Published for SISSA by Springer

Received: *October 21, 2017*
Revised: *January 17, 2018*
Accepted: *March 11, 2018*
Published: *March 14, 2018*

Novel jet observables from machine learning

Kaustuv Datta and Andrew J. Larkoski

Physics Department, Reed College,
Portland, OR 97202, U.S.A.

E-mail: dattak@alumni.reed.edu, larkoski@reed.edu

Abstract: Previous studies have demonstrated the utility and applicability of machine learning techniques to jet physics. In this paper, we construct new observables for the discrimination of jets from different originating particles exclusively from information identified by the machine. The approach we propose is to first organize information in the jet by resolved phase space and determine the effective N-body phase space at which discrimination power saturates. This then allows for the construction of a discrimination observable from the N-body phase space coordinates. A general form of this observable can be expressed with numerous parameters that are chosen so that the observable maximizes the signal vs. background likelihood. Here, we illustrate this technique applied to discrimination of $H \to b\bar{b}$ decays from massive $g \to b\bar{b}$ splittings. We show that for a simple parametrization, we can construct an observable that has discrimination power comparable to, or better than, widely-used observables motivated from theory considerations. For the case of jets on which modified mass-drop tagger grooming is applied, the observable that the machine learns is essentially the angle of the dominant gluon emission off of the $b\bar{b}$ pair.

Keywords: Jets, QCD Phenomenology

ArXiv ePrint: 1710.01305

https://doi.org/10.1007/JHEP03(2018)086

Contents

1 Introduction

Several groups have recently applied promising machine learning techniques to the problem of classifying jets from different originating particles [1–17]. A review of the advances of the field is presented in ref. [18]. These approaches, while demonstrating exceptional discrimination power, often come with the associated costs of utilizing hundreds of low-level input variables with thousands of correlations between them, and lack an immediately accessible physical interpretation. Ref. [19] presented the first application of a bottom-up organizing principle, whereby neural networks were trained and tested on minimal and complete bases of observables sensitive to the phase space of M subjets in a jet. By appropriately identifying M subjets in a jet, this study probed their phase space with sets of $3M - 4$ infrared and collinear (IRC) safe observables. This essentially utilizes the distribution of the M subjets on the phase space to identify useful information for discrimination. By increasing the dimensionality of the phase space in a systematic way, ref. [19] used machine learning to demonstrate that in the case of discriminating boosted hadronic decays of Z bosons from jets initiated by light partons, once 4-body phase space is resolved, no more information is observed to contribute meaningfully to discrimination power. In addition, ref. [20] also presented a first application of this method to developing promising generic anti-QCD taggers that match or outperform the discrimination power of dedicated taggers. Such approaches motivate the development of novel observables that can capture all of the salient information for discrimination of jets as learned by the machine.

While the output of a neural network, boosted decision tree, or other machine learning method is itself an observable, it is in general a highly non-linear function of the input. Additionally, the precise form of the explicit observable constructed by the machine is very sensitive to the assumed parameters; for example, the number of nodes or layers in a neural network. In this paper, we propose a procedure to identify discriminating features of jets learned by machines to then generate novel observables. These observables capture the important physics identified by the machine while at the same time being human-parseable. The general procedure is as follows:

1. Construct a basis of observables that is sensitive to the phase space of subjets in a jet. Measure these basis observables on your signal and background samples.

2. Use machine learning techniques, such as neural networks, to identify the resolved M-body phase space at which signal vs. background discrimination power saturates.

3. Construct a function of the phase space variables (with tunable parameters) at which discrimination power saturates. This function will be a new observable on the jets that can be used individually for discrimination.

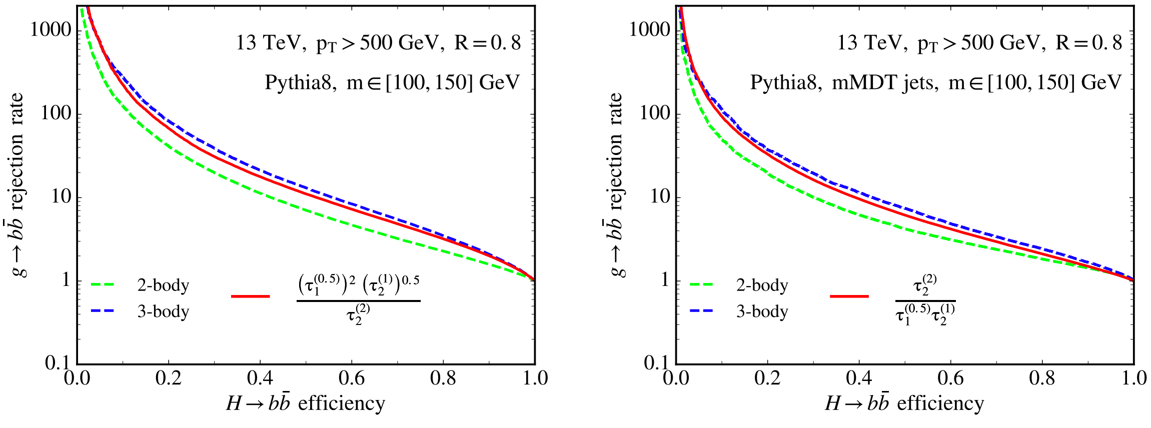
4. Fix the parameters in the new observable by demanding that it maximizes some discrimination metric, such as the area under the signal vs. background efficiency curve (ROC curve).

This algorithm is simple enough that it can be automated with essentially no human input, with a specified basis of observables to span M-body phase space and an appropriate functional form for the observable. We will present and use a particular choice for the phase space basis and functional form of the final observable in this paper, but these may need to be modified and optimized for different studies.

For concreteness, here we apply the above approach to the problem of discriminating highly boosted decays of the Standard Model Higgs boson to a pair of b-quarks from splittings of gluons to b-quarks. We study identification of $H \to b\bar{b}$ decays here as the signal and background jets both have a two-prong substructure, and theoretically-optimized discrimination observables have not been studied in great detail. Recently, ref. [21] utilized jet substructure approaches to propose a promising search strategy for this boosted decay mode of the Higgs, encouraging the possibility of discovery in data from Run-II of the LHC. In order to further increase the probability of discovery, it is necessary to explore new strategies to ensure sensitivity to the specific features of this decay mode.

Using the organizing principle proposed in ref. [19], if discrimination power saturates at M-body phase space then the machine must be learning some function of the corresponding $3M - 4$ phase space variables. To resolve the M-body phase space, we use the N-subjettiness observables [22, 23], as employed in ref. [19]. In the case of discrimination of boosted $H \to b\bar{b}$ decays from $g \to b\bar{b}$ splittings, we find that the discrimination power increases only slightly once 3-body phase space is resolved. Thus, we will only study the resolved 3-body phase space in this paper. The function of the observable on 3-body phase space that we study is a simple product form

$$\beta_3 \equiv \left(\tau_1^{(0.5)}\right)^a \left(\tau_1^{(1)}\right)^b \left(\tau_1^{(2)}\right)^c \left(\tau_2^{(1)}\right)^d \left(\tau_2^{(2)}\right)^e . \tag{1.1}$$



(a) **(b)**

Figure 1. $H \to b\bar{b}$ jet efficiency vs. $g \to b\bar{b}$ jet rejection rate plots for ungroomed (a) and groomed (b) jets. For both, the M-body curves determined by a neural network demonstrate a large increase in discrimination power between 2- and 3-body phase space. Also shown (in red) is the new observable which captures the majority of information important for discrimination identified by resolving 3-body phase space.

