This can make the substructure look more like a heavy QCD jet than a top quark decay, and subsequently the jet may not be properly identified.

To demonstrate this point, in Fig. 19 we plot the distribution in z for jets with mass within a window around the top quark mass. The data represent simulated $t\bar{t}$ events as described in the Appendix. In this sample, the top quarks have a p_T between 500–700 GeV, so that many are expected to be reconstructed in a single jet.

The distribution for CA jets is very different from the parton-level distribution (Fig. 17). The excess at small values of z arises from soft recombinations in the CA algorithm, which make the distribution similar to that for QCD jets [Figs. 9(a) and 9(b)]. For the k_T algorithm, there are rarely soft recombinations late in the algorithm, because the metric orders according to z as well as ΔR.

The k_T algorithm distorts the dynamics of jet substructure less than does CA, but it has a serious drawback. The k_T algorithm tends to yield a much broader mass distribution for reconstructed tops than the CA algorithm, since soft particles that dominate the periphery of the jet are recombined early in the algorithm. This means that soft

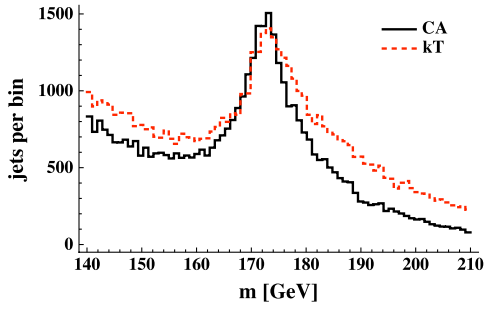


FIG. 20 (color online). Distribution in jet mass for jets in the neighborhood of the top mass in $t\bar{t}$ events for the CA (solid black line) and k_T (dotted red line) algorithms.

energy depositions in the calorimeter near the decay products of a top quark have a higher probability of being included in the jet and broadening the reconstructed top mass distribution. (This is essentially the statement that k_T jets have larger and more irregular ''areas'' than CA jets [24].) In Fig. 20, we plot the jet mass distribution in the neighborhood of the top mass for jets in the same $t\bar{t}$ sample as in Fig. 19 for both algorithms.

The top mass peak is broadened for the k_T algorithm relative to CA. From the point of view of the jet substructure, we cannot identify vertex-specific variables (such as z and ΔR) that characterize this broadening, because it is due to recombinations early in the algorithm. However, we will find that techniques used to remove the systematic effects of the algorithm from the substructure of jets are effective in narrowing mass distributions.

V. IDENTIFYING RECONSTRUCTED HEAVY PARTICLES WITH JET SUBSTRUCTURE

In the previous two sections we examined several kinematic distributions for QCD splittings and for heavy particle decays. We saw that while at the parton level the two processes have distinct kinematic features, these features are biased by the effects of the parton shower and the jet algorithm. The algorithm attempts to undo the showering but introduces its own biases. We would like to understand these effects and if possible remove them.

Our parton-level studies can be briefly summarized. In Sec. III, we used a toy model for QCD splittings in jets that contained the dominant soft and collinear physics of QCD, and studied the kinematics of the first splitting. In Sec. IV, we looked at one- and two-step decays with fixed boost. For the two-step top quark decay, requiring full reconstruction of the top (including the W as a subjet) from the three final-state quarks imposed kinematic restrictions that depended on the algorithm used. These studies led to the z and ΔR_{12} distributions seen in Figs. 6 and 7 (QCD), Figs. 13 and 15 (one-step decays), and Figs. 17 and 18 (two-step decays). We can see that the distributions in ΔR_{12} are quite similar, but that QCD splittings tend to have smaller z values than heavy particle decays for fixed mass and p_T.

Observing these parton-level differences is difficult because the QCD shower and the jet algorithm shape the jet substructure. The ordering of recombinations for the k_T and CA algorithms imposes significant kinematic constraints on the phase space for the last recombinations in a jet. This leads to kinematic distributions for the last recombination in a jet that depend as much on the algorithm as the underlying physics of the jet. For instance, in Fig. 9, we find that the kinematics of the last recombination in QCD jets are very different between the k_T and CA algorithms. In particular, we can compare Figs. 9(a) and 9(b), the distribution in z of the last recombination for QCD jets, with Fig. 19, the distribution in z of the last

recombination for jets in a $t\bar{t}$ sample that reconstruct the top quark mass. For the k_T algorithm, the differences reflect the different physics of QCD splittings and decays. However, the CA algorithm has shaped the distributions to have a large enhancement at small z for both processes. We cannot identify the physics of the jet simply from the value of z in the last recombination for CA. For the k_T algorithm, the final recombinations better discriminate between decays and QCD, but the mass resolution is poorer than for CA (Fig. 20).

There is one more important contribution to jet substructure common to QCD jets and heavy particle decays that we have not yet discussed. This is the combined effect of splash-in from several sources: soft radiation from other parts of the hard scattering, the underlying event (UE, the rest of the pp interaction), and pileup (other pp collisions that occur in the same time bin). All of these sources add particles to jets that are typically soft and approximately uncorrelated. Splash-in particles will mostly be located at a large angle to the jet core, simply because there is more area there. How these particles affect jet substructure depends on the algorithm used. We expect them to contribute similarly to soft radiation from the QCD shower, discussed at the ends of Secs. III and IV. For concreteness, we now examine briefly the effect of adding UE to our Monte Carlo events. We expect other splash-in effects to be similar.

In Fig. 21, we show the effect of adding UE on jet masses. The effect here is simple: adding extra energy to jets pushes the mass distribution higher. Note that for top jets, the mass peak has also broadened, making it harder to find the signal mass bump over the background distribution. In Fig. 22, we show how distributions in z and ΔR_{12} are affected by the UE. Because of the extra radiation at large angles from the UE, the distribution in the angle of

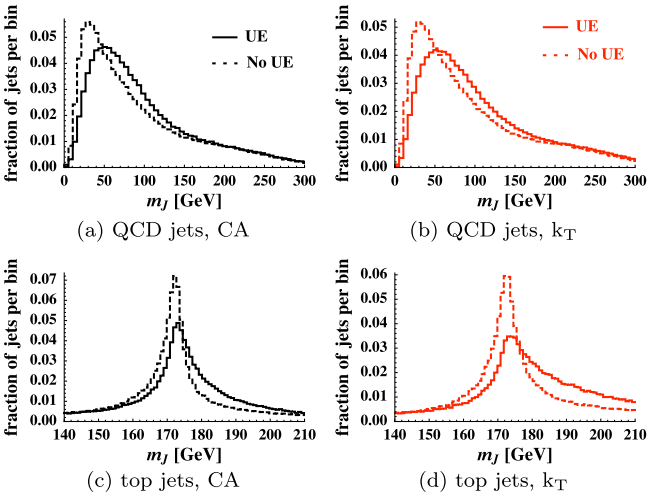


FIG. 21 (color online). Distribution in m_J with and without underlying event for QCD and top jets, using the CA and k_T algorithms. The jets have p_T between 500 and 700 GeV, and $D = 1.0$. The samples are described further in the Appendix.

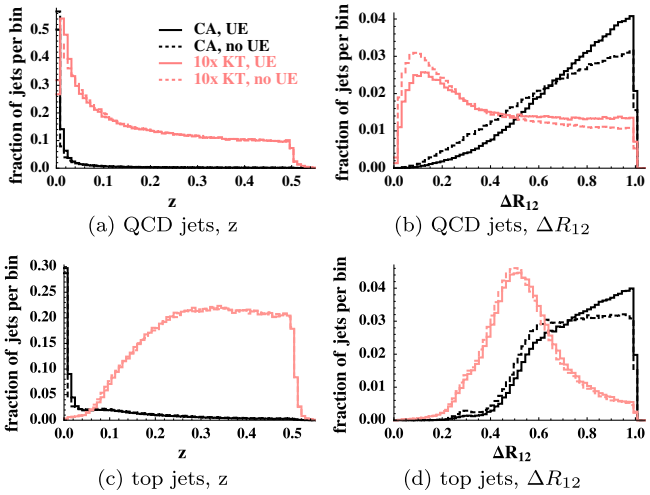


FIG. 22 (color online). Distributions in z and ΔR_{12} with and without underlying event for QCD and top jets, using the CA and k_T algorithms. The jets have p_T between 500 and 700 GeV, and $D = 1.0$. The samples are described further in the Appendix.

the last recombination, ΔR_{12}, is systematically shifted to larger values. The UE populates the same region in the jet as soft radiation from the hard partons, meaning the distribution in z is not significantly altered by the UE.

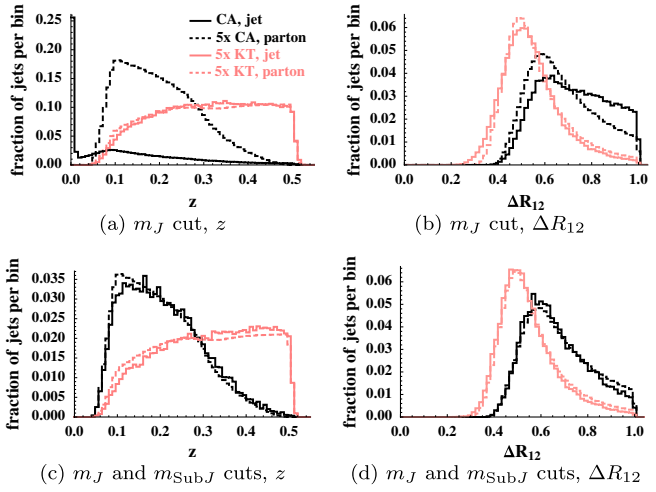
We have seen numerous examples that the kinematics of the jet substructure in the last recombination for CA is a poor indicator for the physics of the jet. However, we can characterize the aberrant substructure very simply. For the CA algorithm, late recombinations (necessarily at large ΔR) with small z are more likely to arise from systematics effects of the algorithm than from the dynamics of the underlying physics in the jet. For the k_T algorithm, the poor mass resolution of the jet arises from earlier recombinations of soft protojets. The last recombination for k_T is representative of the physics of the jet, but the degraded mass resolution makes it difficult to efficiently discriminate between jets reconstructing heavy particle decays and QCD. While small-z, large-ΔR recombinations are not as frequent late in the k_T algorithm as in CA, they do contribute the most to the poor mass resolution of k_T.

As a simple example of the sensitivity of the mass to small-z, large-ΔR recombinations, consider the recombination $i, j \rightarrow p$ of two massless objects in the small-angle approximation. The mass of the parent p is given by $m_p^2 = p_{T_p}^2 z(1-z)\Delta R_{ij}^2$, as in Eq. (9). Suppose the value of the k_T recombination metric, $\rho_{ij}(k_T) = p_{T_p} z \Delta R_{12}$, is bounded below by a value ρ_0 (say by previous recombinations), and the recombination $i, j \rightarrow p$ occurs at $\rho_{ij}(k_T) = \rho_0$. Then the mass of the parent is $m_p^2 = \rho_0^2(1-z)/z$, which is maximized for small z. Therefore, at a given stage of the algorithm, small-z recombinations have a large effect on the mass of the jet.

When we can resolve the mass scales of a decay in a jet, the distribution of kinematic variables matches closely

what we expect from the parton-level kinematics of the decay. For the example of the top quark decay, if we select jets with the top mass that have a daughter subjet with the W mass, the kinematic distributions of z and ΔR_{12} closely match the distributions from the parton-level decay of the top quark. We show this in Fig. 23, where we make a top quark "hadron-parton" comparison for z and ΔR_{12}. The specifics of the mass cuts are described in Sec. VII. In the parton-level events, we simply require that the top quark decay to three partons be fully reconstructed by the algorithm in a single jet, namely, that the W is correctly recombined first from its decay products before recombination with the b quark to make the top. The parton-level events have the same distribution of top quark boosts as the top jets in the hadron-level events. It is clear that simply requiring the hadron-level jet to have the top mass, which makes no cut on the substructure, leads to kinematic distributions in z and ΔR_{12} for CA that do not match the parton-level distributions, although the distributions do match quite well for the k_T algorithm. The excess of small-z recombinations for CA in the hadron-level jet with only a jet mass cut arises from jet algorithm effects discussed previously. After the subjet mass cut, these are removed and the distribution of z in the jet matches the reconstructed parton-level decay very well.

Therefore, when we can accurately reconstruct the mass scales of a decay in a jet, the kinematics of the jet substructure tend to reproduce the parton-level kinematics of the decay. This suggests that if we can reduce systematic effects that generate misleading substructure, we can improve heavy particle identification and separation from background. Reducing these systematic effects can also improve the mass resolution of the jet, which will aid in identifying a heavy particle decay reconstructed in a jet and in rejecting the QCD background.



