Simulation-assisted decorrelation for resonant anomaly detection

Kees Benkendorfer,[1,3,*] Luc Le Pottier,[2,3,†] and Benjamin Nachman[3,‡]

[1]*Physics Department, Reed College, Portland, Oregon 97202, USA*
[2]*Physics Department, University of Michigan, Ann Arbor, Michigan 48109, USA*
[3]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

(Received 18 September 2020; accepted 2 July 2021; published 5 August 2021)

A growing number of weak and unsupervised machine learning approaches to anomaly detection are being proposed to significantly extend the search program at the Large Hadron Collider (LHC) and elsewhere. One of the prototypical examples for these methods is the search for resonant new physics, where a bump hunt can be performed in an invariant mass spectrum after applying a classifier to enhance the presence of a potential signal. A significant challenge to methods that rely entirely on data is that they are susceptible to sculpting artificial bumps from the dependence of the machine learning classifier on the invariant mass. We explore two solutions to this challenge by minimally incorporating simulation into the learning. In particular, we study the robustness of simulation assisted likelihood-free anomaly detection to correlations between the classifier and the invariant mass. Next, we propose a new approach that only uses the simulation for decorrelation but uses the classification without labels approach for achieving signal sensitivity. Both methods are compared using a full background fit analysis on simulated data from the LHC Olympics and are robust to correlations in the data.

DOI: 10.1103/PhysRevD.104.035003

I. INTRODUCTION

Despite compelling experimental (e.g., dark matter) and theoretical (e.g., the hierarchy problem) evidence for new phenomena at the electroweak scale, experiments at the Large Hadron Collider (LHC) have not yet discovered any physics beyond the Standard Model (BSM). There are major search efforts across LHC experiments [1–7], where most analyses target a particular class of BSM models. While this work is well motivated and is continuing to improve in sensitivity (in part due to machine learning [8–11]), there is also a growing need for new search strategies capable of discovery in unexpected scenarios.

A variety of automated anomaly detection techniques using innovative machine learning methods are being proposed to cover the unexpected [12–40]. An important subset of these proposals targets resonant new physics, where sideband methods can be used to estimate the Standard Model background directly from data after applying a classifier to enhance the presence of a potential signal. A key challenge facing such methods is that the machine learning classifiers must be relatively independent from the resonant feature, for otherwise artificial bumps can be formed. Many automated decorrelation methods have been proposed to ensure that classifiers are decorrelated from particular features by construction [41–52], but they may not apply in all cases. In particular, weakly supervised approaches that learn directly on the signal region cannot be simply combined with a decorrelation scheme because such an approach could degrade the performance in the presence of a signal. A localized signal would manifest as a dependence between the resonant feature and other features for classification, so forcing independence could eliminate signal sensitivity.

In this paper, two weakly supervised approaches are studied: classification without labels (CWOLA) [13–15,53] and simulation assisted likelihood-free anomaly detection (SALAD) [27]. CWOLA is a method that does not depend on simulation and achieves signal sensitivity by comparing a signal region with nearby sideband regions in the resonance feature. As a result, CWOLA is particularly sensitive to dependencies between the classification features and the resonant feature. SALAD uses a reweighted simulation to achieve signal sensitivity. Since it never directly uses the sideband region, SALAD is expected to be more robust than CWOLA to dependencies. In order to recover the performance of CWOLA in the presence of signific ant dependence between the classification features and the resonant feature, a new method called simulati

*kebenkend@reed.edu
†luclepot@umich.edu
‡bpnachman@lbl.gov

on-augmented CWOLA (SA-CWOLA) is introduced. The SA-CWOLA approach augments the CWOLA loss function to penalize the classifier for learning differences between the signal region and the sideband region in simulation, which is signal free by construction. All of these methods will be investigated using the correlation test proposed in Ref. [28].

This paper is organized as follows. Section II reviews the SALAD and CWOLA methods and introduces the simulation-augmented CWOLA search strategy. Furthermore, the sideband analysis is set up in Sec. II. These methods are illustrated with a Gaussian example in Sec. III, and a physics example is presented in Sec. IV. The paper ends with the conclusions and outlook in Sec. V.

II. METHODS

For a set of features $(m, x) \in \mathbb{R}^{n+1}$, let $f: \mathbb{R}^n \to [0, 1]$ be parametrized by a neural network. The observable m is special, for it is the resonance feature that should be relatively independent from $f(x)$. The signal region (SR) is defined by an interval in m and the sidebands (SB) are neighboring intervals.