Here, the $\tau_N^{(\beta)}$ are the N-subjettiness observables, 3-body phase space is 5 dimensional, and the parameters a, b, c, d, e will be chosen to maximize discrimination power. We emphasize that while this product form is simple, there may be a better choice for the form of function on phase space.

We show in figure 1 the results of this analysis. We consider jets in simulation on which no grooming has been applied and on which the modified mass-drop tagger (mMDT) [24, 25] has been applied. We then measure the mass m_J of these ungroomed or groomed jets (as appropriate) and make a cut of $m_J \in [100, 150]$ GeV, in the range of the Higgs peak. On these jets, we then measure the 2- and 3-body phase space variables and determine the single function of the 3-body phase space variables as described above. The signal and background efficiencies in figure 1 do not include the effects of the mass cut, so that the curves end at 100% signal and background efficiency. The ROC curve for the new observable is shown as the solid line on these plots, and exhibits significant improvement over the 2-body phase space observables and nearly captures all discrimination power of 3-body phase space. In the case of mMDT jets, we will show that this new observable effectively corresponds to the angle of the dominant gluon emission off of the $b\bar{b}$ pair.

The outline of this paper is as follows. In section 2 we review and define the minimal and complete observable bases that are used to identify the coordinates of M-body phase space, and discuss the mMDT groomer. In section 3, we describe our event simulation and machine learning implementation, and demonstrate the application of our procedure for developing the new observable in the case of $H \to b\bar{b}$ vs. $g \to b\bar{b}$. Further, we explore the physics implications of the functional form of the observable, and compare its discrimination power to standard observables motivated from QCD considerations. We conclude in section 4 and discuss other possible applications of this procedure.

2 Observable basis

In this section, we discuss the basis of IRC safe observables that we use to identify structure in the jet, following the approach presented in ref. [19]. For our analysis, we exclusively use the N-subjettiness observables [22, 23, 26]. This is without loss of generality and the analysis can, for example, be equivalently implemented with the N-point energy correlation functions [27] or the four-momentum of subjets from the exclusive k_T algorithm. This specific choice for each M-body basis is only to ensure that the set of observables minimally and completely span the phase space of emissions in a jet.

The N-subjettiness observable $\tau_N^{(\beta)}$ provides a measure of the radiation about N axes in the jet, specified by the angular exponent $\beta > 0$:

$$\tau_N^{(\beta)} = \frac{1}{p_{TJ}} \sum_{i \in \text{Jet}} p_{Ti} \min\left\{R_{1i}^\beta, R_{2i}^\beta, \ldots, R_{Ni}^\beta\right\} . \tag{2.1}$$

Here, p_{TJ} is the transverse momentum of the jet of interest, p_{Ti} is the transverse momentum of particle i in the jet, and R_{Ki}, for $K = 1, 2, \ldots, N$, is the angle in pseudorapidity and azimuth between particle i and axis K in the jet. In our analyses, we choose to define the N axes in the jet according to the exclusive k_T algorithm [28, 29] with E-scheme recombination [30].

To identify structure in the jet, we use the organizing principle proposed in ref. [19] so that our choice of basis of observables is complete and minimal. We first identify the set of N-subjettiness observables that completely specify the coordinates of M-body phase space. Since M-body phase space is $3M-4$ dimensional, we only measure $3M-4$ N-subjettiness observables, as follows:

$$\left\{\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(0.5)}, \tau_2^{(1)}, \tau_2^{(2)}, \ldots, \tau_{M-2}^{(0.5)}, \tau_{M-2}^{(1)}, \tau_{M-2}^{(2)}, \tau_{M-1}^{(1)}, \tau_{M-1}^{(2)}\right\} . \tag{2.2}$$

For further details on this method, we ask the reader to refer to ref. [19]. Note that when all particles have non-zero energy and are at a finite angle to one another, the $3(M-2)+2 = 3M-4$ observables span the space of phase space variables for generic momenta configurations. Following from the above, we list the sets of observables that were used in our analysis for resolving particular M-body phase space:

2-body: $\tau_1^{(1)}, \tau_1^{(2)}$

3-body: $\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(1)}, \tau_2^{(2)}$

4-body: $\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(0.5)}, \tau_2^{(1)}, \tau_2^{(2)}, \tau_3^{(1)}, \tau_3^{(2)}$

5-body: $\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(0.5)}, \tau_2^{(1)}, \tau_2^{(2)}, \tau_3^{(0.5)}, \tau_3^{(1)}, \tau_3^{(2)}, \tau_4^{(1)}, \tau_4^{(2)}$

6-body: $\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(0.5)}, \tau_2^{(1)}, \tau_2^{(2)}, \tau_3^{(0.5)}, \tau_3^{(1)}, \tau_3^{(2)}, \tau_4^{(0.5)}, \tau_4^{(1)}, \tau_4^{(2)}, \tau_5^{(1)}, \tau_5^{(2)}$

2.1 mMDT grooming algorithm

In the analysis of the next section, we measure the aforementioned observables on samples of both ungroomed jets and jets groomed with the modified mass-drop tagger (mMDT) [24, 25]. Given a set of constituents of a jet with radius R, and for a fixed transverse momentum fraction parameter z_{cut}, the mMDT grooming algorithm proceeds as follows:

1. Recluster the jet with the Cambridge/Aachen (C/A) algorithm [31–33].
2. Sequentially step through the branching history of the reclustered jet. At each branching with daughter branches i and j, check the mMDT criterion

$$\frac{\min(p_{Ti}, p_{Tj})}{p_{Ti} + p_{Tj}} > z_{\text{cut}}. \tag{2.3}$$

   If the condition fails, drop the softer of two daughter branches and follow through to the next branching in the rest of the clustering history.
3. This continues until the mMDT criterion is passed. At this point the algorithm terminates and the final jet is groomed of all branches that fail to pass the test. This ensures that the softest emissions at wide angles from the hard subjets, and contamination from the underlying event (UE) and initial state radiation (ISR), are effectively removed from the final groomed jet.

In the groomed case, the observable bases are measured after grooming, thus the collection of particles in the jet effectively contributing to the phase space are different than for ungroomed jets.

