

Parton shower uncertainties in jet substructure analyses with deep neural networks

James Barnard,^{*} Edmund Noel Dawe,[†] Matthew J. Dolan,[‡] and Nina Rajcic[§]
*ARC Centre of Excellence for Particle Physics at the Terascale, School of Physics,
 University of Melbourne, Victoria 3010, Australia*
 (Received 13 October 2016; published 18 January 2017)

Machine learning methods incorporating deep neural networks have been the subject of recent proposals for new hadronic resonance taggers. These methods require training on a data set produced by an event generator where the true class labels are known. However, this may bias the network towards learning features associated with the approximations to QCD used in that generator which are not present in real data. We therefore investigate the effects of variations in the modeling of the parton shower on the performance of deep neural network taggers using jet images from hadronic W bosons at the LHC, including detector-related effects. By investigating network performance on samples from the Pythia, Herwig and Sherpa generators, we find differences of up to 50% in background rejection for fixed signal efficiency. We also introduce and study a method, which we dub zooming, for implementing scale invariance in neural-network-based taggers. We find that this leads to an improvement in performance across a wide range of jet transverse momenta. Our results emphasize the importance of gaining a detailed understanding of what aspects of jet physics these methods are exploiting.

DOI: [10.1103/PhysRevD.95.014018](https://doi.org/10.1103/PhysRevD.95.014018)

I. INTRODUCTION

The past decade has seen an explosion of interest in understanding and exploiting the distribution of energy (substructure) within hadronic jets and boosted resonances at the Large Hadron Collider (LHC) [1–4]. The study of jet substructure and the ability to identify (“tag”) the hadronic decay products of a wide variety of such resonances—the Higgs, W and Z bosons; top quarks; supersymmetric particles; and other beyond-the-Standard-Model (BSM) states—is crucial in the analysis of both Standard Model processes and in searches for BSM physics, which will only become more important now that the LHC is running at high energy and with future colliders on the horizon.

Since the foundational work in Ref. [5] on studying jet substructure in Higgs-boson-associated production, a multitude of taggers and variables related to substructure have been proposed [5–14] (further discussion of which can be found in the BOOST proceedings [1–4]). These generally exploit our knowledge of QCD to construct functions which effectively discriminate between signal and background. Some of these techniques have already been applied to the problem of identifying boosted massive vector bosons and top quarks by the ATLAS and CMS collaborations in run 1 of the LHC [15–25].

Another approach currently under development involves the application of machine learning (ML) techniques to

hadronic resonance tagging and searches for new physics. The ML community has made large strides in problems related to image recognition and computer learning, which may now also be applied to particle physics. Signals produced by the LHC detectors may be processed into pixelated jet images [26], and ML algorithms can be adapted to discriminate between a signal (such as $h \rightarrow b\bar{b}$ or boosted hadronic top decays [27]) and background. These algorithms have also been proposed as classifiers in neutrino experiments [28,29].

The use of ML, and neural networks in particular, has a long history in particle physics and the idea of using neural networks for quark-gluon discrimination [30–32], Higgs tagging [33] and track identification [34] goes back over 25 years. However, the development of efficient deep neural networks (DNNs) and the computing power associated with graphics processing units (GPUs) means that image recognition technology has become extremely powerful, driving the resurgence of interest in these techniques.

Recent work has seen the application of neural networks with two hidden layers to hadronic top-quark tagging [27], and deep convolutional neural networks (known to have excellent performance in image classification) to the problem of identification of hadronic W decays [35]. These initial papers focussed on demonstrating and understanding the network performance, and used truth-level Monte Carlo (MC). The effects of pileup and detector resolution were explored in Ref. [36], which showed that despite the loss of resolution when these are taken into account the neural network is still somewhat superior to traditional techniques.

^{*}james.barnard@unimelb.edu.au

[†]edmund.dawe@unimelb.edu.au

[‡]dolan@unimelb.edu.au

[§]n.rajcic@student.unimelb.edu.au

There has also been work on extending these methods to jet flavor classification [37].

The theoretical study of these techniques and their utility in high-energy particle physics is still in its infancy, and there are a number of issues still to be clarified in how deep learning methods may be applied at the LHC. Some of these are related to the robustness of these techniques: How can we guarantee that a network is learning about the physics differences between signal and background, and not details particular to a specific MC event generator? How robust are taggers based on these networks against detector effects such as smearing and how do they degrade in the presence of pileup? A particular concern is that the network achieves a substantial fraction of its discriminatory power from soft features in the spectrum which are modeled phenomenologically rather than via perturbative QCD. This paper provides a study of some of these issues.

We study the behavior of neural networks over a number of different event generators, and hence different parton showers and models for hadronization. For simplicity, we will usually just refer to these collectively as the parton shower.

We find that varying the parton shower leads to changes in the background rejection efficiency of up to 50%, depending on the shower model and selected signal efficiency. We consider this to be large, and perhaps more than would be expected from perturbative uncertainties from the parton shower. We believe that caution is therefore required before these methods are applied on data, and our results emphasize the necessity of understanding what features of the jet images the neural networks are relying on to achieve their discriminatory power. We also find changes in the factorization and renormalization scales lead to negligible differences, while the addition of pileup leads to an overall degradation in network performance (in agreement with Ref. [36]) but not to a change in our conclusions.

There has also been interest recently in the development of scale-invariant jet and substructure taggers [38–41], and we discuss how similar ideas may be implemented in DNN-based taggers by applying a p_T -dependent “zooming” factor on the jet images. The addition of zooming leads to a slight improvement of around 10%–20% in the network performance over a wide range of jet transverse momenta. While in this article we focus on discriminating between hadronically decaying W bosons and QCD jets as a “standard candle,” these methods should be applicable to a wide variety of tagging and substructure issues.