All neural networks are implemented in KERAS [54] with the TENSORFLOW backend [55] and are optimized with ADAM [56]. Each network is composed of three hidden layers with 64 nodes each and uses the rectified linear unit activation function. The sigmoid function is used after the last layer. Training proceeds for 20 epochs with a batch size of 200. All other parameters use the KERAS and ADAM defaults. None of the parameters are optimized; it is likely that improved performance can be achieved with an in-situ optimization based on a validation set.

A. Simulation assisted likelihood-free anomaly detection (SALAD)

The SALAD network [27] is optimized using the following loss:

$$\mathcal{L}_{\text{SALAD}}[f] = -\sum_{i\in\text{SR,data}} \log(f(x_i)) - \sum_{i\in\text{SR,sim.}} w(x_i, m)\log(1 - f(x_i)), \quad (2.1)$$

where $w(x_i, m) = g(x_i, m)/(1 - g(x_i, m))$ is a set of weights using the classification for tuning and reweighting (DCTR) [57] method. The function g is a parametrized classifier [58,59] trained to distinguish data and simulation in the sideband:

$$\mathcal{L}[g] = -\sum_{i\in\text{SB,data}} \log(g(x_i, m)) - \sum_{i\in\text{SB,sim.}} \log(1 - g(x_i, m)). \quad (2.2)$$

The above neural networks are optimized with binary cross entropy, but one could use other functions as well, such as the mean-squared error. Intuitively, the idea of SALAD is to train a classifier to distinguish data and simulation in the SR. However, there may be significant differences between the background in the data and the background simulation, so a reweighting function is learned in the sidebands that makes the simulation look more like the background in the data.

B. Simulation-augmented classification without labels (CWoLa)

The idea of CWOLA [53] is to construct two mixed samples of data that are each composed of two classes. Using CWOLA for resonant anomaly detection [13,14], one can construct the mixed samples using the SR and SB. In the absence of signal, the SR and SB should be statistically identical and therefore the CWOLA classifier does not learn anything useful. However, if there is a signal, then it can detect the presence of a difference between the SR and SB. In practice, there are small differences between the SR and SB because there are dependencies between m and x and so CWOLA will only be able to find signals that introduce a bigger difference than already present in the background. The CWOLA anomaly detection strategy was recently used in a low-dimensional application by the ATLAS experiment [15].

We propose a modification of the usual CWOLA loss function in order to construct a simulation-augmented CWOLA classifier,

$$\mathcal{L}_{\text{SA-CWola}}[f] = -\sum_{i\in\text{SR,data}} \log(f(x_i)) - \sum_{i\in\text{SB,data}} \log(1 - f(x_i)) - \lambda\left(\sum_{i\in\text{SR,sim.}} \log(1 - f(x_i)) + \sum_{i\in\text{SB,sim.}} \log(f(x_i))\right), \quad (2.3)$$

where $\lambda > 0$ is a hyperparameter. The limit $\lambda \to 0$ is the usual CWOLA approach, and for $\lambda > 0$ the classifier is penalized if it tags the signal region in simulation as signal-like and the sideband region in simulation as unlike the signal. This is equivalent, in the limit $\lambda \lesssim 1$, to penalizing the classifier for being able to distinguish the SR from the

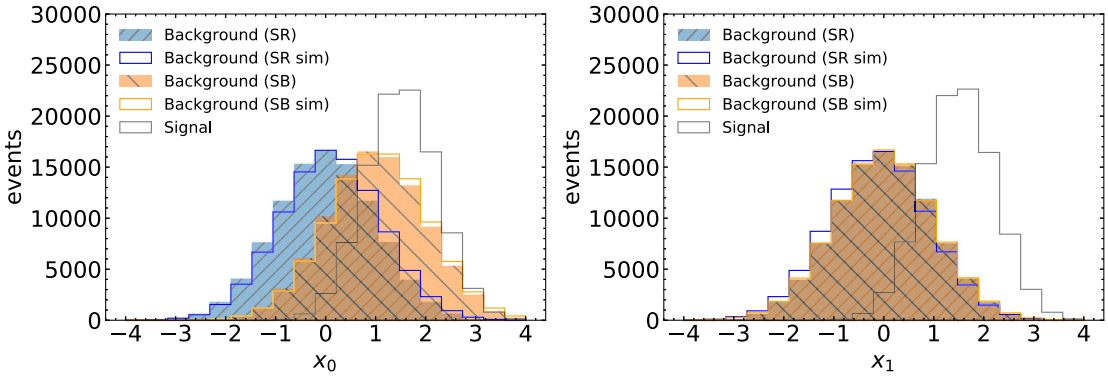

FIG. 1. The Gaussian distributions generated to test the classification methods. Left: The first feature x_0. Right: The second feature x_1.

SB in the (background-only) simulation.[1] In order to help the learning process, the upper and lower sidebands are given the same total weight as each other and together, the same weight as the SR.