3 How to make an observable

In this section, we describe our event simulation and implementation of machine learning to the N-subjettiness basis of observables described in the previous section. We generate background $pp \to Z + b\bar{b}$ and signal $pp \to Z(H \to b\bar{b})$ events at the 13 TeV LHC with MadGraph5 v2.5.4 [34]. The Z bosons were used as a control and decayed exclusively to neutrinos. These tree-level events are then showered in Pythia v8.226 [35, 36] with default settings. Later in this section we will also show results obtained by applying the observable learned from Pythia to events showered with Herwig v7.1.1 [37, 38]. We use FastJet v3.2.1 [39, 40] to cluster the jets. On the clustered anti-k_T [41] jets with radius $R = 0.8$ and minimum p_T of 500 GeV, we then measure the basis of N-subjettiness observables using the code provided in FastJet contrib v1.026. The observables are measured at the particle level and we do not apply any detector simulation.

We then study these jets without grooming and with mMDT grooming with $z_{\text{cut}} = 0.1$. On these samples, we measure the jet mass and apply a cut of $100 < m_J < 150$ GeV which selects the Higgs signal region. Additionally, we measure the sufficient collection of N-subjettiness observables to completely determine up through 6-body phase space. We then proceed to develop novel observables learned from the machine that further discriminate signal and background.

To do this, we use the approach of ref. [19]. This enables us to identify the resolved phase space that captures the vast majority of the discrimination power. To calculate the

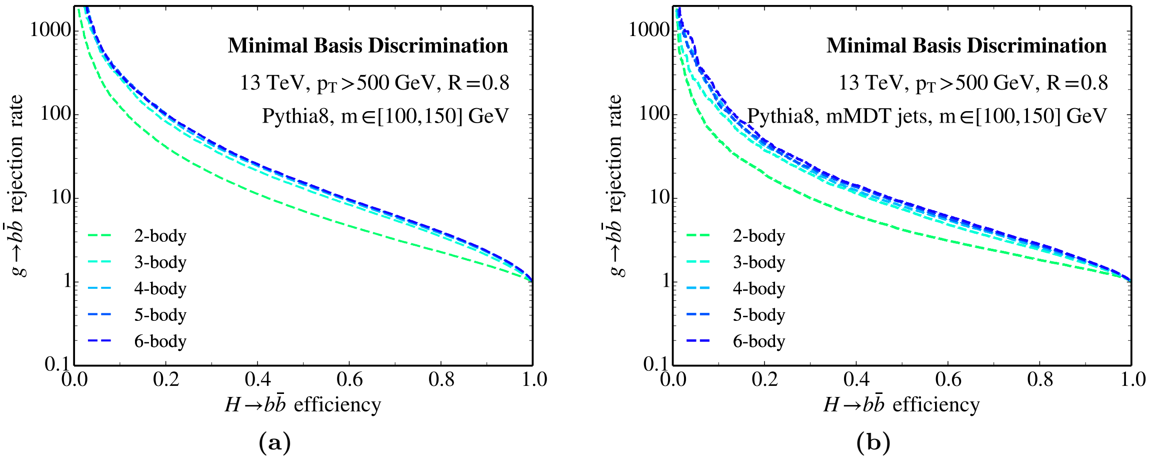


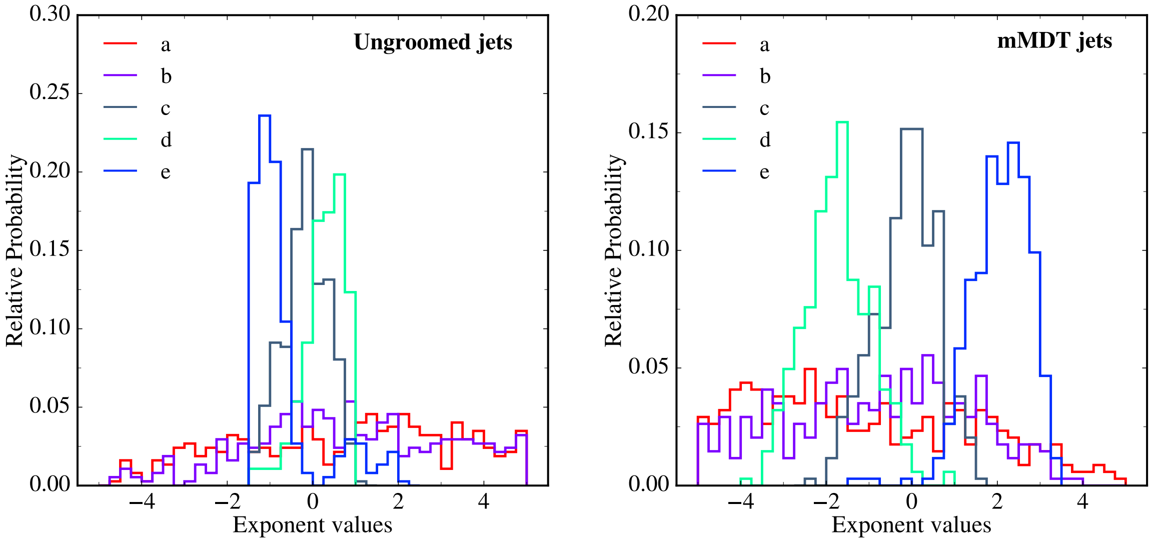
Figure 2. $H \to b\bar{b}$ jet efficiency vs. $g \to b\bar{b}$ jet rejection rate plot for the ungroomed (a) and groomed (b) jets, as determined by a neural network. The curves effectively demonstrate saturation of discrimination power on the resolution of 3-body phase space.

ROC curves for M-body phase space shown in figure 2, we trained deep neural networks with fully-connected layers on the bases of N-subjettiness observables discussed in the previous section. Discrimination power is seen to dramatically increase on going from 2- to 3-body phase space, and higher phase space improves discrimination only slightly. All networks were trained using the Keras [42] deep learning libraries. However, it is important to note here that this procedure, as justified in ref. [19], is agnostic of the specific machine learning method used and, equivalently, other machine learning methods (like a boosted decision tree) could have been used to identify the point of saturation. From these results, in what follows we will exclusively study 3-body phase space.

From the assumption that 3-body phase space effectively saturates discrimination power, our goal is to define a single observable that captures this discrimination power. As previously discussed, this proposed observable must be a function of the phase space variables at the point of saturation. Determining this function thus requires parametrizing the possible functions of the phase space variables somehow. Our approach will be illustrative and demonstrate the procedure for doing so. However, there may be smarter or more effective ways to optimize this process. Here, we will just consider the observable formed from the product of 3-body (5 dimensional) phase space variables, raised to powers a, b, c, d and e:

$$\beta_3 = \left(\tau_1^{(0.5)}\right)^a \left(\tau_1^{(1)}\right)^b \left(\tau_1^{(2)}\right)^c \left(\tau_2^{(1)}\right)^d \left(\tau_2^{(2)}\right)^e . \tag{3.1}$$

At this stage, the optimal values of these powers are undetermined, and there is no guarantee that this form of the observable actually includes all discrimination power of 3-body phase space. We leave the problem of a complete observable basis to future work.