In Sec. II we outline the architecture and training of the neural networks and in Sec. III discuss how we construct jet images, and present an idea of how to implement a scale-invariant tagger. In Sec. IV we show the variability in the DNN performance across multiple event generators and parton shower models.

II. NETWORK ARCHITECTURE, TRAINING, AND PERFORMANCE EVALUATION

We follow Ref. [35] in our choice of network architecture, who have already investigated the performance of a variety of different neural networks. While we have investigated convolutional networks, all results we present here have been produced using the MaxOut [42] architecture. The network input consists of 625 units, equal to the number of pixels (25×25) present in each jet image. The input layer is followed by two dense MaxOut layers consisting of 256 and 128 units each. The next two layers are fully connected with 64 and 25 units and use a ReLU activation function [43]. The output layer consists of two nodes and a sigmoid activation. Further discussion of network choices can be found in Ref. [35].

We used the Keras Deep Learning library [44] and the Adam algorithm [45] to train our networks on four NVIDIA Tesla K80 GPUs. After selecting jet images within a window on the jet mass, $50 < m < 110$ GeV, and transverse momentum, $200 < p_T < 500$ GeV, networks were trained with approximately 3M signal and 3M background images where the signal and background images have been weighted to produce flat p_T distributions. A portion (10%) of the training images were set aside to evaluate a cross-entropy loss function after each epoch and the network training terminated after 100 epochs or after 10 epochs without an improvement in the loss function. The Adam algorithm learning rate parameter was initially set to 0.001 and then reduced by 2% after each epoch. We obtained reasonable performance with a batch size of 100. We also implemented and tested a cross-validated Bayesian optimization procedure to determine optimal parameter values but did not observe performance that was significantly better and so we have left such investigations for future work. Further optimizing the DNN should anyway not affect our conclusions here as we probe the general variability of a DNN with reasonable performance over different parton shower models.

Finally, we evaluate the performance of a network by computing the inverse background efficiency as a function of signal efficiency across a binned likelihood ratio of the signal-to-background output of the network. This variant of the standard receiver operating characteristic (ROC) curve better displays differences in background rejection at low signal efficiency and can be constructed from arbitrary jet observables or combinations of observables through a (possibly multidimensional) binned likelihood ratio.

III. CONSTRUCTING A JET IMAGE

This section provides a complete description of how we construct jet images, from event generation to image output, along with the reasoning behind many of our choices. Our process closely follows the one described

in Ref. [35] with the addition that we have also tested an optional zooming step to reduce p_T dependence.

We have developed a new framework for jet image construction, network training and performance evaluation. Low-level Cython [46] wrappers have been developed for PYTHIA [47], Delphes [48], and FastJet [49] that allow these tools to be connected in the Python programming language, where particles, calorimeter towers, jets, and jet images are stored as structured NumPy [50] arrays and optionally written to files on disk in the HDF5 [51] format. This design provides the potential to train networks on jet images generated on the fly. For the studies presented here, however, we have created large HDF5 data sets of jet images once that are then split and used for network training and performance studies. Aside from the direct interface with PYTHIA, the framework is able to study output from other event generators by reading intermediate HepMC [52] files.

Following event generation, particles are given to the Delphes detector simulator configured with ATLAS-like settings where calorimeter towers extend to a maximum absolute pseudorapidity of 4.9. Jets are then reconstructed from the calorimeter towers (referred to as jet constituents below) using the anti- k_r algorithm [53] as implemented by FastJet 3.1.3. We have selected a jet clustering size of $R = 1.0$ for all studies presented here. For boosted W bosons with two-body decays the characteristic maximal separation of the subjects scales according to

$$\Delta R = \frac{2m_W}{p_T^{\min}} \quad (3.1)$$

where p_T^{\min} is the minimum transverse momentum of the jets to be considered in the analysis. We have studied jets with transverse momenta above 200 GeV, making $R = 1.0$ a reasonable choice.

The highest p_T jet is selected and subjects are formed in a jet trimming [7] stage, which also serves to lessen contributions from soft radiation in the underlying event. Using the k_t algorithm we recluster the jet constituents into subjects with a fixed size of $r = 0.3$ and then discard all subjects with less than 5% of the original jet momentum to form a trimmed jet. All jet observables are computed with the trimmed jet.

The next stages are designed to remove spatial symmetries. First, all constituents of the trimmed jet are translated in $\eta - \phi$ space to place the leading subject at the origin. We then define a grid of pixels with a resolution of 0.1×0.1 in $\eta - \phi$ space and a jet image is formed by taking the total transverse energy measured within each pixel,

$$E_{T,i} = \sum_j \frac{E_j}{\cosh \eta_j}, \quad (3.2)$$

for all constituents j in pixel i , with energy E_j and original pseudorapidity η_j . This image is rotated, either to put the subleading subject directly below the leading subject or to align the principle component of the jet image along the vertical axis if only one subject is present. It is then reflected, either to put the third-leading subject on the right-hand side of the image or to ensure that the total image intensity is highest on the right-hand side if there are only two subjects.

After the reflection stage above we are left with an image in which the leading subject is centered and the subleading subject (if present) is directly below. The separation between the two subjects is not constant, but varies linearly with $2m/p_T$. By standardizing this separation we can potentially improve the DNN performance over a wide range of jet p_T . The aim is to pick a scaling factor that enhances and standardizes features in signal images, i.e. those from boosted W decays, without artificially creating similar features in background QCD images. This optional step is in addition to those detailed in Ref. [35].