Another way of implementing Eq. (2.3) is to reverse the order of the second term and its sign, i.e., it is a standard SR (label 1) versus SB (label 0) classifier in simulation, but with an overall minus sign to penalize classifier performance. This has similar properties to Eq. (2.3), but is numerically unstable as $\lambda \to 1$ due to a near cancellation of the constraint on the background.

C. Bump hunt analysis

In addition to quantifying performance with receiver operating characteristic (ROC) curves, it is also useful to emulate a proper background estimation based on a bump hunt. A histogram of the m_{jj} spectrum, possibly after applying a threshold on one of the classifiers described above, is fit to the following parametric function

$$\frac{d\sigma}{dm_{jj}} = \frac{p_0(1-x)^{p_1}}{x^{p_x+p_3\log(x)}}, \tag{2.4}$$

where $x = m_{jj}/\sqrt{s}$ and p_i are fit parameters. This function has a long history and has also been recently used by the ATLAS and CMS Collaborations (see, e.g., [60,61]). Alternative nonparametric functions are also possible (such as Gaussian processes [62]), but these are not needed for the demonstration considered here. The SR is masked during the fit and then a p-value of the observed data is computed in the usual way. In particular, a test statistic is formed from the profile likelihood ratio

$$\lambda_0 = \frac{\max_\theta p(n|\mu=0,\theta)}{\max_{\mu,\theta} p(n|\mu,\theta)}, \tag{2.5}$$

where n is the number of observed events in the SR and θ is a nuisance parameter from the sideband fit

$$p(n|\mu,\theta) = \text{Poisson}(n|b+\theta+\mu)e^{-\theta^2/2\sigma^2}, \tag{2.6}$$

where b and σ are the number of events and the uncertainty from the sideband fit, respectively. The test statistic itself is $q_0 = -2\log(\lambda_0)$ when the extracted signal strength, $(\mu,\theta) = \text{argmax}_{\mu',\theta'} p(n|\mu',\theta')$, is $\mu > 0$ and 0 otherwise. Asymptotic formulae from Wald and Wilks then give the significance $Z = \sqrt{q_0}$ [63–65].

In practice, one would scan the signal region across the m_{jj} spectrum. In this analysis, we will focus on a single region with or without the signal injected. The signal region is defined by $m_{jj} \in [3.3, 3.7]$ TeV and the sideband for CWOLA training is defined by $m_{jj} \in [3.1, 3.3] \cup [3.7, 3.9]$ TeV. Long sidebands extended by 300 GeV in either direction are used to train the SALAD reweighting function. The background fit is performed between 2.6 and 5 TeV using 30 equally spaced bins.

III. GAUSSIAN EXAMPLE

To demonstrate the behavior of the methods in a controlled context, we generated a simple dataset of events drawn from two-dimensional Gaussian distributions $\mathcal{N}(\mu,\sigma)$. The "data" in the SR and SB was drawn from distributions with mean and covariance matrix

$$\mu_{\text{SR}} = (0,0), \qquad \mu_{\text{SB}} = (1,0),$$
$$\Sigma_{\text{SR}} = \Sigma_{\text{SB}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.1}$$

Signal events had mean and covariance

[1] One could also use the SALAD-reweighted background simulation. In practice, we found little difference between using and not using the weights as the data/sim differences were a subleading correction to the mass dependence. However, this may be more useful in other applications. We thank Jesse Thaler for this interesting idea.

$$\mu = (1.5, 1.5), \qquad \Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}. \tag{3.2}$$

In order to simulate a slightly imperfect background simulation, "simulation" events had mean and covariance

$$\mu_{\mathrm{SR}} = (0.1, -0.1), \qquad \mu_{\mathrm{SB}} = (1.05, 0),$$
$$\Sigma_{\mathrm{SR}} = \Sigma_{\mathrm{SB}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.3}$$

Plots of these distributions are provided in Fig. 1. By design, the signal events are much more sidebandlike than signal-region-like; this simulates a strong correlation between a resonant feature and the classification features.

To establish performance bounds, fully supervised classifiers were trained to distinguish between signal and data events using both available features and using only the second feature (x_1). This second feature has the same probability density in the SR and SB. For every other test, 500 signal events were injected into signal region samples of 50000 background events, corresponding to a significance of about 2σ. Each classifier was trained for 20 epochs; for nonsupervised methods, 20 classifiers were trained, and their performance is exhibited in the mean and standard deviation.

The results are displayed in Fig. 2. The solid purple line corresponds to the SALAD classifier, while the solid orange line corresponds to the SA-CWOLA classifier with hyperparameter $\lambda = 1$. The dashed red line is standard CWOLA. The solid green line corresponds to the fully supervised classifier, while the dashed green line corresponds to the supervised classifier trained only on the second feature, x_1.