(a) **(b)**

Figure 3. Histograms for the values of exponents of the product observable. For ungroomed (a) and groomed (b) jets exponent values in these histograms were accepted when the generated AUCs for the binned signal and background likelihood distributions were above 0.81 and 0.73 respectively.

3.1 Determining optimal parameters

Utilizing the measurements of N-subjettiness observables from our datasets of groomed and ungroomed jets, we run a Monte Carlo simulation whereby uniform random numbers in the range $[-5, 5]$ were assigned to the exponents a, b, c, d and e.[1] In each run, values of the resultant product observable were measured on samples of 200,000 signal and background jets from Pythia that passed the mass cut of $m_J \in [100, 150]$ GeV. We then construct the 1 dimensional binned likelihood distributions of the observable, which is the optimal discriminant for a given functional form of the observable by the Neyman-Pearson lemma [43]. The likelihood distributions of the observable, for each set of exponent values, were then used to calculate the area under the ROC curve (AUC) to estimate the discrimination power. For each run, the values of the exponents and the calculated AUC were stored only when the AUC crossed a threshold value of 0.5. This was necessary to exclude binning effects on the measured discrimination power of the observable.

We apply this procedure to jets that have been groomed with mMDT and those that have not. In the groomed case, due to the exclusion of soft emissions and contamination from initial state radiation (ISR) or underlying event, it is relatively straightforward to extract a useful physical understanding from the obtained functional form of the observable. In figure 3, we plot the distributions of the exponents a, b, c, d and e with the requirement

[1] We also attempted to identify the set of exponent values that maximizes the AUC using stochastic gradient descent. However, due to the finite binning necessary to calculate the likelihood and therefore the AUC, we were unable to demonstrate satisfactory convergence to maxima. Additionally, the Monte Carlo approach enables a direct study of correlations of the exponents on the discrimination power. This will be demonstrated shortly.

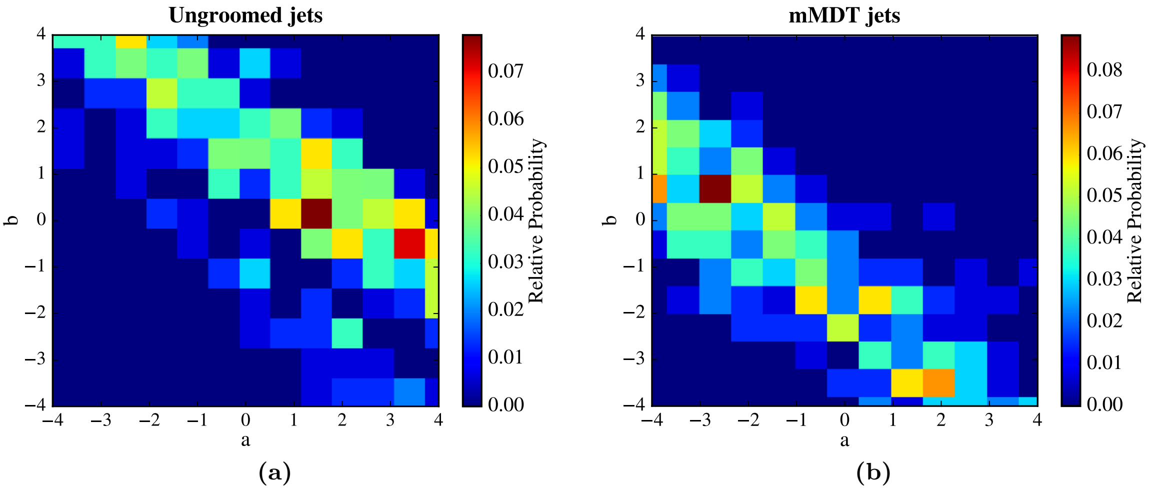


Figure 4. Heat maps of the correlation between a and b exponents of the product observable for ungroomed (a) and groomed (b) jets.

that the AUC for the corresponding product observable is greater than 0.81 or 0.73, for ungroomed and groomed jets, respectively. These distributions will enable us to extract the exponent values for which the AUC is maximized for the binned likelihood distributions of the product observable measured on signal and background.

By studying the histograms of the exponents one can make the following conclusions:

- For the ungroomed jets, AUC is maximized when $c = 0$, $d = 0.5$, $e = -1$ as the distributions for these exponents are very narrow. Since the distributions for a and b are both approximately uniform on $[-5, 5]$, further interpretation is required. This will be done shortly.

- For the groomed jets, AUC is maximized when $c = 0$, $d = -2$, $e = 2$. Again, since a and b are both approximately uniformly distributed over $[-5, 5]$, further analysis is required to determine the values that maximize the AUC.

To determine the values for a and b for both ungroomed and groomed jets, we work to understand the correlation between the exponents.

To determine the correlation between the exponents a and b, we plot their joint probability distribution from the uniform sampling on $[-5, 5]$ with the same cuts on the resulting observables' AUC. For both ungroomed and groomed jets, this is shown in figure 4. These plots demonstrate a strong correlation between these exponents, which to very good approximation is:

$$\text{Ungroomed}: \; a + b = 2\,, \qquad\qquad \text{Groomed}: \; a + b = -2\,. \tag{3.2}$$

These relationships can be used to fix b, for example, as a function of exponent a. To determine the value of the exponent a, we then fix b, c, d, and e as earlier, and calculate the AUC for $a \in [-5, 5]$. The results of this scan are shown in figure 5 for ungroomed

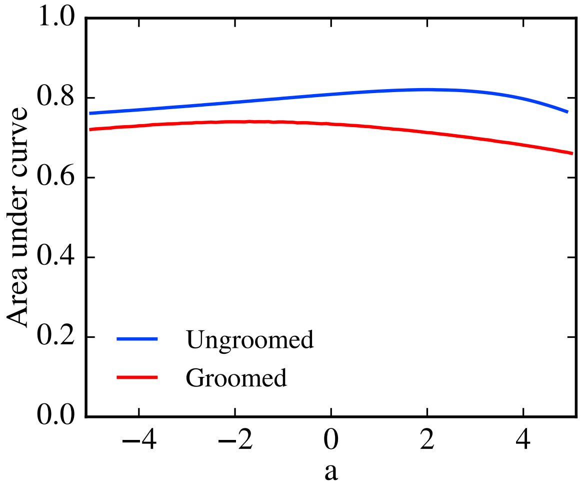


Figure 5. Variation of area under the ROC curve for the observable when the a exponent is scanned over the range $[-5, 5]$, keeping the c, d and e exponents fixed and varying b with a as per eq. (3.2).

and groomed jets. In particular, we note from the plot that AUC is maximized for the ungroomed case when $a = 2$ and for the groomed case when $a = -2$. This implies that the exponent $b = 0$ for both cases, using eq. (3.2).

