Jet substructure classification in high-energy physics with deep neural networks

Pierre Baldi,[1] Kevin Bauer,[2] Clara Eng,[3] Peter Sadowski,[1] and Daniel Whiteson[2]

[1]*Department of Computer Science, University of California, Irvine, California 92697, USA*
[2]*Department of Physics and Astronomy, University of California, Irvine, California 92697, USA*
[3]*Department of Chemical and Biomolecular Engineering, University of California,*
Berkeley California 94270, USA
(Received 15 April 2016; published 27 May 2016)

At the extreme energies of the Large Hadron Collider, massive particles can be produced at such high velocities that their hadronic decays are collimated and the resulting jets overlap. Deducing whether the substructure of an observed jet is due to a low-mass single particle or due to multiple decay objects of a massive particle is an important problem in the analysis of collider data. Traditional approaches have relied on expert features designed to detect energy deposition patterns in the calorimeter, but the complexity of the data make this task an excellent candidate for the application of machine learning tools. The data collected by the detector can be treated as a two-dimensional image, lending itself to the natural application of image classification techniques. In this work, we apply deep neural networks with a mixture of locally connected and fully connected nodes. Our experiments demonstrate that without the aid of expert features, such networks match or modestly outperform the current state-of-the-art approach for discriminating between jets from single hadronic particles and overlapping jets from pairs of collimated hadronic particles, and that such performance gains persist in the presence of pileup interactions.

I. INTRODUCTION

Collisions at the LHC occur at such high energies that even massive particles are produced at large enough velocities that their decay products become collimated. In the case of a hadronic decay of a boosted W boson ($W \to qq'$), the two jets produced from these two quarks then overlap in the detector, creating a single merged jet. The substructure of the jet's energy deposition can distinguish between jets which are due to a single hadronic particle or due to the decay of a massive object into multiple hadronic particles; this classification is known as jet "tagging" and is critical for understanding the nature of the particles produced in the collision [1].

This classification task has been the topic of intense research activity [2–5]. The difficult nature of the problem has lead physicists to reduce the dimensionality of the problem by designing expert features [6–15] which incorporate their domain knowledge. In the current state-of-the-art applications, jets are either classified based on one of these features alone or by combining multiple designed features with shallow machine learning classifiers such as boosted decision trees (BDTs). It is possible, however, that these designed expert features do not capture all of the available information [16–18], as the data are very high dimensional. Despite extensive theoretical progress in the microphysics of jet formation [19–21] and development of theoretically motivated tools to mitigate the impact of unrelated interactions [22] which makes the expert features robust and physically

meaningful, there exists no complete analytical model for classification directly from theoretical principles, though see Ref. [23]. On the other hand, the existence of effective simulation tools [24,25] allows for the generation of large simulated samples. Therefore, approaches that use the higher-dimensional but lower-level detector information to learn this classification function may outperform those which rely on fewer high-level expert-designed features.

Measurements of the emanating particles can be projected onto a cylindrical detector and then unwrapped and considered as two-dimensional images, enabling the natural application of computer vision techniques. Recent work demonstrates encouraging results with shallow classification models trained on jet images [26–28]. Deep networks have shown additional promise in particle-level studies [29]. However, deep learning has not yet been applied to more realistic scenarios which include simulation of the detector response and resolution, and most importantly, the effect of unrelated simultaneous pp interactions, known as *pileup* which contributes significant energy depositions unrelated to the particles of interest.

In this paper, we perform jet classification on images built from simulated detector response using deep neural network models with a combination of locally connected and fully connected layers. Our results demonstrate that deep networks can distinguish between detector clusters due to single or multiple jets without using domain knowledge, matching or exceeding the performance of shallow classifiers used to combine many expert features.

II. THEORY

A typical application of jet classifiers is to discriminate single jets produced in quark or gluon fragmentation from two overlapping jets produced when a high-velocity W boson decays to a collimated pair of quarks. The goal is then to learn the classification function, or equivalently, the likelihood ratio:

$$\frac{P_{W \to qq}(\text{jet})}{P_{q/g}(\text{jet})}.$$

In practice, there are two significant obstacles to calculating and applying this ratio.

