Reports of My Demise Are Greatly Exaggerated: \( N \)-subjettiness Taggers Take On Jet Images

Liam Moore\(^1\), Karl Nordström\(^2,3,\)\(^b\), Sreedevi Varma\(^4,\)\(^c\), Malcolm Fairbairn\(^4,\)\(^d\)

\(^1\) Center for Cosmology, Particle Physics and Phenomenology (CP3), Université Catholique de Louvain, Chemin du Cyclotron 2, B-1348 Louvain-la-Neuve, Belgium
\(^2\) National Institute for Subatomic Physics (NIKHEF) Science Park 105, 1098 XG Amsterdam, Netherlands
\(^3\) Laboratoire de Physique Théorique et Hautes Énergies (LPTHE), UMR 7589 CNRS & Sorbonne Université, 4 Place Jussieu, F-75252, Paris, France
\(^4\) Theoretical Particle Physics and Cosmology, Physics, King’s College London, London WC2R 2LS, United Kingdom

\(^a\) liam.moore@uclouvain.be  \(^b\) karl.nordstrom@nikhef.nl  \(^c\) sreedevi.varma@kcl.ac.uk  \(^d\) malcolm.fairbairn@kcl.ac.uk

July 16, 2018

Abstract

We compare the performance of a convolutional neural network (CNN) trained on jet images with dense neural networks (DNNs) trained on \( n \)-subjettiness variables to study the distinguishing power of these two separate techniques applied to boosted hadronic \( Z \) boson and top quark decays. We find that they perform almost identically once jet mass information is included in a consistent manner, which suggests they are accessing the same underlying information which can be intuitively understood as being contained in 4-, 5-, and 6-body kinematic phase spaces depending on the sample. This suggests both of these methods are highly useful for heavy object tagging and provides a tentative answer to the question of what the image network is actually learning.

Contents

1 Introduction 2

2 Sample Generation and Machine Learning Techniques 3
  2.1 Signal and background sample generation 4
  2.2 Jet image generation 1
  2.3 Analysing jet images with Convolutional Neural Networks 5
  2.4 Using \( n \)-subjettiness variables to learn the \( m \)-body kinematic phase space using Dense Neural Networks 0

3 \( Z \) Boson Tagging Results 7
1 Introduction

New particles created above the electroweak scale are often expected to decay into the most massive members of the Standard Model family, namely the $W$ and $Z$ bosons, the (Brout-Englert-)Higgs boson $h$ and the top quark. Distinguishing between different heavy particles is challenging if they subsequently decay into lighter coloured particles since these form jets as the initial partons radiate a ‘shower’ of hadrons. At a hadron collider such as the LHC, the multiplicity of jets which can be produced in a single collision by simple QCD processes is very large, and pile-up and minimum bias hadronic activity further complicate the final state. It can therefore be extremely difficult to map hadronic activity in the final state to the hadronic decay of a heavy particle. This makes it difficult to separate signal from backgrounds in many searches for new physics, and is the dominant issue for measurements of some important Standard Model predictions such as the $h \to b \bar{b}$ branching ratio \cite{1}.

The situation can be improved if one considers the substructure of jets which contain within them all of the decay products of a boosted heavy object. In this case the inherent scales inside the jets allow them to be separated from largely scaleless light quark and gluon jets, alleviating the combinatorial backgrounds which make an analysis using fully resolved decay products challenging. Following the early work of Seymour \cite{2}, the most successful applications have been at the LHC where for example the BDRS method \cite{3}, John Hopkins tagger \cite{4}, shower deconstruction \cite{5,6}, and HEPTopTagger \cite{7,8,9} all have seen use by the experiments as well as considerable phenomenological attention.

As the LHC approaches its high luminosity runs, it is also necessary to consider ways to look for new physics in as many directions as possible to make sure we make the most of the available data. It is therefore additionally clear that efficient algorithms for identifying heavy particle decays which can handle large data volumes and don’t require hand-tuning for specific decays will become increasingly important in the future, which suggests machine learning techniques as an obvious avenue to consider.