Thus, the product observable takes on the following forms for the two kinds of jets:

$$\text{Ungroomed}: \ \beta_3 = \frac{\left(\tau_1^{(0.5)}\right)^2 \left(\tau_2^{(1)}\right)^{0.5}}{\tau_2^{(2)}}, \qquad \text{Groomed}: \ \beta_3^{(g)} = \left(\frac{\tau_2^{(2)}}{\tau_1^{(0.5)}\tau_2^{(1)}}\right)^2. \tag{3.3}$$

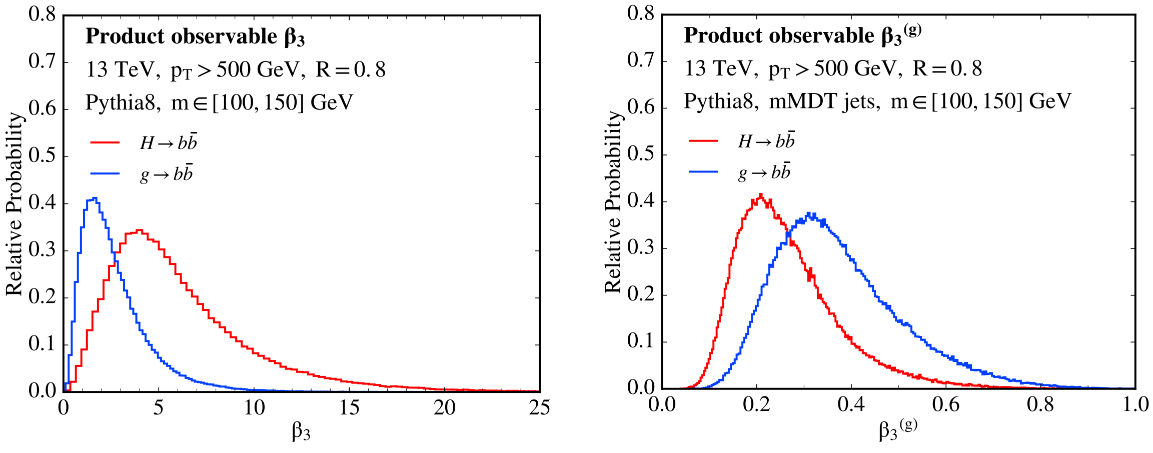
Any monotonic function of the observable will produce the same discrimination power, and so we can simplify the expression for the groomed product observable. For the product observable for groomed jets, we use the expression:

$$\beta_3^{(g)} = \frac{\tau_2^{(2)}}{\tau_1^{(0.5)}\tau_2^{(1)}}. \tag{3.4}$$

It is interesting to note that the observables from this method are Sudakov safe [44, 45] because they are formed from ratios of IRC safe observables.

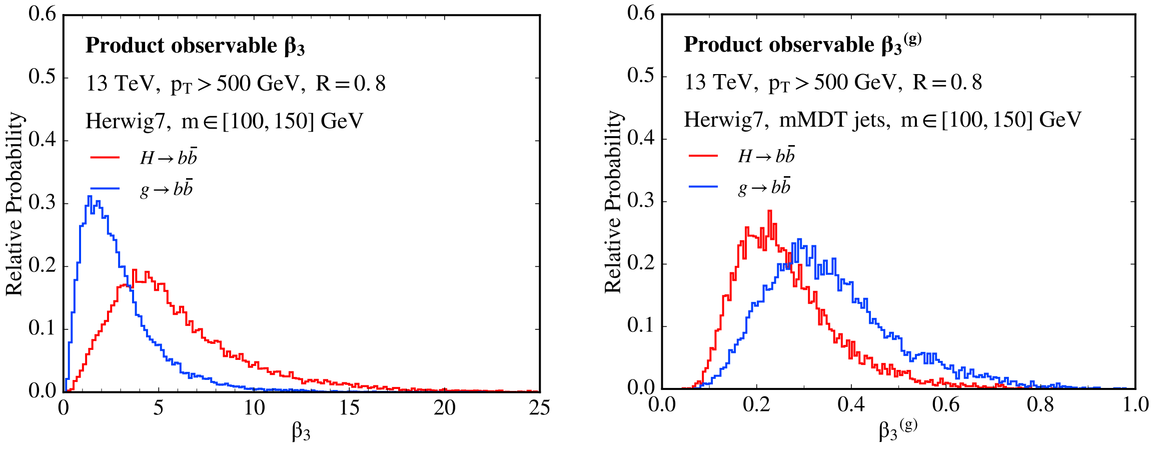
3.2 Physical interpretation

In figure 6, we plot the distribution of these new product observables measured on signal and background jets showered in Pythia. This shows that these product observables on ungroomed and groomed jets effectively separate signal from background. Additionally, in figure 7, we measure these product observables determined from the Pythia signal and background samples on the jets showered with Herwig. We observe a similar relative separation between the distributions, although the absolute scale is different, in the Herwig samples suggesting that these observables are sensitive to real physics, and not idiosyncrasies of the parton shower programs.



(a) **(b)**

Figure 6. Distributions of the product observable for signal (red) and background (blue), measured on the samples of ungroomed (a) and groomed (b) jets showered with Pythia, within a mass cut of $m_J \in [100, 150]$ GeV.



(a) **(b)**

Figure 7. Distributions of the product observable for signal (red) and background (blue), measured on the samples of ungroomed (a) and groomed (b) jets showered with Herwig.

Especially for groomed jets, these simple forms for the product observables enable a nice interpretation of the physics to which they are sensitive. In the case of ungroomed jets, there are multiple sources of radiation (final state, initial state, underlying event, etc.) that makes an interpretation a bit more challenging, so we won't discuss it more here. When the jets are groomed with mMDT, however, contamination radiation from the initial state or underlying event is dominantly removed, and so a picture of the jet exclusively with radiation from the final state is accurate. In this case, the mMDT jet with resolved 3-body phase space consists of the b and $\bar{b}$ pair, and the dominant gluon emitted off of them. The 3-body phase space configuration is shown in figure 8, with transverse momentum fractions

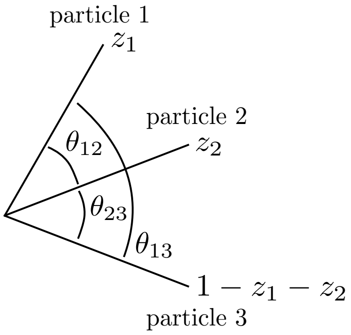


Figure 8. Illustration of the momentum fraction and pairwise angle variables that describe 3-body phase space.

z_i and pairwise angles θ_{ij}. In what follows, we will let particles 1 and 2 be the b and $\bar{b}$, and particle 3 be the gluon.

