Decorrelated jet substructure tagging using adversarial neural networks

Chase Shimmin, Peter Sadowski, Pierre Baldi, Edison Weik, Daniel Whiteson, Edward Goul, and Andreas Søgaard

Phys. Rev. D **96**, 074034 — Published 30 October 2017

DOI: 10.1103/PhysRevD.96.074034

Decorrelated Jet Substructure Tagging using Adversarial Neural Networks

Chase Shimmin
Department of Physics and Astronomy, UC Irvine, Irvine, CA 92627 and
Department of Physics, Yale University, New Haven, CT

Peter Sadowski and Pierre Baldi
Department of Computer Science, UC Irvine, Irvine, CA 92627

Edison Weik and Daniel Whiteson
Department of Physics and Astronomy, UC Irvine, Irvine, CA 92627

Edward Goul
Department of Physics, MIT, Cambridge, MA 02139

Andreas Søgaard
Department of Physics and Astronomy, University of Edinburgh, Edinburgh UK

We describe a strategy for constructing a neural network jet substructure tagger which powerfully discriminates boosted decay signals while remaining largely uncorrelated with the jet mass. This reduces the impact of systematic uncertainties in background modeling while enhancing signal purity, resulting in improved discovery significance relative to existing taggers. The network is trained using an adversarial strategy, resulting in a tagger that learns to balance classification accuracy with decorrelation. As a benchmark scenario, we consider the case where large-radius jets originating from a boosted resonance decay are discriminated from a background of nonresonant quark and gluon jets. We show that in the presence of systematic uncertainties on the background rate, our adversarially-trained, decorrelated tagger considerably outperforms a conventionally trained neural network, despite having a slightly worse signal-background separation power. We generalize the adversarial training technique to include a parametric dependence on the signal hypothesis, training a single network that provides optimized, interpolatable decorrelated jet tagging across a continuous range of hypothetical resonance masses, after training on discrete choices of the signal mass.

I. INTRODUCTION

The enormous center-of-mass energy of the Large Hadron Collider (LHC) enables the production of particles at such extreme velocities that the decay products of even massive particles can become collimated. Rather than producing distinct deposits of energy in the calorimeter, hadronic decay products of such boosted objects can overlap, creating a single large jet. Distinguishing between jets originating from a single particle (such as a quark or gluon), and those which contain two or three hadronic decay products, is known as jet tagging, and has become an essential component of searches for new physics at the LHC [1–5].

However, optimizing the LHC discovery potential requires balancing the competing constraints of signal discrimination and systematic uncertainties. We consider the case posed in Ref. [6] in which a spectrum of jet masses is examined for the presence of a signal-like resonance peak. The background is dominated by QCD jets, while the hypothetical signal is produced via the hadronic decay of a boosted resonance.

On one hand, there has been intense theoretical work to develop jet substructure tagging tools [7, 8] with powerful discrimination between these types of jets. On the other hand, the processes that produce backgrounds to these searches are often not well understood or are poorly modeled by simulation tools. As a result, experiments in practice rely on the assumption of a smooth background spectrum in jet mass which can be interpolated under a signal peak from observed sidebands in data. This allows the background to be estimated without incurring large systematic effects that would be difficult to control due to the limited understanding of the background processes. The existence of well-developed techniques designed to search for a localized signal over a smooth continuum background gives the jet mass a special importance as an observable; however, these techniques are only effective to the extent that the background may be well described by a simple functional form. Unfortunately, the jet-

tagging discrimination quantities may be correlated with jet mass, resulting in a distortion of the background shape [9] when used in the analysis selection. Hence, the desire for optimal discrimination and reduced sensitivity to systematic uncertainties in general (and jet mass interpolation in this particular example) are naturally at tension with each other.

One solution, Designing Decorrelated Taggers (DDT) [9], uses a simple parametric function to construct a modified version of one tagging variable (e.g. τ_{21}), adjusted specifically to avoid distorting the mass spectrum. This has been shown [10] to effectively balance the issues of discrimination and systematic uncertainty for the quantity τ_{21}.

However, a multivariate classifier (such as a neural network) utilizing the full suite of tagging variables will have considerably greater discrimination power than any individual variable, or pair of variables [11]. In principle, the DDT approach could be generalized to handle multiple variables, or even the output of a machine-learning-based combination of these variables, but the more complex and non-linear response will require increasingly complex and non-linear corrections.

In this paper, we incorporate the decorrelation requirement directly into the machine learning strategy by modifying the learning rule to include a constraint which attempts to penalize solutions that distort the background mass spectrum. The training strategy is adversarial [12–15], in which a pair of networks, a classifier and an adversary, are trained simultaneously with different objectives. The classifier is trained in the traditional manner to maximize classification accuracy. As proposed by Ref. [16], the adversary is trained to infer the value of one of the classifier inputs from the classifer response. In this scheme, the two networks together perform a constrained optimization which maximizes classification accuracy while minimizing the dependence of the classifier response on the selected input. Here, one network performs jet substruture classification, while the adversary attempts to infer the jet mass solely from the classifier response.

Lastly, we generalize the adversarial decorrelation technique to include the case where both the classifier and its adversary are parameterized by some external quantity, such as a theoretical hypothesis for the mass of a new particle or a field coupling strength. This is motivated by the fact that resonance searches, such as the one described here, are often performed as scan over a range of potential masses. Generally the optimal classifier for each hypothesis will differ. However, the signal simulations used for training can usually only be sampled for a small number of hypotheses values due to the computational expense of producing them.