One of the most widely studied problems in the larger machine learning community is the task of extracting classification information from large image datasets. It is therefore natural to consider the possibility of converting information from LHC events into images which we can train a machine learning algorithm on in order to teach it the mapping from hadronic activity to various classifications such as heavy particle decay and ‘QCD background’ or light quark and gluon jets \cite{10,11,12}. These kind of techniques have seen attention for example in quark/gluon classification \cite{13}, top-tagging \cite{14} and $W$ tagging \cite{10}.

There have also been initial studies of the sensitivity of these algorithms to the modeling uncertainties involved in the use of Monte Carlo event generators which (just to mention one issue among many) rely entirely on phenomenological models to hadronise the final state after
the (well-understood) parton shower has increased the coloured particle multiplicity considerably \cite{15}. Such questions of the extent to which we might be teaching the machine spurious modeling details rather real physics become more pertinent as more advanced algorithms such as jet images squeeze more information out of the radiation patterns. It is therefore interesting to compare the performance of jet image networks to ones with a more intuitive physical interpretation.\footnote{The usual reply is of course that these algorithms should ultimately be trained on purified data samples and not simulation, but a good understanding of the physics they are learning is still necessary and identifying issues with the modeling could also potentially help the Monte Carlo generator development community.}

\(N\)-subjettiness variables were first discussed in \cite{16} as a jet substructure application of the ideas introduced by \(n\)-jettiness \cite{17}. These quantify how much the radiation contained within a jet (event) is aligned along different (sub)jet axes within it. By analysing the bunching and spatial distribution of jets and subjets in an event using this procedure, one can powerfully distinguish between different decay and event topologies. The \(n\)-subjettiness variables were first applied to distinguishing \(W\) bosons and top quarks from QCD backgrounds in \cite{16,18}, and have subsequently seen wide use in the phenomenological literature and by the ATLAS and CMS experiments (see, e.g., \cite{19,21}). Later on similar ideas led to the development of Energy Correlation Functions \cite{22,23} and most recently Energy Flow Polynomials (EFPs) \cite{24}. The \(n\)-subjettiness variables can also be used as inputs for a machine learning algorithm and this approach can perform very well at heavy object discrimination \cite{24,26}. Such an approach is additionally attractive since the small number of input variables here allows for a mapping to the \(m\)-body kinematic phase space of the jet constituents, which provides an intuitive understanding of what the machine is actually learning.

In this work, we set out to compare the performance of the \(n\)-subjettiness approach to the jet image approach for machine learned top quark tagging at three different jet \(p_T\) ranges, roughly corresponding to the mostly resolved, mixed, and highly boosted regions. Since such comparisons have only (to our knowledge) been presented in \cite{24} previously, we begin by also presenting a study of \(Z\) tagging which follows \cite{25} but clarifies the role the jet mass information has in the ultimate performance of the algorithms. Since many preprocessing steps (including ones we use here) used in the jet image literature smear this information away from the images on which the algorithm is actually trained, it is useful to confirm that adding this information back in works exactly as you would expect by comparing to our \(n\)-subjettiness results.

This paper is structured as follows: in Section 2 we describe the details of the event generation and machine learning implementations of our heavy object taggers. In Section 3 we present the results of our \(Z\) boson tagging study, and in Section 4 we present the results of our top quark tagging study. We then discuss the results and conclude in Section 5.

2 Sample Generation and Machine Learning Techniques

2.1 Signal and background sample generation

In each case, we generate matrix elements at leading order (LO) in QCD and electroweak couplings with \textsc{MadGraph5 \_aMC@NLO} v2.60 \cite{27} with the NNPDF 2.3 PDF set \cite{28} interfaced via \textsc{LHAPDF} \cite{29}. Idealised signal and background samples are chosen to ensure each has very little hadronic activity outside of the signal and background jets.

For our \(Z\) boson tagging study we generate \(pp \rightarrow Z_1Z_2, Z_1 \rightarrow jj, Z_2 \rightarrow \nu\bar{\nu}\) as our signal
sample and $pp \to Zj$, $Z \to \nu\bar{\nu}$ as our background sample. We require $p_T(Z/j) \geq 500$ GeV via generator-level cuts on the appropriate partons.