Because we make a cut on the jet mass and there is no soft singularity for $g \to b\bar{b}$ splitting, we assume that the emitted gluon is relatively soft and/or collinear with respect to the b and the $\bar{b}$. With this assumption, then the value of $\tau_1^{(0.5)}$ is completely determined by the b and $\bar{b}$. Then, the value of $\tau_1^{(0.5)}$ is approximately

$$\begin{aligned}\tau_1^{(0.5)} &\simeq z((1-z)\theta_{12})^{0.5} + (1-z)(z\theta_{12})^{0.5} \\ &= (z(1-z)\theta_{12}^2)^{0.25}\left(z^{0.75}(1-z)^{0.25} + z^{0.25}(1-z)^{0.75}\right) .\end{aligned} \tag{3.5}$$

Here, z is the transverse momentum fraction of the b quark subjet, for example. The combination $z(1-z)\theta_{12}^2$ is just the ratio of the jet mass to the jet transverse momentum to this order, m_J^2/p_{TJ}^2, and is approximately constant because the mass cut is relatively narrow. The term in parentheses on the right, $\left(z^{0.75}(1-z)^{0.25} + z^{0.25}(1-z)^{0.75}\right)$, is typically an order-1 number, as there is no soft singularity for $g \to b\bar{b}$ splitting nor for $H \to b\bar{b}$ decays. So, to good approximation, $\tau_1^{(0.5)}$ on these jets with a mass cut is just some constant value.

The remaining factor in $\beta_3^{(g)}$ however, contains significantly interesting physics. The two other N-subjettinesses that appear in that observable can be expressed as (see ref. [19] for more details):

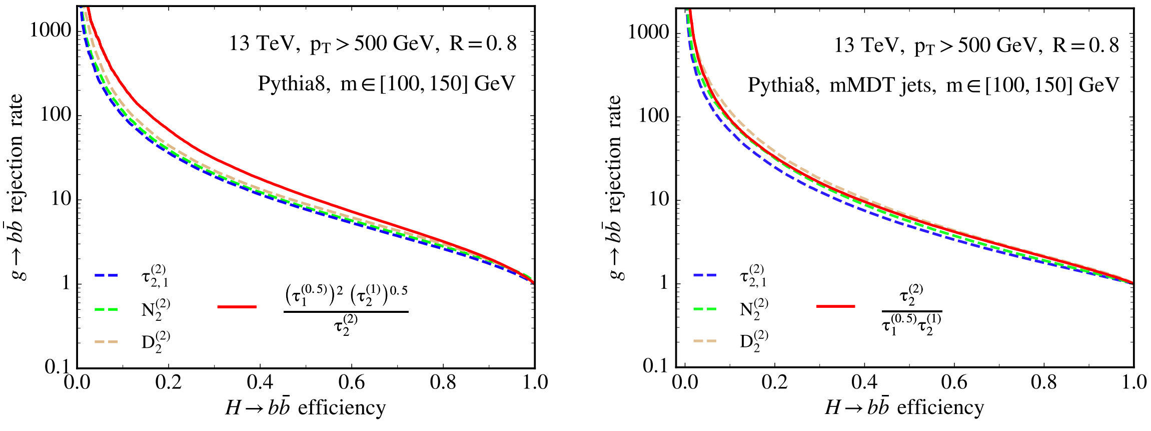
$$\tau_2^{(1)} = \frac{2z_3 z_i}{z_i + z_3}\theta_{i3} , \qquad \tau_2^{(2)} = \frac{z_3 z_i}{z_i + z_3}\theta_{i3}^2 . \tag{3.6}$$

In writing this, we assume that the gluon, with transverse momentum fraction $z_3 = 1 - z_1 - z_2$, is the first particle clustered by the k_T algorithm, and therefore sets the value of these τ_2 observables. z_i is the transverse momentum fraction of the closer in angle of particles 1 or 2 (the b or $\bar{b}$), with θ_{i3} this angle. The ratio that appears in $\beta_3^{(g)}$ is therefore

$$\frac{\tau_2^{(2)}}{\tau_2^{(1)}} = \frac{\min[\theta_{13}, \theta_{23}]}{2} . \tag{3.7}$$

Therefore, with these assumptions, the groomed jet product observable is approximately proportional to the angle between the dominant gluon emission and the closer of the b or $\bar{b}$:

$$\beta_3^{(g)} \propto \min[\theta_{13}, \theta_{23}] . \tag{3.8}$$



(a) **(b)**

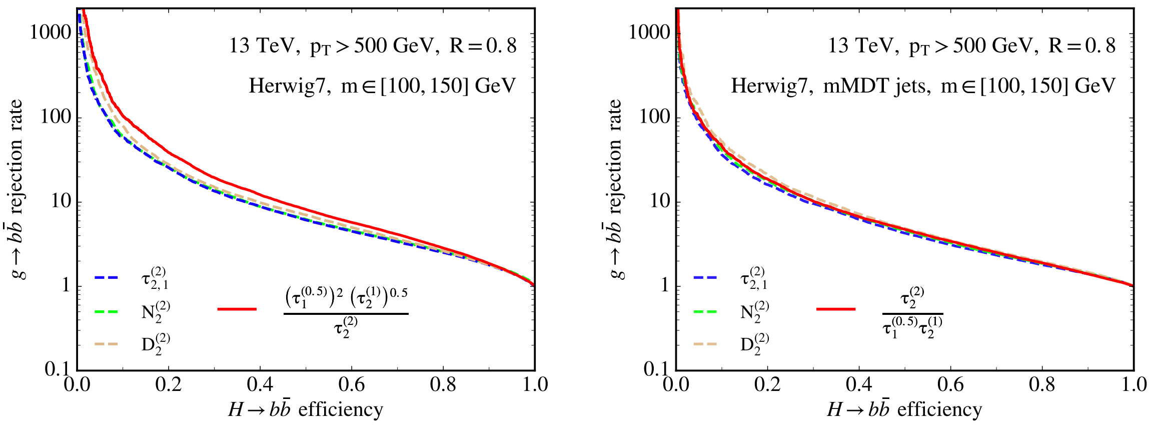
Figure 9. Signal efficiency versus background rejection rate for N-subjettiness ratio $\tau_{2,1}^{(2)}$, $N_2^{(2)}$, and $D_2^{(2)}$, measured on ungroomed (a) and groomed (b) jets showered using Pythia, compared to the discrimination power of the product observable β_3 or $\beta_3^{(g)}$. The discrimination power of the product observable is comparable to that of the standard observables.

As color octets, gluons preferably emit at wide angles, while singlet Higgs bosons emit at small angles, and so we do expect this observable to provide discrimination power.

3.3 Comparison to standard observables

To demonstrate that these observables learned by the machine are indeed powerful, we compare their discrimination power to that of a collection of standard observables. For comparison, we use N-subjettiness ratio $\tau_{2,1}^{(2)}$ with winner-take-all axes [46–48] and (generalized) energy correlation function ratios $D_2^{(2)}$ [49] and $N_2^{(2)}$ [50]. While these and related observables have been used for identification of boosted $H \to b\bar{b}$ decays, they are not necessarily optimized for this purpose. Nevertheless, they provide a useful benchmark. The signal efficiency versus background rejection rates for jets showered in Pythia are shown in figure 9 and for jets from Herwig, in figure 10. Most interestingly, for ungroomed jets, the new product observable β_3 outperforms each of these standard observables. On groomed jets, the discrimination power of all of the observables is much closer and $D_2^{(2)}$ apparently slightly outperforms $\beta_3^{(g)}$. Nevertheless, this demonstrates that, with very little human input, powerful discrimination observables can be constructed from what the machine learns.