First, while theoretical understanding of the processes involved has made significant progress, a formulation of this likelihood ratio from fundamental QCD principles is not yet available. However, there do exist effective models which have been successfully incorporated into widely used tools capable of generating simulated samples. Such samples can then be used to deduce the likelihood ratio, but the task is very difficult due to its high dimensionality. Expert features with solid theoretical grounding exist to reduce the dimensionality of this problem, but it is unlikely that they capture all of the information, as the theoretical understanding is not complete and the concepts which motivate them do not include the detector effects or the impact of pileup interactions. The goal of this paper is to attempt to capture as much of the information as possible and learn the classification function from simulated samples which include these effects, without making the simplifying theoretical assumptions necessary to construct expert features.

Second, the effective models used in simulation tools do not provide a perfectly accurate description of observed collider data. A classification function learned from simulated samples is limited by the validity of those samples. While deep networks may provide a powerful method of deducing the classification function, expert features which encapsulate theoretical understanding of the process of jet formation are valuable in assessing the success and failure of these models. In this paper, we use expert features as a benchmark to measure the performance of learning tools which access only the higher-dimensional lower-level data. We expect that deep networks may provide additional classification power in concert with the insight offered by expert features, and perhaps motivate the development of modifications to such features rather than blindly replacing them.

III. DATA

Training samples for both classes were produced using realistic simulation tools widely used in particle physics.

Samples of boosted $W \to qq'$ were generated with a center of mass energy $\sqrt{s} = 14$ TeV using the diboson production and decay process $pp \to W^+W^- \to qqqq$ leading to two pairs of quarks; each pair of quarks are collimated and lead to a single jet. Samples of jets originating from single quarks and gluons were generated using the $pp \to qq, qg, gg$ process. In both cases, jets are generated in the range of $p_T \in [300, 400]$ GeV.

Collisions and immediate decays were simulated with MADGRAPH5 [30] v2.2.3, showering and hadronization simulated with PYTHIA [24] v6.426, and response of the detectors simulated with DELPHES [31] v3.2.0. The jet images are characterized by the energies deposited at different points on the approximately cylindrical calorimeter surface.

The classification of jets as due to $W \to qq'$ or single quarks and gluons is sensitive to the presence of additional in-time pp interactions, referred to as *pileup* events. We overlay such interactions in the simulation chain, with an average number of interactions per event of $<\mu> = 50$, as an estimate of future ATLAS Run 2 data with the LHC delivering collisions at a 25 ns bunch crossing interval. The impact of pileup events on jet reconstruction can be mitigated using several techniques. After reconstructing jets with the anti-k_T [32] clustering algorithm using distance parameter $R = 1.2$, we apply a jet-trimming algorithm [22] which is designed to remove pileup while preserving the two-pronged jet substructure characteristic of boson decay. Jet trimming reclusters the jet constituents using the k_T [33] algorithm into subjets of radius 0.2 and discards subjets with p_T less than 3% of the original jet. Then the final trimmed jet is built using the remaining subjets. Trimmed jets with 300 GeV $< p_T < 400$ GeV are selected, in order to ensure the minimum W boson velocity needed for collimated decays. In principle, the machine learning algorithms may be able to classify jets without such filtering; we leave this for future studies.

To compare our approach to the current state-of-the-art approach, we calculate six high-level jet variables commonly used in the literature; calculations are performed using FastJet [34] v3.1.2. First, the invariant mass of the trimmed jet is calculated. Then, the trimmed jet's constituents are used to calculate the other substructure variables, N-subjettiness [9,35] $\tau_{21}^{\beta=1}$, and the energy correlation functions [10,36] $C_2^{\beta=1}$, $C_2^{\beta=2}$, $D_2^{\beta=1}$, and $D_2^{\beta=2}$. A comprehensive summary of these six jet substructure variables can be found in Ref. [2]. Figure 1 shows the distribution of the variables for the two classes of jets, both with and without pileup conditions.