(a) m_J cut, z (b) m_J cut, ΔR_{12}

(c) m_J and $m_{\mathrm{Sub}J}$ cuts, z (d) m_J and $m_{\mathrm{Sub}J}$ cuts, ΔR_{12}

FIG. 23 (color online). Distributions in z and ΔR_{12} comparing for top quark decays at the parton level and from Monte Carlo events. The jets have p_T between 500 and 700 GeV, and have $D = 1.0$. The parton-level top decays have the same distribution of boosts as the Monte Carlo top jets. Jets in the upper plots have a mass cut on the jet; the lower plots include a subjet mass cut. The details of these cuts are described in Sec. VII.

VI. THE PRUNING PROCEDURE

In this section we define a technique that modifies the jet substructure to reduce the systematic effects that obscure heavy particle reconstruction. In general, we will think of a *pruning procedure* as using a criterion on kinematic variables to determine whether or not a branching is likely to represent accurate reconstruction of a heavy particle decay. This takes the form of a cut: if a branching does not pass a set of cuts on kinematic variables, that recombination is vetoed. This means that one of the two branches to be combined (determined by some test on the kinematics) is discarded and the recombination does not occur.

In Sec. V, we identified recombinations that are unlikely to represent the reconstruction of a heavy particle. These can be characterized in terms of the variables z and ΔR: recombinations with large ΔR and small z are much more likely to arise from systematic effects of the jet algorithm and in QCD jets rather than heavy particle reconstruction (compare the upper and lower figures in Fig. 23). We expect that removing (pruning) these recombinations will tend to improve our ability to measure jet substructure, including subjet masses. We also expect that this procedure will systematically shift the QCD mass distribution lower, reducing the background in the signal mass window. Finally this procedure is expected to reduce the impact of uncorrelated soft radiation from the underlying event and pileup. We therefore define the following pruning procedure:

(0) Start with a jet found by any jet algorithm, and collect the objects (such as calorimeter towers) in the jet into a list L. Define parameters D_{cut} and z_{cut} for the pruning procedure.

(1) Rerun a jet algorithm on the list L, checking for the following condition in each recombination $i, j \to p$:

$$z = \frac{\min(p_{Ti}, p_{Tj})}{p_{Tp}} < z_{\mathrm{cut}} \quad \text{and} \quad \Delta R_{ij} > D_{\mathrm{cut}}.$$

This algorithm must be a recombination algorithm such as the CA or k_T algorithms, and should give a "useful" jet substructure (one where we can meaningfully interpret recombinations in terms of the physics of the jet).

(2) If the conditions in 1 are met, do not merge the two branches 1 and 2 into p. Instead, discard the softer branch, i.e., veto on the merging. Proceed with the algorithm.

(3) The resulting jet is the *pruned jet*, and can be compared with the jet found in step 0.

This technique is intended to be generically applicable in heavy particle searches. It generalizes analysis techniques suggested by other authors, including filtering

[12] and top tagging [9], in that these methods also modify the jet substructure to assist separate a particular signal from backgrounds. In particular, the use of the variables z and ΔR_{ij} follows the use of δ_p and δ_r in [9], with the significant difference that δ_p measures softness relative to the total jet, and we define z to be a "local" variable that only depends on the two protojets being recombined. A more important distinction is that filtering and top tagging are designed to find a specific number of subjets to map onto a specific decay, whereas pruning is intended to be applied to an entire jet with no bias toward a specific substructure configuration. While we think this generality is novel, we emphasize that pruning is an evolution from earlier methods and relies on the same physical effects. We have endeavored to justify our claim for generality with the discussions in Secs. III, IV, and V, which demonstrate that the interpretation of jet substructure is subject to generic systematic effects that can be well characterized. Pruning is not the only option, but offers some advantages which we explore in further studies below.

In the analysis of pruning, we will explore the dependence of the pruned jets on the value of D from the jet algorithm. When reconstructing a boosted heavy particle in a single jet, without pruning the reconstruction is optimized if the value of D is fit to the expected opening angle of the decay. However, this angle depends on the mass of the particle (which is not known in a search) and its p_T. We will show that pruning reduces the sensitivity to D and allows one to use large-D jets over a broad range in p_T to search for heavy particles.

Values for the two parameters of the pruning procedure, $z_{\rm cut}$ and $D_{\rm cut}$, can be well motivated. In the following studies, we will show that the results of pruning are rather insensitive to the parameters, and that the optimal parameters are similar for different searches. That is, it is not necessary to tune the pruning procedure for individual searches.

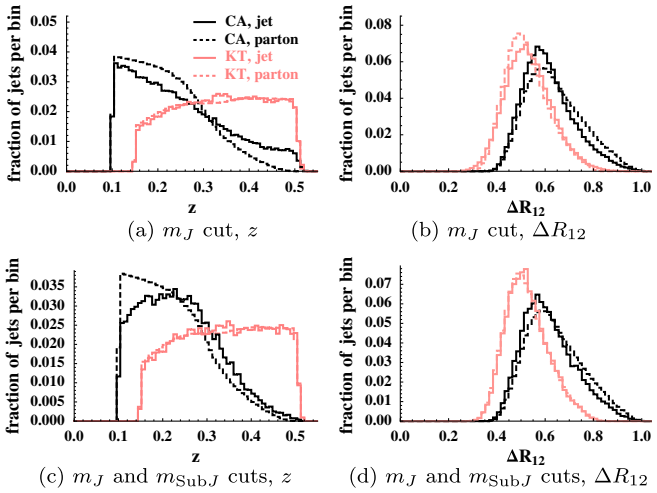
The parameter $z_{\rm cut}$ can be chosen based on the analysis of single-step and multistep decays in Sec. IV. Near the limit in boost where decays are reconstructed in a single jet, the value of z is typically large. It is only at large boosts, where the production rate of heavy particles is much smaller, that small values of z are allowed for reconstructed decays (see Fig. 15). Therefore, we can choose a value of $z_{\rm cut}$ that will keep all reconstructed parton-level decays at small boost, and only remove a small fraction of decays at larger boosts. We expect that a $z_{\rm cut} \sim 0.10$ will be a reasonable compromise. Note that Fig. 23(a) indicates that much of the soft radiation distorting the substructure for CA jets has $z \lesssim 0.02$, so that at least for CA a $z_{\rm cut}$ not much bigger than this should be effective.

The parameter $D_{\rm cut}$ can be determined on a jet-by-jet basis, allowing pruning to be more adaptive than a fixed-parameter procedure. $D_{\rm cut}$ determines how much of the jet substructure can be pruned, with smaller values allowing for more pruning. $D_{\rm cut}$ should be sufficiently small so that if a decay is "hidden" inside the jet substructure by late recombinations of, say, UE particles, the substructure can be pruned and the decay can be found. A value that is too small, however, will result in overpruning. A natural scale for $D_{\rm cut}$ is the opening angle of the jet. However, this is an infrared unsafe quantity, as soft radiation can change the opening angle. Instead, the dimensionless ratio m_J/p_{T_J} for the jet is related to the opening angle: typically, $\Delta R_{12} \approx 2m_J/p_{T_J}$. Therefore, we choose $D_{\rm cut}$ to scale with $2m_J/p_{T_J}$. $D_{\rm cut} = m_J/p_{T_J}$ is a reasonable starting value.

Effects of pruning

Having defined the pruning procedure, we now wish to study its effects. In this study, we use the parameters $D_{\rm cut} = m_J/p_{T_J}$ for both algorithms, and $z_{\rm cut} = 0.10$ for the CA algorithm and 0.15 for the k_T algorithm. We will motivate these parameters in Sec. VIII A. First, in Fig. 24, we reproduce the hadron-parton comparison from Fig. 23, using pruning at both the hadron and parton level. The parton-level pruning is implemented in the same way as defined above, treating the three partons of the reconstructed top quark as the jet.

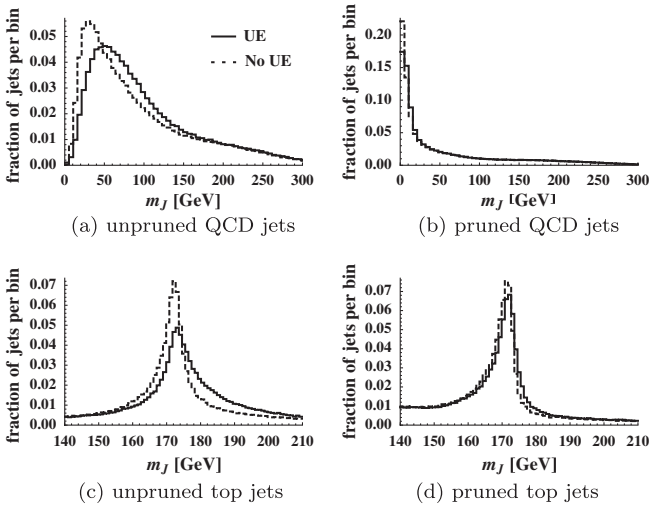
It is clear by comparing Figs. 23 and 24 that pruning has removed much of the systematic effects in the CA algorithm; when only a jet mass cut is made, the distributions in z and ΔR_{12} for pruned jets match the parton-level distribution much better than unpruned jets. When both mass and subjet mass cuts are made, pruning shows a slightly poorer agreement to the parton-level kinematics than the unpruned case. Note however that for pruned jets, the efficiency of the subjet mass cut is considerably greater since we more often identify one of the daughter subjets as a W (see the discussion of Fig. 30 in Sec. VIII A).



(a) m_J cut, z (b) m_J cut, ΔR_{12}

(c) m_J and $m_{{\rm Sub}J}$ cuts, z (d) m_J and $m_{{\rm Sub}J}$ cuts, ΔR_{12}

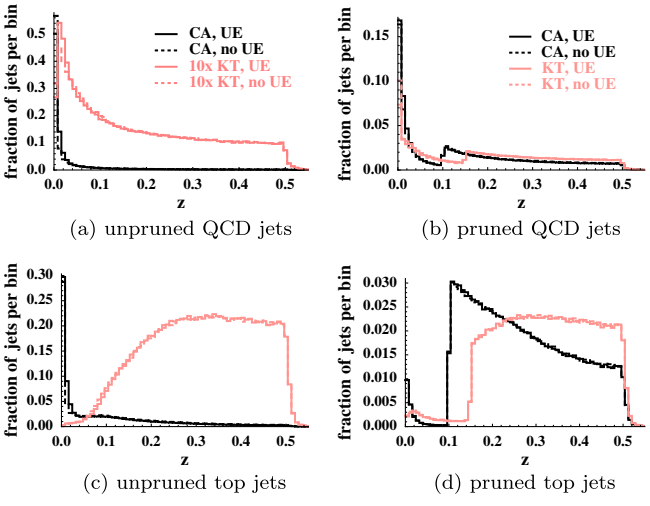
FIG. 24 (color online). Distributions in z and ΔR_{12} comparing for top quark decays at the parton level and from Monte Carlo events after implementing pruning. This figure uses the same samples and cuts as Fig. 23.



(a) unpruned QCD jets (b) pruned QCD jets

(c) unpruned top jets (d) pruned top jets

FIG. 25. Distributions in m_J with and without underlying event, for QCD and top jets, using the CA algorithm, with and without pruning. The jets have p_T between 500 and 700 GeV, and $D = 1.0$.



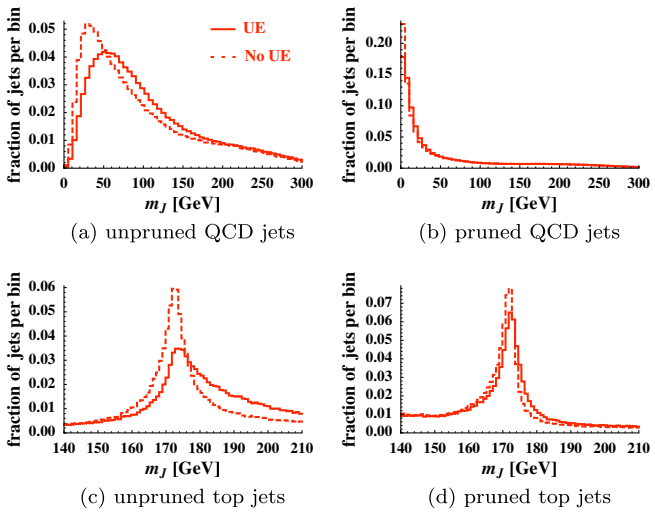
(a) unpruned QCD jets (b) pruned QCD jets

(c) unpruned top jets (d) pruned top jets

FIG. 27 (color online). Distribution in z with and without underlying event, for QCD and top jets, using the CA and k_T algorithms, with and without pruning. The legends for plots (c) and (d) correspond to (a) and (b), respectively. The jets have p_T between 500 and 700 GeV, and $D = 1.0$.

In addition to improving the kinematics of the jet substructure, pruning reduces the contribution of the underlying event and improves the mass resolution of reconstructed decays. In Figs. 25 and 26 we give the mass distribution of jets with and without the UE in both the QCD and $t\bar{t}$ samples for the CA and k_T algorithms, but now with and without pruning. In Figs. 27 and 28 we show the effect of UE on distributions in z and ΔR_{12}, with and without pruning.