Networks parameterized in this way [17, 18] can interpolate to provide optimal classification for hypotheses which were not included in the training, allowing sensitivity to be evaluated without generating simulations at those points. We show that a single adversarially-trained classifier, parameterized in the hypothesis signal mass, remains decorrelated over the range of values upon which it is trained.

II. BENCHMARK DATA

Simulated samples are used to model the kinematics of the signal and background processes. As a benchmark signal, we use the Z' model from Ref. [6], which produces a hadronically-decaying resonance boosted by its recoil against an initial state photon (Fig. 1). The same model can be used to study recoil against initial-state gluons or W bosons; we choose the photon channel due to the simpler event topology.

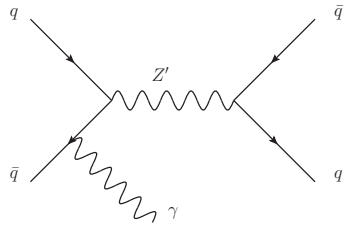

FIG. 1. Diagram of a hadronically-decaying resonance (Z') produced recoiling against an initial state photon (γ).

Signal events in which a hypothetical Z' boson decays to quarks are simulated at parton level with MADGRAPH5 [19] v2.2.3, with PYTHIA [20] v6.4.28 for showering and hadronization, and with DELPHES [21] v3.1.2 in the ATLAS-style configuration for primitive detector simulation. The primary background is due to γ+jets production, which is generated with SHERPA [22] v.2.2.0 requiring one photon and one to three additional hard partons.

The measurement of jet masses is sensitive to the presence of additional in-time pp interactions, referred to as *pile-up* events. We overlay such interactions in the simulation chain, with an average number of interactions per event of $\langle\mu\rangle = 15$, which is comparable to the level observed in ATLAS 2015 data, with the LHC delivering collisions at a 25ns

bunch crossing interval. Effects due to out-of-time pile-up are not modeled or accounted for.

To mitigate the impact of pile-up events on large-radius jet reconstruction, we apply a jet-trimming algorithm [24] which is designed to remove contituents of the jet cluster originating from pile-up interactions, while preserving the two-pronged substructure characteristic of boson decay. Jets are trimmed by reclustering into k_{T} subjets, with $R_{\mathrm{trim}} = 0.2$, and dropping subjets with less than 3% of the original jet p_{T}. Only jets reconstructed with $m^{trim} > 20$ GeV are considered in this analysis.

As the angular separation of the quarks may be quite small in the case of a high-p_{T} Z', we reconstruct a single large-radius jet with distance parameter $R = 1.0$. To reflect the thresholds imposed by the ATLAS trigger, we require $p_{\mathrm{T}}^{\gamma} > 150$ GeV and $p_{\mathrm{T}}^{\mathrm{jet}} > 150$ GeV. In the case of multiple large-R jets, the one with greatest p_{T} is selected.

For the large-radius jets, we calculate various jet substructure variables such as the N-subjettiness ratio τ_{21} [7, 25], and the Energy Correlation Functions [8, 26]. Recent studies have shown that deep neural networks applied to lower-level calorimeter information can match the performance of several of these higher-level variables in combination [11], but these higher-level variables capture most of the discriminative information and are theoretically well understood.

Distributions of the various kinematic quantities for jets selected in signal and background processes are shown in Fig. 2. The neural networks described below use eleven variables:

- Jet pseudo-rapidity, azimuthal angle, transverse momentum, and invariant mass;
- Jet energy correlation variables, C_2 and D_2 [8];
- Jet N-subjettiness (τ_{21}) [7]; and
- Photon energy, pseudo-rapidity, azimuthal angle, transverse momentum.

For comparison with Ref. [9], we additionally apply the DDT procedure to produce a modified variable, τ_{21}', which has reduced correlation with jet mass. However, no simple linear relationship was seen between the profile of τ_{21} and the jet mass, and a linear correction does not remove the dependence; this may be due to the application of jet trimming, which differs from the treatment in Ref. [9]. To provide a fair comparison, we extend the DDT-style approach to use a second-order correction, producing a variable τ_{21}'', which demonstrates reasonable indepenence from the jet mass (Fig. 5).

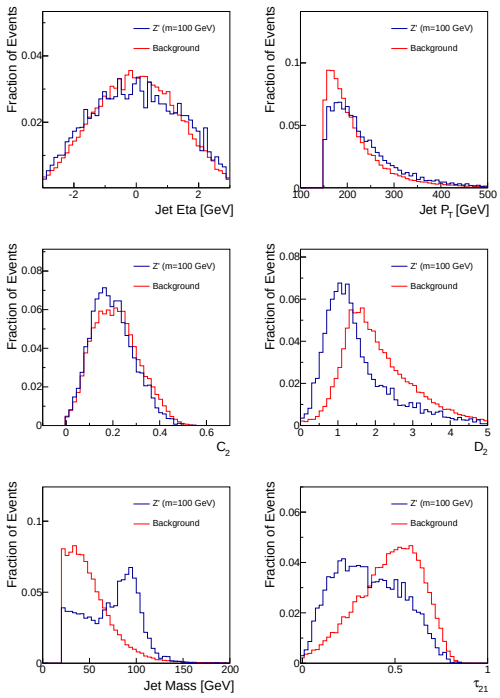

FIG. 2. Distributions of jet variables in simulated $Z' + \gamma$ signal events, with $m_{Z'} = 100$ GeV, as well as γ+jet background events. From top left to bottom right are shown the jet pseudorapidity, transverse momentum, energy correlation variables C_2 and D_2 [8], jet invariant mass, and N-subjettiness(τ_{21}) [7]. There are five additional input variables described in the text (not shown).