In this paper, we investigate the power of classification of the jets directly from the lower-level but higher-dimensional calorimeter data, without the dimensional reduction provided by the variables above. The strategy follows that of well-established image classification tools by treating the distribution of energy in the calorimeter as an image. The images were preprocessed as in previous work by centering and rotating into a canonical orientation. The origin of the coordinate axis was set at the center of
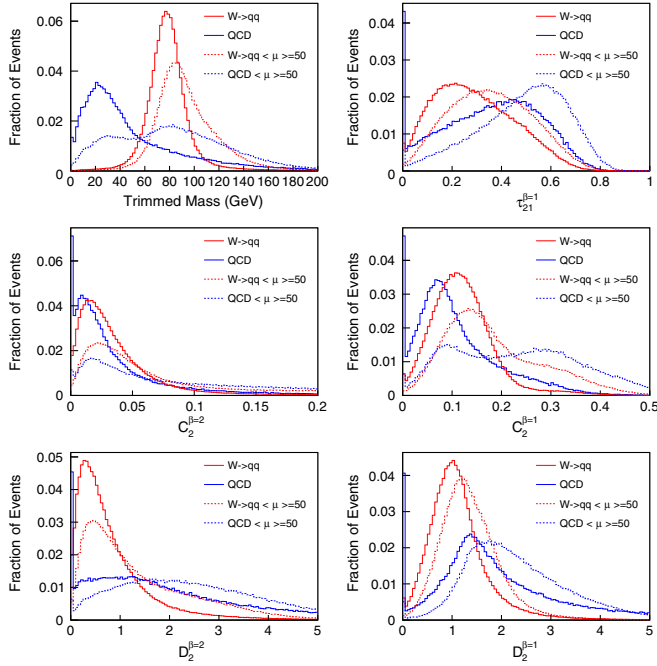
FIG. 1. Distributions in simulated samples of high-level jet substructure variables widely used to discriminate between jets due to collimated decays of massive objects ($W \to qq$) and jets due to individual quarks or gluons (QCD). Two cases are shown: with and without the presence of additional in-time pp interactions, included at the level of an average of 50 such interactions per collision.

energy of each jet, then the image was rotated so that the principal axis θ is in the same direction for each jet, where θ is defined as

$$\tan(\theta) = \sum_i \frac{\phi_i \times E_i}{R_i} \bigg/ \sum_i \frac{\eta_i \times E_i}{R_i} \qquad (1)$$

$$R_i = \sqrt{\eta_i^2 + \phi_i^2}. \qquad (2)$$

Images are then reflected so that the maximum energy value is always in the bottom half of the image.
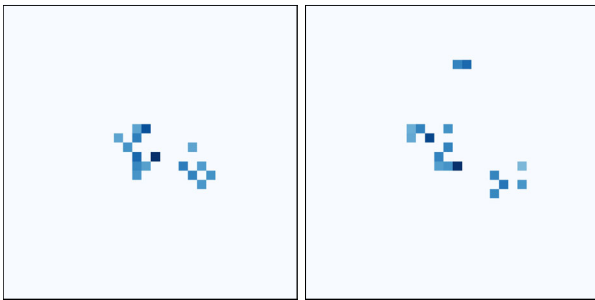

FIG. 2. Typical jet images from class 1 (single QCD jet from q or g) on the left, and class 2 (two overlapping jets from $W \to qq'$) on the right, after preprocessing as described in the text.
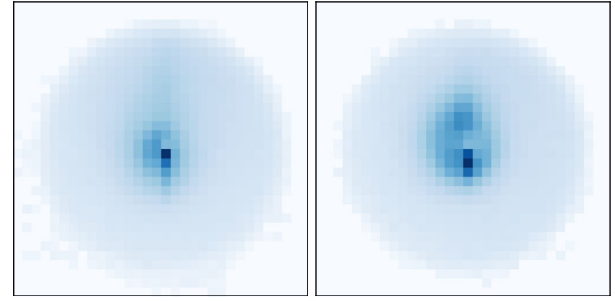

FIG. 3. Average of 100,000 jet images from class 1 (single QCD jet from q or g) on the left, and class 2 (two overlapping jets from $W \to qq'$) on the right, after preprocessing.

The jet energy deposits were centered and cropped to within a 3.0×3.0 radian window, then binned into pixels to form a 32×32 image, approximating the resolution of the calorimeter cells. When two calorimeter cells were detected within the same pixel, their energies were summed. Example individual jet images from each class are shown in Fig. 2, and averages over many jets are shown in Fig. 3.