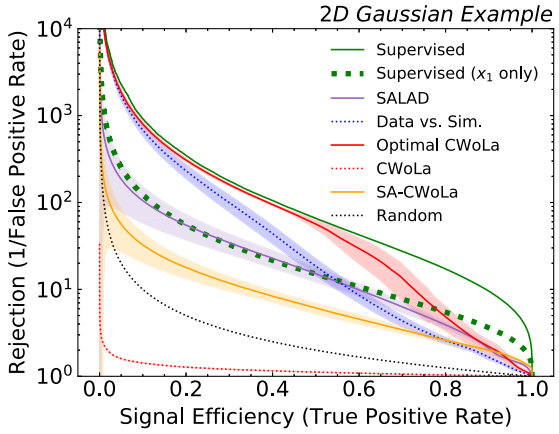

FIG. 2. ROC curves for anomaly detection methods described in the text. Better performance is to the upper right. With the exception of the supervised classifiers, the solid/dashed lines correspond to the mean performance of 20 classifiers; the shaded bands display the standard deviation in performance. The SA-CWOLA hyperparameter was set to $\lambda = 1$.

The dashed black line corresponds to a classifier which randomly assigns values to inputs.

The solid red line labeled "Optimal CWOLA" displays the results of a classifier that was trained to distinguish a mixed sample of signal region data and signal events from a pure sample of signal region data events. It is optimal in the sense that the only difference between the two datasets is the presence of signal in one of them; the only way for the classifier to minimize its loss is to learn to tag signal events. In principle, no weakly supervised classifier should be able to outperform the optimal CWOLA classifier.

The dashed blue line labeled "Data vs. Sim." presents the results of a classifier trained to distinguish the mixed sample of signal region data and signal events from a pure sample of signal region simulation events. Since the simulation is statistically similar, but not identical, to the data, the performance is slightly worse than that of optimal CWOLA. In the x_0 observable, the simulated background is signal-like while, in the x_1 observable, the simulated background is background-like. This is why the performance of the "Data vs Sim." is good at low signal efficiency and poor at high signal efficiency. This curve can be made arbitrarily bad by shifting the means and variances of the background simulation toward the signal values.

A couple of observations are worth mentioning. First, while the strong similarity between the sideband region and the signal causes CWOLA to consistently antitag the signal, the SA-CWOLA classifier is able to recover significant signal tagging ability. Additional tests showed that varying the SA-CWOLA hyperparameter λ had a large impact on performance; in some cases, for high values of λ, SA-CWOLA was able to surpass the signal tagging performance of the x_1-only supervised classifier. The trade-off is that higher values of λ make it more likely that the classifier will antitag signal region background events while tagging sideband background events as signal-like, thereby carving a large deficit in the signal region during a bump hunt. In our case, setting $\lambda = 1$ balances the competing priorities of the classifier, though this may not be desirable in other applications. Note also that, in this example, the SALAD classifier performs strictly better than the SA-CWOLA classifier. This is not always true, as will be seen in Sec. IVB.

IV. PHYSICS EXAMPLE

A. Simulation

The simulations used for this study were produced for the LHC Olympics 2020 community challenge [66]. In particular, the background process is composed of generic dijet events with a requirement for at least one such jet with $p_T > 1.3$ TeV. Signal events are $W' \to XY$ for $m_{W'} = 3.5$ TeV and hypothetical particles X and Y of mass 500 and 100 GeV, each decaying into pairs of quarks. Due to the mass hierarchy between the W' boson and its decay products, the final state is characterized by two large-radius