III. NEURAL NETWORKS

The strategy outlined in Ref. [16] describes how to train a classifier which is uncorrelated with a nuisance parameter. Here, we apply this strategy to the closely-related problem of decorrelating the classifier with respect to the jet invariant mass, as the nuisance parameter is not well defined; further discussion of this issue is found below in Sec. V. In Sec. VII, we extend this strategy to a problem requiring a parameterized solution.

Two neural networks — a jet classifier and an adversary — constitute two distinct segments of the feedforward architecture shown in Fig. 3. The loss of the tagger is defined as

$$L_{\mathrm{tagger}} = L_{\mathrm{classification}} - \lambda L_{\mathrm{adversary}},$$

where λ is a positive constant, and $L_{\mathrm{classification}}$ and $L_{\mathrm{adversary}}$ are the standard classification-error loss

functions for each segment. The two neural networks are trained concurrently; the tagger's objective is to minimize L_{tagger}, while adversary minimizes only $L_{\text{adversary}}$. The hyperparameter λ represents a tradeoff between the two objective terms; we found that a value of $\lambda = 100$ was a good tradeoff for our task, but in general this hyperparameter can be optimized like any other.

The classifier network in this experiment consisted of eleven input features, three fully-connected hidden layers each with 300 nodes having hyperbolic tangent activation functions, and a single logistic output node with the binomial cross-entropy classification objective. The adversarial network consisted of a single input, 50 nodes with hyperbolic tangent activation functions, and a softmax output layer with 10 classes corresponding to binned values of the jet invariant mass (each bin representing one decile of the background), and the multi-class cross-entropy classification objective.

Because the adversary is challenged with adapting to an ever-changing input as the classifier is trained, and also because its task is relatively easy, two strategies were used to train the adversary faster than the classifier. First, the adversary was given a head start at the beginning of training with 100 updates while the classifier was fixed. Second, the adversary was trained with a larger learning rate of 1.0 compared to 10^{-3} for the tagger objective.

The data set used for experiments was divided into training (80%), validation (10%, used for hyperparameter tuning), and testing (10%) subsets. Each classifier input feature was log-scaled if the empirical skew estimate was greater than 1.0, then standardized to zero mean and unit variance. Model parameters were initialized from a scaled normal distribution [27].

Training was performed using stochastic gradient descent, applied to mini-batches of 100 examples from each class. During training, the event weights were scaled so that the average weight for each class was 1.0. However, in the adversarial loss function $L_{\text{adversary}}$, the signal events were given zero weight, rendering them invisible to the adversary.

Updates were made using a training momentum term of 0.5; the learning rate decayed by a factor of 10^{-5} after each update. Training was stopped after 100 epochs, where an epoch was defined as a single pass through the background samples ($\approx$ 400k training events). Models were implemented in Keras [28] and Theano [29], and hyperparameters were optimized on a cluster of Nvidia Titan Black processors.

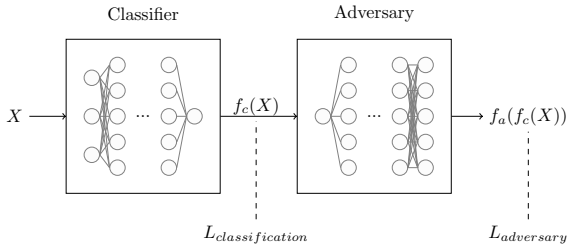

FIG. 3. Architecture of the neural networks in the adversarial training strategy. The classifying network distinguishes signal from background using the eleven variables (X) described in the text. The adversarial network attempts to predict the invariant mass using only the output of the classifier, $f_c(X)$; note that the adversary has multiple binary classification outputs, corresponding to bins in jet invariant mass, rather than a single regression output.

IV. PERFORMANCE

We compare the discrimination power of five candidate classifiers: the NN trained without an adversary, the adversarially-trained NN, the unmodified τ_{21}, and the two DDT-modified variables τ'_{21}, and τ''_{21}. The performance can be characterized by measuring the signal efficiency and background rejection of various thresholds on these discriminators (Fig. 4).

The variable τ'_{21}, which is modified to reduce correlation with the mass, results in a modest decrease in its classification power relative to the unmodified τ_{21} at $m_{Z'} = 100$ GeV, though note that these effects are mass-dependent for both τ'_{21} and τ''_{21}. Similarly, the adversarial network does not match the discrimination power of the traditional classification network, due to the additional constraint imposed in its optimization. However, both NNs are clearly able to take advantage of the combined power of the substructure variables, and offer a large improvement in background rejection for similar signal efficiencies compared to classification based on τ_{21} alone.

The focus of this study, however, is to look beyond the pure discriminatory power of these tools and study their effect on the jet mass spectrum. In Fig. 5, it can be seen that the adversarial network output for background events has a profile which is largely independent of jet mass, while the classifying network is strongly dependent on jet mass. Similarly, τ'_{21} and τ''_{21} have a lessened dependence on jet mass, compared to τ_{21}. Figure 7 shows the effect on the jet mass distribution of successively

stricter requirements on these variables. Note that the adversarial network's dependence on jet mass is diminished, but not eliminated, as can be seen in the contour plot of Fig. 5. This is a reflection of the trade-off inherent in balancing classification power with jet mass dependence.

In Fig. 5, we also show the profile of the neural network output versus jet mass, for various thresholds on the jet p_{T}, which shows some small p_{T}-dependent effects, but no large features. As an alternative strategy, we trained a network using an adversarial strategy with respect to $\log(m/p_{\mathrm{T}})$, which more closely mimics the approach used in Ref. [9]; the training succeeded in finding a network with a flat response in $\log(m/p_{\mathrm{T}})$, but the distortion in jet mass was much more significant. In principle, it is possible to use the adversary to enforce a two-dimensional decorrelation, but since the p_{T}-dependence is not severe here, we leave this for future study.