In our top quark tagging study we use $pp \to W^- t$, $t \to W^+_2 b$, $W^+_2 \to jj$, $W^-_1 \to e^- \bar{\nu}_e$ as our signal process, and $pp \to W^- j$, $W \to e^- \bar{\nu}_e$ as our background. We examine three different $p_T$ ranges to capture differences in signal and background kinematics as the three-pronged top quark decay becomes more collinear: $[350, 400]$ GeV, $[500, 550]$ GeV and $[1300, 1400]$ GeV, with these cuts enforced similarly at the parton-level.

To facilitate reproduction of our results we provide the gluon to light quark ratios of our background samples in Table 1.

The Les Houches events [30] are matched to the parton shower Pythia 8.2.26 [31,32] to obtain hadron-level events in the HEPMC format [33]. The input datasets for each NN are then extracted from these by a tailored Rivet [34] analysis.

We cluster jets therein with the anti-$k_T$ algorithm throughout [35] and using the FASTJET library [36]. In our $Z$-boson tagging application, we define jets originating from both signal and background processes with a radius parameter of $R = 0.8$ and require $|\eta| \leq 5.0$.

In our top-tagging study we cluster jets into two groups, requiring $R = 1.5$ "fat" jets within $|\eta| < 1$ in the mild and moderately boosted $p_T$ ranges $[350, 400]$ and $[500, 550]$ GeV, and standard $R = 0.8$ jets within $|\eta| < 2.5$ for the ultra-boosted top jets with $p_T$ in the range $[1300, 1400]$ GeV.

From each sample we extract from the leading $p_T$ jet both a jet image and a set of $N$-subjettiness variables as defined in Sec 2.2 and Sec 2.4 respectively, which serve as the raw pseudodata to be fed to the classifiers after further preprocessing.

These datasets consist of approximately 2M jets each divided evenly between signal and background, of which 100k are used for validation and 200k for testing.

## 2.2 Jet image generation

In order to facilitate their use with the CNN detailed below the jets are first converted into images. Each jet is turned into a $33 \times 33$ pixel jet image where energy deposits in a 'calorimeter' (here approximated by summing up the $p_T$ inside each pixel) are treated as the pixel intensity. The pixels span the $\eta - \phi$ space covering the entire region of the jet with a particular radius.

The images span the $\eta - \phi$ space covering the entire region of the jet with a particular radius. The images generated undergo preprocessing steps before being fed to the CNN, following closely the procedure set out in [13]. The applied preprocessing steps are briefly described here. The pixel with maximum $p_T$ is brought to the center of the jet image. This is equivalent to subtracting the $p_T$ weighted mean of $\eta$ and $\phi$. The image is then cropped to $33 \times 33$ pixels in the range $\eta, \phi \in [-R, R]$. Each jet image is then normalized so that the summed intensity of all pixels is 1. In the final step the entire dataset is standardised using the pixel-by-pixel mean $\mu_{i,j}$ and standard deviation $\sigma_{i,j}$ of the training dataset, $I_{i,j} \rightarrow (I_{i,j} - \mu_{i,j})/(\sigma_{i,j} + \nu)$, where $i,j$
are pixel indices, and \( r = 10^{-5} \) is used to suppress noise [13]. We use only grayscale images without any additional information in the colour channels. The average images of the Z boson signal and background samples after normalisation and the full preprocessing are shown in [1].

We stress that while these preprocessing steps in general help the training process converge faster, it will also smear out the jet mass entirely from the images fed to the network, which necessitates that we add this information.

### 2.3 Analysing jet images with Convolutional Neural Networks

Convolutional neural networks have been used in the phenomenological literature for quark/gluon classification [13], by ATLAS [37] and CMS [38], and for top quark [11] and W-boson tagging [10,12,39]. The architecture of the CNN used here is identical to Refs. [13,24]. The network contains three convolutional layers and two fully connected layers, where He-uniform weights are used for the initialisation. The full architecture is presented in Figure 2. ReLu is used as the activation function except at the Softmax activated output [40,41]. A batch size of 100 is used with a learning rate (\( \alpha \)) of 0.001. A pixel dropout of 0.1 is applied to each CNN layer.