Denoting the physical separation between the two leading subjects as ΔR_{act} , enlarging all jet images by a factor $R/\Delta R_{\text{act}}$ for some fixed R gives a standardized jet image in which the separation (in pixels) between the two leading subjects is fixed for all images. The downside of this approach is that this is true for both signal and background images, so an improvement in isolation of the subleading subject is tempered by an enhancement of signal-like features in background images. For this reason we use the characteristic size assuming the W mass, $s = 2m_W/p_T$, and enlarge all jet images by a scaling factor $\max(R/s, 1)$, where R is the original jet clustering size. Images are enlarged by performing bicubic interpolation [54] at a higher resolution. For signal images $\Delta R_{\text{act}} \approx s$, so this rescaling is very similar to, if a little less effective than, using the actual separation of subjects to define the scaling factor. For background images this scaling is not strongly correlated with the subject separation so the subleading subject tends to be smeared out.

Jet images are then cropped at 25×25 pixels (whether they have been zoomed or not) and are normalized such that the sum of the squared pixel intensities is 1. As discussed in Ref. [35] this does not preserve the jet mass that can be calculated from the original jet image, but our zooming step destroys this information anyway.

In summary, the full jet image construction and preprocessing steps are as follows:

- (i) *Jet clustering and trimming*: Reconstruct jets from all calorimeter towers using the anti- k_r algorithm with a jet size $R = 1.0$ and select the leading jet. Trim the jet using the k_t algorithm with a subject size $r = 0.3$.
- (ii) *Translation*: All jet constituents are translated in $\eta - \phi$ space to put the leading subject at the origin.
- (iii) *Pixelization*: Pixelize the transverse energy of the jet using pixels of size $(0.1, 0.1)$ in $\eta - \phi$ space. This produces a jet image.

- (iv) *Rotation*: Rotate the jet image to put the subleading subjet directly below the leading subjet. If no subjets are present rotate to align the principle component of the jet image along the vertical axis.
- (v) *Reflection*: Reflect the jet image horizontally to put the third-leading subjet on the right-hand side. If there are only two subjets, reflect to ensure that the summed image intensity is highest on the right-hand side.
- (vi) *Zooming*: Optionally zoom the jet image by a factor that reduces dependence on the jet momentum.
- (vii) *Cropping and normalization*: Crop the jet image at 25×25 pixels and normalize pixel intensities to make the sum of their squares equal to 1.

In Fig. 1 we show the average jet images for boosted W and QCD jets in the range $200 < p_T < 500$ GeV for the default PYTHIA shower using the standard preprocessing in the top panels, and using the zooming procedure in the bottom panels. For the W -jets we note that the zooming procedure results in a more regular and compact average shower shape, and that the second (lower p_T) subjet becomes better spatially defined as expected. While the average image of the QCD jets becomes more compact, the

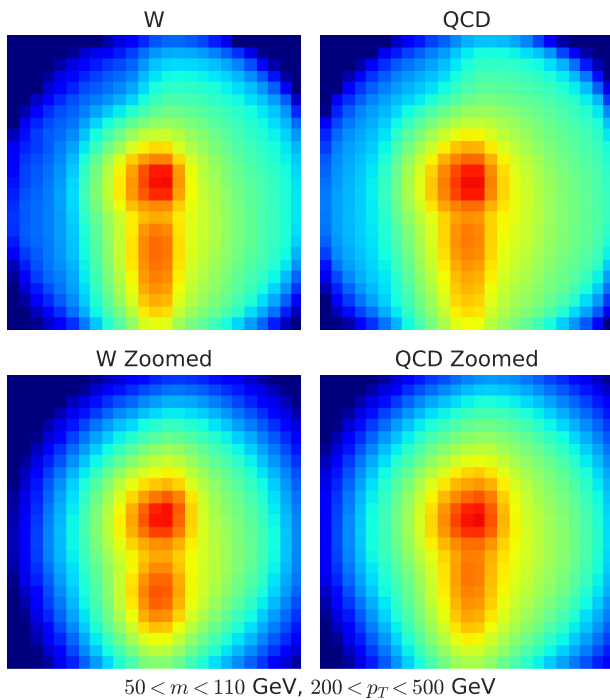


FIG. 1. We show the average jet images obtained for hadronic W bosons and QCD as modeled by the PYTHIA default shower. The images on the top have been preprocessed in the standard way, while those on the bottom have also undergone the zooming procedure outlined in Sec. III. The axes are left unlabeled since they do not correspond to the physical η and ϕ dimensions following image rotations, reflections, and zooming. Pixels are colored according to higher (red) and lower (blue) average normalized pixel intensities.

subjets remain somewhat smeared compared with the W -jets. Since the subjets do not originate in the decay of a heavy resonance and hence are not associated with a specific mass scale, this is not a surprise.

An obvious conceptual advantage of using the zooming technique is that it makes the construction of scale-invariant taggers easier. Scale-invariant searches [38–41] which are able to interpolate between the boosted and resolved parts of phase space have the advantage of being applicable over a broad range of masses and kinematics, allowing a single search or analysis to be effective where previously more than one may have been necessary.

We show in Fig. 2 the ROC curves for two different neural networks: the first (the solid blue line) was trained without zooming, while the second (the green dashed line) used zooming. Both networks were trained and tested on samples of jet images in the mass window $50 < m < 110$ GeV and a large p_T range, $200 < p_T < 500$ GeV. As predicted, the zoomed network outperforms the unzoomed one, particularly at low signal efficiency, where the background rejection rises by around 20%. We obtain similar results when we do not restrict the sample of jet images within a mass window. We find that the zooming has the greatest effect at high p_T . For less boosted W decays the enhancement in background rejection is around 10%, which rises to just over 20% for $300 < p_T < 500$ GeV.