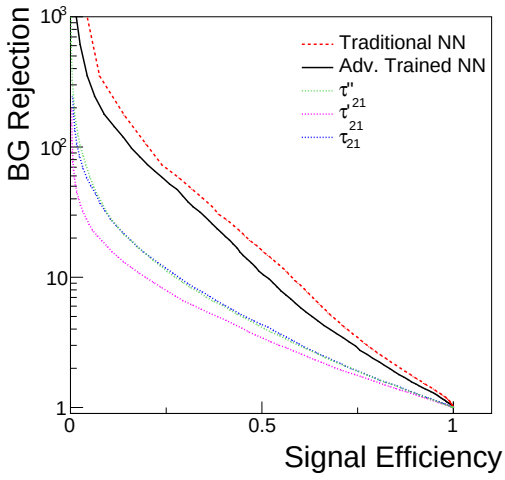

FIG. 4. Signal efficiency and background rejection (1/efficiency) for varying thresholds on the outputs of several jet-tagging discriminants: traditional networks trained to optimize classification, networks trained with an adversarial strategy to optimize classification while minimizing impact on jet mass, the unmodified τ_{21}, and the two DDT-modified variables τ'_{21}, and τ''_{21}. The signal samples have $m_{Z'} = 100$ GeV for this example. Generalization to other masses is shown in Sec. VII.

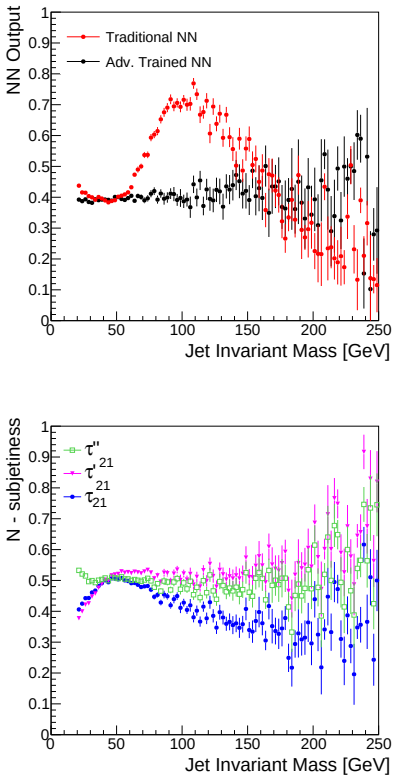

FIG. 5. Top, relationship between jet mass and neural network output in background events for a network trained to optimize classification compared to an adversarial network trained to optimize classification while minimizing dependence on jet mass. Bottom, relationship between jet mass and jet substructure variable τ_{21} and the DDT-modified τ'_{21} and τ''_{21} which attempt to minimize dependence on jet mass.

V. STATISTICAL INTERPRETATION

The ability to discriminate jets due the hadronic decay of a boosted object from those due to a quark or gluon is an important feature of a jet substructure tagging tool, but as discussed above it is not the only requirement. Due to the necessity of accurately modeling the background, it is desirable that the jet tagger avoid distortion of the background distribution. Simpler background shapes are especially preferred because they allow for robust estimates that are constrained by the sidebands; backgrounds that can be modeled with fewer parameters and inflections avoid degeneracy with signal features, such as a peak.

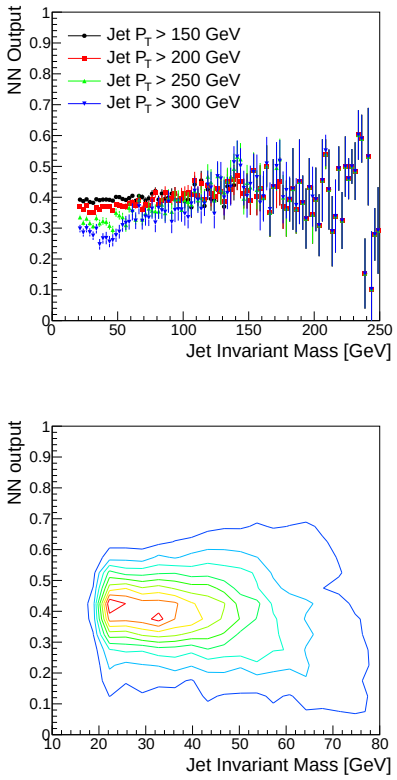

FIG. 6. Top, profile of neural network output versus jet mass for the adversarial trained network with varying jet p_{T} thresholds. Bottom, contour plot of neural network output versus jet mass in background events for the adversarially-trained network. The signal sample used in training has $m_{Z'} = 100$ GeV; generalization to other masses is shown in Sec. VII.

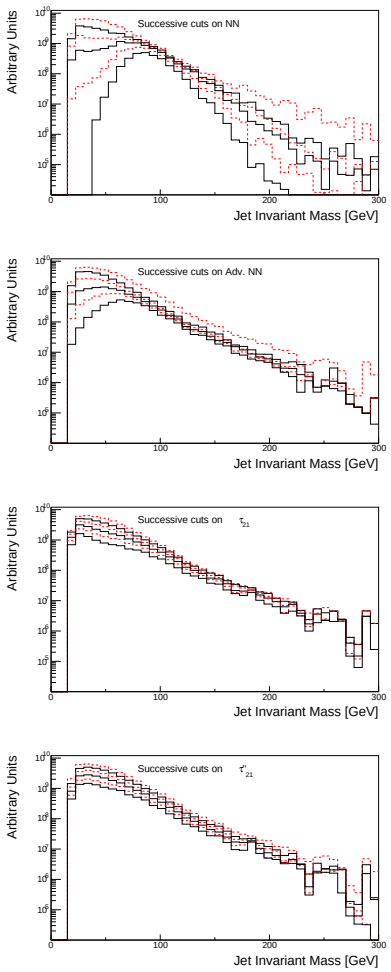

FIG. 7. Jet mass distributions for background events with successively stricter requirements on different substructure discrimination strategies, giving signal efficiencies of $\varepsilon_{\mathrm{sig}} = 50, 60, 70, 80, 90, 100\%$, shown alternating red and black lines. Shown are the impact of threshold requirements on a neural network output trained to optimize classification, an adversarial network which attempts to minimize depenence on jet mass, τ_{21} and τ_{21}''.