IV. TRAINING

Deep neural networks were trained on the jet images and compared to the standard approach of BDTs trained on expert-designed variables that capture domain knowledge [2]. All classifiers were trained on a balanced training data set of 10 million examples, with 500 thousand of these used as a validation set. The best hyperparameters for each method were selected using the Spearmint Bayesian optimization algorithm [37] to optimize over the supports specified in Tables I and II. The best models were then tested on a separate test set of 5 million examples.

Neural networks consisted of hidden layers of *tanh* units and a logistic output unit with cross-entropy loss. Weight updates were made using the ADAM optimizer [38] ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e - 08$) with mini-batches of size 100. Weights were initialized from a normal distribution with the standard deviation suggested by Ref. [39]. The learning rate was initialized to 0.0001 and decreased by a factor of 0.9 every epoch. Training was stopped when the

TABLE I. Hyperparameter support for Bayesian optimization of deep neural network architectures. For the no-pileup case, networks with a single hidden layer were allowed to have up to 1000 units per layer, in order to remove the possibility of the deep networks performing better simply because they had more tunable parameters.

Hyperparameter	Range		Optimum	
	Min	Max	No Pileup	Pileup
Hidden units per layer	100	500	425	500
Fully connected layers	1	5	4	5
Locally connected layers	0	5	4	3

TABLE II. Hyperparameter support for BDTs trained on 6 high-level features, and the best combinations in 110 and 140 experiments, respectively, for the no-pileup and pileup tasks. Minimum leaf percent was constrained to be one fourth of the minimum split percent in all cases.

	Range		Optimum	
Hyperparameter	Min	Max	No Pileup	Pileup
Tree depth	15	75	49	49
Learning rate	0.01	1.00	0.07	0.07
Minimum split percent	0.0001	0.1000	0.0021	0.0021

validation error failed to improve or after a maximum of 50 epochs. All computations were performed using Keras [40] and Theano [41,42] on NVidia Titan X processors.

We explore the use of locally connected layers, where each neuron is only connected to a distinct 4-by-4 pixel region of the previous layer. This local connectivity constrains the network to learn spatially localized features in the lower layers without assuming translational invariance, as in convolutional layers where the weights of the receptive fields are shared. Fully connected layers were stacked on top of the locally connected layers to aggregate information from different regions of the detector image. The network architecture—the number of layers of each type, plus the width of the fully connected layers—was optimized using Spearmint. Out of the 25 network architectures explored on the no-pileup task, the best had four locally connected layers followed by four fully connected layers of 425 units. This network has roughly 750,000 tunable parameters, while the best shallow network (one hidden layer of 1000 units) had over 1 million parameters. On the pileup data, 19 different network architectures were tested; the best was again an 8-hidden-layer architecture, with 3 locally connected layers, five fully connected layers, and 500 hidden units in each layer. Convolutional networks were also explored, but as expected, the translational invariance provided by these architectures did not provide any performance boost.

BDTs were trained on the six high-level variables using Scikit-Learn [43]. The maximum depth of each estimator, the minimum number of examples required to constitute an internal node (parametrized as a fraction of the training set), and the learning rate were separately optimized for the data sets with and without pileup using Spearmint (110 and 140 experiments, respectively). The number of estimators was fixed to 500; when evaluating the marginal improvement of performance with the addition of each estimator, we observed that in the best model, performance plateaued after inclusion of less than 100 estimators. This suggests that the number of estimators was not limiting. The minimum number of examples required to form a leaf node was fixed to be one fourth of that required to constitute an internal node. In both cases, the best BDT classifier had a maximum tree depth of 49, a minimum split

TABLE III. Performance results for BDT and deep networks. Shown for each method are both the signal efficiency at background rejection of 10, as well as the AUC, the integral of the background efficiency versus signal efficiency. For the neural networks, we report the mean and standard deviation of three networks trained with different random initializations.

	Performance	
Technique	Signal efficiency at background rejection = 10	AUC
	No pileup	
BDT on derived features	86.5%	95.0%
Deep NN on images	87.8%$_{(0.04\%)}$	95.3%$_{(0.02\%)}$
	With pileup	
BDT on derived features	81.5%	93.2%
Deep NN on images	84.3%$_{(0.02\%)}$	94.0%$_{(0.01\%)}$

requirement of 0.0021, and a learning rate of 0.07. The best BDT trained on the no-pileup data had approximately 700,000 tunable parameters, while the best BDT trained on the pileup data had approximately 750,000.