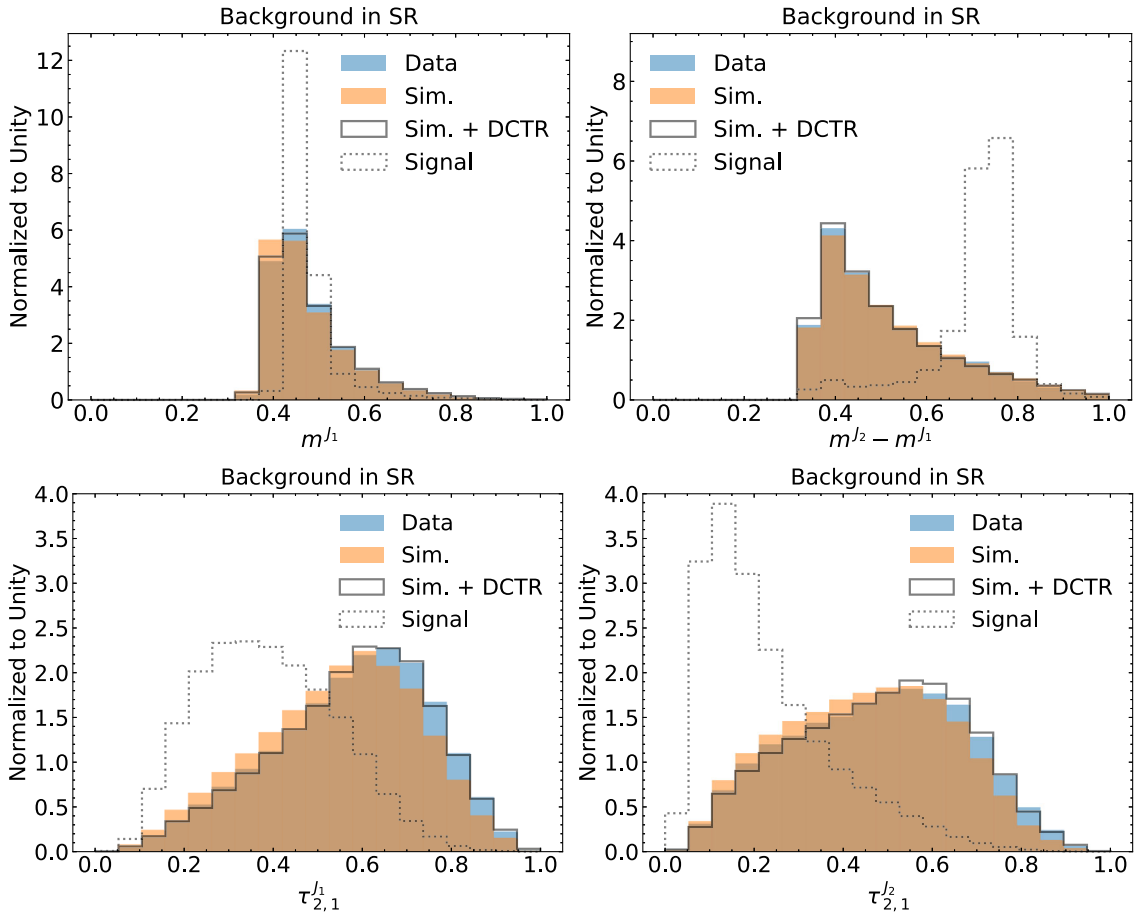

FIG. 3. Left: The jet mass and τ_{21} of the jet with a smaller mass. Right: The difference between the heavier and lighter jet masses and τ_{21} of the heavier jet. In addition to showing the data, simulation, and signal, the histogram labeled "Sim. + DCTR" is the simulation with weights derived from one of twenty parametrized reweighting functions trained on long sidebands used by the SALAD method.

jets with two-prong substructure. The background and signal are simulated using PYTHIA8 [67,68] and an alternative background sample is simulated using HERWIG++ [69]. A detector simulation is performed with Delphes 3.4.1 [70–72] using the default CMS detector card. Particle flow objects are the input to jet clustering, implemented using FastJet [73,74] and the anti-k_t algorithm [75] and using $R = 1.0$ for the radius parameter. In what follows, PYTHIA will play the role of "data" and the HERWIG sample will be used as the "simulation." There are one million events for both background samples, corresponding to an integrated luminosity of about 100 fb^{-1}. In order to simplify the analysis, the dataset is divided in half for training and testing. More complicated procedures based on k-folding to use the entire dataset for both training and testing are also possible, but are not considered here [13,14].

Both the CWOLA and SALAD methods have been demonstrated on the unmodified LHC Olympics dataset. Following Ref. [28], the dependence between the jet masses and m_{jj} is artificially strengthened by adding in a linear relationship between m_j and m_{jj}. In particular, we redefine $m_{j_i} \mapsto m_{j_i} + \alpha_i m_{jj}$ for $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$. As shown in Ref. [28], shifts like this are sufficient to reduce the efficacy of the unmodified CWOLA method.

In addition to the dijet invariant mass, four features are used for the anomaly detection: the invariant mass of the lighter jet, the mass difference of the leading two jets, and the τ_{21} [76,77] of the leading two jets. The N-subjettiness τ_{21} quantifies the extent to which a jet is characterized by two subjets or one subjet. Histograms of the four input features for the background are shown in Fig. 3. The signal jet masses are localized at the X and Y masses (shifted by $\alpha m_{W'}$) and the τ_{21} are shifted to lower values, indicating two-prongness. In addition to presenting the data and simulation histograms, Fig. 3 also shows the reweighted background simulation using parametrized weights learned from a long sideband.