Figs. 5 and 6 shows qualitatively that the adversarial network's response is not strongly dependent on jet mass. But a quantitative assessment is more difficult. Mass-independence is not in itself the goal; instead, we seek reduced dependence on knowledge of the background shape and reduced sensitivity to the systematic uncertainties that tend to dilute the statistical significance of a discovery.

However, our lack of knowledge of the true background model in general also makes it non-trivial to rigorously define and estimate the background uncertainty. In practice, experimentalists use an assumed functional form, with parameters constrained by background-dominated sidebands to predict the background in the signal region. These assumptions may be validated by examining control regions in which the signal is not present, and the background processes are expected to exhibit physically similar properties. For example, the tagger selection may be inverted to yield a sample with high background purity which may be used as a template. If the tagger selection induces a distortion of the spectrum, these techniques are ineffective. Moreover, when tagger-induced distortion depletes data from the sidebands (as is typically the case), any background model becomes more difficult to constrain. To demonstrate these effects on the overall statistical performance

of a search, we construct a simplified statistical test which has the desired behavior of penalizing discriminators which yield excessive distortion of the background shape.

A threshold is placed on the discriminator output, after which a likelihood fit is performed, binned in the distribution of reconstructed large-radius jet masses using signal and background templates from simulated samples[1]. An uncertainty on the rate of the background is included in order to model our lack of knowledge of the background. We calculate expected discovery significance using a profile likelihood ratio [30] with the CLs technique [31, 32], marginalizing over the unknown background rate.

Any background model used (whether a template or functional form) will necessarily incorporate nuisance parameters corresponding to unknown properties of the background; what is important in practice is that these parameters can be effectively constrained in the observed data. Though the shape of the background model considered here is fixed via the template, the uncertainty on the rate provides the statistical behavior we seek. Specifically, if the uncertainty in the rate of the background is large enough, then the discovery significance is sensitive also to the shape of the background distribution as follows. In the case that the background is fairly flat, there are background-dominated sidebands which can constrain the rate uncertainty. In the opposite case that the background is distorted to mimic the signal, these sideband constraints have reduced power, and the signal and background are more difficult to distinguish statistically. Hence, the presence of rate uncertainties penalizes a solution which distorts the background spectrum as desired. Although this simple approach likely underestimates the true impact of more realistic systematics, it is sufficient to illustrate the effect on sensitivity. In the following, we take for the small (large)-uncertainty case a relative uncertainty of 5% (50%) on the overall background rate.

Examples of the final jet mass distribution are shown in Figs. 8 and 9 for thresholds on the discriminants which result in signal efficiency of 90% and 50% respectively.

[1] In principle, the most powerful approach is a likelihood directly on the output of the discriminator, but this requires a valid model of the background, which is lacking in this case.

VI. RESULTS

The discovery significance is measured for varying thresholds on the discriminator outputs. While all of the discriminators exhibit some degree of classification power, this study explores the question of whether they provide additional discovery significance.

Figure 10 shows the discovery significance as a function of the signal efficiency of the discriminator threshold, for two choices of background uncertainty. In the case of the small uncertainty (5% relative), applying a tighter threshold on the discriminator improves the discovery significance, despite lowering the signal efficiency, due to the heightened background suppression. Even at fairly low signal efficiencies of 50%, where the background is sculpted to look like the signal (see Fig. 9), the discovery significance is improved. This is as expected; if the background rate and shape are well known, then the lack of constraining sidebands is not detrimental.

For the case of the larger background rate uncertainty, thresholds on τ_{21} provide a smaller boost to the significance. The large relative uncertainty on the background will penalize configurations in which the background is sculpted to resemble the signal, preventing the data from constraining the background rate in the sidebands. Thresholds on τ_{21}' and τ_{21}'' are slightly stronger, as expected, due to their decreased correlation with jet mass. Thresholds on the output of the classifier network, which has the strongest discrimination power, only weakens the discovery significance, due to the background mass distortion. However, the adversarial network is still capable of powerful discrimination which improves the discovery power at high signal efficiency, around 90%. Table I shows the maximal discovery significance for each case. The qualitative results persist for other signal-to-background ratios.

VII. PARAMETERIZED NEURAL NETWORKS

The studies above demonstrate the application for the case of a single example value of the hypothetical Z' mass. In this section, we show that the same approach can be generalized to solve a set of closely related problems, jet classification for different Z' masses, using a single neural network parameterized in $m_{Z'}$.

