Title

Robust Jet Classifiers through Distance Correlation

Permalink

https://escholarship.org/uc/item/2k41j4x2

Journal

Physical Review Letters, 125(12)

ISSN

0031-9007

Authors

Kasieczka, Gregor
Shih, David

Publication Date

2020-09-18

DOI

10.1103/physrevlett.125.122001

Peer reviewed

Robust Jet Classifiers through Distance Correlation

Gregor Kasieczka[1,*] and David Shih[2,3,4,†]

[1]*Institut für Experimentalphysik, Universität Hamburg, 22761 Hamburg, Germany*
[2]*NHETC, Dept. of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854 USA*
[3]*Theory Group, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[4]*Berkeley Center for Theoretical Physics, University of California, Berkeley, California 94720, USA*

(Received 26 February 2020; accepted 11 August 2020; published 14 September 2020)

While deep learning has proven to be extremely successful at supervised classification tasks at the LHC and beyond, for practical applications, raw classification accuracy is often not the only consideration. One crucial issue is the stability of network predictions, either versus changes of individual features of the input data or against systematic perturbations. We present a new method based on a novel application of "distance correlation," a measure quantifying nonlinear correlations, that achieves equal performance to state-of-the-art adversarial decorrelation networks but is much simpler and more stable to train. To demonstrate the effectiveness of our method, we carefully recast a recent ATLAS study of decorrelation methods as applied to boosted, hadronic W tagging. We also show the feasibility of regularization with distance correlation for more powerful convolutional neural networks, as well as for the problem of hadronic top tagging.

DOI: 10.1103/PhysRevLett.125.122001

Introduction.—Recent breakthroughs in deep learning have begun to revolutionize many areas of high energy physics. One area that has received considerable focus is the problem of classifying different types of jets at the LHC. Deep neural networks have been applied, for example, to distinguishing top quarks from light quark and gluon jets. For this problem, a large number of architectures based on fully connected neural networks [1,2], image-based methods [3,4], recursive clustering [5,6], physics variables [7–10], sets [11], and graphs [12,13] have been studied [14–16]. Related challenges of identifying vector bosons [17,18], b-quarks [19,20], and Higgs bosons [13,21] and of distinguishing light quark from gluon jets [22–25] have seen similar progress. Beyond classifying single particles in an event, there is also work on developing holistic methods that classify full events according to the likely physics process that produced them [26,27]. Finally, some of these novel deep learning methods are beginning to be applied to concrete experimental analyses (see, e.g., [28–30]).

So far, the recent activity in developing better jet classifiers with deep learning has focused on maximizing their raw performance. However, the most accurate classifier is often not the best one for actual experimental applications. Instead, what is often desired is the most accurate classifier *given the constraint that it is decorrelated with one or more auxiliary variables*.

The underlying reason for this requirement is that classifiers are trained on Monte Carlo (MC) simulated examples (for which perfect truth labels are available) but are applied to (unlabeled) collision data. While the simulated events are of high fidelity, they do not perfectly reproduce the real data, and this gives rise to systematic differences between training and testing data. Understanding and mitigating these systematic differences is essential in any experimental analysis, and having a decorrelated classifier has many applications in this regard. For example, if the sources of systematic uncertainty are known, one can attempt to explicitly decorrelate a classifier against them in order to reduce or eliminate their effects [31–34]. Or, one can attempt to control for these systematic differences using data-driven methods such as sidebanding in the invariant mass (Although different auxiliary variables can be used in experimental analyses, one of the most common choices is invariant mass. So for concreteness, and without loss of generality, we will focus on the case of invariant mass for the remainder of this paper). If the signal is localized but the background is smooth in mass, the sideband method allows one to calculate MC vs data correction factors, define control samples, and estimate backgrounds. But if the classifier sculpts features (e.g., bumps) into the background mass distribution, it cannot be relied on for sidebanding. A classifier that is decorrelated with mass is sufficient (although not necessary) to guarantee smoothness of the background mass distribution.

The issue is especially acute for powerful multivariate classifiers such as neural networks, which will have a strong incentive to "learn the mass" when building the optimal discriminant. Even if one excludes mass from the list of inputs to the machine learning algorithm, it may not be enough to achieve a decorrelated classifier—many of the other inputs may be correlated with mass, and machine learning methods in general are flexible enough to exploit correlations of inputs. Such improvements will be especially relevant for (but not limited to) searches for new resonances with unknown mass. The identification of resonances in invariant mass distributions is historically the main avenue to discovery in experimental particle physics and relies on robust background estimates. Therefore, an important and significant challenge is to design classifiers that are as fully decorrelated from mass as possible while using maximal information.