A simple perceptron neural network is used to combine jet mass information with the CNN output. This additional network has two input nodes, one given by the jet mass and the other...
Figure 2: Network architecture of the CNN. This figure was generated by adapting the code from [https://github.com/gwding/draw_convnet](https://github.com/gwding/draw_convnet).

by the corresponding CNN output. This network has a single hidden layer with 300 nodes. Random normal weights are used for initialisation. ReLu is used as the activation function. The learning rate and the dropout rate are same as for the CNN.

The CNN is trained for 50 epochs and the additional network is trained for 500 epochs. As we will show and discuss, this improves the performance of the CNN approach for all of our samples, as our preprocessing steps remove the jet mass information from the CNN input.

TensorFlow 1.3.0 [42] is used to build the networks. An Nvidia GeForce GTX 1080Ti with CUDA 9.0 [43] is used for training.

2.4 Using \( n \)-subjettiness variables to learn the \( m \)-body kinematic phase space using Dense Neural Networks

Our alternative heavy object tagger is based on the proposal in [25] to use a basis of \( n \)-subjettiness variables [16] spanning an \( m \)-body phase space to teach a Dense Neural Network to separate signal and background using a relatively small set of input variables. The same idea was later used in [44] to build a tagger intended to generically distinguish QCD jets from heavy object jets, and to construct new simple observables with improved discrimination power in [26]. It was also used in a similar manner as here as a benchmark for comparing different ML implementations of heavy object tagging in [24]; the difference to the present study is that we here focus specifically on top tagging and investigate the effect of adding jet mass information in two different implementations. The input variables are given by:

\[
\left\{ r_1^{(0.5)}, r_1^{(1)}, r_1^{(2)}, r_2^{(0.5)}, r_2^{(1)}, r_2^{(2)}, \ldots, r_{M-2}^{(0.5)}, r_{M-2}^{(1)}, r_{M-2}^{(2)}, r_{M-1}^{(1)}, r_{M-1}^{(2)} \right\}
\]

(1)

where

\[
r_N^{(\beta)} = \frac{1}{p_{T,J}} \sum_{i \in J} p_{T,i} \min \left\{ R_{1i}^{\beta}, R_{2i}^{\beta}, \ldots, R_{Ni}^{\beta} \right\}
\]

(2)

and \( R_{ni} \) is the distance in the \( \eta - \phi \) plane of the jet constituent \( i \) to the axis \( n \). We choose the \( N \) axes using the \( k_T \) algorithm [45] with \( E \)-scheme recombination [46].

\[^2\text{This choice is identical to that in [25] and we have not attempted to optimise it any further, since (as we...}]}$
There are several benefits to this choice of input variables: for one, it allows an intuitive understanding of what the network actually learns since enlarging the input data set corresponds to allowing the network to access higher body kinematic phase space information, and the point where the network saturates can therefore be given a physical interpretation.

Additionally, \(n\)-subjettiness variables are very well-understood and are some of the most studied and used jet substructure observables with for example ATLAS detector-level and unfolded distributions for \(\tau_2^{(1)}/\tau_1^{(1)}\) and \(\tau_3^{(1)}/\tau_2^{(1)}\) in a dijet sample having been published in [19].

We train simple DNNs using these as input nodes. They consists of four fully connected hidden layers, the first two with 300 nodes and a dropout regularisation [47] of 0.2, and the last two with 100 nodes and a dropout regularisation of 0.1. The output layer are two nodes. We use the ReLu activation function [40] throughout and minimise the cross-entropy using Adam optimisation [48]. To add jet mass information we simply include it among the input parameters in Equation 1.

We implement this DNN using CUDA 9.0 [43] and TensorFlow 1.8.0 [42]. An Nvidia GeForce GTX 1080 is used for training. The number of epochs trained depends on the input data used and ranges from 500-1500 to ensure convergence.