These parameterized neural networks [18] address a common problem in physics: solving a classifica-

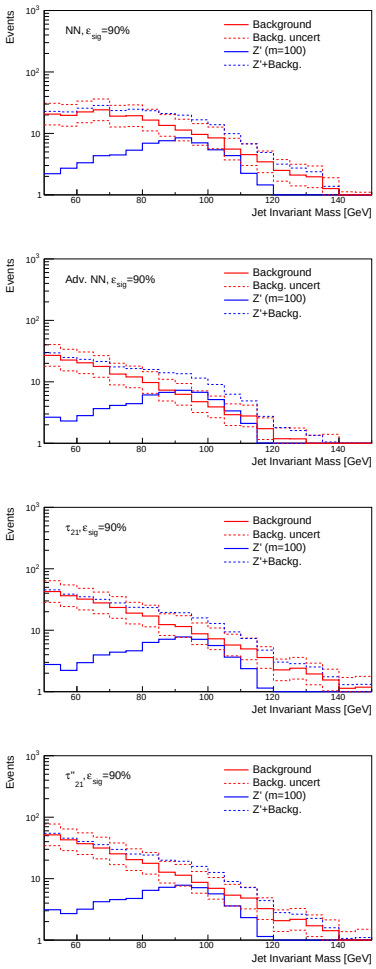

FIG. 8. Distributions of jet mass after selection with signal efficiency of 90% using the NN classifier, the adversarial network, τ_{21} or τ''_{21}. Background distributions are shown with 50% uncertainty.

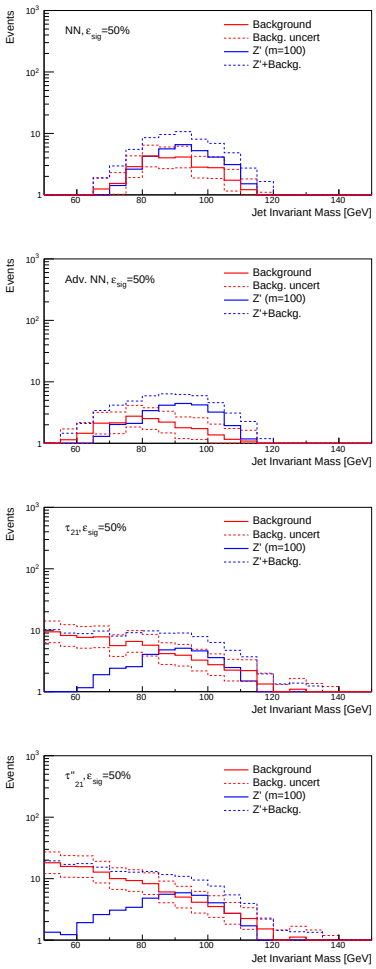

FIG. 9. Distributions of jet mass after selection with signal efficiency of 50% using the NN classifier, the adversarial network, τ_{21} or τ''_{21}. Background distributions are shown with 50% uncertainty.

tion task multiple times for different values of an unknown latent variable, like $m_{Z'}$. Simulations used to train jet classifiers are generally performed for a small set of fixed Z' mass values. In the traditional approach, a separate neural network classifier is trained for each Z' mass value. However, by treating $m_{Z'}$ as just another input feature, a single parameterized neural network can learn to solve the related classification tasks all at once (Fig. 11). Furthermore, the classifier can interpolate to other values of $m_{Z'}$ if the function is smooth.

For this experiment, some hyperparameters were tuned to this more complex task. The classifier had three hidden layers of 300 tanh nodes, with a learning rate of 10^{-4}, a momentum of 0.95, and an L2 weight decay factor of 10^{-3} in each layer. The adversary consisted of two hidden layers of 100 tanh nodes each, with a learning rate of 10^{-2}, a momentum of 0.95, and an L2 weight decay factor of 10^{-4} in each layer. The parameter λ was set to 10.

The adversary was also parameterized by including the Z' mass as an input along with the classifier output. The resulting classifier predictions for background events are mostly independent of mass when conditioned on each theory mass (Fig. 12). Without this parameterization of the adversary, the marginalized classifier predictions are independent of mass, but not the conditional classifier predictions.

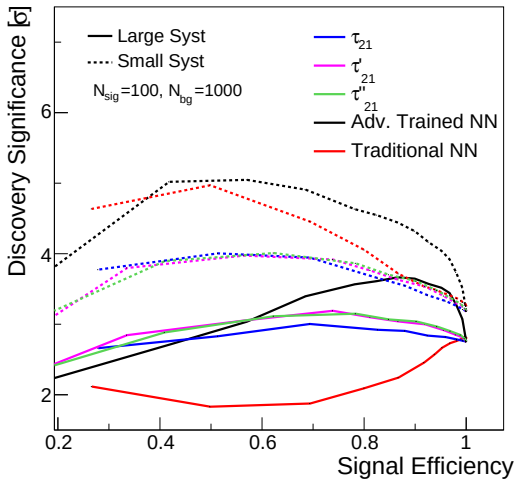

FIG. 10. Statistical significance of a hypothetical signal for varying thresholds on the outputs of networks trained to optimize classification compared to adversarial networks trained to optimize classification while minimizing impact on jet mass. Shown are two scenarios, in which the uncertainty on the background level is negligible or large, both with $N_{\rm sig} = 100, N_{\rm bg} = 1000$.

TABLE I. Signal and background efficiencies at maximal discovery significance at $m_{Z'} = 100$ GeV for each method and for scenarios of large (50%) or small (5%) relative systematic uncertainty on the background rate. Uncertainties are approximately 0.01 in all cases.

Method	Signal Eff.	Background Eff.	Discovery Signif. (σ)
5% background uncertainty			
Adv. Trained NN	0.44	0.06	5.05
Traditional NN	0.39	0.03	4.97
τ_{21}	0.44	0.19	4.00
τ'_{21}	0.50	0.29	3.97
τ''_{21}	0.52	0.26	4.01
50% background uncertainty			
Adv. Trained NN	0.82	0.48	3.67
Traditional NN	1.00	1.00	2.82
τ_{21}	0.60	0.32	3.00
τ'_{21}	0.70	0.50	3.19
τ''_{21}	0.70	0.45	3.15

As expected, the resulting classifier demonstrates better performance than the single input features τ_{21}, τ'_{21} or τ''_{21} at all signal mass hypotheses tested (Fig. 13). As in the non-parameterized case, the traditional NN trained to maximize classification accuracy achieves the best separation.