In this Letter, we will present a new method for training decorrelated classifiers that achieves performance comparable to state-of-the-art methods while being much easier to train. The key observation is that a statistical measure called "distance correlation" (DisCo) [35–38] is sensitive to general, nonlinear correlations between two random variables and can be efficiently computed from finite samples. DisCo is well-known in statistics and has been applied to various fields, including data science [39] and biology [40]. To our knowledge, this is the first application of DisCo to particle physics.

By including DisCo as an additive regularizer term in the loss function, we demonstrate that we can achieve a state-of-the art decorrelated classifier with just one additional hyperparameter (the coefficient of the DisCo regularizer). By varying this coefficient, we can control the tradeoff between classification performance and decorrelation, interpolating between a fully decorrelated tagger and a fully performant one.

To validate our methods and rigorously demonstrate that they are state of the art, we will carefully reproduce the results of a recent ATLAS study of decorrelated taggers for identifying boosted W bosons [41]. This study includes a comprehensive set of decorrelation methods, including [31,42–44]. The most promising technique so far (in terms of achieving the highest classifier performance for a given level of decorrelation) has been adversarially training a pair of neural networks: a classifier distinguishing different classes and an adversary predicting the mass [31,44] for a given classifier output.

The downside of the adversarial method has been that it is extremely difficult to implement in practice. Not only does one have to essentially train two separate neural networks, each with its own set of hyperparameters, but one has to carefully tune these two neural networks against each other. This stems from the nature of adversarial training: the objective is not to minimize a loss function but rather to find a saddle point where the classifier loss is minimized but the adversary loss is maximized. Without careful tuning of learning rate schedules, number of epochs, minibatch sizes, etc., the training easily becomes unstable (since the loss is unbounded from below) and can quickly run away to a meaningless result.

By contrast, DisCo regularization maintains the convex objective of the original loss function (i.e., the DisCo term is a positive measure of nonlinear correlations), making it much more stable to train. And since it only has one additional hyperparameter, no additional tuning is required. We will show, in the context of the ATLAS W-tagging study, that the result of DisCo decorrelation is comparable to that of adversarial decorrelation. In the Supplemental Material [45], we will also demonstrate the state-of-the-art performance for top tagging with jet images and convolutional neural networks (CNNs).

Distance correlation.—Given a sample of paired vectors $(\vec{x}_i, \vec{y}_i)$ (where the index i runs over the sample) drawn randomly from some distribution, we would like a function that measures the extent to which they are drawn from *independent* distributions, i.e., the extent to which $P_{\mathrm{joint}}(\vec{X}, \vec{Y}) = P_X(\vec{X})P_Y(\vec{Y})$. In order for this function to be applicable in a deep learning context, we also require that this function be differentiable and that it can be computed directly from the sample.

In our case, the vectors are one dimensional and correspond to mass $X = m$ and classifier output $Y = y$ but clearly one can imagine many more applications of such a measure at the LHC and beyond.

The usual Pearson correlation coefficient R only measures linear dependencies, so it is not suitable for our purposes. Specifically, features can have nonlinear dependencies and still exhibit zero Pearson R (While the Pearson correlation coefficient is nonzero only if features are correlated, it can, however, be used to actively *correlate* features (see, e.g., [53]). There are many information-theoretic measures of similarity of distributions such as KL divergence, Jensen-Shannon distance, and mutual information. These are difficult to compute directly from the sample without binning. One can approximate these measures by training a classifier and using the likelihood ratio trick, but this again leads to adversarial methods (see, e.g., [33,54–57]).

One measure that seems to fit the bill perfectly is "distance correlation", which originated in the works of [35–38]. It can be computed from the sample, and it has the key property of being zero iff X and Y are independent.

The definition of distance covariance is as follows:

$$\mathrm{dCov}^2(X, Y) = \int d^p s d^q t |f_{X,Y}(s,t) - f_X(s) f_Y(t)|^2 w(s,t), \tag{1}$$

where $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, f_X and f_Y are the characteristic functions for the random variables X and Y, and $f_{X,Y}$ is the joint characteristic function for X and Y. Finally,

$$w(s,t) \propto |s|^{-(p+1)} |t|^{-(q+1)} \tag{2}$$

is a weight function that is uniquely determined up to an overall normalization by the requirement that dCov is invariant under constant shifts and orthogonal transformations and equivariant under scale transformations [58]. Since $f_{X,Y} = f_X f_Y$ iff X and Y are independent random variables, Eq. (1) makes it clear that distance covariance is a measure of the independence of X and Y that is zero iff X and Y are independent.