While the machine-learning-based bump hunt is nearly model independent, the choice of features does introduce some model dependence. In particular, the four selected features provide sensitivity to a broad class of models that have jets with particular masses and/or two-prong substructure.

B. Sensitivity

As a benchmark, 1000 signal events corresponding to a fitted significance of about 2σ is injected into the data for training. For evaluation, the entire signal sample (except for the small number of injected events) is used. Figure 4 shows the performance of various configurations. The fully supervised classifier uses a high statistics signal and background samples in the SR with full label information. Since the data are not labeled, this is not achievable in practice. An additional analysis of supervised classifiers trained on every individual feature in the dataset is provided in the Appendix (Fig. 10) to show that all four features are important for achieving signal sensitivity. A fully unsupervised classifier using an autoencoder is also provided for comparison. The encoder and decoder have one hidden layer each, with 64 nodes/layer. The latent space has two dimensions. These parameters were chosen to approximately match the number of parameters of the other methods. Deeper networks were found to have worse performance. The autoencoder was trained on pure background in the signal region (an idealized setup) using the mean-squared error loss (MSE). The per-event MSE is used as an anomaly score for Fig. 4. In agreement with Ref. [35,78], the unsupervised approach is not effective for the signal studied in this paper.

A solid red line in Fig. 4 labeled "Optimal CWOLA" corresponds to a classifier trained using two mixed samples, one composed of pure background in the single region and the other composed of mostly background (independent from the first sample) in the SR with the 1000 signal events. The Optimal CWOLA line is far below the fully supervised classifier because the neural network needs to identify a small difference between the mixed samples over the natural statistical fluctuations in both sets. The actual CWOLA method is shown with a dotted red line. By construction, there is a significant difference between the phase space of the SR and SB and so the classifier is unable to identify the signal. At low efficiency, the CWOLA classifier actually antitags because the SR-SB differences are such that the signal is more SB-like then SR-like. Despite this drop in performance, the simulation augmenting modification (solid orange) with $\lambda = 1$ nearly recovers the full performance of Optimal CWOLA.

For comparison, a classifier trained using simulation directly is also presented in Fig. 4. The line labeled "Data vs Sim." directly trains a classifier to distinguish the data and simulation in the SR without reweighting. Due to the differences between the background in the data and the simulated background, this classifier is not effective. In fact, the signal is more like the background simulation than the data background and so the classifier is worse than random (it preferentially removes signal). The performance is significantly improved by adding in the parametrized reweighting, as advocated by Ref. [27]. With this reweighting, the SALAD classifier is significantly better than random and is comparable to SA-CWOLA. The Optimal CWOLA line also serves as the upper bound in performance for SALAD because it corresponds to the case where the background simulation is statistically identical to the background in the data. The means and standard deviations of these models are provided in the Appendix (Fig. 11).

The SA-CWOLA method has one free parameter that must be tuned. Figure 5 quantifies the performance of the SA-CWOLA classifier as a function of λ. The performance of SA-CWOLA is strong and relatively stable for $0.3 < \lambda < 0.6$. For $\lambda \gtrsim 0.2$, the classifier is effectively blinded to differences between the SR and SB as illustrated by the orange lines in Fig. 5 approaching 0.5 in the left plot.

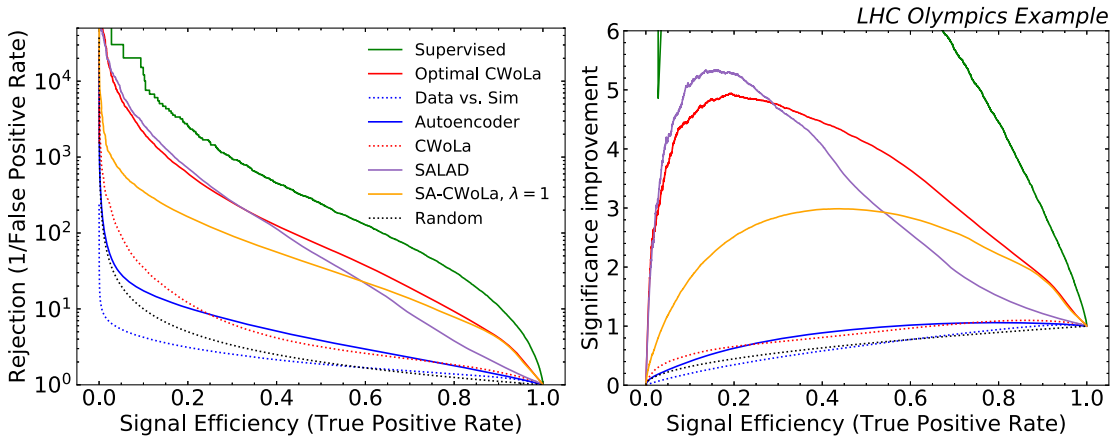

FIG. 4. An ROC curve (left) and significance improvement curve (right) for various anomaly detection methods described in the text. Curves represent the mean of 20 model performances. The significance improvement is defined as the ratio of the signal efficiency to the square root of the background efficiency. A significance improvement of 2 means that the initial significance would be amplified by about a factor of 2 after employing the anomaly detection strategy. The supervised line is unachievable unless there is no mismodeling and one designed a search for the specific W' signal used in this paper. The curve labeled "Random" corresponds to equal efficiency for signal and background.