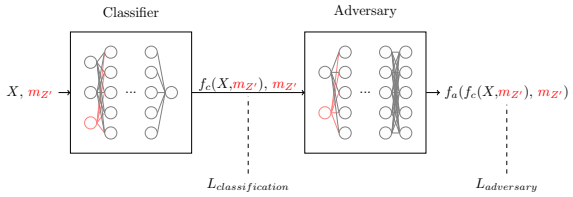

FIG. 11. Architecture of the neural networks in the parameterized adversarial training strategy. The classifying network distinguishes signal from background using the eleven variables described in the text (X) plus $m_{Z'}$. The classifying network output is then a function of $m_{Z'}$. The adversarial network attempts to predict the invariant mass using the output of the classifier, $f_c(X, m_{Z'})$ as well as $m_{Z'}$.

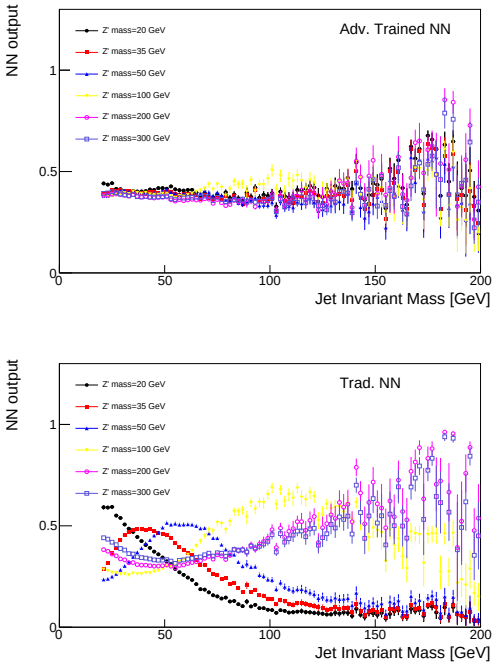

FIG. 12. Profile of the paramterized NN responses to background versus jet mass, where the parameterized network was evaluated at different Z' mass hypotheses. Top shows the response of the adversarially-trained classifier, which minimizes correlation with jet mass; bottom shows the response of a network trained in the traditional manner, to optimize classification accuracy.

Moreover, the lack of background distortion by the adversarially-trained network preserves the ability to distinguish the background and signal mass distributions, leading to improved discovery signifi-

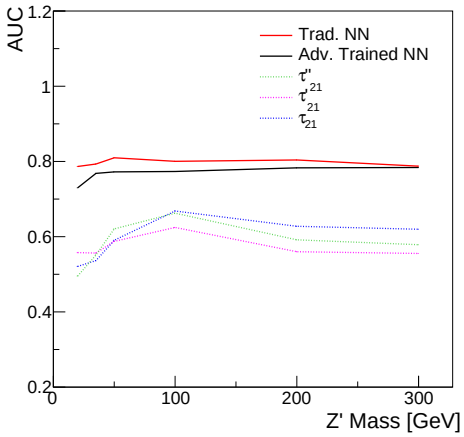

FIG. 13. The AUC metric (Area Under the Curve) for NNs parameterized in $m_{Z'}$ and tested at several values (both traditional and adversarial training techniques), compared to the discrimination of the individual features τ_{21}, τ'_{21}, and τ''_{21}.

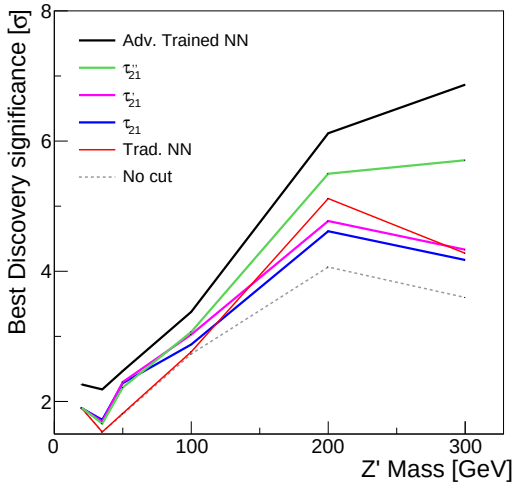

FIG. 14. Discovery significance for a hypothetical signal after optimizing thresholds on the output of networks parameterized in $m_{Z'}$ trained with an adversarial or traditional approaches, compared to thresholds on τ_{21}, τ'_{21} and τ''_{21} or to placing no threshold. Significance is evaluated for the case of 50% background uncertainty.

cance; see Fig. 14. The statistical test is performed as for the previous case, fitting a binned likelihood on the jet mass distribution after applying a threshold on the discriminator output. As before, the improved separation of the traditional NN does not translate to improved discovery significance.

We note that while the performance shown here is evaluated on hypothesized mass values used for training, Ref. [18] demonstrates this architecture is able to successfully interpolate to other values of $m_{Z'}$.

VIII. DISCUSSION

We have demonstrated that an adversarial training strategy may yield a jet classification tagger which leverages the powerfully discriminating information obtained by combining several input features, while decorrelating its output from the variable of interest, the jet mass. This allows the classifier to enhance signal to noise ratio while minimizing the tendency of the background distribution to morph into a shape which is degenerate with the observable signal. When the background cannot be reliably predicted *a priori*, as is often the case, it is important to be able to constrain its rate in sidebands surrounding the signal region. Therefore, avoiding such degeneracy is critical to performing successful measurements.