Using the definition of the characteristic function it is straightforward to verify that we can also express dCov as

$$\mathrm{dCov}^2(X,Y) = \langle|X-X'||Y-Y'|\rangle + \langle|X-X'|\rangle\langle|Y-Y'|\rangle - 2\langle|X-X'||Y-Y''|\rangle \tag{3}$$

where $|\cdot|$ refers to the Euclidean vector norm (In fact, there is a family of distance covariance measures parameterized by $0 < \alpha < 2$ where one uses $|X-X'|^\alpha$ instead of $|X-X'|$. These relax the requirement of strict equivariance under rescalings. In this Letter, we will focus on $\alpha = 1$, but in principle this would be another hyperparameter to explore) and (X,Y), (X',Y'), (X'',Y'') are independently identically distributed from the joint distribution of (X,Y) [X'' is not used in Eq. (3)]. Using this alternative form of dCov^2, it is straightforward to compute a sampling estimate of dCov^2 from a dataset of (x_i, y_i) (In the following we will be reweighting by p_T. So we actually need a *weighted* form of distance correlation. That follows easily from the sample Eq. (3)).

Finally, we normalize the distance covariance by the individual distance variances to obtain *distance correlation*:

$$\mathrm{dCorr}^2(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\mathrm{dCov}(X,X)\mathrm{dCov}(Y,Y)} \tag{4}$$

The distance correlation is bounded between 0 and 1. Normalizing ensures equally strong decorrelation independent of the overall scale.

We will add dCorr^2 as a regularizer term to the usual classifier loss function in the following (In principle, another hyperparameter is the exact power of dCorr that one adds to the loss function. We have not explored this in much detail). In detail,

$$L = L_{\mathrm{classifier}}(\vec{y}, \vec{y}_{\mathrm{true}}) + \lambda \mathrm{dCorr}^2_{y_{\mathrm{true}}=0}(\vec{m}, \vec{y}), \tag{5}$$

where λ is a single hyperparameter that controls the tradeoff between classifier performance and decorrelation, $\vec{y}$ is the output of the neural network on a single minibatch, and $\vec{y}_{\mathrm{true}}$ and $\vec{m}$ are the true labels and masses, respectively (Our implementation of DisCo is available at [59]). The subscript $y_{\mathrm{true}} = 0$ indicates that the distance correlation is only calculated for the subset of the minibatch that is background; this is the appropriate mode for W tagging. Of course, for other applications it may be more appropriate to apply the decorrelation to all events or even to signal events only.

Sample.—As discussed in the Introduction, we will focus in this paper on W tagging, for which there is a detailed study of existing decorrelation methods by the ATLAS collaboration [41]. (See the Supplemental Material [45] for a brief demonstration of DisCo decorrelation for top tagging.) By recasting the ATLAS study as closely as possible, we will be able to validate our methods and rigorously demonstrate that our method of distance correlation is state of the art.

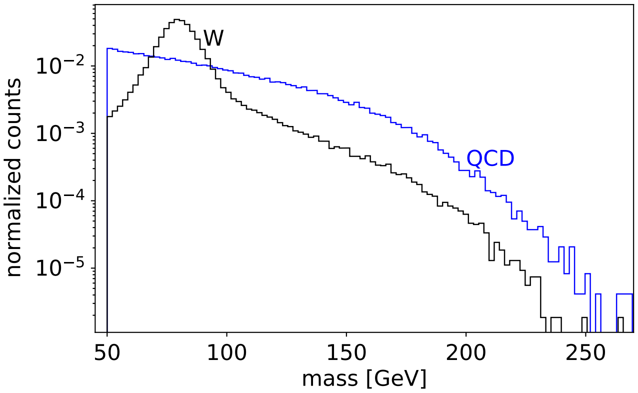

FIG. 1. Invariant mass distribution for the inclusive W and QCD samples.

Following the ATLAS study, we generate the standard model processes $pp \to WW$ and $pp \to jj$ in PYTHIA8.219 [60] at $\sqrt{s} = 13$ TeV with a generator level cut of $p_T >$ 250 GeV on the initial particles. We use DELPHES3.4.1 with the default detector card for detector simulation [61]. We also use the built-in functionality of DELPHES to simulate pileup with $\langle N_{PU}\rangle = 24$ as per the ATLAS study [41].