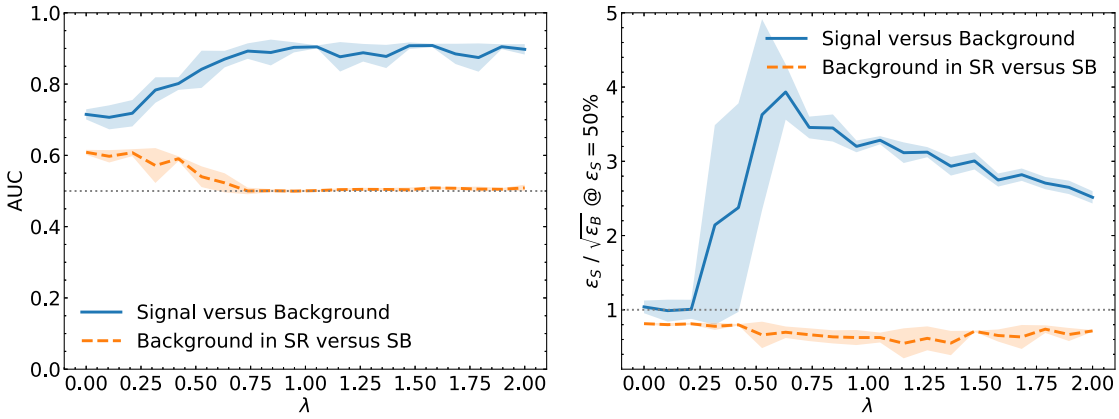

FIG. 5. The area under the ROC curve (AUC) (left) and the significance improvement at 50% signal efficiency (right) using the SA-CWOLA method for a scan in the hyperparameter λ introduced in Eq. (2.3). Each curve is an average over five classifiers, with the standard deviation displayed in the shaded region. When $\lambda = 0$, SA-CWOLA is the same as the original CWOLA method. For comparison, the performance of the classifier for distinguishing between the signal and background is shown in blue and the performance for distinguishing between the SR and SB is shown in orange. Ideally, the latter would have an AUC of 0.5.

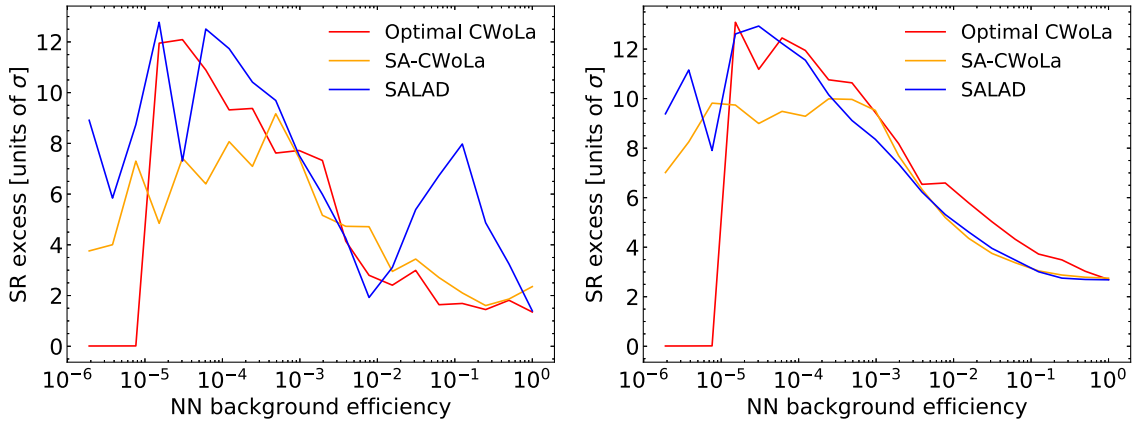

FIG. 6. Fit excess with signal injected using the statistical procedure described in Sec. IIC. For each method, event output scores have been averaged over five separate classifiers to improve stability. There is a small local deficit in the simulation. The left plot shows the fitted excess without modifying the background while the right plot corrects for the initial deficit by subtracting the residuals of the background-only fit before performing the signal + background fit. In the latter case, the significances are still not $S/\sqrt{B}$ due to the uncertainty from the sideband fit.