We note that, from Fig. 9, it is clear that applying sufficiently tight cuts to the adversarial classifier causes significant background morphing, particularly when compared to the τ_{21}-based discriminants. However, the solid lines of Fig. 10 illustrate the case where the background rate is uncertain and hence benefits from sideband constraints. We see that the optimal significance is realized for the adversarial classifier at a relatively high signal efficiency of roughly 90%, where the background morphing is quite limited (Fig. 8). Hence, the adversarial classifier achieves its goal of optimizing the tradeoff between correlation and discrimination power.

We also note that the decorrelation could potentially be improved. The contour plot in Fig. 6 shows that while the average NN output is independent of mass, there is certainly still structure that results in the background sculpting still observed. The residual p_{T} dependence could also be removed, possibly with a more sophisticated adversary that is trained to predict multiple variables simultaneously. These improvements we leave for future work.

Finally, we extend the strategy to the case of a parameterized network wherein the NN classifier is trained to tag specific signal hypotheses, useful for scanning a range of theoretical parameter space with a search. The resulting combined approach should

be readily applicable to experimental measurements and searches, boosting their discovery significance or search sensitivity.

IX. ACKNOWLEDGMENTS

The authors acknowledge useful conversations with Kyle Cranmer, Jesse Thaler, Kevin Bauer, and Dan Guest, helpful comments from Sal Rappoccio, Derek Soeder and Michela Paganini, and are grateful to the Aspen Center for Physics, where much useful discussion occured.

[1] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Phys.Rev.Lett. **100**, 242001 (2008), arXiv:0802.2470 [hep-ph].
[2] D. Adams, A. Arce, L. Asquith, M. Backovic, T. Barillari, *et al.*, (2015), arXiv:1504.00679 [hep-ph].
[3] A. Abdesselam *et al.*, *Boost 2010 Oxford, United Kingdom, June 22-25, 2010*, Eur. Phys. J. **C71**, 1661 (2011), arXiv:1012.5412 [hep-ph].
[4] A. Altheimer *et al.*, *BOOST 2011 Princeton , NJ, USA, 2226 May 2011*, J. Phys. **G39**, 063001 (2012), arXiv:1201.0008 [hep-ph].
[5] A. Altheimer *et al.*, *BOOST 2012 Valencia, Spain, July 23-27, 2012*, Eur. Phys. J. **C74**, 2792 (2014), arXiv:1311.2708 [hep-ex].
[6] C. Shimmin and D. Whiteson, Phys. Rev. **D94**, 055001 (2016), arXiv:1602.07727 [hep-ph].
[7] J. Thaler and K. Van Tilburg, JHEP **03**, 015 (2011), arXiv:1011.2268 [hep-ph].
[8] A. J. Larkoski, G. P. Salam, and J. Thaler, JHEP **06**, 108 (2013), arXiv:1305.0007 [hep-ph].
[9] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, JHEP **05**, 156 (2016), arXiv:1603.00027 [hep-ph].
[10] CMS Collaboration, CMS-PAS-EXO-16-030 (2016).
[11] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, Phys. Rev. **D93**, 094034 (2016), arXiv:1603.09349 [hep-ex].
[12] J. Schmidhuber, Neural Computation **4**, 863 (1991).
[13] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, J. Mach. Learn. Res. **17**, 2096 (2016).
[14] H. Edwards and A. J. Storkey (2016) arXiv:1511.05897 [cs.LG].
[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014) pp. 2672–2680.
[16] G. Louppe, M. Kagan, and K. Cranmer, (2016), arXiv:1611.01046 [stat.ME].
[17] K. Cranmer, J. Pavez, and G. Louppe, (2015), arXiv:1506.02169 [stat.AP].
[18] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, Eur. Phys. J. **C76**, 235 (2016), arXiv:1601.07913 [hep-ex].
[19] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, JHEP **07**, 079 (2014), arXiv:1405.0301 [hep-ph].
[20] T. Sjostrand, S. Mrenna, and P. Z. Skands, JHEP **0605**, 026 (2006), arXiv:hep-ph/0603175 [hep-ph].
[21] J. de Favereau *et al.* (DELPHES 3), JHEP **1402**, 057 (2014), arXiv:1307.6346 [hep-ex].
[22] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter, JHEP **02**, 007 (2009), arXiv:0811.4622 [hep-ph].
[23] M. Cacciari, G. P. Salam, and G. Soyez, Eur.Phys.J. **C72**, 1896 (2012), arXiv:1111.6097 [hep-ph].
[24] D. Krohn, J. Thaler, and L.-T. Wang, JHEP **02**, 084 (2010), arXiv:0912.1342 [hep-ph].
[25] J. Thaler and K. Van Tilburg, JHEP **02**, 093 (2012), arXiv:1108.2701 [hep-ph].
[26] A. J. Larkoski, I. Moult, and D. Neill, JHEP **12**, 009 (2014), arXiv:1409.6298 [hep-ph].
[27] X. Glorot and Y. Bengio, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010* (2010) pp. 249–256.
[28] F. Chollet, *Keras* (GitHub, 2015).
[29] Theano Development Team, arXiv e-prints **abs/1605.02688** (2016).
[30] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Eur.Phys.J. **C71**, 1554 (2011), arXiv:1007.1727 [physics.data-an].
[31] A. L. Read, J.Phys. **G28**, 2693 (2002).
[32] T. Junk, Nucl.Instrum.Meth. **A434**, 435 (1999), arXiv:hep-ex/9902006 [hep-ex].