Jets are reconstructed using FASTJET3.0.1 [62] and the anti-k_T algorithm [63] with $R = 1$ distance parameter. Jets are required to have $|\eta| < 2$ and to be within $\Delta R < 0.75$ or the original parton. The daughters of the W are also required to be within $\Delta R < 0.75$ of the original W. Finally, jets are trimmed [64] with parameters $R_{\mathrm{sub}} = 0.2$ and $f_{\mathrm{cut}} = 5\%$. For the final sample, jets are required to have $m \in [50, 300]$ GeV and $p_T \in [300, 400]$ GeV; the mass distributions for signal and background are shown in Fig. 1. Apart from the very last requirement on p_T, these are all following the ATLAS study. Here we choose to focus on a more narrow range in p_T for simplicity.

From this sample of jets, we compute the complete list of high-level kinematic variables shown in Table 1 of the ATLAS study (see [41] for more details and original references). These form the inputs for all the methods in the ATLAS study. We will also use them as inputs for the dense neural network (DNN) plus distance correlation.

Since we will also study the decorrelation of CNN classifiers (see below), we will also form jet images in the same way as [65]. We form images with $\Delta\eta = \Delta\phi = 2$ and 40×40 pixel resolution. For simplicity, we stick to gray-scale images (with pixel intensity equal to p_T) for this study. Figure 2 shows the average of 100 000 W and QCD jet images.

For all methods, we reweight the training samples so that the p_T distributions of signal and background are flat,

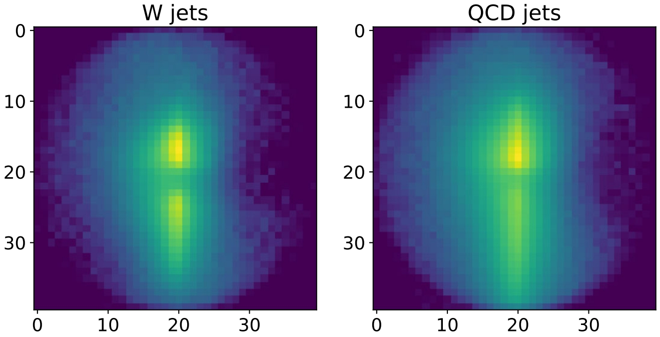

FIG. 2. Average of 100 000 jet images for W jets (left) and QCD jets (right).

following the ATLAS study. We use 50 evenly spaced p_T bins between 300 and 400 GeV. For evaluation, ATLAS also reweights the signal p_T distribution to look like background. But since we are taking such a narrow p_T slice, our p_T distributions are basically identical, so we skip this step.

All of the data samples used for this study will be made publicly available here [66].

Methods.—Following [41], we measure the tagging performance by the rejection factor R_{50} corresponding to the inverse of the false positive rate (the probability of misidentifying a QCD jet as a W jet) at a true positive rate (the probability of correctly identifying a W jet) of 50%. The decorrelation is quantified by the inverse of the Jensen-Shannon divergence (JSD) $1/\mathrm{JSD}_{50}$ between the inclusive background distribution and the background distribution passing the selection corresponding to a true positive rate of 50%. The JSD is calculated from histograms with 50 bins between lowest and highest value. The binned entropy is measured in *bits*.

We have implemented the following pairs of (W tagging, decorrelation) methods in our work. From the ATLAS study: [τ_{21}, designed decorrelated tagger (DDT)] [42,67], [D_2, k-Nearest Neighbors regression (kNN)] [68–70], [Adaboost boosted decision tree (BDT), uBoost] [71], and (DNN, adversary) [31]. We will additionally include the simplest and possibly oldest decorrelation method, namely "planing", or reweighting events so that the mass histograms of signal and background are identical. As this approach is relatively simple to implement and does not add much computational cost, it is a good baseline procedure (See [72] for a recent comparison study of planing against other methods.). Finally, to all of this we will add our new method (DNN, DisCo regularization) for comparison. For details on all these methods, see the Supplemental Material [45].