3 \(Z\) Boson Tagging Results

The receiver operating characteristic (ROC) curves for the image network and the \(n\)-subjettiness networks with up to 5-body phase space information (to show that the information saturates at 4-body phase space) are presented in Figure 3, with results without mass information on the left and with mass information on the right. The area-under-curve values for the ROC curves for a selection of network configurations are presented in Table 2. In general the performance of the two methods is very comparable both with and without jet mass information included, which suggests the image network ultimately is probing very similar information as the \(n\)-subjettiness one. This is a non-trivial test of whether or not our image network (as we have implemented it here) accesses information which can not be considered safe from a modeling perspective: since it saturates at 4-body kinematic phase space information, we can be fairly certain it is learning features of the hard splittings and first few splittings in the parton shower rather than low-energy features which should not be trusted.

The similar performance of the image network and saturated \(n\)-subjettiness network without mass information and the large and comparable improvements that can be seen by including the mass information (which is itself also included in the plot with mass information) can be understood by considering the information included in the network inputs without mass: in the CNN case, the pre-processing steps we take smear out the mass information considerably [12,14], and in the \(n\)-subjettiness one we effectively remove all jet scales from the input information by using the normalised definition of \(\tau_N^{(j)}\) in Equation 2 smearing any jet mass information the network can deduce from the \(n\)-subjettiness variables themselves.

will show) we get very competitive performance ‘out of the box’.
Figure 3: ROC curves for Z boson tagging without mass on the left and with mass on the right for $p_T > 500 \text{ GeV}$. Without mass information the image network underperforms the saturated $n$-subjettiness network, but once mass information is added it beats the $n$-subjettiness network at very high signal efficiency and is competitive at lower signal efficiencies.

Table 2: The area-under-curve (AUC) values for a selection of our ROC curves. Larger values are better and AUC=1 corresponds to perfect signal and background discrimination.

| Sample                      | mass + jet image | mass + 3-body | mass + 5-body |
|-----------------------------|------------------|---------------|---------------|
| $Z$ $p_T \geq 500 \text{ GeV}$ | 0.9407           | 0.9067        | 0.9207        |
| Top $p_T \in [350 - 400] \text{ GeV}$ | 0.9654           | 0.9490        | 0.9613        |
| Top $p_T \in [500 - 550] \text{ GeV}$ | 0.9660           | 0.9506        | 0.9648        |
| Top $p_T \in [1300 - 1400] \text{ GeV}$ | 0.9775           | 0.9580        | 0.9711        |

4 Top Quark Tagging Results

The ROC curves for the top quark results are presented in Figures 4, 5, and 6 with networks not including mass information explicitly on the left and the ones that do on the right. The area-under-curve values for the ROC curves for a selection of network configurations are presented in Table 2. We again find comparable performance for the two methods, suggesting our results from the Z case above generalise to more complicated decays.

Compared to the Z results, the improvement from adding mass information is relatively speaking larger for the image network than the $n$-subjettiness one. This can be understood as we limit the jet $p_T$ ranges of the samples tightly to isolate specific kinematic regions, which has the unintended consequence of allowing the normalised $n$-subjettiness variables to retain more information about the mass scale of the jet before adding it explicitly.

This suggests only the results with mass information included should be considered when answering the question of what the image network is learning with reference to the $n$-subjettiness results. Again the image network therefore shows excellent agreement with the saturated $n$-subjettiness network, which this time happens at 5-body kinematic phase space for $p_T \in [350, 400] \text{ GeV}$ and $p_T \in [500, 550] \text{ GeV}$ and 6-body for $p_T \in [1300, 1400] \text{ GeV}$. The $p_T$ dependence of the point where the networks saturate should not be oversold, however: the differences are small and could potentially be removed by improving the network architecture and hyperparameter optimisation.
Figure 4: ROC curves for top quark tagging without mass on the left and with mass on the right, for $p_T \in [350, 400]$ GeV. Without mass information the $n$-subjettiness network beats the image network, but once mass information is added the differences become very small.

Figure 5: ROC curves for top quark tagging without mass on the left and with mass on the right, for $p_T \in [500, 550]$ GeV. In this case the performance after adding mass information is very similar.