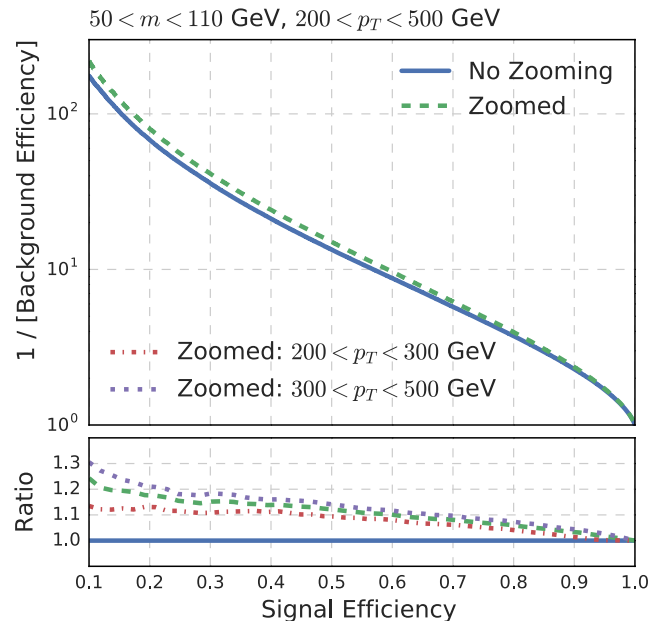


FIG. 2. The ROC curves for the zoomed (solid blue) and unzoomed (dashed green) jet images for the PYTHIA default shower. The lower panel shows the ratio of the zoomed to unzoomed efficiencies, also showing the efficiency sliced in bins from $200 < p_T < 300$ GeV (dotted-dashed red) and $300 < p_T < 500$ GeV (short dashed blue).

IV. EVENT GENERATOR DEPENDENCY

The networks require supervised training prior to being applied on unlabeled data. Since it is difficult to isolate regions of very high signal purity, training on simulated data is necessary before application to real LHC events. However, all MC event generators and parton showers are only approximations of the full Standard Model. Understanding what features of QCD a DNN is learning about, and whether it is learning event-generator-dependent approximations is thus an important question. Furthermore, there are features of real-world QCD such as color reconnection which, while modeled in the parton shower, are in reality poorly understood. We will not attempt to quantify those effects in this work.

To gain an understanding of the systematic uncertainties in using networks trained on simulated data, we study the behavior of networks across a variety of different generators and parton showers which all provide an adequate description of current LHC data. We assume that given a number of different ROC curves derived from different generators and parton showers, the envelope of these curves provides an approximate uncertainty band associated with training the network on simulated, rather than real, data.

Recently, Bellm *et al.* [55] studied parton shower uncertainties in HERWIG7. They divided the uncertainties into a number of classes: numerical, parametric, algorithmic, perturbative and phenomenological. Numerical uncertainties can be decreased by increasing the number of events, while parametric uncertainties are those external to the MC generator: masses, couplings, PDFs and so forth. The focus of our work in this section is on algorithmic uncertainties, those due to different choices of parton shower algorithm. Bellm *et al.* [55] focused on perturbative and phenomenological uncertainties, which are from truncation of expansion series and parameters deriving from nonperturbative models. Our work is more in the spirit of that of Andersen *et al.* [56]. Previous studies also exist within the HERWIG framework on the implications of MC uncertainties on jet substructure in the context of Higgs searches [57].

We generate background and signal events with three of the most widely used MC generators: PYTHIA8.219 [47], SHERPA2.0 [58,59] and HERWIG7.0 [60,61]. For PYTHIA8 we study both the default shower and the VINCIA shower [62,63], and for HERWIG we include both the default (angular ordered) and dipole showers [64,65], giving us five different parton shower models to study.

The default HERWIG shower (known as Qtilde) is based on $1 \rightarrow 2$ splittings using the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations, with an angular ordering criterion [66]. The SHERPA shower is based on a Catani-Seymour dipole formalism [67]. In this case one particle of the dipole is the emitter which undergoes the splitting, while the other is a spectator which compensates

for the recoil from the splitting and ensures that all particles remain on their mass shells throughout the shower, leading to easier integration with matching and merging techniques. The default shower in PYTHIA8 is also a dipole-style shower [68], ordered in transverse momentum.

While parton showers have traditionally been based upon partonic DGLAP splitting functions, another possibility is to consider color-connected parton pairs which undergo $2 \rightarrow 3$ branchings (note that this is distinct from Catani-Seymour dipoles used in SHERPA, where one parton is still an emitter, and the other recoils). In these so-called antenna showers, the two-parton antenna is described with a single radiation kernel. This has the advantage, for instance, of explicitly including both the soft and collinear limits. We use the recently released VINCIA [62,63] plug-in for PYTHIA8 as a representative antenna shower.

These event generators also provide different treatments of the soft radiation from the underlying event which accompanies each hard partonic scattering. They also possess different implementations of the parton-to-hadron fragmentation process being based either around cluster fragmentation ideas (HERWIG and SHERPA) or the Lund string model (PYTHIA), giving us a wide range of QCD-related effects to probe. To incorporate detector effects such as smearing we pass all events through the Delphes 3 detector simulator [48]. In the studies presented here, our baseline shower is PYTHIA8 with its default settings.