In addition, we will go beyond the ATLAS study and examine a CNN classifier acting on jet images, together with adversarial and DisCo decorrelation. This will demonstrate that DisCo regularization is effective enough to decorrelate more powerful deep learning classifiers that use low-level, high-dimensional features. For the CNN classifier, we use a scaled down version of the classifier in [65]. There are four convolutional layers with 64, 32, 32, 32 filters (size 4 × 4) with 2 × 2 Max pooling after the second and fourth layer. This is followed by three hidden layers with 32, 64, and 64 nodes. All activations are rectified linear units. Finally, we output to SOFTMAX.

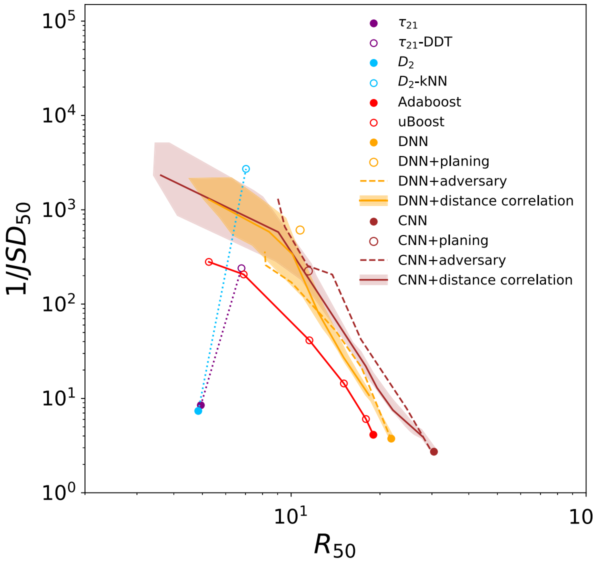

FIG. 3. Decorrelation against background rejection for different approaches.

For both CNN and DNN with DisCo regularization, we used the Adam optimizer with minibatch size of 2048 and a fixed learning rate of 10^{-4}. We found that the relatively large batch size of 2048 helped with the numerical stability of the DisCo regularizer. We note that the sampling estimate (3) for distance covariance is known to be statistically biased, and an unbiased estimator was given in [73]. The bias goes to zero as $\sim 1/n$ where n is the size of the sample (the minibatch size in our case). We have verified that, as our minibatch size is sufficiently large, there is no practical benefit to using the unbiased estimate of distance covariance in our case.

For the DNN (CNN), we performed a scan in DisCo parameter λ in the range 0–600 (0–250). All classifiers were trained for 200 epochs; no early stopping was used. We have checked that 200 epochs is enough to ensure convergence in the sense that training for more epochs does not improve things. Then, for each λ and training instance, the model with the best validation loss is selected. This procedure is repeated six times with different random seeds to obtain a sense of the variability in the training outcomes.

In all of the machine learning-based methods we use 250 000/80 000/80 000 signal jets and 110 000/330 000/770 000 background jets for training/validation/testing. We use so many background jets in order to minimize the statistical error on the JSD calculation (which is calculated only for the background).

The deep learning algorithms were implemented with PyTorch and trained on an NVIDIA P100GPU.

Results.—Our final result is shown in Fig. 3, where the performance of various decorrelation methods on the test

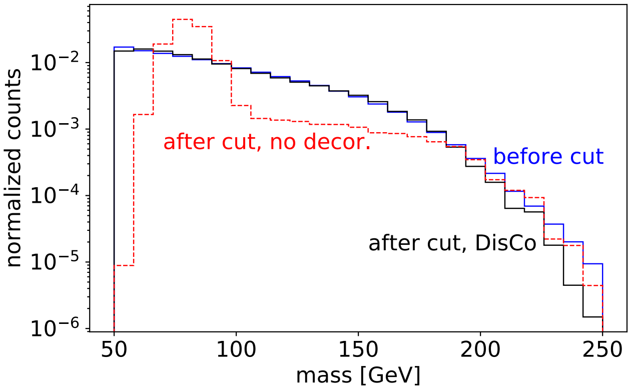

FIG. 4. QCD mass distribution before and after a cut on CNN plus DisCo (W tagging) with signal efficiency of 50% and JSD $\sim 10^{-3}$.

set is summarized in the plane of $1/\mathrm{JSD}_{50}$ (which measures decorrelation) vs R_{50} (which measures classifier performance). For DNN + DisCo and CNN + DisCo, the envelopes of the six independent trainings per λ are shown, together with lines connecting the median-decorrelated points for different values of rejection. For the other machine learning methods, a representative result is shown per decorrelation parameter. (We have checked that the envelopes for DNN + adversary and CNN + adversary are comparable to their DisCo counterparts).