Figure 6: ROC curves for top quark tagging without mass on the left and with mass on the right, for $p_T \in [1300, 1400]$ GeV. The image network once again underperforms before adding mass information, but is slightly better at high signal efficiency after mass information is added.
5 Discussion and Conclusion

From our results it is clear that the performance of a CNN image network at heavy object tagging can be closely matched by a simpler DNN using $n$-subjettiness variables for both $Z$ bosons and top quarks at a number of $p_T$ ranges, once mass information is added to both networks in a consistent manner. This clarifies the question of what the image network is actually learning, as the saturated $n$-subjettiness network gives an intuitive mapping to $m$-body kinematic phase space which suggests the information accessed is mostly contained in 4- and 5-body (and in the case of highly boosted top quarks, 6-body) kinematic phase space, which can be expected to be well-modeled by modern Monte Carlo event generators. The small differences that remain are better explained by differences in hyperparameter optimisation and overall architecture of the networks than differences in the underlying information accessed.

It is interesting to note that [24] found comparable performance between jet image networks, $n$-subjettiness networks, and a linear ML algorithm making use of EFPs at a number of different tagging applications as well. Due to the simpler structure and training of linear ML algorithms and EFPs forming a complete linear basis of jet substructure observables, this suggests all of these methods are doing a good job at accessing the full substructure information available in a generic jet. This conclusion is also supported by the performance of the ML tagger operating directly on the kinematic information of individual final state particles presented in [49]: when taking detector granularity into account, such information only becomes necessary at very high boosts where the calorimeters are too granular to resolve the hard decay, but the ultimate performance of the tagger matches a comparable image network to the one presented here (with mass information taken into account by only taking pre-processing steps which don’t smear it out) for top quarks with $p_T \in [350, 400]$ GeV. We note that the $n$-subjettiness network presented here also can take particle flow information into account, and therefore should remain competitive at large boosts.

Having established that heavy object tagging using jet image and $n$-subjettiness variables offer comparable performance in an idealised setting, it is interesting to also ask the question of to what extent this agreement survives under a more realistic setting where there are additional coloured particles and pile-up in the final state, background samples are more complex, and detector effects are taken into account. We leave such a study to future work.

Acknowledgements

We would like to thank Kalliopi Petraki, Andy Buckley, Michael Russell, and Anja Butter for useful discussions. SV and MF would like to thank Robert Hogan for many related discussions and support. The work of MF was supported partly by the STFC Grant ST/L000326/1. MF is also funded by the European Research Council under the European Union’s Horizon 2020 programme (ERC Grant Agreement no.648680 DARKHORIZONS). SV is the recipient of a Rick Trainor PhD Scholarship. KN is supported by the NWO Vidi grant "Self-interacting asymmetric dark matter". LM is supported by the FNRS under the IISN convention "Maximising the LHC physics potential by advanced simulation tools and data analysis methods".

---

\[ We note that we find slightly better agreement between the jet image and $n$-subjettiness results than [24], especially for top quark tagging, however this could be due to the manner in which we include the jet mass information. \]
References

[1] M. Aaboud et al., *Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector*, JHEP 12, 024 (2017), doi:10.1007/JHEP12(2017)024, 1708.03299

[2] M. H. Seymour, *Searches for new particles using cone and cluster jet algorithms: A Comparative study*, Z. Phys. C62, 127 (1994), doi:10.1007/BF01559532

[3] J. M. Butterworth, A. R. Davison, M. Rubin and G. P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, Phys. Rev. Lett. 100, 242001 (2008), doi:10.1103/PhysRevLett.100.242001, 0802.2470

[4] D. E. Kaplan, K. Rehermann, M. D. Schwartz and B. Tweedie, *Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks*, Phys. Rev. Lett. 101, 142001 (2008), doi:10.1103/PhysRevLett.101.142001, 0806.0848

[5] D. E. Soper and M. Spannowsky, *Finding physics signals with shower deconstruction*, Phys. Rev. D84, 074002 (2011), doi:10.1103/PhysRevD.84.074002, 1102.3480

[6] D. E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, Phys. Rev. D87, 054012 (2013), doi:10.1103/PhysRevD.87.054012, 1211.3140