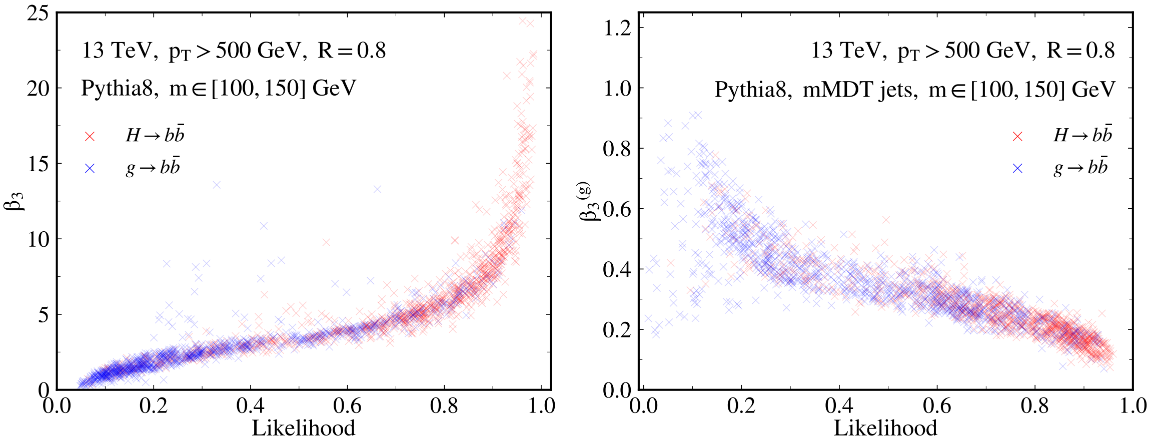
It is instructive to also directly compare the value of our β_3 observables directly to the output likelihood ratio as determined on the 3-body phase space observables from the neural network. In figure 11, we have made scatter plots of the value of β_3 or $\beta_3^{(g)}$ (as appropriate) versus the likelihood ratio as measured on 1000 of both signal and background jets. If the β_3 observables perfectly captured the information in the likelihood, these scatter plots should reduce to a monotonic curve. The deviation from monotonicity is a measure of the information that β_3 misses with respect to the likelihood. Broadly, these plots demonstrate a monotonic relationship between β_3 and the likelihood, but there is some spread. The relative size of the spread is less for β_3 measured on ungroomed jets, which may reflect that in this case, β_3 captures more of the information in the likelihood than $\beta_3^{(g)}$.



(a)

(b)

Figure 10. Signal efficiency versus background rejection rate for N-subjettiness ratio $\tau_{2,1}^{(2)}$, $N_2^{(2)}$, and $D_2^{(2)}$, measured on ungroomed (a) and groomed (b) jets showered using Herwig, compared to the discrimination power of the product observable β_3 or $\beta_3^{(g)}$.



(a)

(b)

Figure 11. Scatter plots of 1000 signal (red) and background (blue) jets in the plane of β_3 versus the 3-body phase space likelihood ratio from the neural network. Ungroomed jets are shown in (a), and mMDT groomed jets in (b). The gross monotonic relationship between β_3 and the likelihood indicates that most of the information in the likelihood is captured in β_3, while the spread indicates that there is some discrimination information that is missed in β_3.

4 Conclusions

Previous deep learning studies in jet physics have shown immense promise. While it has been shown that appropriately designed deep learning techniques can outperform standard observables, such studies have not effectively probed what more information the machines are identifying. Building on ref. [19], we propose a procedure that develops powerful new observables from the knowledge of the information contained in jets that contributes to an observable's discrimination power. By systematically controlling the information fed to neural networks, it is possible to identify the minimal amount of information required to effectively discriminate between highly boosted decays of different massive particles and light QCD partons. The method of planing introduced in ref. [17] may also be useful in identifying the minimal information necessary for powerful discrimination.

Here, we have proposed an algorithm that can, in principle, be automated to construct new observables for any discrimination problem. While the $H \to b\bar{b}$ application shows promise, such a procedure might also be applied to other specific problems like identifying q vs. g jets, top quarks, or even to develop new observables that work effectively as generic anti-QCD taggers. The most improvement to this method would be accomplished by construction of an optimal basis of functions with parameters that can be tuned to maximize discrimination power. Of course, at this point, one is faced with a trade-off between simplicity and efficacy that must be taken into account. By utilizing constructive deep learning techniques that are sensitive to exotic configurations within jets, this approach is presented with an intention to open the door to a whole new class of powerful substructure observables that can be tailored to specific or generic classification problems, while also providing further physics insight regarding the jets being studied.

Acknowledgments

We thank Gregor Kasieczka, Ian Moult, and Ben Nachman for comments on the manuscript. The work of K.D. was supported by the Reed College Physics Department's Delord-Mockett Fund and Reed College Summer Scholarship Funds.

References

[1] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, *JHEP* **02** (2015) 118 [arXiv:1407.5675] [INSPIRE].

[2] L.G. Almeida, M. Backović, M. Cliche, S.J. Lee and M. Perelstein, *Playing Tag with ANN: Boosted Top Identification with Pattern Recognition*, *JHEP* **07** (2015) 086 [arXiv:1501.05968] [INSPIRE].

[3] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, *Jet-images — deep learning edition*, *JHEP* **07** (2016) 069 [arXiv:1511.05190] [INSPIRE].

[4] P. Baldi, K. Bauer, C. Eng, P. Sadowski and D. Whiteson, *Jet Substructure Classification in High-Energy Physics with Deep Neural Networks*, *Phys. Rev.* **D 93** (2016) 094034 [arXiv:1603.09349] [INSPIRE].

[5] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban and D. Whiteson, *Jet Flavor Classification in High-Energy Physics with Deep Neural Networks*, *Phys. Rev.* **D 94** (2016) 112002 [arXiv:1607.08633] [INSPIRE].

[6] J.S. Conway, R. Bhaskar, R.D. Erbacher and J. Pilot, *Identification of High-Momentum Top Quarks, Higgs Bosons and W and Z Bosons Using Boosted Event Shapes*, Phys. Rev. **D 94** (2016) 094027 [arXiv:1606.06859] [INSPIRE].

[7] J. Barnard, E.N. Dawe, M.J. Dolan and N. Rajcic, *Parton Shower Uncertainties in Jet Substructure Analyses with Deep Neural Networks*, Phys. Rev. **D 95** (2017) 014018 [arXiv:1609.00607] [INSPIRE].

[8] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, JHEP **01** (2017) 110 [arXiv:1612.01551] [INSPIRE].

[9] L. de Oliveira, M. Paganini and B. Nachman, *Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis*, Comput. Softw. Big Sci. **1** (2017) 4 [arXiv:1701.05927] [INSPIRE].