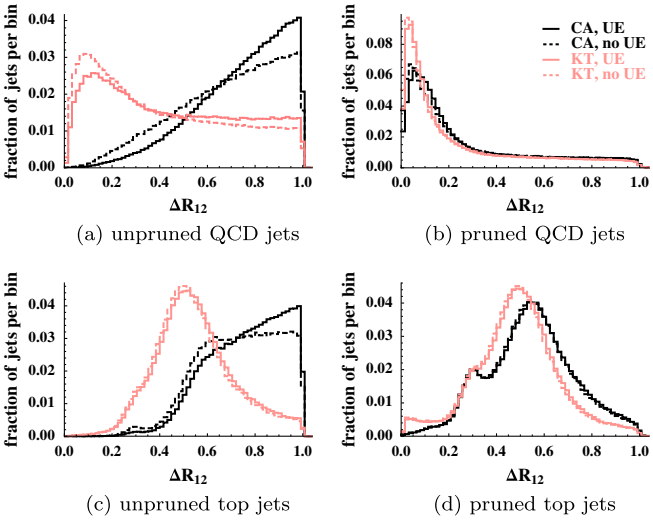
Three distinctions between pruned and unpruned jets are clear. First, the distributions with and without the UE are very similar for pruned jets, while they differ noticeably for unpruned jets. This shows that pruning has drastically reduced the contribution of the underlying event. Second, the mass peak of jets near the top quark mass in the $t\bar{t}$ sample is significantly narrowed by the introduction of pruning (especially when the UE is included). This is evidence of the improved mass resolution of pruning, and will contribute to the improvement in heavy particle identification with pruning. And finally, the mass distribution of QCD jets is pushed significantly downward by pruning. The QCD jet mass is dominantly built from the soft, large-angle recombinations—most recombinations are soft, and for fixed p_T, larger-angle recombinations contribute more to the jet mass. Removing these by pruning the jets reduces



(a) unpruned QCD jets (b) pruned QCD jets

(c) unpruned top jets (d) pruned top jets

FIG. 26 (color online). Distributions in m_J with and without underlying event, for QCD and top jets, using the k_T algorithm, with and without pruning. The jets have p_T between 500 and 700 GeV, and $D = 1.0$.



(a) unpruned QCD jets (b) pruned QCD jets

(c) unpruned top jets (d) pruned top jets

FIG. 28 (color online). Distribution in ΔR_{12} with and without underlying event, for QCD and top jets, using the CA and k_T algorithms, with and without pruning. The jets have p_T between 500 and 700 GeV, and $D = 1.0$.

the QCD mass distribution in the large-mass range and will contribute to the reduction of the QCD background.

We move on to examine pruning through a set of studies using Monte Carlo simulated events. We will investigate the parameter dependence of pruning, motivating the parameters used above. We will extensively study both top and W reconstruction with pruning, and quantify the improvements from pruning in terms of basic statistical measures. These studies will provide evidence of the insensitivity of pruning to the value of D in the jet algorithm.

VII. MONTE CARLO STUDIES

A. Study layout

The parameter space for questions about pruning procedures is very large. We will not be able to answer all possible questions in this paper, but we will attempt to answer the most important. We use Monte Carlo samples to study W reconstruction and the rejection of W + jets backgrounds, as well as top quark reconstruction and the rejection of QCD multijet backgrounds. To test the usefulness of pruning across a range of jet m/p_T, and hence the heavy particle boost, we study both signals in four p_T bins. We will also be able to compare a signal with a single mass scale (the W) to one with two (the top). The details of the Monte Carlo samples and their generation are described in the Appendix.

In the following sections, we define a particular method to identify the heavy particles using jet substructure, and examine pruning in this context. We are more concerned with the *improvements* provided by pruning than its absolute performance. Therefore, we compare pruning to an analysis procedure where the jets are left unpruned. This comparison removes dependence on quantities that have large uncertainties, such as signal and background cross sections, or are not specified, such as the integrated luminosity. Instead, the performance of pruning is quantified in terms of how much *better* pruning resolves the physically relevant substructure of the jet and separates signal and background processes versus using the substructure from unpruned jets.

Additionally, we test the performance of pruning as parameters of the jet algorithm and the pruning procedure are varied, including D. We expect the D dependence to be closely correlated with the jet p_T, as it is a direct measure of the boost of the heavy particle. We aim to draw some basic conclusions about how pruning should be applied in a search.

B. Measures used to quantify pruning

Mass variables are by far the strongest discriminator between QCD jets and jets reconstructing heavy particle decays. QCD jets have a smooth mass distribution set by the jet p_T (see Sec. III), while a decaying particle can have multiple intrinsic mass scales. We define simple criteria to identify a jet as coming from a top quark: if the jet mass is in the top mass window and one of the two subjets has a mass in the W mass window, then we tag the jet as a *top jet*. The top and W mass windows are defined by fitting the relevant mass peaks of the signal sample, which we describe in detail below. The W study proceeds analogously with only a jet mass cut. In a real search for a particle of unknown mass, one obviously cannot fit a ''signal sample.'' However, we employ this method to demonstrate two effects of pruning: sharpening the signal mass peak and reducing the QCD background in this region. These two effects will determine how well pruning improves our ability to find bumps in jet mass distributions.

We use a common set of variables to measure the difference between a jet algorithm and its pruned version. Let $N_S(A)$ be the number of jets in the signal sample identified as a reconstructed heavy particle for algorithm A, and $N_B(A)$ the analogous number of jets in the background sample. Use pA to denote the pruning procedure run on jets found with algorithm A. Then the variables we use are

$$\epsilon = \frac{N_S(pA)}{N_S(A)},$$

$$R = \frac{N_S(pA)/N_B(pA)}{N_S(A)/N_B(A)}, \quad \text{and} \quad S = \frac{N_S(pA)/\sqrt{N_B(pA)}}{N_S(A)/\sqrt{N_B(A)}}. \tag{30}$$

ϵ is the relative efficiency of pruning in identifying heavy particles in the signal sample, while R and S are the relative signal-to-background and signal-to-noise ratios for the pruned and unpruned algorithms. We also evaluate the relative mass window widths, which we label w_{rel}. For the W study, this is the ratio of the W mass window width for pruning relative to not pruning; for the top study it is the ratio in the top mass window width. Note that in the top study, a W subjet mass cut is also used. A value of $w_{\text{rel}} < 1$ means pruning has improved the mass resolution of the jets. These ratios are independent of the integrated luminosity and the total cross sections, and are representative of the improvements that pruning would provide in an analysis.

To determine the mass window for a particular signal sample, we fit the mass peak to determine the window width. In these studies, a skewed Breit-Wigner is sufficient to fit the peak, with a power law continuum background. These functions used to fit mass peaks are

$$\text{peak: } f(m) = \frac{M^2\Gamma^2}{(m^2 - M^2)^2 + M^2\Gamma^2}(a + b(m - M));$$

$$\text{continuum: } g(m) = \frac{c}{m} + \frac{d}{m^2}. \tag{31}$$

M is the location of the mass peak; Γ is the width of the peak. A sample fit is shown in Fig. 29. The mass window $[M - \Gamma, M + \Gamma]$ is found to be nearly optimal, given this

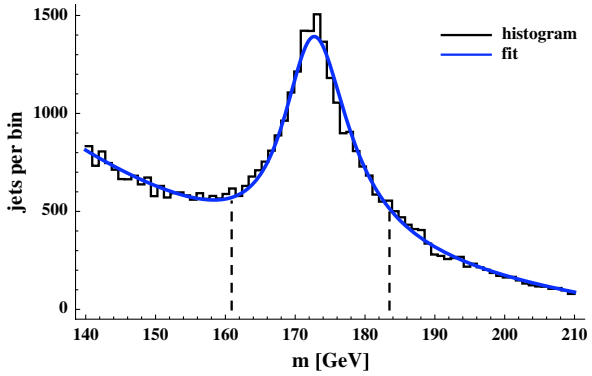


FIG. 29 (color online). A sample fit showing the jet mass distribution (black histogram) and sample fit (blue curve) for CA jets from $t\bar{t}$ events.

functional form, in measures similar to ϵ, R, and S: the area in the window ($\sim \epsilon$), the ratio of area to the window width ($\sim R$), and the ratio of area to the square root of the width ($\sim S$).

VIII. STUDY RESULTS

In this section we present results comparing analyses with pruned jets to unpruned jets. We demonstrate two main points: first, pruning is useful and broadly applicable, and second, its parameters do not need fine-tuning for it to provide significant improvement.

The natural starting point is to investigate the parameters particular to the pruning procedure, D_{cut} and z_{cut}. The most important question is whether these need to be tuned to the signal. To answer this, in Sec. VIII A we study the performance of pruning as we vary its parameters for two different signals across the full p_T range for the samples. We find that optimal choices of z_{cut} and D_{cut} vary slowly with m/p_T, but that our choice of parameters is not far from optimal in all cases.

After fixing z_{cut} and D_{cut}, we consider the effect of varying D in the jet algorithm. In Sec. VIII B we study pruning with D fixed at 1.0 over all p_T bins. This type of analysis is like a search where the mass (and hence m/p_T) of the new heavy particle is not known. For comparison, in Sec. VIII C we redo the analysis, but with D adjusted for each bin to fit the expected angular size of the decay in that bin. In this case, the unpruned jet algorithm performs better than with a constant D, as expected, but pruning still shows improvements in finding W's and tops. In all cases, pruned jets are a better way to identify heavy particles than unpruned. In Sec. VIII D we compare the results of Secs. VIII B and VIII C. Significantly, if jets are pruned, we find that it does not make much difference what the initial D value was, indicating that searches with large fixed D do not suffer in power compared to searches with D tuned to known or suspected m/p_T.

In Sec. VIII E we give some absolute measures of top finding with pruned jets for comparison to other methods. In Sec. VIII F we directly compare the CA and k_T algorithms, before and after pruning. Finally, in Sec. VIII G we consider the effect of a crude detector model where we smear the energies of all particles in the calorimeter. We find that the performances of the pruned and unpruned algorithms are degraded, but that pruning still provides significant improvement.

A. Dependence on pruning parameters

The pruning procedure we have defined has two free parameters (in addition to those of the jet algorithms themselves). In introducing the procedure, we argued that $z_{\text{cut}} = 0.10$ and $D_{\text{cut}} = m_J/p_{T_J}$ were sensible choices. We now investigate how pruning performs when each of these parameters is varied while the other is held fixed, for both (W and top) signals and across the four p_T bins for each signal.

We will look at the values of the metrics w_{rel}, ϵ, R, and S defined in Sec. VII B. The priority in choosing particular values for z_{cut} and D_{cut} should be in optimizing S, as it is the criterion for discovery. That being said, ϵ and R are still important measures as they determine the total size of the signal and remaining fraction relative to the background. We also evaluate w_{rel} because the mass window width drives the other three metrics. As the relative width decreases, in general the measures R and S will increase because the heavy particle is better resolved and more of the background is rejected, but ϵ will tend to decrease simply because the narrower width selects fewer signal jets. ϵ can, however, increase with decreasing mass window width if enough high-mass signal jets are being pruned into the mass window.

In Fig. 30, we show all four metrics for top and W jets, for both CA and k_T jets. D_{cut} is set to m_J/p_{T_J} throughout, and z_{cut} is varied in [0, 0.25]. $z_{\text{cut}} = 0$ represents no pruning and we can see that all metrics are 1 here. With increasing pruning, the mass window width initially decreases rapidly, then levels out. In all but the smallest p_T bin, the relative signal efficiency ϵ increases as the width narrows, suggesting that signal jets that had ''vacuumed up'' too much UE or soft radiation are being pruned back into the mass window. Note that for the top quark sample with the k_T algorithm, ϵ merely flattens out for a range in z_{cut}, and does not increase as it does for the other samples. Once the window stops shrinking significantly (around $z_{\text{cut}} = 0.05$), the relative signal efficiency starts decreasing; now the dominant effect is overpruning signal jets out of the mass window. Note, however, that even though the relative signal efficiency is *decreasing*, the relative signal-to-background ratio R is *increasing* over the full range. So even as signal jets are being removed from the mass window, background jets are being removed even faster. If we look at signal-to-noise, S, there appears to be a broad optimal range in z_{cut} that depends somewhat on the signal, on the p_T bin, and on the jet algorithm.