The qualitative (and even quantitative) agreement with Fig. 11(a) of [41] is excellent, and we see a clear tradeoff between classifier performance and the amount of decorrelation.

Comparing DNN + DisCo to the other methods, we find that it has comparable performance to DNN + adversary. Meanwhile, DNN + DisCo is much easier to train—whereas DisCo adds exactly one hyperparameter and no additional neural network parameters to the DNN, the adversary more than doubles the number of hyperparameters and adds an entire second neural network to the story. See the Supplemental Material [45] for a complete list of hyperparameters for the adversarial training. These were found through manual tuning and their sheer complexity nicely illustrates the need for a simpler method of decorrelation.

We see that DisCo regularization is equally capable of decorrelating the more powerful CNN classifier and again achieves comparable performance to CNN + adversary. One concern could have been that a more powerful deep learning method such as the CNN could overpower the DisCo regularizer, but our result demonstrates that this is not the case. At the highest levels of decorrelation, we note that both DNN and CNN performances are comparable.

In Fig. 4, we indicate more directly the level of decorrelation in the background mass distribution for the pure CNN case (no decorrelation) and for the CNN + DisCo method at a working point that achieves $1/\mathrm{JSD}_{50} \sim 10^3$. We see that DisCo is quite effective at stabilizing the background mass distribution against a cut on the classifier.

Finally, let us also comment briefly on the performance of planing. Unlike DisCo regularization and some of the other methods studied here, planing yields a single working point instead of a tunable tradeoff between decorrelation and classifier performance. Since its performance depends on the joint probability distribution for mass and the other observables (Planing replaces $p(x,m)$ with $p(x,m)/p(m)$, which does not guarantee independence), planing is not guaranteed to achieve strong results. But it is interesting to see that in this case (and in many of the cases studied in [72]), planing the DNN and CNN classifiers achieves very good performance. The performance lies on the DisCo regularization curve, and DisCo is capable of further decorrelation.

Conclusion.—Deep learning is greatly increasing the classification performance for a wide number of reconstruction problems in particle physics. With the increasing adoption of these powerful machine learning solutions, a thorough understanding of their stability is needed.

In this Letter, it was shown how a simple regularization term based on the distance correlation metric can achieve state-of-the-art decorrelation power. Training is easier to set up, has far fewer hyperparameters to optimize, and is more stable than adversarial networks, while simultaneously being more powerful than simpler approaches.

DisCo regularization is an effective and promising new method for decorrelation that should have a host of immediate experimental applications at the LHC. At the same time, the potential use cases are much wider and include problems of fairness and bias of decision algorithms in social applications. This will be an extremely interesting direction for future exploration.

We thank Joern Bach, Ben Nachman, Bryan Ostdiek, Tilman Plehn, Matt Schwartz, and Mike Williams for helpful discussions. We are grateful to Chris Delitzsch, Steven Schramm, and especially Andreas Sogaard for help with details of the ATLAS decorrelation study. G. K. is supported by the Deutsche Forschungsgemeinschaft under Germany‘s Excellence Strategy—EXC 2121 “Quantum Universe—390833306.” G. K. is grateful for the generous support and hospitality of the Rutgers NHETC Visitor Program where this work was initiated. D. S. is supported by DOE Grant No. DOE-SC0010008 and by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under the Contract No. DE-AC02-05CH11231. D. S. thanks LBNL, BCTP, and BCCP for their generous support and hospitality during his sabbatical year.

[*]Corresponding author.
gregor.kasieczka@uni-hamburg.de
[†]Corresponding author.
dshih@physics.rutgers.edu

[1] L. G. Almeida, M. Backović, 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.

[2] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, Jet constituents for deep neural network based top quark tagging, arXiv:1704.02124.

[3] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, Deep-learning top taggers or the end of QCD?, J. High Energy Phys. 05 (2017) 006.

[4] S. Macaluso and D. Shih, Pulling out all the tops with computer vision and deep learning, J. High Energy Phys. 10 (2018) 121.

[5] G. Louppe, K. Cho, C. Becot, and K. Cranmer, QCD-aware recursive neural networks for jet physics, J. High Energy Phys. 01 (2019) 057.

[6] S. Egan, W. Fedorko, A. Lister, J. Pearkes, and C. Gay, Long short-term memory (LSTM) networks with jet constituents for boosted top tagging at the LHC, arXiv:1711.09059.