[7] T. Plehn, G. P. Salam and M. Spannowsky, *Fat Jets for a Light Higgs*, Phys. Rev. Lett. 104, 111801 (2010), doi:10.1103/PhysRevLett.104.111801, 0910.5472

[8] T. Plehn, M. Spannowsky, M. Takeuchi and D. Zerwas, *Stop Reconstruction with Tagged Tops*, JHEP 10, 078 (2010), doi:10.1007/JHEP10(2010)078, 1006.2833

[9] G. Kasieczka, T. Plehn, T. Schell, T. Strebler and G. P. Salam, *Resonance Searches with an Updated Top Tagger*, JHEP 06, 203 (2015), doi:10.1007/JHEP06(2015)203, 1503.05921

[10] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, JHEP 02, 118 (2015), doi:10.1007/JHEP02(2015)118, 1407.5675

[11] 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, 086 (2015), doi:10.1007/JHEP07(2015)086, 1501.05968

[12] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, *Jet-images — deep learning edition*, JHEP 07, 069 (2016), doi:10.1007/JHEP07(2016)069, 1511.05190

[13] P. T. Komiske, E. M. Metodiev and M. D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, JHEP 01, 110 (2017), doi:10.1007/JHEP01(2017)110, 1612.01551

[14] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning Top Taggers or The End of QCD?*, JHEP 05, 006 (2017), doi:10.1007/JHEP05(2017)006, 1701.08784
[15] J. Barnard, E. N. Dawe, M. J. Dolan and N. Rajcic, Parton Shower Uncertainties in Jet Substructure Analyses with Deep Neural Networks, Phys. Rev. D95(1), 014018 (2017), doi:10.1103/PhysRevD.95.014018, 1609.00607

[16] J. Thaler and K. Van Tilburg, Identifying Boosted Objects with N-subjettiness, JHEP 03, 015 (2011), doi:10.1007/JHEP03(2011)015, 1011.2268

[17] I. W. Stewart, F. J. Tackmann and W. J. Waalewijn, N-Jettiness: An Inclusive Event Shape to Veto Jets, Phys. Rev. Lett. 105, 092002 (2010), doi:10.1103/PhysRevLett.105.092002, 1004.2489

[18] J. Thaler and K. Van Tilburg, Maximizing Boosted Top Identification by Minimizing N-subjettiness, JHEP 02, 093 (2012), doi:10.1007/JHEP02(2012)093, 1108.2701

[19] G. Aad et al., Jet mass and substructure of inclusive jets in $\sqrt{s} = 7$ TeV pp collisions with the ATLAS experiment, JHEP 05, 128 (2012), doi:10.1007/JHEP05(2012)128, 1203.4606

[20] A. Altheimer et al., Boosted objects and jet substructure at the LHC. Report of BOOST2012, held at IFIC Valencia, 23rd-27th of July 2012, Eur. Phys. J. C74(3), 2792 (2014), doi:10.1140/epjc/s10052-014-2792-8, 1311.2708

[21] L. Asquith et al., Jet Substructure at the Large Hadron Collider : Experimental Review (2018), 1803.06991

[22] A. J. Larkoski, G. P. Salam and J. Thaler, Energy Correlation Functions for Jet Substructure, JHEP 06, 108 (2013), doi:10.1007/JHEP06(2013)108, 1305.0007

[23] I. Moult, L. Necib and J. Thaler, New Angles on Energy Correlation Functions, JHEP 12, 153 (2016), doi:10.1007/JHEP12(2016)153, 1609.07483

[24] P. T. Komiske, E. M. Metodiev and J. Thaler, Energy flow polynomials: A complete linear basis for jet substructure, JHEP 04, 013 (2018), doi:10.1007/JHEP04(2018)013, 1712.07124

[25] K. Datta and A. Larkoski, How Much Information is in a Jet?, JHEP 06, 073 (2017), doi:10.1007/JHEP06(2017)073, 1704.08249

[26] K. Datta and A. J. Larkoski, Novel Jet Observables from Machine Learning, JHEP 03, 086 (2018), doi:10.1007/JHEP03(2018)086, 1710.01305