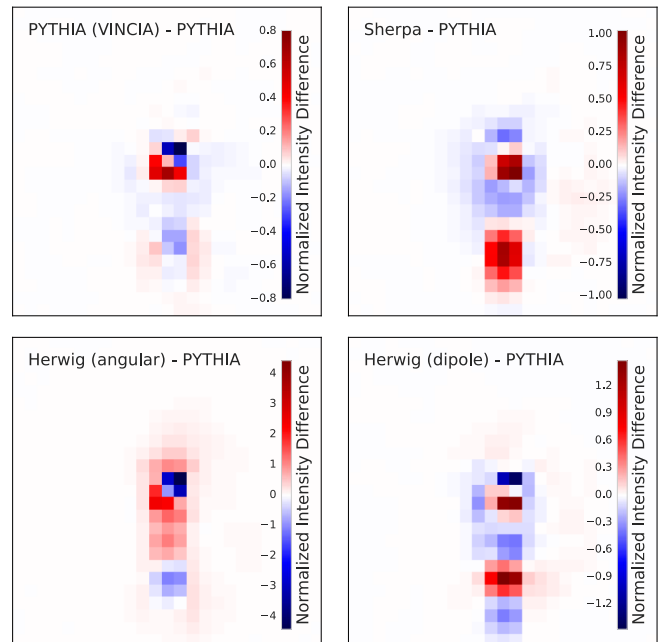


FIG. 3. This figure shows the W-jet image differences between the default PYTHIA shower and the alternate VINCIA shower in PYTHIA (top left), the default SHERPA shower (top right), the default HERWIG angular shower (bottom left) and the HERWIG dipole shower (bottom right). The plots have been individually normalized.

We construct average jet images for all five different generators and showers under investigation, and then subtract the default PYTHIA average jet image in order to see the differences in the average radiation patterns. The results are shown in Fig. 3 for the W-jet signal. We have normalized the intensity differences of the pixels so that red indicates a region of excess and blue a deficit relative to the PYTHIA default. While the VINCIA is roughly similar to the PYTHIA default, the SHERPA and HERWIG dipole showers exhibit more intense radiation in the resolved subjects and a substantial deficit in the region between the subjects. The HERWIG angular shower shows the opposite, with less radiation in the subject cores and more diffuse radiation. QCD radiation exhibits similar features.

Next we show ROC curves for the different showers in Fig. 4. We used the same network discussed in Sec. III trained on the default PYTHIA shower (without zooming), and then used events from the other generators and parton showers as input; e.g. we ask a neural network trained on the PYTHIA shower to discriminate between QCD and W-jets from SHERPA.

We do not extend the ROC curves down to zero signal efficiency since they are more statistically limited there. The PYTHIA ROC is higher than all other shower efficiency

curves. While both the SHERPA and HERWIG dipole images exhibit superficial similarities in Fig. 3, the network is better at discriminating the SHERPA events. At a fixed low signal efficiency the HERWIG angular and dipole showers have the lowest background rejection, smaller than that obtained using the PYTHIA default by a factor of 2. The VINCIA and SHERPA showers have a slightly lower rejection rate than the PYTHIA one. For signal efficiency of 50% the uncertainty from changing the event generator is around 40%.

For a large background rejection rate we note that the network trained on the PYTHIA events has a lower efficiency for selecting signal events generated from the other showers; i.e. it is maximally efficient for the shower it was trained on. This may be due to the network learning some features associated specifically with the PYTHIA shower and thus performing well on PYTHIA-like events.

We also show in Fig. 4 the ROC curves we obtain for the trimmed jet mass and the n -subjettiness ratio $\tau_{21} \equiv \tau_2/\tau_1$ [11] which is often used as a discriminating variable in studies of jet substructure [69]. We see that the neural network consistently outperforms these variables (in agreement with the conclusions already reached in Ref. [35]). This result stands independent of the uncertainty induced