[7] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, Deep-learned top tagging with a Lorentz layer, SciPost Phys. **5**, 028 (2018).

[8] M. Erdmann, E. Geiser, Y. Rath, and M. Rieger, Lorentz boost networks: Autonomous physics-inspired feature engineering, J. Instrum. **14**, P06006 (2019).

[9] L. Moore, K. Nordström, S. Varma, and M. Fairbairn, Reports of my demise are greatly exaggerated: N-subjettiness taggers take on jet images, SciPost Phys. **7**, 036 (2019).

[10] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, Uncovering latent jet substructure, Phys. Rev. D **100**, 056002 (2019).

[11] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, J. High Energy Phys. 01 (2019) 121.

[12] H. Qu and L. Gouskos, ParticleNet: Jet tagging via particle clouds, Phys. Rev. D **101**, 056019 (2020).

[13] E. A. Moreno, T. Q. Nguyen, J.-R. Vlimant, O. Cerri, H. B. Newman, A. Periwal, M. Spiropulu, J. M. Duarte, and M. Pierini, Interaction networks for the identification of boosted $H \to b\bar{b}$ decays, Phys. Rev. D **102**, 012010 (2020).

[14] CMS Collaboration, Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment, Report No. CMS-PAS-JME-18-002, CERN, 2019.

[15] ATLAS Collaboration, Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC, Eur. Phys. J. C **79**, 375 (2019).

[16] G. Kasieczka *et al.*, The machine learning landscape of top taggers, SciPost Phys. **7**, 014 (2019).

[17] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-images—deep learning edition, J. High Energy Phys. 07 (2016) 069.

[18] Y.-C. J. Chen, C.-W. Chiang, G. Cottin, and D. Shih, Boosted W/Z tagging with jet charge and deep learning, Phys. Rev. D **101**, 053001 (2020).

[19] ATLAS Collaboration, Identification of jets containing b-hadrons with recurrent neural networks at the ATLAS experiment, CERN, Geneva, Technical Report no. ATL-PHYS-PUB-2017-003, 2017.

[20] CMS Collaboration, Performance of the deepjet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13 TeV with phase 1 CMS detector, Report No. CMS-DP-2018-058, CERN, 2018.

[21] J. Lin, M. Freytsis, I. Moult, and B. Nachman, Boosting $H \to b\bar{b}$ with machine learning, J. High Energy Phys. 10 (2018) 101.

[22] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, Deep learning in color: towards automated quark/gluon jet discrimination, J. High Energy Phys. 01 (2017) 110.

[23] G. Kasieczka, N. Kiefer, T. Plehn, and J. M. Thompson, Quark-gluon tagging: Machine learning vs detector, SciPost Phys. **6**, 069 (2019).

[24] H. Luo, M.-x. Luo, K. Wang, T. Xu, and G. Zhu, Quark jet versus gluon jet: Fully-connected neural networks with high-level features, Sci. China Phys. Mech. Astron. **62**, 991011 (2019).

[25] K. Fraser and M. D. Schwartz, Jet charge and machine learning, J. High Energy Phys. 10 (2018) 093.

[26] M. Erdmann, B. Fischer, and M. Rieger, Jet-parton assignment in $t\bar{t}$H events using deep learning, J. Instrum. **12**, P08020 (2017).

[27] S. Diefenbacher, H. Frost, G. Kasieczka, T. Plehn, and J. M. Thompson, CapsNets continuing the convolutional quest, SciPost Phys. **8**, 023 (2020).

[28] M. Aaboud *et al.* (ATLAS Collaboration), Search for pair production of heavy vector-like quarks decaying into hadronic final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Rev. D **98**, 092005 (2018).

[29] A. M. Sirunyan *et al.* (CMS Collaboration), Search for pair production of vectorlike quarks in the fully hadronic final state, Phys. Rev. D **100**, 072001 (2019).

[30] CMS Collaboration, Search for direct top squark pair production in events with one lepton, jets, and missing transverse momentum at 13 TeV with the CMS experiment, J. High Energy Phys. 05 (2020) 032.

[31] G. Louppe, M. Kagan, and K. Cranmer, Learning to pivot with adversarial networks, arXiv:1611.01046.

[32] C. Englert, P. Galler, P. Harris, and M. Spannowsky, Machine learning uncertainties with adversarial neural networks, Eur. Phys. J. C **79**, 4 (2019).