[10] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning Top Taggers or The End of QCD?*, JHEP **05** (2017) 006 [arXiv:1701.08784] [INSPIRE].

[11] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics*, arXiv:1702.00748 [INSPIRE].

[12] L.M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly Supervised Classification in High Energy Physics*, JHEP **05** (2017) 145 [arXiv:1702.00414] [INSPIRE].

[13] J. Pearkes, W. Fedorko, A. Lister and C. Gay, *Jet Constituents for Deep Neural Network Based Top Quark Tagging*, arXiv:1704.02124 [INSPIRE].

[14] T. Cohen, M. Freytsis and B. Ostdiek, *(Machine) Learning to Do More with Less*, JHEP **02** (2018) 034 [arXiv:1706.09451] [INSPIRE].

[15] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned Top Tagging with a Lorentz Layer*, arXiv:1707.08966 [INSPIRE].

[16] E.M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, JHEP **10** (2017) 174 [arXiv:1708.02949] [INSPIRE].

[17] S. Chang, T. Cohen and B. Ostdiek, *What is the Machine Learning?*, arXiv:1709.10106 [INSPIRE].

[18] A.J. Larkoski, I. Moult and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning*, arXiv:1709.04464 [INSPIRE].

[19] K. Datta and A. Larkoski, *How Much Information is in a Jet?*, JHEP **06** (2017) 073 [arXiv:1704.08249] [INSPIRE].

[20] J.A. Aguilar-Saavedra, J.H. Collins and R.K. Mishra, *A generic anti-QCD jet tagger*, JHEP **11** (2017) 163 [arXiv:1709.01087] [INSPIRE].

[21] CMS collaboration, *Inclusive search for the standard model Higgs boson produced in pp collisions at $\sqrt{s} = 13\,TeV$ using $H \to b\bar{b}$ decays*, CMS-PAS-HIG-17-010.

[22] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, JHEP **03** (2011) 015 [arXiv:1011.2268] [INSPIRE].

[23] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, JHEP **02** (2012) 093 [arXiv:1108.2701] [INSPIRE].

[24] M. Dasgupta, A. Fregoso, S. Marzani and G.P. Salam, *Towards an understanding of jet substructure*, JHEP **09** (2013) 029 [arXiv:1307.0007] [INSPIRE].

[25] M. Dasgupta, A. Fregoso, S. Marzani and A. Powling, *Jet substructure with analytical methods*, *Eur. Phys. J.* **C 73** (2013) 2623 [arXiv:1307.0013] [INSPIRE].

[26] I.W. Stewart, F.J. Tackmann and W.J. Waalewijn, *N-Jettiness: An Inclusive Event Shape to Veto Jets*, *Phys. Rev. Lett.* **105** (2010) 092002 [arXiv:1004.2489] [INSPIRE].

[27] A.J. Larkoski, G.P. Salam and J. Thaler, *Energy Correlation Functions for Jet Substructure*, *JHEP* **06** (2013) 108 [arXiv:1305.0007] [INSPIRE].

[28] S. Catani, Y.L. Dokshitzer, M.H. Seymour and B.R. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, *Nucl. Phys.* **B 406** (1993) 187 [INSPIRE].

[29] S.D. Ellis and D.E. Soper, *Successive combination jet algorithm for hadron collisions*, *Phys. Rev.* **D 48** (1993) 3160 [hep-ph/9305266] [INSPIRE].

[30] G.C. Blazey et al., *Run II jet physics*, in *QCD and weak boson physics in Run II. Proceedings*, Batavia, U.S.A., March 4–6, June 3–4, November 4–6, 1999, pp. 47–77 (2000), hep-ex/0005012, [INSPIRE].

[31] Y.L. Dokshitzer, G.D. Leder, S. Moretti and B.R. Webber, *Better jet clustering algorithms*, *JHEP* **08** (1997) 001 [hep-ph/9707323] [INSPIRE].

[32] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross-sections in deep inelastic scattering*, in *Monte Carlo generators for HERA physics. Proceedings, Workshop*, Hamburg, Germany, 1998–1999, pp. 270–279 (1998), hep-ph/9907280 [INSPIRE].

[33] H1 collaboration, C. Adloff et al., *Measurement and QCD analysis of jet cross-sections in deep inelastic positron-proton collisions at $\sqrt{s}$ of 300 GeV*, *Eur. Phys. J.* **C 19** (2001) 289 [hep-ex/0010054] [INSPIRE].

[34] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[35] T. Sjöstrand, S. Mrenna and P.Z. Skands, *PYTHIA 6.4 Physics and Manual*, *JHEP* **05** (2006) 026 [hep-ph/0603175] [INSPIRE].

[36] T. Sjöstrand et al., *An Introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[37] M. Bahr et al., *HERWIG++ Physics and Manual*, *Eur. Phys. J.* **C 58** (2008) 639 [arXiv:0803.0883] [INSPIRE].

[38] J. Bellm et al., *HERWIG 7.0/HERWIG++ 3.0 release note*, *Eur. Phys. J.* **C 76** (2016) 196 [arXiv:1512.01178] [INSPIRE].

[39] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J.* **C 72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[40] M. Cacciari and G.P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, *Phys. Lett.* **B 641** (2006) 57 [hep-ph/0512210] [INSPIRE].

[41] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[42] F. Chollet, *Keras*, https://github.com/fchollet/keras, (2015).

[43] J. Neyman and E.S. Pearson, *On the problem of the most efficient tests of statistical hypotheses*, *Phil. Trans. Roy. Soc. Lond.* **A 231** (1933) 289.

[44] A.J. Larkoski and J. Thaler, *Unsafe but Calculable: Ratios of Angularities in Perturbative QCD*, *JHEP* **09** (2013) 137 [arXiv:1307.1699] [INSPIRE].

[45] A.J. Larkoski, S. Marzani and J. Thaler, *Sudakov Safety in Perturbative QCD*, *Phys. Rev.* **D 91** (2015) 111501 [arXiv:1502.01719] [INSPIRE].

[46] D. Bertolini, T. Chan and J. Thaler, *Jet Observables Without Jet Algorithms*, *JHEP* **04** (2014) 013 [arXiv:1310.7584] [INSPIRE].

[47] A.J. Larkoski, D. Neill and J. Thaler, *Jet Shapes with the Broadening Axis*, *JHEP* **04** (2014) 017 [arXiv:1401.2158] [INSPIRE].

[48] A.J. Larkoski and J. Thaler, *Aspects of jets at 100 TeV*, *Phys. Rev.* **D 90** (2014) 034010 [arXiv:1406.7011] [INSPIRE].

[49] A.J. Larkoski, I. Moult and D. Neill, *Power Counting to Better Jet Observables*, *JHEP* **12** (2014) 009 [arXiv:1409.6298] [INSPIRE].

[50] I. Moult, L. Necib and J. Thaler, *New Angles on Energy Correlation Functions*, *JHEP* **12** (2016) 153 [arXiv:1609.07483] [INSPIRE].