V. RESULTS

Deep networks with locally connected layers showed the best performance. For example, the best network with 5 hidden layers has two locally connected layers followed by three fully connected layers of 300 units each; this architecture performs better than a network of five fully connected layers of 500 units each.

Final results are shown in Table III. The metric used is the area under the curve (AUC), calculated in signal efficiency versus background efficiency, where a larger AUC indicates better performance. In Fig 4, the signal efficiency is shown versus background rejection, the inverse of background efficiency. In the case without pileup, as studied in Ref. [29], the deep network modestly outperforms the physics domain variables, demonstrating first that successful classification can be performed without expert-designed features and that there is some loss of information in the dimensional reduction such features provide. See the discussion below, however, for comments on the continued importance of expert features.

Our results also demonstrate for the first time that such performance holds up under the more difficult and realistic conditions of many pileup interactions; indeed, the gap between the deep network and the expert variables in this case is more pronounced. This is likely due to the fact that the physics-inspired variables rest on arguments motivated by idealized pictures.

VI. INTERPRETATION

Current typical use in experimental analysis is the combination of the jet mass feature with τ_{21} or one of
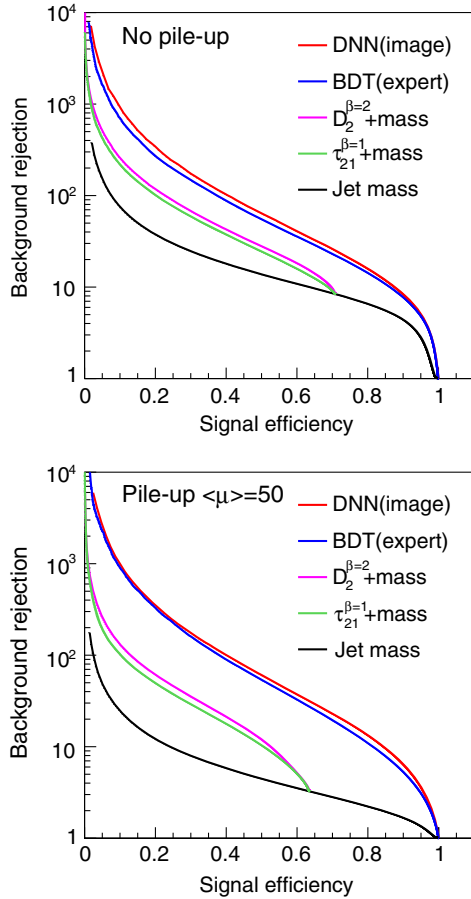
FIG. 4.   Signal efficiency versus background rejection (inverse of efficiency) for deep networks trained on the images and boosted decision trees trained on the expert features, both with (bottom) and without pileup (top). Typical choices of signal efficiency in real applications are in the 0.5–0.7 range. Also shown are the performance of jet mass individually as well as two expert variables in conjunction with a mass window.

FIG. 5.   Distributions in simulated samples without pileup of high-level jet substructure variables for pure signal ($W \to qq$) and pure background (QCD) events. To explore the decision surface of the ML algorithms, also shown are background events with various levels of rejection for deep networks trained on the images and boosted decision trees trained on the expert features. Both algorithms preferentially select jets with values near the peak signal values. Note, however, that while the BDT has been supplied with these features as an input, the DNN has learned this on its own.

the energy correlation variables. Our results show that even a straightforward BDT combination of all six of the high-level variables provides a large boost in comparison. In probing the power of deep learning, we then use as our benchmark this combination of the variables provided by the BDT.

The deep network has clearly managed to match or slightly exceed the performance of a combination of the state-of-the-art expert variables. Physicists working on the underlying theoretical questions may naturally be curious as to whether the deep network has learned a novel strategy for classification which could inform their studies, or rediscovered and further optimized the existing features.