[33] P. Windischhofer, M. Zgubić, and D. Bortoletto, Preserving physically important variables in optimal event selections: A case study in Higgs physics, J. High Energy Phys. 07 (2020) 001.

[34] S. Wunsch, S. Jörger, R. Wolf, and G. Quast, Reducing the dependence of the neural network function to systematic uncertainties in the input space, Comput. Softw. Big Sci. **4**, 5 (2020).

[35] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Stat. **35**, 2769 (2007).

[36] G. J. Székely and M. L. Rizzo, Brownian distance covariance, Ann. Appl. Stat. **3**, 1236 (2009).

[37] G. J. Székely and M. L. Rizzo, The distance correlation t-test of independence in high dimension, J. Multivariate Anal. **117**, 193 (2013).

[38] G. J. Székely and M. L. Rizzo, Partial distance correlation with methods for dissimilarities, Ann. Stat. **42**, 2382 (2014).

[39] R. Li, W. Zhong, and L. Zhu, Feature screening via distance correlation learning, J. Am. Stat. Assoc. **107**, 1129 (2012).

[40] A. Villaverde and J. Banga, Reverse engineering and identification in systems biology: Strategies, perspectives and challenges, J. R. Soc. Interface **11**, 20130505 (2014).

[41] ATLAS Collaboration, Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS, Report No. ATL-PHYS-PUB-2018-014, CERN, 2018.

[42] 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.

[43] I. Moult, B. Nachman, and D. Neill, Convolved substructure: Analytically decorrelating jet substructure observables, J. High Energy Phys. 05 (2018) 002.

[44] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Søgaard, Decorrelated jet substructure tagging using adversarial neural networks, Phys. Rev. D **96**, 074034 (2017).

[45] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.125.122001 for more details on the decorrelation methods implemented in this Letter, and for a comparison of CNN + DisCo vs CNN + adversary for top tagging, which includes Refs. [46–52].

[46] CMS Collaboration, Top tagging with new approaches, Report No. CMS-PAS-JME-15-002, CERN, 2016.

[47] 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).

[48] ATLAS Collaboration, Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector, J. High Energy Phys. 06 (2016) 093.

[49] S. Chang, T. Cohen, and B. Ostdiek, What is the machine learning?, Phys. Rev. D **97**, 056009 (2018).

[50] ATLAS Collaboration, Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS, CERN, Geneva, Technicl Report no. ATL-PHYS-PUB-2018-014, 2018.

[51] A. Rogozhnikov *et al.*, hep_ml: Machine learning for high energy physics, version 0.6, https://github.com/arogozhnikov/hep_ml.

[52] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv:1502.03167.

[53] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, Correlational neural networks, arXiv:1504.07225.

[54] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, Mine: Mutual information neural estimation, arXiv:1801.04062.

[55] S. Nowozin, B. Cseke, and R. Tomioka, f-gan: Training generative neural samplers using variational divergence minimization, arXiv:1606.00709.

[56] S. Mohamed and B. Lakshminarayanan, Learning in implicit generative models, arXiv:1610.03483.

[57] K. Cranmer, J. Pavez, and G. Louppe, Approximating likelihood Ratios with calibrated discriminative classifiers, arXiv:1506.02169.

[58] G. J. Székely and M. L. Rizzo, On the uniqueness of distance covariance, Stat. Probab. Lett. **82**, 2278 (2012).

[59] See https://github.com/gkasieczka/DisCo.

[60] T. Sjostrand, S. Mrenna, and P. Z. Skands, A brief introduction to PYTHIA8.1, Comput. Phys. Commun. **178**, 852 (2008).

[61] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.

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

[63] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.

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

[65] S. Macaluso and D. Shih, Pulling out all the tops with computer vision and deep learning, J. High Energy Phys. 10 (2018) 121.

[66] G. Kasieczka and D. Shih, Datasets for boosted w tagging (2020) https://doi.org/10.5281/zenodo.3606767.

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

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

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

[70] S. A. Dudani, The distance-weighted k-nearest-neighbor rule, IEEE Trans. Syst. Man, Cybernet. **SMC-6**, 325 (1976).

[71] J. Stevens and M. Williams, uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers, J. Instrum. **8**, P12013 (2013).

[72] L. Bradshaw, R. K. Mishra, A. Mitridate, and B. Ostdiek, Mass agnostic jet taggers, SciPost Phys. **8**, 011 (2020).

[73] G. J. Szekely and M. L. Rizzo, Partial distance correlation with methods for dissimilarities, arXiv:1310.2926.