[27] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, JHEP 07, 079 (2014), doi:10.1007/JHEP07(2014)079, 1405.0301

[28] R. D. Ball, V. Bertone, S. Carrazza, C. S. Deans, L. Del Debbio et al., Parton distributions with LHC data, Nucl.Phys. B867, 244 (2013), doi:10.1016/j.nuclphysb.2012.10.003, 1207.1303

[29] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, M. Rüfenacht, M. Schönherr and G. Watt, LHAPDF6: parton density access in the LHC precision era, Eur. Phys. J. C75, 132 (2015), doi:10.1140/epjc/s10052-015-3318-8, 1412.7420
[30] J. Alwall et al., *A Standard format for Les Houches event files*, Comput. Phys. Commun. **176**, 300 (2007), doi:10.1016/j.cpc.2006.11.010 [hep-ph/0609017].

[31] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, *MadGraph 5: Going Beyond*, JHEP **06**, 128 (2011), doi:10.1007/JHEP06(2011)128 [1106.0522].

[32] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen and P. Z. Skands, *An Introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191**, 159 (2015), doi:10.1016/j.cpc.2015.01.024 [1410.3012].

[33] M. Dobbs and J. B. Hansen, *The HepMC C++ Monte Carlo event record for High Energy Physics*, Comput. Phys. Commun. **134**, 41 (2001), doi:10.1016/S0010-4655(00)00189-2.

[34] A. Buckley, J. Butterworth, L. Lonnblad, D. Grellscheid, H. Hoeth et al., *Rivet user manual*, Comput.Phys.Commun. **184**, 2803 (2013), doi:10.1016/j.cpc.2013.05.021 [1003.0694].

[35] M. Cacciari, G. P. Salam and G. Soyez, *The Anti-k(t) jet clustering algorithm*, JHEP **04**, 063 (2008), doi:10.1088/1126-6708/2008/04/063 [0802.1189].

[36] M. Cacciari, G. P. Salam and G. Soyez, *FastJet User Manual*, Eur.Phys.J. **C72**, 1896 (2012), doi:10.1140/epjc/s10052-012-1896-2 [1111.6097].

[37] *Quark versus Gluon Jet Tagging Using Jet Images with the ATLAS Detector*, Tech. Rep. ATL-PHYS-PUB-2017-017, CERN, Geneva (2017).

[38] *New Developments for Jet Substructure Reconstruction in CMS* (2017).

[39] P. Baldi, K. Bauer, C. Eng, P. Sadowski and D. Whiteson, *Jet Substructure Classification in High-Energy Physics with Deep Neural Networks*, Phys. Rev. **D93**(9), 094034 (2016), doi:10.1103/PhysRevD.93.094034 [1603.09349].

[40] V. Nair and G. E. Hinton, *Rectified linear units improve restricted boltzmann machines*, In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML/10, pp. 807–814. Omnipress, USA, ISBN 978-1-60558-907-7 (2010).

[41] J. S. Bridle, *Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition*, In F. F. Soulié and J. Hérault, eds., *Neurocomputing*, pp. 227–236. Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-642-76153-9 (1990).

[42] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, CoRR abs/1603.04467 (2016), [1603.04467].

[43] J. Nickolls, I. Buck, M. Garland and K. Skadron, *Scalable parallel programming with cuda*, Queue **6**(2), 40 (2008), doi:10.1145/1365490.1365500.

[44] J. A. Aguilar-Saavedra, J. H. Collins and R. K. Mishra, *A generic anti-QCD jet tagger*, JHEP **11**, 163 (2017), doi:10.1007/JHEP11(2017)163 [1709.01087].
[45] S. D. Ellis and D. E. Soper, *Successive combination jet algorithm for hadron collisions*, Phys. Rev. **D48**, 3160 (1993), doi:10.1103/PhysRevD.48.3160, hep-ph/9305266.

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

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *Dropout: A simple way to prevent neural networks from overfitting*, Journal of Machine Learning Research **15**, 1929 (2014).

[48] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, CoRR abs/1412.6980 (2014), 1412.6980.

[49] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned Top Tagging with a Lorentz Layer* (2017), 1707.08966.