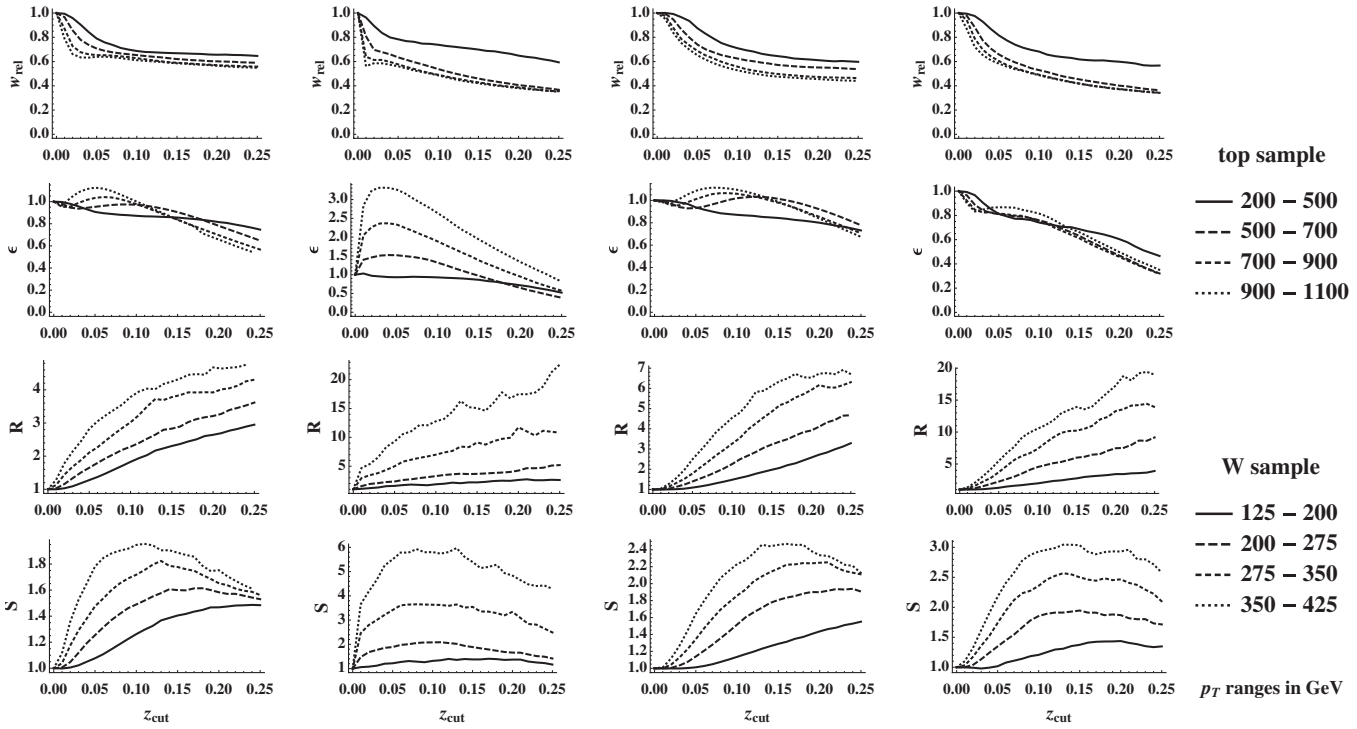
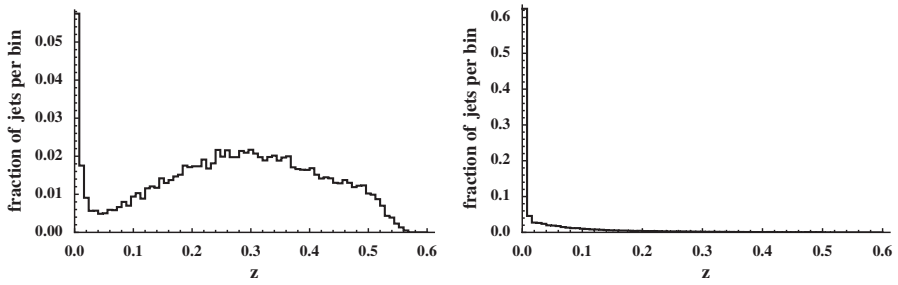


FIG. 30. Relative statistical measures w_{rel}, ϵ, R, and S vs z_{cut} for W's and tops, using CA and k_T jets. Four p_T bins are shown for each sample. Statistical errors (not shown) are $\mathcal{O}(1\%)$ for w_{rel} and ϵ, and $\mathcal{O}(10\%)$ for R and S.

There are two important lessons to be learned from these plots. First, more pruning is required for k_T jets than for CA to achieve similar results. The right two columns (k_T) are similar to the left two (CA) except that features are shifted out in z_{cut}. Second, the peak in S does not depend strongly on the signal or the p_T, in the three largest p_T bins. The dependence on S in the smallest p_T bin, however, is different from the others due to threshold effects of the heavy particle being reconstructed in a single jet. In this bin, the boosts of the W's or tops are small enough that many decays are just at the threshold for being reconstructed. Decays at the reconstruction threshold typically have poor mass resolution, and cutting more aggressively on z reduces these threshold effects and significantly decreases the background, leading to an increase in S over the whole range in z_{cut}. For CA, our ''reasonable choice'' of z_{cut} of 0.10 looks close to optimal for the upper three bins, and not far off for the smallest. For k_T, a larger z_{cut} is needed; 0.15 is close to optimal.

Additionally, these plots offer an interesting perspective on the role of z in jet substructure. The $t\bar{t}$ sample for the CA algorithm is the most instructive. In this case, small values of z_{cut} lead to dramatically increased efficiency for finding top jets in the larger p_T bins. This is due to the improved ability after pruning to find the W as a subjet of the top. At large p_T with a fixed $D = 1.0$, the opening angle of the top quark decay is much smaller than D. This means that the top quark decay is very localized in the jet, and much of the jet area includes soft radiation. For the CA algorithm, which recombines solely by the angle between protojets, this tends to delay recombining the soft peripheral radiation until the end of the algorithm. The result is substructure with small z at the last recombination that is not representative of the top quark decay—neither daughter protojet of the top has the W mass. As an illustration of this point, in Fig. 31 we plot the distribution of z for unpruned jets in the top mass range for the CA algorithm in the largest and smallest p_T bins. Note that in the largest p_T bin, where the top quark decay is highly localized in the jet and the decay angle is much less than D, there is a substantially increased fraction of jets with a small value of z. This does not occur in the smallest p_T bin, where most of the reconstructed tops are at threshold for being just inside the jet. When pruning is implemented, however, much of this soft radiation is removed. In Fig. 32, we plot the same distributions as in Fig. 31, but for pruned jets. In this case, no jets with the top mass have small z, since pruning has removed those recombinations. This leads to a highly enhanced efficiency to resolve the W subjet and identify the jet and a top jet. In Sec. VIII B, we will study pruning when the value of D is matched to the average angle of the heavy particle decay, and we will see that the performance of the unpruned CA algorithm improves.

By contrast, this situation does not occur for the k_T algorithm. Even when the value of D is mismatched with the top quark decay angle, the soft radiation on the periph-



(a) p_T bin 1, 200–500 GeV (b) p_T bin 4, 900–1100 GeV

FIG. 31. Distribution in z for unpruned CA jets in the top mass window for two p_T bins. The small p_T bin distribution (left panel) has only a small enhancement of entries at small z, while the large p_T bin distribution (right panel) is dominated by small z.

ery of the jet is recombined early in the k_T algorithm because of the p_T weighting in the recombination metric. Therefore, there is no increase in efficiency with increasing $z_{\rm cut}$ for large p_T, and the decrease in ϵ comes from the narrower width of the top and W mass distributions. The small variation in the measures R and S for the k_T algorithm at small $z_{\rm cut}$ is evidence of the fact that k_T tends to have many fewer small-z recombinations at the end of the algorithm, and supports the larger value of $z_{\rm cut} = 0.15$ for the k_T algorithm that we will use in the remainder of the study.

We now fix $z_{\rm cut}$ to study the dependence on $D_{\rm cut}$. For the CA algorithm we choose $z_{\rm cut} = 0.1$, and for k_T we choose 0.15. In Fig. 33, we plot $w_{\rm rel}$, ϵ, R, and S as $D_{\rm cut}$ is varied in $[0, 5m_J/p_{T_J}]$. While $z_{\rm cut}$ sets the minimum p_T asymmetry that recombinations can have, $D_{\rm cut}$ sets the minimum opening angle for recombinations that can be pruned. We can think of $D_{\rm cut}$ as determining which recombinations can be pruned, and $z_{\rm cut}$ as determining whether or not that pruning takes place. This difference is clearer when we consider two limiting values of $D_{\rm cut}$ and their impact on the pruned jet substructure.

As $D_{\rm cut}$ grows past $2m_J/p_{T_J}$, any recombination must have a large opening angle between the daughters to be pruned. Note that the limit $D_{\rm cut} \to \infty$ is the limit of no pruning. For both the CA and k_T algorithms, in this limit only very late recombinations in the algorithm can be pruned (if the jet can be pruned at all). In this limit, we expect the statistical measures to tend to one as the amount of pruning decreases.

The second limit is $D_{\rm cut} \to 0$. In this limit any recombination can be pruned, since the minimum opening angle needed is very small. As $D_{\rm cut}$ decreases toward zero, more of the jet substructure can be pruned. In particular, earlier recombinations—those with smaller opening angle on average—can be pruned as $D_{\rm cut}$ decreases. In general, these early recombinations are associated with the QCD shower, and pruning them can degrade the mass resolution of the jet because too much radiation is being removed. Therefore, we expect the performance of pruning to be poor in this region.

Both of these limits are present in Fig. 33, and our expectations about these limits are correct. It is in the intermediate region, where $D_{\rm cut} \approx m_J/p_{T_J}$, that the performance of pruning is optimal, with a maximum in S that is not very sensitive to the p_T bin, sample, or algorithm. This value of $D_{\rm cut} = m_J/p_{T_J}$ is sensible when we recognize that the average opening angle of the jet is approximately $2m_J/p_{T_J}$, and half this value allows for pruning of late recombinations but not the soft, small-angle recombinations associated with the QCD shower.

For the remainder of the study, we fix the pruning parameters $z_{\rm cut} = 0.1$ for the CA algorithm and $z_{\rm cut} = 0.15$ for the k_T algorithm, as well as $D_{\rm cut} = m_J/p_{T_J}$ for both algorithms. With these parameters fixed, we move on to discuss more interesting tests of the pruning procedure.

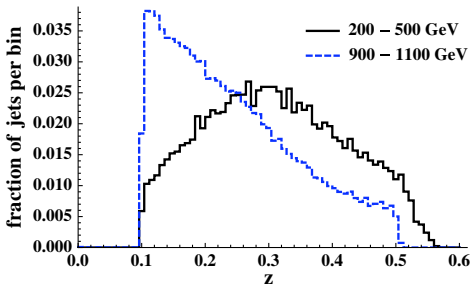
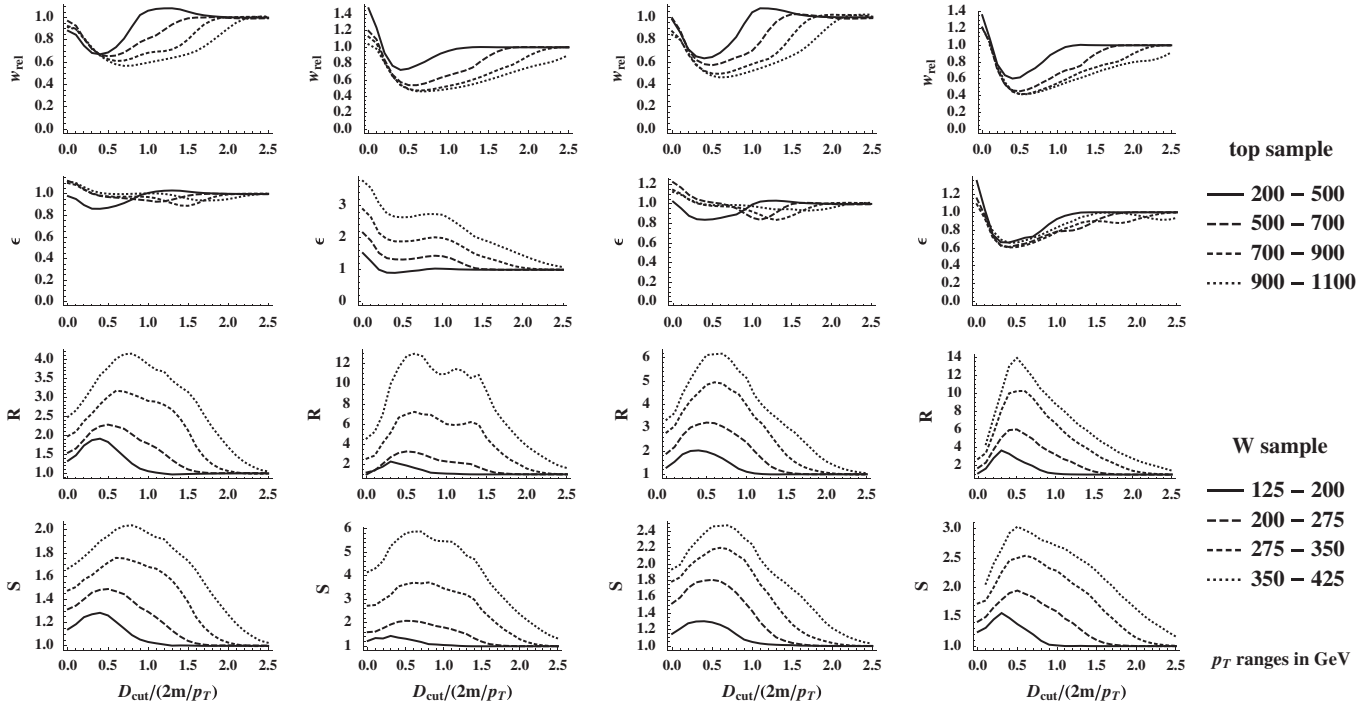


FIG. 32 (color online). Distribution in z for pruned CA jets in the top mass window for two p_T bins, using $z_{\rm cut} = 0.10$.