While one cannot probe the motivation of the machine learning (ML) algorithm, it is possible to compare distributions of events categorized as signal-like by the different algorithms in order to understand how the classification is being accomplished. To compare distributions between different algorithms, we study simulated events with
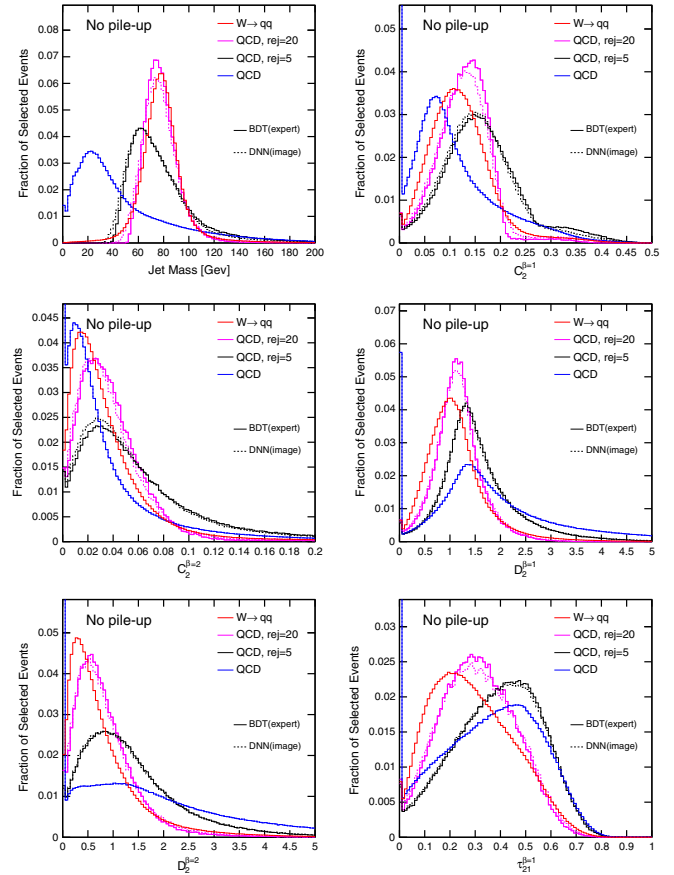
equivalent background rejection, see Figs. 5 and 6 for a comparison of the selected regions in the expert features for the two classifiers. The BDT preferentially selects events with values of the features close to the characteristic signal values and away from background-dominated values. The deep neural network (DNN), which has a modestly higher efficiency for the equivalent rejection, selects events near the same signal values, but in some cases can be seen to retains a slightly higher fraction of jets away from the signal-dominated region. The likely explanation is that the DNN has discovered the same signal-rich region identified by the expert features, but has in addition found avenues to optimize the performance and carve into the background-dominated region. Note that DNNs can also be trained to be independent of mass, by providing a range of mass in training, or training a network explicitly parametrized [44,45] in mass.
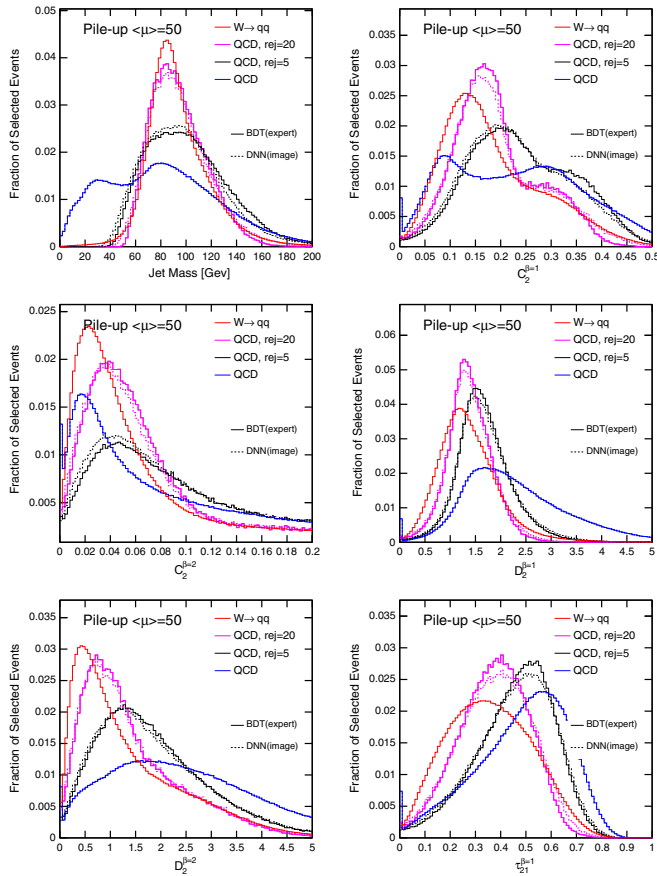
FIG. 6.   Distributions in simulated samples with pileup of high-level jet substructure variables for pure signal ($W \to qq$) and pure background (QCD) events. To explore the decision surface of the ML algorithms, also shown are background events with various levels of rejection for deep networks trained on the images and boosted decision trees trained on the expert features. Both algorithms preferentially select jets with values near the peak signal values. Note, however, that while the BDT has been supplied with these features as an input, the DNN has learned this on its own.