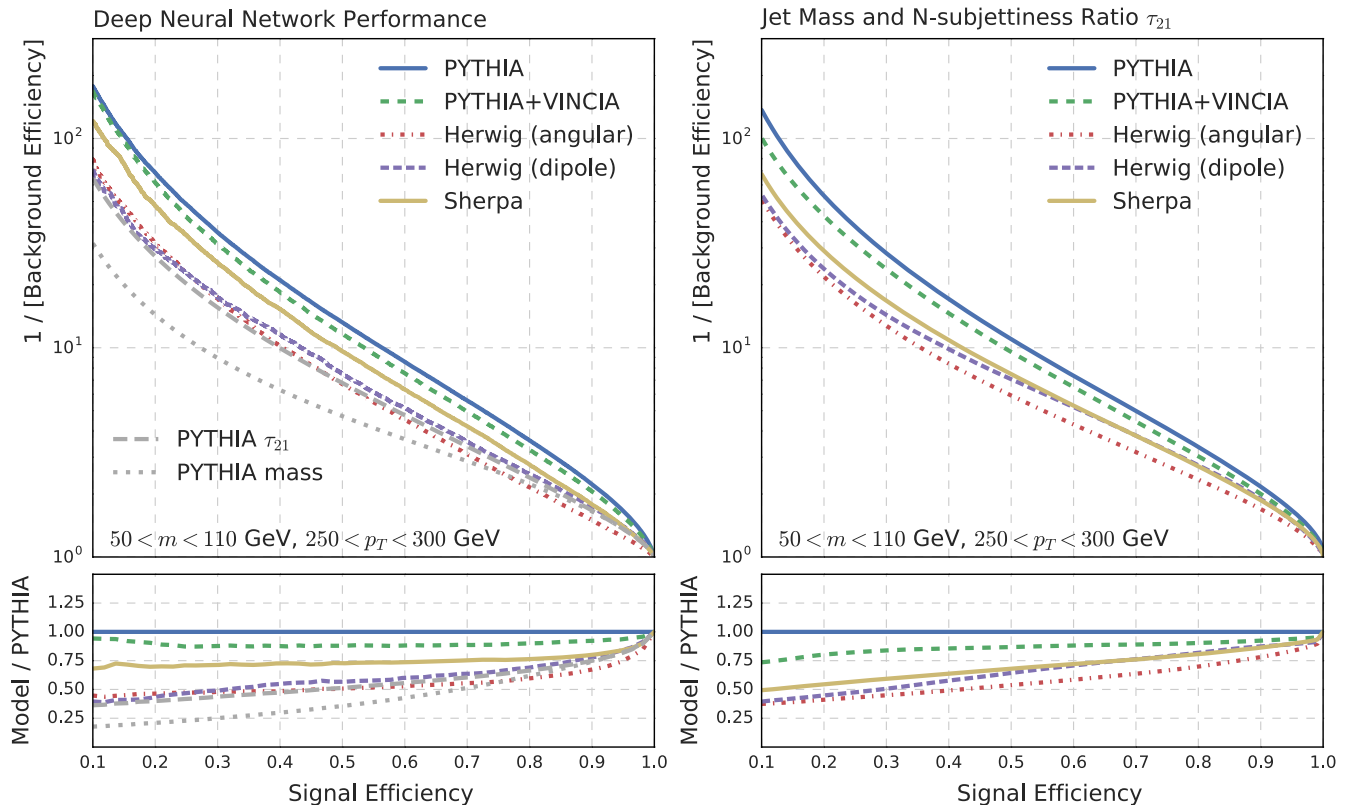


FIG. 4. This figure shows the ROC curves of the PYTHIA (solid blue), VINCIA (dashed green), HERWIG angular (red dotted-dashed) and dipole (dashed purple), and SHERPA (solid gold) showers for the DNN output (left) and the combination of the jet mass and n -subjettiness ratio τ_{21} through a two-dimensional binned likelihood ratio (right). The lower panels show the ratio of the ROCs with the default PYTHIA shower. All ROC curves are computed using jet images within a window on the jet mass, $50 < m < 110$ GeV, and transverse momentum, $250 < p_T < 300$ GeV.

by the choice of event generator, although the results for the HERWIG showers are close to being degenerate with it.

In the right panel of Fig. 4 we show the ROC curves we obtain from the combined jet mass and τ_{21} observables for the different parton showers. We see that the parton shower uncertainties in this case are very similar to those obtained from the jet images. The uncertainties related to varying the parton shower for the jet images are thus of similar size to those associated with other more common variables, such as those found in theoretical studies of the D2 tagger [70,71] and those used by the ATLAS Collaboration in searches for boosted W -bosons [18].

Another possible independent source of uncertainty is the dependence of the shower profile on the common renormalization and factorization scales, μ_R and μ_F respectively, which for our purposes are set to be $\mu = p_T$. As is standard, we vary the scale μ upwards and downwards by a factor of 2 from its default value of $\mu = p_{T,W}$. We find that the changes in ROC curves due to this were negligible.

V. CONCLUSIONS

The use of deep neural networks to construct classifiers for hadronic substructure using jet images is an exciting proposal. However, it is important to quantify the dependence on the training data set, and whether the network is learning the approximations inherent in the MC generator. We trained a network on the default parton shower from the PYTHIA generator and studied its performance on events from HERWIG and SHERPA. We found that the network performed better on test events also from the default PYTHIA shower, indicating that the network may be learning some PYTHIA-specific features. The change in performance through using different parton showers could be up to a factor of 2 in background rejection. Our results thus indicate that care is required to avoid overinterpretation of small changes in ROC curves, given the parton shower uncertainties. These uncertainties are relatively large, and further study is required to ascertain whether the network performance is truly being driven by features in the parton shower (which are under control in perturbative QCD) or by softer physics such as hadronization modeling (which is not). Either way, our results demonstrate that caution is required in the application of machine learning

techniques on simulated data. We intend to return to this issue in the near future.

There are many opportunities for further work in this area. One way to achieve event generator independency is by avoiding the use of training data through using data-driven unsupervised learning algorithms. Will these prove as powerful as the supervised techniques using DNNs proposed thus far? It would also be desirable to incorporate more than just calorimeter information into jet images, in a similar vein to recent work on heavy flavor tagging [37]. We note that tracking information has been proposed in the context of substructure as being particularly important at high energies [72]. Since jet images can be both large and sparse, new algorithms may be required to render this feasible [73]. In an ideal world, it would be possible to use information from the whole detector to classify events into signal or background, as in event deconstruction [74].

The main outcome of our study is to emphasize the importance of what the neural networks are learning and how they use it to discriminate between signal and background. This lesson is true for any application of machine learning in particle physics, from widely used techniques such as boosted decision trees to newer methods like the deep neural network we have studied. However, provided that cautious and detailed studies of the uncertainties involved lead to methods to constrain them in the analysis of real data, these methods may prove to be powerful and reliable analysis tools for future searches and measurements.

ACKNOWLEDGMENTS

We thank Peter Skands, Michael Spannowsky and Phillip Urquijo for helpful discussions and comments. This work was supported in part by the Australian Research Council. We are grateful for the computational support provided by our colleagues in CoEPP (Centre for Excellence in Particle Physics) and Research Compute Services at the University of Melbourne for generously providing access to the GPUs (Graphics Processing Unit). We thank the Aspen Center for Physics, which is supported by National Science Foundation Grant No. PHY-1066293, the Munich Institute for Astro and Particle Physics and the Mainz Institute of Theoretical Physics, where part of this work was completed, for their hospitality and support.