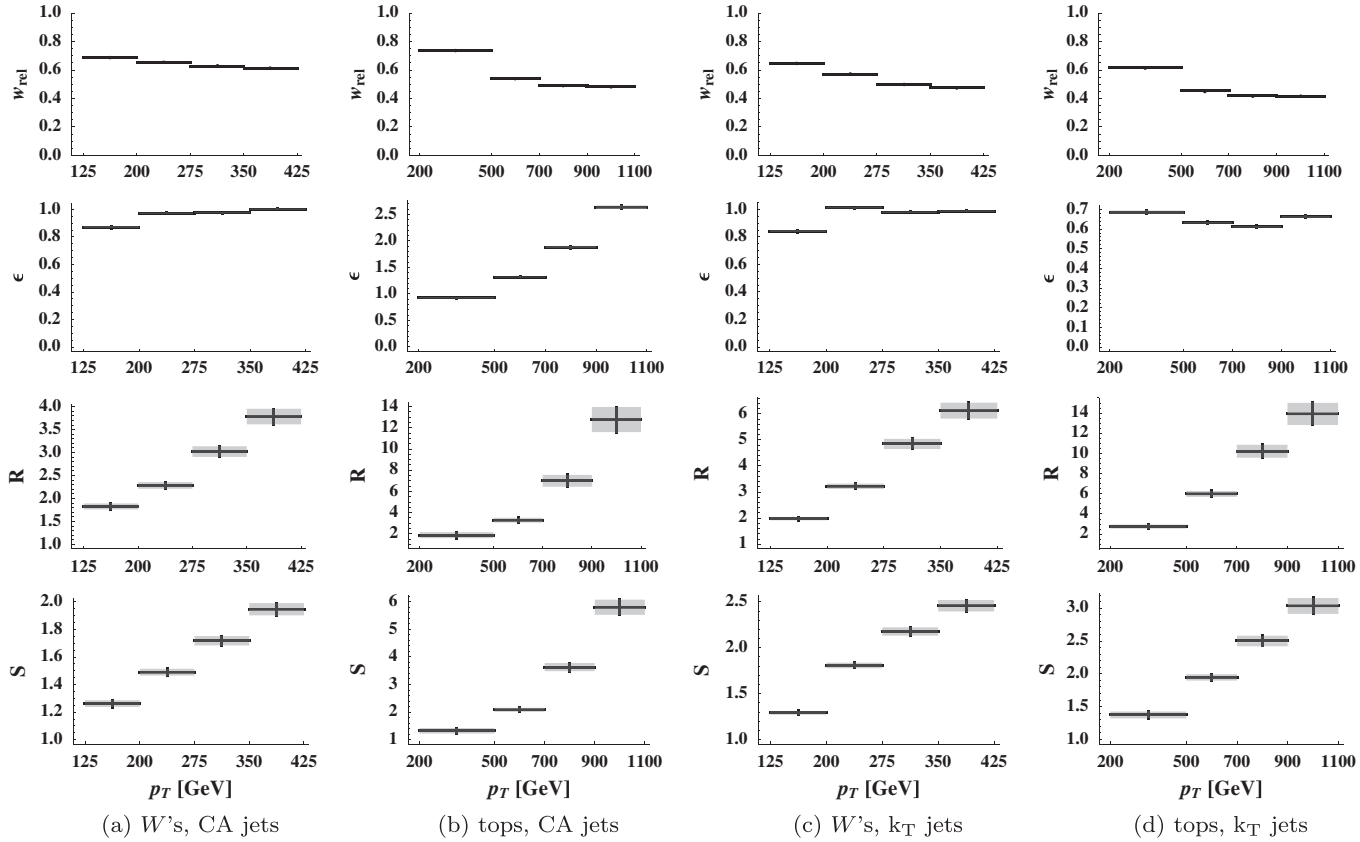
B. Top and W identification with constant D

In a search for heavy particles decaying into jets, it may be unfeasible to divide a sample into p_T bins and use a tailored jet algorithm to look for local excesses in the jet mass distribution in each p_T bin. (A "variable-R" method for avoiding p_T binning, which we do not consider here, has recently been suggested [25]. This still requires knowing or guessing the mass of the new particle, since it is m/p_T that determines the relevant angular size.) For instance, the appropriate angular scale may be unknown because the mass of the heavy particle is not known or the production mechanism is not well understood (so that



(a) W's, CA jets (b) tops, CA jets (c) W's, k_T jets (d) tops, k_T jets

FIG. 33. Relative statistical measures $w_{\rm rel}$, ϵ, R, and S vs $D_{\rm cut}/\frac{2m_J}{p_{T_J}}$ for W's and tops, using CA and k_T jets. Four p_T bins are shown for each sample. Statistical errors (not shown) are $\mathcal{O}(1\%)$ for $w_{\rm rel}$ and ϵ, and $\mathcal{O}(10\%)$ for R and S.



(a) W's, CA jets (b) tops, CA jets (c) W's, k_T jets (d) tops, k_T jets

FIG. 34. Relative statistical measures $w_{\rm rel}$, ϵ, R, and S vs p_T for W's and tops, using CA and k_T jets with $D = 1.0$. Statistical errors are shown.

TABLE I. "Tuned" D values for W and top p_T bins. The fixed-D analysis used $D = 1.0$, so the smallest bin does not change.

			W	
p_T (GeV)	125–200	200–275	275–350	350–425
tuned D	1.0	0.8	0.6	0.4
			Top	
p_T (GeV)	200–500	500–700	700–900	900–1100
tuned D	1.0	0.7	0.5	0.4

the spectrum of heavy particle boosts is not known). In this case, a large-D jet algorithm may be used to search for heavy particles reconstructed in single jets. To mimic such an analysis, and provide a reference point for further tests of pruning, we find our statistical measures for W and top quark jets with a fixed D of 1.0.

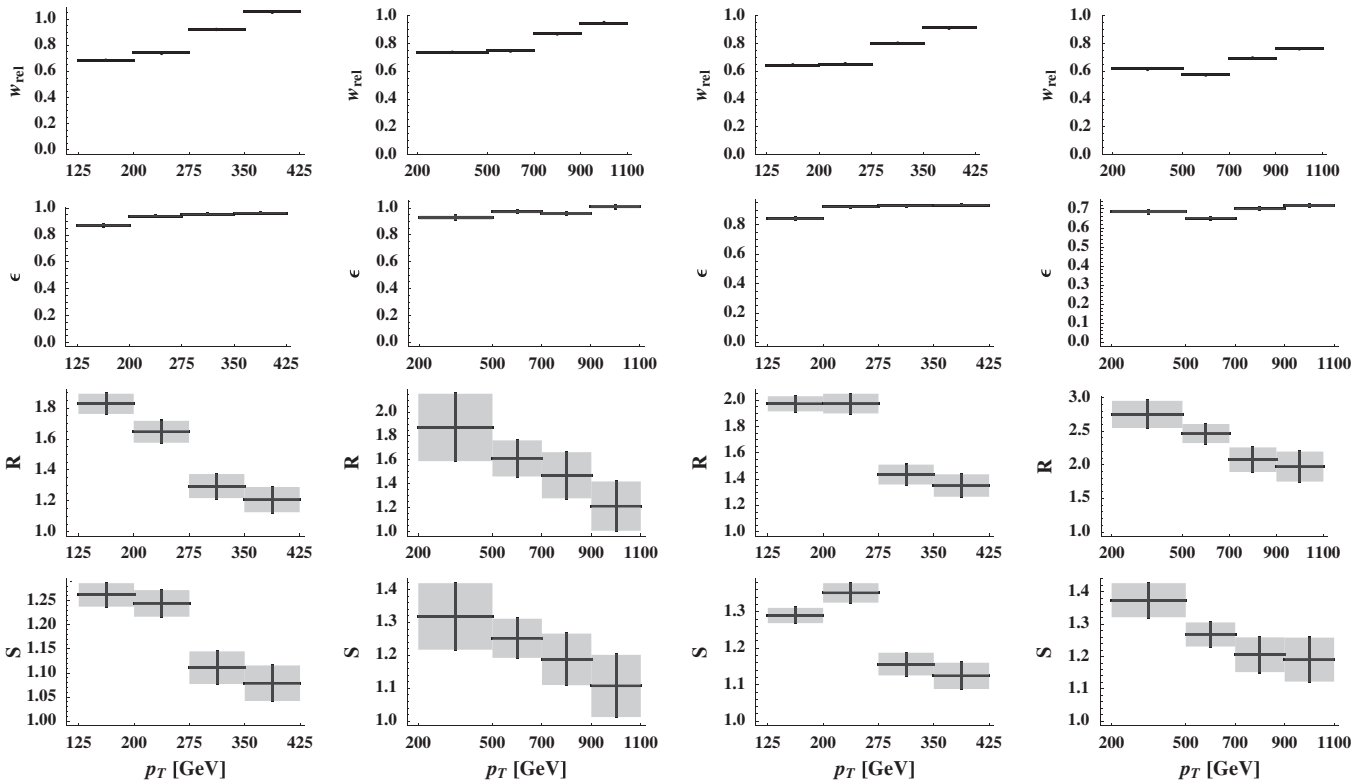
In Fig. 34 we plot the values for w_{rel}, ϵ, R, and S versus p_T bin for W's and tops, using the CA and $\mathrm{k_T}$ algorithms. Pruning improves W and top finding for both algorithms, with substantial improvements for large p_T. The measure S in the smallest p_T bins ranges from 30%–40%, growing to values between 100%–600% in the largest p_T bins. At large p_T in the top quark study, the improvement in signal-to-noise for the CA algorithm is larger than for the $\mathrm{k_T}$ algorithm, as is the relative efficiency to identify tops. This arises because the CA algorithm is poor at reconstructing the W as a subjet of the top jet at large p_T when the value of D is not matched to the opening angle of the decay. We will investigate this case further in the rest of the analysis.

C. Top identification with variable D

For an analysis where the heavy particle mass is known, the jet algorithm can be tailored to the jet p_T. The D value can be chosen using the relation

$$D = \min\left(1.0, 2\frac{m}{p_T}\right), \tag{32}$$

where m is the heavy particle mass and p_T is the transverse momentum of the jet. We take 1.0 to be the maximum allowed value of D. The D values we use are given in Table I. In Fig. 35, we plot w_{rel}, ϵ, R, and S for jets with these D values used for each p_T bin. Note that Eq. (32) neglects the differences between algorithms, which depend on the particular decay. As an example of the fidelity of this relation for D, recall Fig. 18, which plotted the distribution in ΔR for reconstructed parton-level top quark decays with a top boost of $\gamma = 3$. Equation (32) suggests the value $D = 0.7$, while the means of the CA and $\mathrm{k_T}$ distributions for the



(a) W's, CA jets (b) tops, CA jets (c) W's, $\mathrm{k_T}$ jets (d) tops, $\mathrm{k_T}$ jets

FIG. 35. Relative statistical measures w_{rel}, ϵ, R, and S vs p_T for W's and tops, using CA and $\mathrm{k_T}$ jets. Instead of a fixed $D = 1.0$, a tuned D is used for each p_T bin (see Table I). Statistical errors are shown.

reconstructed parton-level decay are 0.75 and 0.65, respectively. Because the distribution in opening angles of the reconstructed decay is broad, by using a smaller, fixed D some decays will not be reconstructed by the jet algorithm.

The difference between the case of constant $D = 1.0$ and variable D is readily apparent. When the D value is matched to the expected opening angle of the decay, the improvements in pruning are flatter over the whole range in p_T, and generally decreasing toward high p_T. The decreased efficiency for pruning, especially for the k_T algorithm, is outweighed by the increases in R and S over the whole range in p_T.

D. Comparing pruning with different D values

In the previous two sections we saw that an unpruned analysis performs much better when D is tuned to the m/p_T of the signal. We now consider whether this is true of a pruned analysis.