VII. DISCUSSION

The signal from massive $W \to qq$ jets is typically obscured by a background from the copiously produced low-mass jets due to quarks or gluons. Highly efficient classification is critical, and even a small relative improvement in the classification accuracy can lead to a significant boost in the power of the collected data to make statistically significant discoveries. Operating the collider is very expensive, so particle physicists need tools that allow them to make the most of a fixed-size data set. However, improving classifier performance becomes increasingly difficult as the accuracy of the classifier increases.

Physicists have spent significant time and effort designing features for jet-tagging classification tasks. These designed features are theoretically well motivated,

but as their derivation is based on a somewhat idealized description of the task (without detector or pileup effects), they cannot capture the totality of the information contained in the jet image. We report the first studies of the application of deep learning tools to the jet substructure problem to include simulation of detector and pileup effects.

Our experiments support two conclusions. First, that machine learning methods, particularly *deep* learning, can *automatically* extract the knowledge necessary for classification, in principle eliminating the exclusive reliance on expert features. The slight improvement in classification power offered by the deep network compared to the combination of expert features is likely due to the fact that the network has succeeded in discovering small optimizations of the expert features in order to account for the detector and pileup effects present in the simulated samples. This marks another demonstration of the power of deep networks to identify important features in high-dimensional problems. In practice, while deep network classification can boost jet-tagging performance, expert features offer powerful insight [23] into the validity of the simulation models used to train these networks. We do not claim that these results make expert features obsolete. However, it suggests that deep networks can provide similar performance on a variety of related problems where the theoretical tools are not as mature. For example, current tools do not always include information from tracking detectors, nor do they offer performance parametrized [44,45] in the mass of the decaying heavy state.

Second, we conclude that the current set of expert features when used in combination (via BDT or other shallow multivariate approach) appear to capture nearly all of the relevant information in the high-dimensional low-level features describe by the jet image. The power of the networks described here is limited by the accuracy of these models, and expert features may be more robust to variation among the several existing simulation models [46]. In experimental applications, this reliance on simulation can be mitigated by using training samples from real collision data, where the labels are derived using orthogonal information.

Data in high energy physics can often be formulated as images. Thus, these results reported on the representative classification task of single q or g jets versus massive jets from $W \to qq'$ are very likely to apply to a broader set of similar tasks, such as classifying jets with three constituents, as in the case of top quark decay $t \to Wb \to qq'b$, or massive jets from other particles such as Higgs boson decays to bottom quark pairs. Note that in more realistic data sets, calorimeter information often contains depth information as well, such that the images are three-dimensional instead of two; however, this does not represent a difficult extrapolation for the machine

learning algorithms. While the fundamental classification problems are very similar from a machine learning standpoint, the literature of expert features is somewhat less mature, further underlining the potential utility of the reported deep learning methods in these areas.

Future directions of research include studies of the robustness of such networks to systematic uncertainties in the input features and to change in the hadronization and showering model used in the simulated events.

Data sets used in this paper containing millions of simulated collisions are available for download [47].

[1] 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).

[2] D. Adams, A. Arce, L. Asquith, M. Backovic, T. Barillari *et al.*, Towards an understanding of the correlations in jet substructure, arXiv:1504.00679.

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

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

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

[6] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, Stop reconstruction with tagged tops, J. High Energy Phys. 10 (2010) 078.

[7] 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).

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

[9] J. Thaler and K. V. Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.

[10] A. J. Larkoski, G. P. Salam, and J. Thaler, Energy correlation functions for jet substructure, J. High Energy Phys. 06 (2013) 108.

[11] J. Thaler, D. Krohn, and L.-T. Wang, Jet Trimming, J. High Energy Phys. 02 (2010) 084.