-
- [1] A. Abdesselam *et al.*, Boosted objects: A probe of beyond the Standard Model physics, *Eur. Phys. J. C* **71**, 1661 (2011).
 [2] A. Altheimer *et al.*, Jet substructure at the Tevatron and LHC: New results, new tools, new benchmarks, *J. Phys. G* **39**, 063001 (2012).

- [3] 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. C* **74**, 2792 (2014).
 [4] D. Adams *et al.*, Towards an understanding of the correlations in jet substructure, *Eur. Phys. J. C* **75**, 409 (2015).

- [5] 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).
- [6] T. Plehn, G. P. Salam, and M. Spannowsky, Fat Jets for a Light Higgs, *Phys. Rev. Lett.* **104**, 111801 (2010).
- [7] D. Krohn, J. Thaler, and L.-T. Wang, Jet trimming, *J. High Energy Phys.* **02** (2010) 084.
- [8] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches, *Phys. Rev. D* **81**, 094023 (2010).
- [9] J. Gallicchio, J. Huth, M. Kagan, M. D. Schwartz, K. Black, and B. Tweedie, Multivariate discrimination and the Higgs + W/Z search, *J. High Energy Phys.* **04** (2011) 069.
- [10] J. Gallicchio and M. D. Schwartz, Seeing in Color: Jet Superstructure, *Phys. Rev. Lett.* **105**, 022001 (2010).
- [11] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, *J. High Energy Phys.* **03** (2011) 015.
- [12] A. Hook, M. Jankowiak, and J. G. Wacker, Jet dipolarity: Top tagging with color flow, *J. High Energy Phys.* **04** (2012) 007.
- [13] A. J. Larkoski, G. P. Salam, and J. Thaler, Energy correlation functions for jet substructure, *J. High Energy Phys.* **06** (2013) 108.
- [14] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft drop, *J. High Energy Phys.* **05** (2014) 146.
- [15] G. Aad *et al.* (ATLAS Collaboration), Search for Dark Matter in Events with a Hadronically Decaying W or Z Boson and Missing Transverse Momentum in pp Collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector, *Phys. Rev. Lett.* **112**, 041802 (2014).
- [16] G. Aad *et al.* (ATLAS Collaboration), Performance of jet substructure techniques for large- R jets in proton-proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector, *J. High Energy Phys.* **09** (2013) 076.
- [17] G. Aad *et al.* (ATLAS Collaboration), Search for high-mass diboson resonances with boson-tagged jets in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector, *J. High Energy Phys.* **12** (2015) 055.
- [18] G. Aad *et al.* (ATLAS Collaboration), Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV, *Eur. Phys. J. C* **76**, 154 (2016).
- [19] G. Aad *et al.* (ATLAS Collaboration), Measurement of the charged-particle multiplicity inside jets from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector, *Eur. Phys. J. C* **76**, 322 (2016).
- [20] S. Chatrchyan *et al.* (CMS Collaboration), Search for anomalous $t\bar{t}$ production in the highly-boosted all-hadronic final state, *J. High Energy Phys.* **09** (2012) 029.
- [21] CMS Collaboration, CERN Technical Report No. CMS-PAS-JME-13-007, 2014.
- [22] V. Khachatryan *et al.* (CMS Collaboration), Identification techniques for highly boosted W bosons that decay into hadrons, *J. High Energy Phys.* **12** (2014) 017.
- [23] V. Khachatryan *et al.* (CMS Collaboration), Search for vector-like charge $2/3$ T quarks in proton-proton collisions at $\sqrt{s} = 8$ TeV, *Phys. Rev. D* **93**, 012003 (2016).
- [24] CMS Collaboration, CERN Technical Report No. CMS-PAS-EXO-15-002, 2015.
- [25] V. Khachatryan *et al.* (CMS Collaboration), Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV, [arXiv:1607.03663](https://arxiv.org/abs/1607.03663).
- [26] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, Jet-images: Computer vision inspired techniques for jet tagging, *J. High Energy Phys.* **02** (2015) 118.
- [27] L. G. Almeida, M. Backovi, M. Cliche, S. J. Lee, and M. Perelstein, Playing tag with ANN: Boosted top identification with pattern recognition, *J. High Energy Phys.* **07** (2015) 086.
- [28] E. Racah, S. Ko, P. Sadowski, W. Bhimji, C. Tull, S.-Y. Oh *et al.*, Revealing fundamental physics from the Daya Bay neutrino experiment using deep neural networks, [arXiv:1601.07621](https://arxiv.org/abs/1601.07621).
- [29] A. Aurisano, A. Radovic, D. Rocco, A. Himmel, M. D. Messier, E. Niner, G. Pawloski, F. Psihas, A. Sousa, and P. Vahle, A convolutional neural network neutrino event classifier, *J. Instrum.* **11**, P09001 (2016).
- [30] L. Lonnblad, C. Peterson, and T. Rognvaldsson, Finding Gluon Jets with a Neural Trigger, *Phys. Rev. Lett.* **65**, 1321 (1990).
- [31] L. Lonnblad, C. Peterson, and T. Rognvaldsson, Using neural networks to identify jets, *Nucl. Phys.* **B349**, 675 (1991).
- [32] C. Peterson, T. Rognvaldsson, and L. Lonnblad, JETNET 3.0: A versatile artificial neural network package, *Comput. Phys. Commun.* **81**, 185 (1994).
- [33] P. Chiappetta, P. Colangelo, P. De Felice, G. Nardulli, and G. Pasquariello, Higgs search by neural networks at LHC, *Phys. Lett. B* **322**, 219 (1994).
- [34] B. H. Denby, Neural networks and cellular automata in experimental high-energy physics, *Comput. Phys. Commun.* **49**, 429 (1988).
- [35] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-images: Deep learning edition, *J. High Energy Phys.* **07** (2016) 069.
- [36] 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**, 094034 (2016).
- [37] 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**, 112002 (2016).
- [38] A. Azatov, M. Salvarezza, M. Son, and M. Spannowsky, Boosting top partner searches in composite Higgs models, *Phys. Rev. D* **89**, 075001 (2014).
- [39] M. Gouzevitch, A. Oliveira, J. Rojo, R. Rosenfeld, G. P. Salam, and V. Sanz, Scale-invariant resonance tagging in multijet events and new physics in Higgs pair production, *J. High Energy Phys.* **07** (2013) 148.
- [40] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure, *J. High Energy Phys.* **05** (2016) 156.
- [41] M. Schlaffer, M. Spannowsky, and A. Weiler, Searching for supersymmetry scalelessly, *Eur. Phys. J. C* **76**, 457 (2016).
- [42] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, Maxout networks, *Journal of Machine Learning Research: Workshop and Conference Proceedings* **28**, 1319 (2013).