In each p_T bin, we can compare the results of pruned jets with $D = 1.0$ with pruned jets using a value of D fit to the expected size of the decay. Because the naive expectation is that the tuned value of D will yield better separation from background, we find the improvements in pruning when D is tuned, relative to pruning with a fixed D of 1.0. Analogous metrics, w_D, ϵ_D, R_D, and S_D, are used, but now they compare the results from pruning with the tuned D value to the results from pruning with $D = 1.0$. For instance,

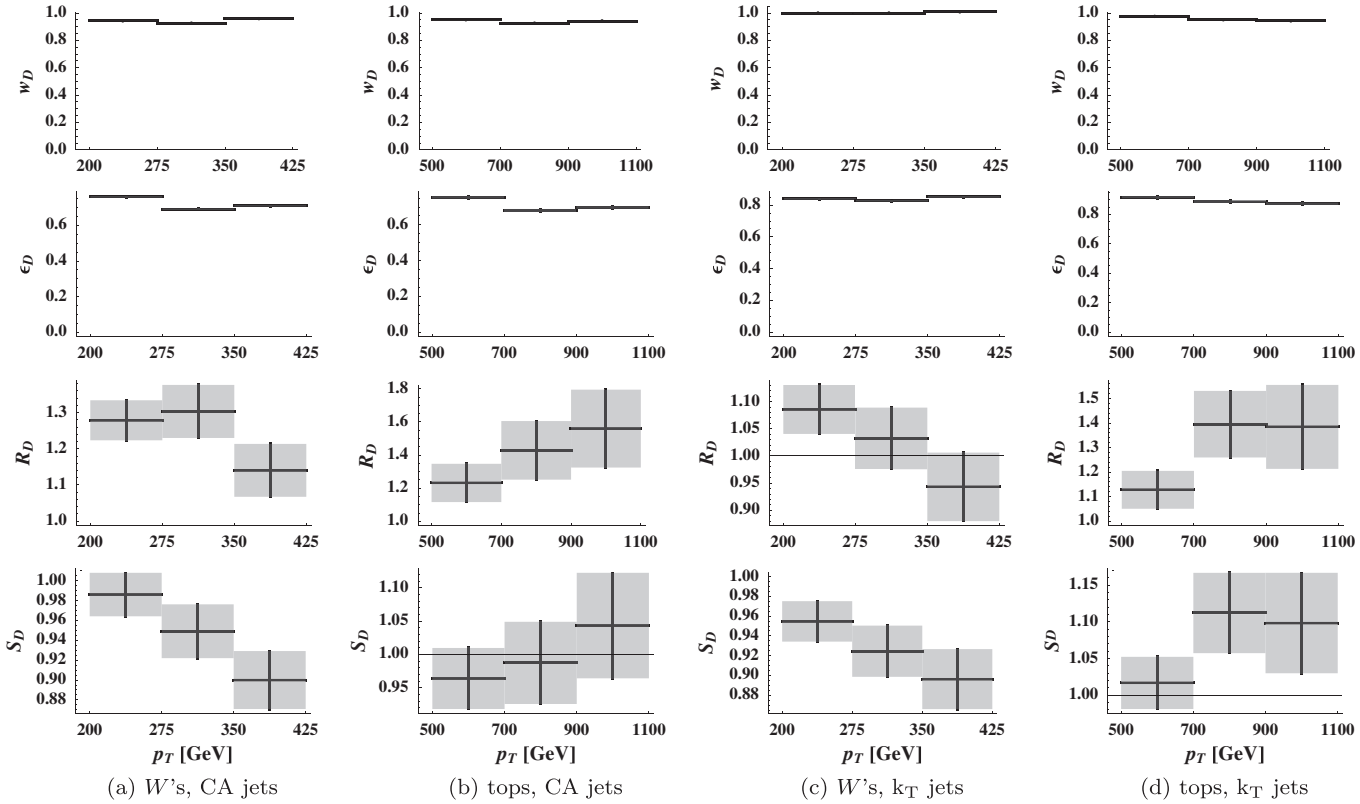
$$R_D \equiv \frac{S/B \text{ from pruning with tuned } D}{S/B \text{ from pruning with } D = 1.0}. \quad (33)$$

Note that $x_D > 1$ indicates that tuning D yields an improvement. The values of these four measures are shown in Fig. 36 over the range of p_T. Note that since the tuned value of D in the smallest p_T bin is 1.0, the comparison there is trivial and so is not shown.

These results show only small improvements in S_D, with the statistical error bars at most data points including the value $S_D = 1$. They indicate that the results after pruning are roughly independent of the value of D used in the jet algorithm, as long as that D is large enough to fit the expected size of the decay in a single jet. From the point of view of heavy particle searches, we can conclude that pruning removes much of the D dependence of the jet algorithm in the search.

E. Absolute measures of pruning

So far, we have only considered measures of pruning relative to a similar analysis without pruning, because this factors out much of the dependence on details of the samples. However, several recent studies report absolute performance metrics for heavy particle identification, so



(a) W's, CA jets (b) tops, CA jets (c) W's, k_T jets (d) tops, k_T jets

FIG. 36. Relative statistical measures w_D, ϵ_D, R_D, and S_D vs p_T for W's and tops, using CA and k_T jets. The measures now compare pruning with a tuned D value in each p_T bin to pruning with a fixed D. Statistical errors are shown.

we examine similar measures here for completeness. In addition, we directly compare the CA and k_T algorithms, with and without pruning.

As can be seen from the plots of w_{rel} in previous sections, pruning reduces the width of the mass distribution for heavy particles. In Fig. 37, we plot the absolute widths of the fitted mass distributions for both the top and W in the $t\bar{t}$ sample and the W in the WW sample, over all p_T bins. We plot this width for the pruned and unpruned versions of the CA and k_T algorithms.

Note that the heavy particle identification method we use in this work selects jets within a range of width 2Γ, with Γ coming from a fit to the signal sample. This gives rise to a mass range cut that is typically much narrower than fixed width ranges used in other studies, and hence the absolute efficiency to identify heavy particles is lower.

In Figs. 38(a) and 38(b), we plot the absolute efficiency to identify tops and W's in the two signal samples for both algorithms, with and without pruning. For the top sample, this efficiency ϵ_{abs} is the ratio

$$\epsilon_{abs} \equiv \frac{\#\text{ of top jets in the signal sample}}{\#\text{ of parton-level tops in the } p_T \text{ range}} \tag{34}$$

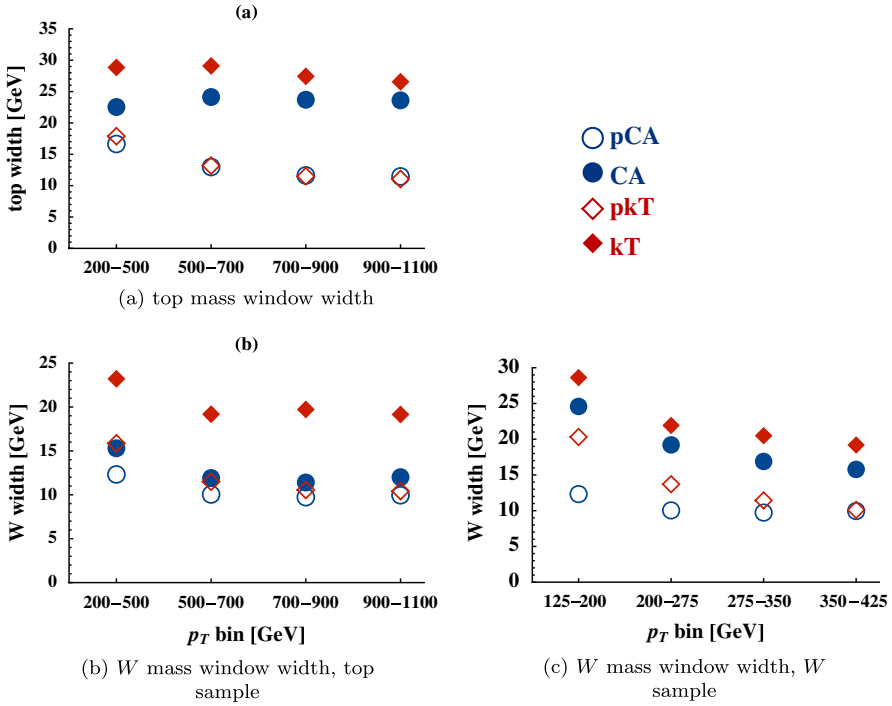
for each p_T bin, with ϵ_{abs} defined analogously for the W sample. Because the substructure of the W decay is much simpler than the top decay, with no secondary mass cut, the absolute identification efficiencies are similar between all algorithms.

The efficiency to find top quarks is only meaningful when compared to the fake rate for QCD jets to be misidentified as a top quark. We define this fake rate as

$$\epsilon_{fake} \equiv \frac{\#\text{ of fake top jets in the background sample}}{\#\text{ of unpruned jets in the } p_T \text{ range}} \tag{35}$$

for each p_T bin, and analogously for the W sample. In Figs. 38(c) and 38(d), we plot ϵ_{fake} for tops and W's in the two background samples for both algorithms, with and without pruning. The fake rate is significantly reduced for pruned jets compared to unpruned jets, for both the top and W studies. The decrease in absolute efficiency arising from using a narrow mass window is compensated by a correspondingly small fake rate for QCD jets.

For top quarks, the efficiencies shown in Fig. 38 can be compared with those given in Table 5 of [26] for several other top-finding methods. Our highest p_T bin is relevant for the comparison. More than a few words of caution are in order, however. Unlike the pruning-to-not-pruning comparisons we have presented so far, comparisons between methods using absolute efficiencies will depend on the details of the signal and background samples, as well as the details of the various cuts included in each analysis. For example, the cuts we have used in this analysis are narrower than fixed mass window cuts used in other top-finding algorithms, and hence our top identification efficiency and background fake rate are both lower than described in other methods. We intend to perform a more

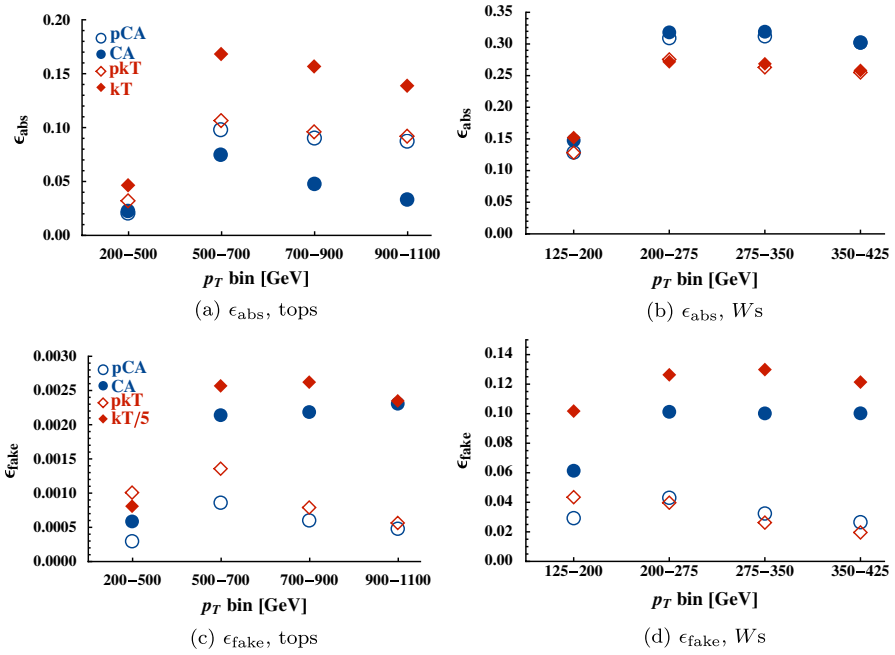


(a) top mass window width

(b) W mass window width, top sample

(c) W mass window width, W sample

FIG. 37 (color online). Widths of the top jet (a), W subjet of the top jet (b), and W jet (c) mass windows for the top and W signal samples.



(a) $\epsilon_{\rm abs}$, tops (b) $\epsilon_{\rm abs}$, Ws

(c) $\epsilon_{\rm fake}$, tops (d) $\epsilon_{\rm fake}$, Ws

FIG. 38 (color online). $\epsilon_{\rm abs}$ and $\epsilon_{\rm fake}$ vs p_T bin, for the CA and k_T algorithms with and without pruning, using $D = 1.0$. A "p" before the algorithm name denotes the pruned version. The legend for (a) applies to (b) and (d)—note the scale difference for k_T jets in (c).

thorough comparison between different substructure approaches in a future work.