[12] C. K. Vermilion, S. D. Ellis, and J. R. Walsh, Recombination algorithms and jet substructure: Pruning as a tool for heavy particle searches, Phys. Rev. D **81**, 094023 (2010).

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

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

[15] M. Dasgupta, A. Powling, and A. Siodmok, On jet substructure methods for signal jets, J. High Energy Phys. 08 (2015) 079.

[16] P. Baldi, P. Sadowski, and D. Whiteson, Searching for exotic particles in high-energy physics with deep learning, Nat. Commun. **5**, 4308 (2014).

[17] P. Baldi, P. Sadowski, and D. Whiteson, Enhanced Higgs Boson to $\tau^+\tau^-$ Search with Deep Learning, Phys. Rev. Lett. **114**, 111801 (2015).

[18] P. Sadowski, J. Collado, D. Whiteson, and P. Baldi, Deep Learning, Dark Knowledge, and Dark Matter, in *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, 2014, pp. 81–87, http://www.jmlr.org/proceedings/papers/v42/sado14.html.

[19] D. E. Soper and M. Spannowsky, Finding physics signals with shower deconstruction, Phys. Rev. D **84**, 074002 (2011).

[20] D. E. Soper and M. Spannowsky, Finding top quarks with shower deconstruction, Phys. Rev. D **87**, 054012 (2013).

[21] I. W. Stewart, F. J. Tackmann, and W. J. Waalewijn, Dissecting Soft Radiation with Factorization, Phys. Rev. Lett. **114**, 092001 (2015).

[22] D. Krohn, J. Thaler, and L.-T. Wang, Jet trimming, J. High Energy Phys. 02 (2010) 084.

[23] A. J. Larkoski, I. Moult, and D. Neill, Analytic boosted boson discrimination, J. High Energy Phys. 05 (2016) 117.

[24] T. Sjöstrand, S. Mrenna, and P. Skands, PYTHIA 6.4 physics and manual, J. High Energy Phys. 05 (2006) 026.

[25] M. Bahr *et al.*, Herwig + + physics and manual, Eur. Phys. J. C **58**, 639 (2008).

[26] J. Cogan, M. Kagan, E. Strauss, and A. Schwarztman, Jet-images: Computer vision inspired techniques for jet tagging, J. High Energy Phys. 02 (2015) 118.

[27] L. G. Almeida, M. Backovic, 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] ATLAS Collaboration, Technical Report No. ATL-PHYS-PUB-2014-004, CERN, Geneva, 2014.

[29] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-Images: Deep Learning Edition, arXiv:1511.05190.

[30] 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, J. High Energy Phys. 07 (2014) 079.

[31] S. Ovyn, X. Rouby, and V. Lemaitre, DELPHES, a framework for fast simulation of a generic collider experiment, arXiv:0903.2225.

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

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

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

[35] J. Thaler and K. V. Tilburg, Maximizing boosted top identification by minimizing N-subjettiness, J. High Energy Phys. 02 (2012) 093.

[36] A. J. Larkoski, I. Moult, and D. Neill, Power counting to better jet observables, J. High Energy Phys. 12 (2014) 009.

[37] J. Snoek, H. Larochelle, and R. P Adams, Practical bayesian optimization of machine learning algorithms, in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., New York, 2012) pp. 2951–2959.

[38] D. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[39] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on Image-Net classification, arXiv:1502.01852.

[40] F. Chollet, Keras, GitHub, 2015, https://github.com/fchollet/keras/.

[41] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, Theano: A CPU and GPU math expression compiler, in Proceedings of the Python for Scientific Computing Conference (SciPy), Austin, TX, June 2010 (unpublished).

[42] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio, Theano: New features and speed improvements, arXiv:1211.5590.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).

[44] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, Parameterized neural networks for high-energy physics, Eur. Phys. J. **C76**, 235 (2016).

[45] K. Cranmer, Approximating Likelihood Ratios with Calibrated Discriminative Classifiers, arXiv:1506.02169.

[46] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, Report No. FERMILAB-PUB-16-046-PPD, 2016, http://inspirehep.net/record/1425540/export/hx.

[47] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, Data for jet substructure for high-energy physics, http://www.igb.uci.edu/ pfbaldi/physics/.