- [43] X. Glorot, A. Bordes, and Y. Bengio, Deep sparse rectifier neural networks, *Journal of Machine Learning Research* **15** (2011).
- [44] F. Chollet, Keras, <https://github.com/fchollet/keras>, 2015.
- [45] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [46] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. Seljebotn, and K. Smith, Cython: The best of both worlds, *Comput. Sci. Eng.* **13**, 31 (2011).
- [47] 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).
- [48] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [49] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, *Eur. Phys. J. C* **72**, 1896 (2012).
- [50] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, The numpy array: A structure for efficient numerical computation, *Comput. Sci. Eng.* **13**, 22 (2011).
- [51] HDF Group, Hierarchical data format, version 5, 1997–2016.
- [52] M. Dobbs and J. B. Hansen, The HepMC C++ Monte Carlo event record for high energy physics, *Comput. Phys. Commun.* **134**, 41 (2001).
- [53] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k(t) jet clustering algorithm, *J. High Energy Phys.* **04** (2008) 063.
- [54] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, New York, 1992), 2nd ed.
- [55] J. Bellm, G. Nail, S. Platzer, P. Schichtel, and A. Sidmök, Parton shower uncertainties with Herwig 7: Benchmarks at leading order, *Eur. Phys. J. C* **76**, 665 (2016).
- [56] J. R. Andersen *et al.*, Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report, in 9th Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2015) Les Houches, France, 2015, 2016.
- [57] P. Richardson and D. Winn, Investigation of Monte Carlo uncertainties on Higgs boson searches using jet substructure, *Eur. Phys. J. C* **72**, 2178 (2012).
- [58] T. Gleisberg and S. Hoeche, Comix: A new matrix element generator, *J. High Energy Phys.* **12** (2008) 039.
- [59] T. Gleisberg, S. Höche, F. Krauss, M. Schönherr, S. Schumann, F. Siegert, and J. Winter, Event generation with SHERPA 1.1, *J. High Energy Phys.* **02** (2009) 007.
- [60] M. Bahr *et al.*, Herwig++ physics and manual, *Eur. Phys. J. C* **58**, 639 (2008).
- [61] J. Bellm *et al.*, Herwig 7.0 / Herwig++ 3.0 release note, *Eur. Phys. J. C* **76**, 196 (2016).
- [62] N. Fischer, S. Prestel, M. Ritzmann, and P. Skands, Vincia for hadron colliders, [arXiv:1605.06142](https://arxiv.org/abs/1605.06142).
- [63] M. Ritzmann, D. A. Kosower, and P. Skands, Antenna showers with hadronic initial states, *Phys. Lett. B* **718**, 1345 (2013).
- [64] S. Platzer and S. Gieseke, Coherent parton showers with local recoils, *J. High Energy Phys.* **01** (2011) 024.
- [65] S. Platzer and S. Gieseke, Dipole showers and automated NLO matching in Herwig++, *Eur. Phys. J. C* **72**, 2187 (2012).
- [66] S. Gieseke, P. Stephens, and B. Webber, New formalism for QCD parton showers, *J. High Energy Phys.* **12** (2003) 045.
- [67] S. Schumann and F. Krauss, A parton shower algorithm based on Catani-Seymour dipole factorisation, *J. High Energy Phys.* **03** (2008) 038.
- [68] T. Sjöstrand and P. Z. Skands, Transverse-momentum-ordered showers and interleaved multiple interactions, *Eur. Phys. J. C* **39**, 129 (2005).
- [69] The ATLAS Collaboration, CERN Technical Report No. ATL-PHYS-PUB-2014-004, 2014.
- [70] A. J. Larkoski, I. Moulton, and D. Neill, Power counting to better jet observables, *J. High Energy Phys.* **12** (2014) 009.
- [71] A. J. Larkoski, I. Moulton, and D. Neill, Analytic boosted boson discrimination, *J. High Energy Phys.* **05** (2016) 117.
- [72] M. Spannowsky and M. Stoll, Tracking new physics at the LHC and beyond, *Phys. Rev. D* **92**, 054033 (2015).
- [73] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, Recurrent models of visual attention, [arXiv:1406.6247](https://arxiv.org/abs/1406.6247).
- [74] D. E. Soper and M. Spannowsky, Finding physics signals with event deconstruction, *Phys. Rev. D* **89**, 094005 (2014).