F. Algorithm comparison

Throughout this paper, we have studied how pruning compares to not pruning for the CA and k_T algorithms. However, it is also of interest to study how the CA and k_T algorithms compare, with and without pruning. To do this, we use statistical measures w_A, ϵ_A, R_A, and S_A analogous to $w_{\rm rel}$, ϵ, R, and S. For instance,

$$R_A \equiv \frac{S/B \text{ from the CA algorithm with } D = 1.0}{S/B \text{ from the } k_T \text{ algorithm with } D = 1.0}. \quad (36)$$

We will change the subscript to pA to compare the pruned versions of the algorithms, e.g.,

$$R_{pA} \equiv \frac{S/B \text{ from pruned CA with } D = 1.0}{S/B \text{ from pruned } k_T \text{ with } D = 1.0}. \quad (37)$$

In Fig. 39, we plot the measures comparing CA to k_T and pruned CA to pruned k_T for both the WW and $t\bar{t}$ samples.

These comparisons illustrate many of the effects that we have observed throughout this paper. For the unpruned algorithm comparison, CA tends to have a much lower efficiency to identify tops than k_T. As p_T increases, CA performs more poorly relative to k_T, with the efficiency decreasing significantly. This arises because the CA has a decreasing efficiency to identify the W at high p_T, when the top quark becomes more localized in the fixed D jet. Pruning corrects for this, though the performance of CA relative to k_T still decreases at high p_T.

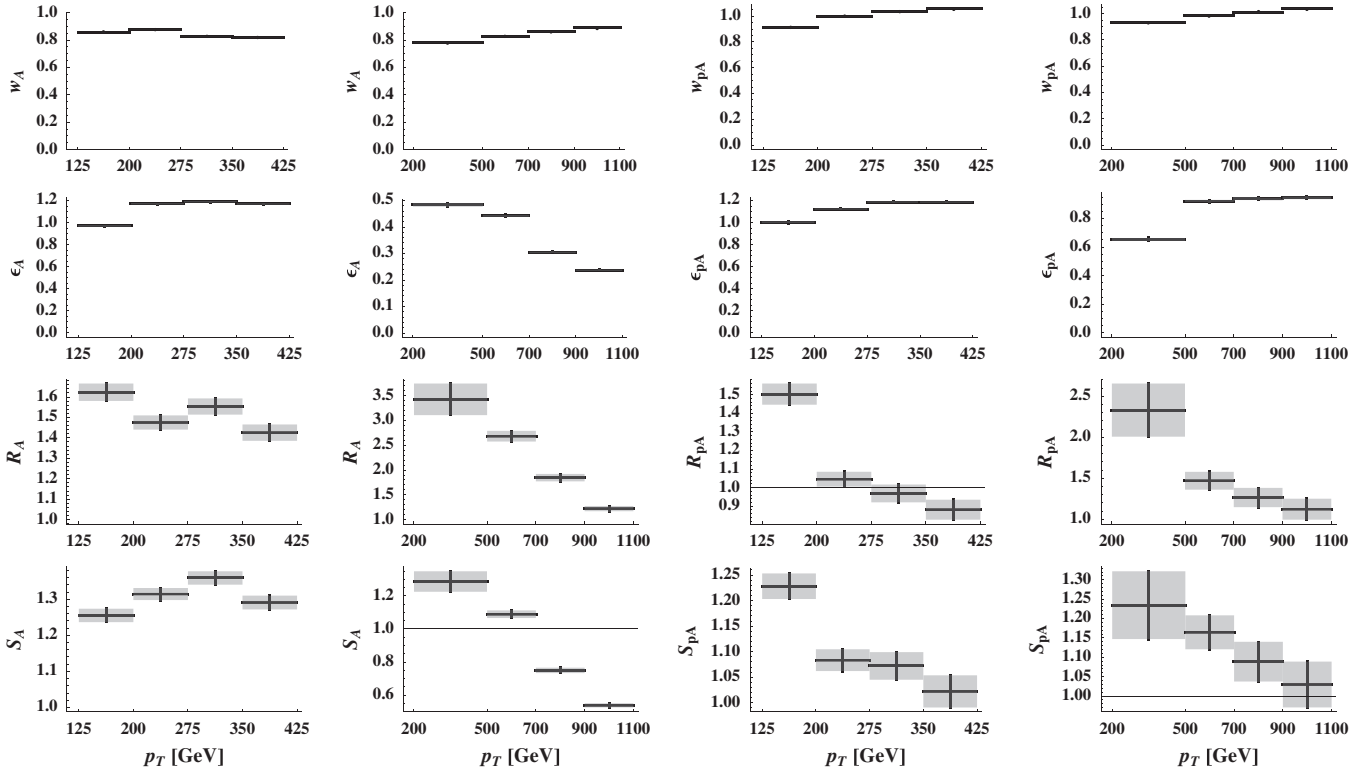
The WW sample is instructive because it lets us compare the effectiveness of pruning between CA and k_T across a wide range in p_T. For the unpruned algorithms, the performance of CA relative to k_T is fairly consistent over all p_T, reflecting the fact that W identification is simpler than top identification, with accurate mass reconstruction the only requirement. However, when the jets are pruned, the performance of pruned CA relative to pruned k_T improves in the smallest p_T bin and worsens in the largest p_T bin, as compared to the performance of CA versus k_T for unpruned jets. This skewing indicates that pruning is more effective for CA than k_T at small p_T, where threshold effects are important, and more effective for k_T than CA at large p_T.

G. Detector effects

So far, no detector simulation has been applied to our events aside from clustering particles into massless calorimeter cells. We now consider a technique that approximates the impact that detector resolution has on the effectiveness of pruning. We modify our top and W jet analyses by smearing the energy E of each calorimeter cell with a factor sampled from a Gaussian distribution with mean E and standard deviation σ given by

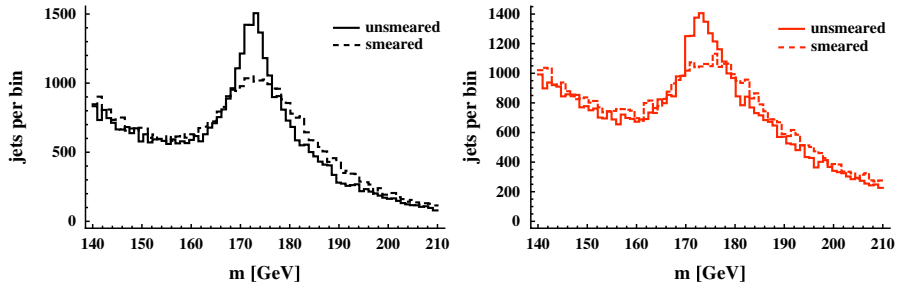
$$\sigma(E) = \sqrt{a^2 E + b^2 + c^2 E^2}. \quad (38)$$

We consider a parameter set motivated by the expected ATLAS hadronic calorimeter resolution [27], $\{a, b, c\} = \{0.65, 0.5, 0.03\}$. One obvious effect of the detector smearing is degraded mass resolution. In Fig. 40, we show this effect by plotting the jet mass distribution for the $t\bar{t}$ sample



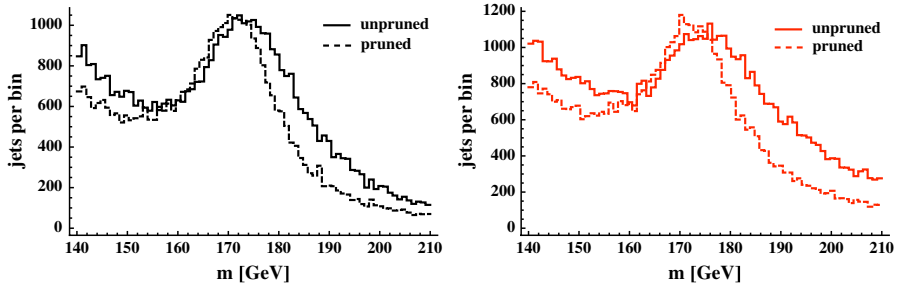
(a) W's, CA vs. k_T (b) tops, CA vs. k_T (c) W's, pCA vs. pk_T (d) tops, pCA vs. pk_T

FIG. 39. Relative statistical measures comparing CA to k_T jets and pruned CA to pruned k_T jets vs p_T for W's and tops, using $D = 1.0$. Statistical errors are shown.



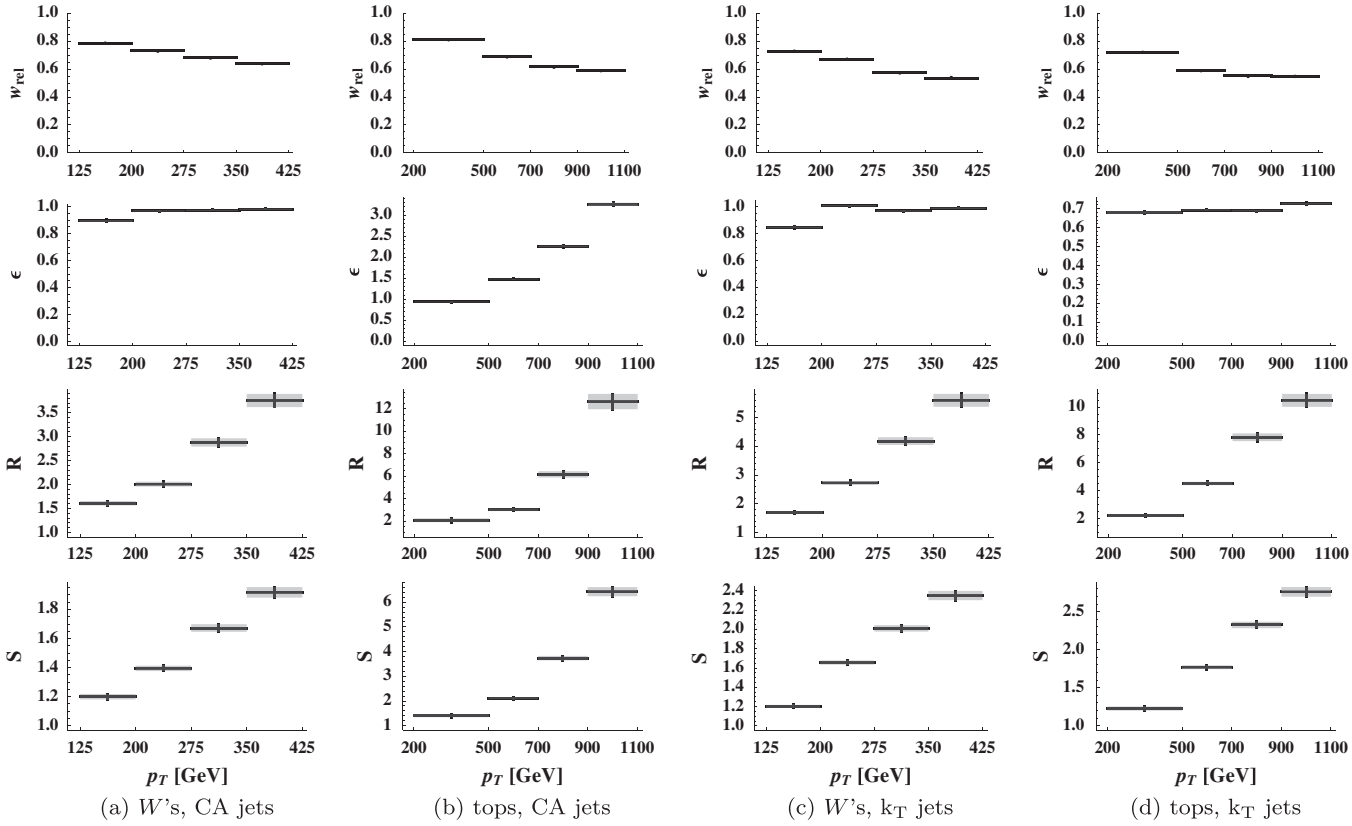
(a) tops, CA jets (b) tops, k_T jets

FIG. 40 (color online). Distribution in jet mass for $t\bar{t}$ events, with (dashed line) and without (solid line) energy smearing. The jets have p_T of 200–500 GeV and $D = 1.0$, and there is no pruning.



(a) tops, pruned CA jets (b) tops, pruned k_T jets

FIG. 41 (color online). Distribution in jet mass for pruned (dashed line) and unpruned (solid line) jets, for $t\bar{t}$ events with energy smearing. The jets have p_T of 200–500 GeV and $D = 1.0$.



(a) W's, CA jets (b) tops, CA jets (c) W's, k_T jets (d) tops, k_T jets

FIG. 42. Relative statistical measures $w_{\rm rel}$, ϵ, R, and S vs p_T for W's and tops, using CA and k_T jets. Calorimeter cell energies are smeared as described in the text. Statistical errors are shown.

in the first p_T bin. Even after smearing, however, pruning improves the jet mass resolution. In Fig. 41, we plot the pruned and unpruned jet mass distributions for the $t\bar{t}$ sample in the first p_T bin. Note that because the QCD jet mass distribution is smooth, only the overall size of the sample in the mass window changes, so we do not plot these distributions.

In Fig. 42, we repeat the basic analysis of Sec. VIII B, applying the detector smearing described above. This figure can be compared to Fig. 34 from the previous analysis, which plots the same measures when no energy smearing is used. The improvements are very similar to those for unsmeared jets, good evidence that pruning may retain its utility in a more realistic detector simulation or in real data.

IX. CONCLUSIONS AND FUTURE PROSPECTS

In this work, we have demonstrated that recombination jet algorithms shape the substructure of heavy particles reconstructed in single jets. We have identified regions in the variables z and ΔR where individual recombinations are unlikely to represent the kinematics of a reconstructed heavy particle. Specifically, soft, large-angle recombinations are unlikely to arise from the accurate reconstruction of a heavy particle decay, and are likely to come from QCD jets, uncorrelated radiation, or systematic effects of the jet algorithm. For the CA algorithm, we have demonstrated that these soft, large-angle recombinations are a key systematic effect that shapes the substructure of the jet, in particular, the final recombinations.

We have presented a procedure, called pruning, that eliminates soft, large-angle recombinations from the substructure of the jet. Using hadronically decaying top quarks and W bosons as test cases, we have demonstrated that the pruning procedure improves the separation between heavy particle decays and a QCD multijet background. We have motivated the parameters of the pruning procedure and demonstrated that they roughly optimize the improvements from pruning in our study for both top quarks and W bosons.

Our studies on pruning have demonstrated many positive results of the procedure. In a heavy particle search, the jet is sensitive to the parameter D, and if the value of D is not well matched to the decay of a heavy particle then the ability to identify that particle in single jets is greatly reduced. Our results indicate that pruning removes much of the jet algorithm's dependence on D. Pruning shows improvements even when D is adjusted to fit the expected decay of the heavy particle. We have demonstrated that pruning largely removes the effects of the underlying event, as the underlying event mainly contributes soft,

uncorrelated radiation that can be pruned away. Additionally, we have shown that the results of pruning are robust to a basic energy smearing applied to the calorimeter cells used to seed the jet algorithm. Finally, we have quantified absolute measures of the pruning procedure that can be used to compare to other jet substructure methods.

It should be reiterated that pruning systematizes methods that have been proposed by other authors for specific searches. Pruning should be applicable to a wide range of searches, and is intended to be a generic jet analysis tool. We have detailed the ideas behind why pruning works and why it should be used, and presented an in-depth discussion of many of the physics issues arising when studying jet substructure.

Future prospects

The conclusions in this paper, like those for any analysis technique not demonstrated on real data, must be taken cautiously. This is especially true for studies like this one on jet substructure, where a majority of the work has been in exploring techniques that may—or may not—actually be useful in an experiment. However, new techniques like jet substructure offer great promise. All studies thus far indicate that jet substructure, and in general a more innovative approach to jets, will be a useful tool for understanding the physics in events with jets at collider experiments.

The most obvious and immediate application of pruning, and jet substructure tools in general, is in rediscovery of the standard model at the LHC. As the LHC collects data from high-energy collisions, there will be an abundant sample of high-p_T top quarks, and W and Z bosons with fully hadronic decays. As these channels are observed using standard analyses, jet substructure techniques can be applied and tested. These channels can also serve as key calibration tools for jet substructure methods applied in the search for new physics.

From the theoretical side, improvements in jet-based analyses can come from a variety of sources. As calculations in perturbative QCD progress, they can be used to improve predictions for jet-based observables in QCD. Improved Monte Carlo tools, such as the continued implementation of next-to-leading order matrix elements and better parton showers, will lead to more accurate studies and a better understanding of jet physics. Additionally, the framework of soft-collinear effective theory (SCET) [28–32] can improve the understanding of QCD jets. As SCET is adapted to describe a wider variety of event topologies and realistic jet algorithms are implemented in the effective theory, it can be used to calculate resummed predictions [33–35] for jet-based observables and accurately describe processes that are difficult to access with fixed-order perturbative QCD. Jets will likely play a central role in new physics searches at the LHC, and a better understanding of jets and jet substructure can aid in the discovery process.

ACKNOWLEDGMENTS

We would like to thank Matt Strassler, Jacob Miner, and Andrew Larkoski for collaboration in early stages of this work. We thank Johan Alwall for help with MADGRAPH/MADEVENT, and acknowledge useful discussions with Steve Mrenna, Gavin Salam, Tilman Plehn, Karl Jacobs, Peter Loch, Michael Peskin, and others in the context of the Joint Theoretical-Experimental Terascale Workshops at the Universities of Washington and Oregon, supported by the U.S. Department of Energy under Task TeV of Grant No. DE-FG02-96ER40956. This work was supported in part by the U.S. Department of Energy under Grant No. DE-FG02-96ER40956. J. R. W. was also supported in part by an LHC Theory Initiative Graduate Program.

APPENDIX: COMPUTATIONAL DETAILS

We give a brief summary of the computational tools employed to do the studies in this paper. We generate LHC (14 TeV) events using MADGRAPH/MADEVENT V4.4.21 [36] interfaced with PYTHIA V6.4 [37]. We employ MLM-style matching, implemented in MADGRAPH (see, e.g., [38]), on the backgrounds. We have checked that our matching parameters are reasonable using the tool MATCHCHECKER [39]. We use the DWT tune [40] in PYTHIA to give a ''noisy'' underlying event (UE). For the hadron-level studies in Secs. III and IV, we exclude the underlying event by setting the PYTHIA parameter MSTP (81) to zero, turning off multiple interactions. The UE comparisons in Sec. V compare samples with this parameter set at 0 or 1. No detector simulation is performed so we can isolate the ''best case'' effects of our method. In Sec. VIII G, we examine the effects of Gaussian smearing on the energies of final-state particles from PYTHIA to get a sense for how much the results may change with a detector.

For the W study, the signal sample is W^+W^- pair production, with exactly one W required to decay leptonically. The background is a matched sample of a W and one or two light partons (gluons and the four lightest quarks) before showering. These partons must be in the central region, $|\eta| < 2.5$. η is the pseudorapidity, $\eta \equiv \ln(\cot(\theta_b/2))$, with θ_b the polar angle with respect to the beam direction ($\eta = y$ for massless particles). Signal and background samples are divided into four p_T bins: [125, 200], [200, 275], [275, 350], and [350, 425] (all in GeV). Each bin is defined by a p_T cut that is applied to single jets in the analysis. These bins confine the W boost to a narrow range and allow us to study the performance of pruning as the jet p_T (or W boost) varies.

For each p_T bin $[p_T^{\min}, p_T^{\max}]$, both samples are generated with a p_T cut on the leptonic W of $p_T^{\min} - 25$ GeV. For the

background, we set the matching scales ($Q_{\mathrm{cut}}^{\mathrm{ME}}$, Q_{match}) to be (10, 15) GeV in all four bins.

For the top quark reconstruction study, the signal sample is $t\bar{t}$ production with fully hadronic decays. The background is a matched sample of QCD multijet production with two, three, or four light partons, with the same cut on parton centrality as in the W study. Samples are again divided into four p_T bins: [200, 500], [500, 700], [700, 900], and [900, 1100] (all in GeV).

We generate signal and background samples with a parton-level h_T cut for generation efficiency, where h_T is the scalar sum of all p_T in the event. For each p_T bin $[p_T^{\mathrm{min}}, p_T^{\mathrm{max}}]$, the parton-level h_T cut is $p_T^{\mathrm{min}} - 25\ \mathrm{GeV} \leq h_T/2 \leq p_T^{\mathrm{max}} + 100\ \mathrm{GeV}$. For the background, we use matching scales (20, 30) GeV for the smallest p_T bin and (50, 70) GeV in the other three bins.

From the hadron-level output of PYTHIA, we group final-state particles into "cells" based on the segmentation of the ATLAS hadronic calorimeter ($\Delta\eta = 0.1$, $\Delta\phi = 0.1$ in the central region). We sum the four-momenta of all particles in each cell and rescale the resulting three-momentum to make the cell massless. After a threshold cut on the cell energy of 1 GeV, cells become the inputs to the jet algorithm. Our implementation of recombination algorithms uses FASTJET [41], with a pruning plugin we have written [42].

Several of the plots in early sections involve mass cuts on jets. The details of these cuts are provided in Sec. VII B.

[1] S. D. Ellis, J. Huston, K. Hatakeyama, P. Loch, and M. Tonnesmann, Prog. Part. Nucl. Phys. **60**, 484 (2008).
[2] M. J. Strassler and K. M. Zurek, Phys. Lett. B **651**, 374 (2007).
[3] D. E. Acosta *et al.* (CDF Collaboration), Phys. Rev. D **71**, 112002 (2005).
[4] R. Akers *et al.* (OPAL Collaboration), Z. Phys. C **63**, 197 (1994).
[5] R. Akers *et al.* (OPAL Collaboration), Z. Phys. C **63**, 363 (1994).
[6] J. M. Butterworth, B. E. Cox, and J. R. Forshaw, Phys. Rev. D **65**, 096014 (2002).
[7] G. Brooijmans, Report No. ATL-PHYS-CONF-2008-008, 2008.
[8] J. Thaler and L.-T. Wang, J. High Energy Phys. 07 (2008) 092.
[9] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, Phys. Rev. Lett. **101**, 142001 (2008).
[10] L. G. Almeida, S. J. Lee, G. Perez, G. Sterman, I. Sung, and J. Virzi, Phys. Rev. D **79**, 074017 (2009).
[11] T. Plehn, G. P. Salam, and M. Spannowsky, Phys. Rev. Lett. **104**, 111801 (2010).
[12] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Phys. Rev. Lett. **100**, 242001 (2008).
[13] J. M. Butterworth, J. R. Ellis, and A. R. Raklev, J. High Energy Phys. 05 (2007) 033.
[14] J. M. Butterworth, J. R. Ellis, A. R. Raklev, and G. P. Salam, Phys. Rev. Lett. **103**, 241803 (2009).
[15] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Phys. Rev. D **80**, 051501 (2009).
[16] M. H. Seymour, Z. Phys. C **62**, 127 (1994).
[17] M. Cacciari, G. P. Salam, and G. Soyez, J. High Energy Phys. 04 (2008) 063.
[18] T. Aaltonen *et al.* (CDF Collaboration), Phys. Rev. D **78**, 052006 (2008).
[19] S. Catani, Yu. L. Dokshitzer, and B. R. Webber, Phys. Lett. B **285**, 291 (1992).
[20] S. Catani, Yu. L. Dokshitzer, M. H. Seymour, and B. R. Webber, Nucl. Phys. **B406**, 187 (1993).
[21] S. D. Ellis and D. E. Soper, Phys. Rev. D **48**, 3160 (1993).
[22] Yu. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, J. High Energy Phys. 08 (1997) 001.
[23] Z. Kunszt and D. E. Soper, Phys. Rev. D **46**, 192 (1992).
[24] M. Cacciari, G. P. Salam, and G. Soyez, J. High Energy Phys. 04 (2008) 005.
[25] D. Krohn, J. Thaler, and L.-T. Wang, J. High Energy Phys. 06 (2009) 059.
[26] G. P. Salam, arXiv:0906.1833.
[27] G. Aad *et al.* (ATLAS Collaboration), J. Instrumentation **3**, S08003 (2008).
[28] C. W. Bauer, S. Fleming, D. Pirjol, and I. W. Stewart, Phys. Rev. D **63**, 114020 (2001).
[29] C. W. Bauer, S. Fleming, and M. E. Luke, Phys. Rev. D **63**, 014006 (2000).
[30] C. W. Bauer, D. Pirjol, and I. W. Stewart, Phys. Rev. D **65**, 054022 (2002).
[31] C. W. Bauer, S. Fleming, C. Lee, and G. Sterman, Phys. Rev. D **78**, 034027 (2008).
[32] C. W. Bauer, A. Hornig, and F. J. Tackmann, Phys. Rev. D **79**, 114013 (2009).
[33] C. W. Bauer and M. D. Schwartz, Phys. Rev. Lett. **97**, 142001 (2006).
[34] I. W. Stewart, F. J. Tackmann, and W. J. Waalewijn, arXiv:0910.0467.
[35] W. M.-Y. Cheung, M. Luke, and S. Zuberi, Phys. Rev. D **80**, 114021 (2009).
[36] J. Alwall, P. Demin, S. de Visscher, R. Frederix, M. Herquet, F. Maltoni, T. Plehn, D. L. Rainwater, and T. Stelzer, J. High Energy Phys. 09 (2007) 028.
[37] T. Sjöstrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026.
[38] J. Alwall, S. de Visscher, and F. Maltoni, J. High Energy Phys. 02 (2009) 017.
[39] P. Demin and S. de Visscher, MATCHCHECKER, 2007, http://cp3wks05.fynu.ucl.ac.be/twiki/bin/view/Software/MatchChecker.
[40] M. G. Albrow *et al.* (TeV4LHC QCD Working Group), arXiv:hep-ph/0610012.
[41] M. Cacciari and G. P. Salam, Phys. Lett. B **641**, 57 (2006).
[42] C. K. Vermilion, FASTPRUNE, 2009, http://bit.ly/pruning.