While the ROC and significance improvement curves are effective for quantifying performance, they do not communicate the complete story because they ignore the impact of background estimation. Figures 6 and 7 show the results of the sideband fit and statistical test [see Sec. IIC]. To improve stability, event output scores have been averaged over five classifiers.[2] The fit quality is excellent when considering all bins (see Fig. 8), but there happens to be a small local deficit in the SR. The right plot of Fig. 6 removes this effect by subtracting the fitted residuals in the background-only case for each value of the Neural Network (NN) background efficiency. The spectra after applying the nominal CWoLa classifier cannot be fit to the same shape and are thus not included—see Fig. 9.

V. CONCLUSIONS

This paper has investigated the impact of dependencies between m_{jj} and classification features for the resonant anomaly detection methods SALAD and CWOLA. It has been shown that while SALAD, because it does not compare events across different bins in m_{jj}, is relatively insensitive to the energy scaling of classifier features and is therefore naturally robust, CWOLA suffers both in classifier

[2]This was observed to be particularly important in the no-signal case, where SALAD and SA-CWOLA sometimes minimize their loss function by assigning the same score, near 0.5, to many events. In one test of SA-CWOLA, for example, the classifier assigned a value of 0.49911904 to 97.7% of events. High-efficiency cuts on the neural network output would then remove more events than desired.

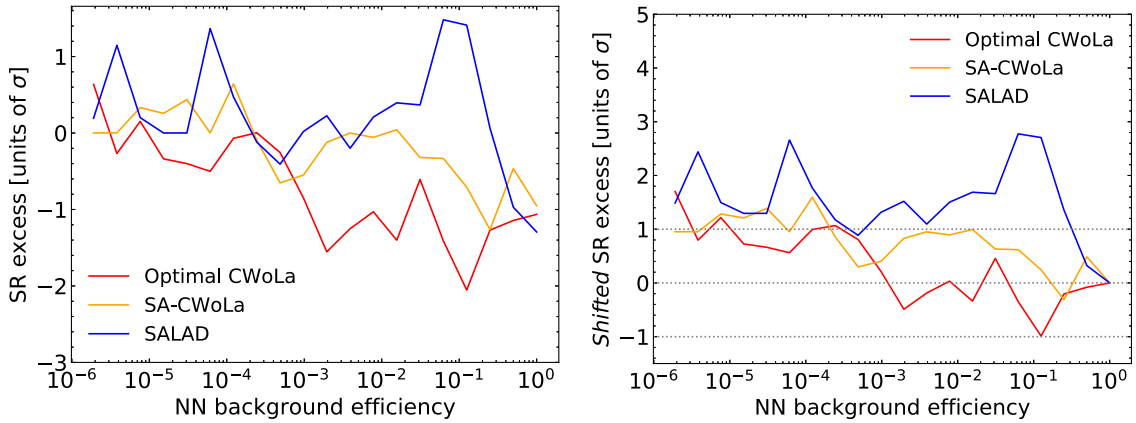

FIG. 7. Fit excess without signal injected using the statistical procedure described in Sec. IIC. For each method, event output scores have been averaged over five separate classifiers to improve stability. Without any signal injected, there is a small ($\sim 1\sigma$) deficit in the simulation. The right plot shifts the curves so that the 100% efficiency point corresponds to 0σ.

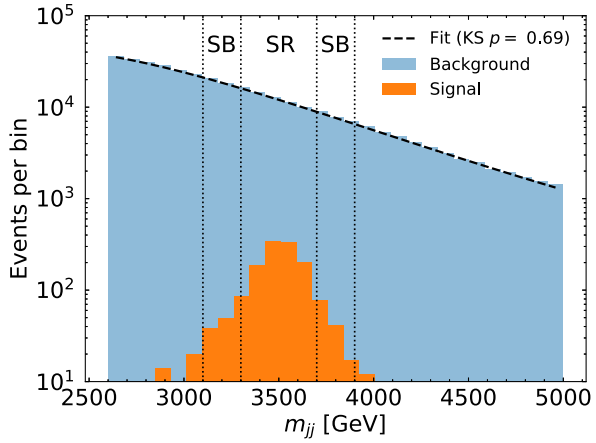

FIG. 8. A fit to the m_{jj} distribution in the background-only case with no selection on any neural networks. The 500 signal events used for training is super-imposed for illustration. Vertical dashed lines indicate the SR and SB regions used for training. A Kolmogorov-Smirnov test using only bins outside of the SR yields a p-value of 0.69.

performance and in the context of a resonance search when large dependencies are present. A new simulation-augmented approach has been proposed to remedy these challenges with the CWOLA method. This modification, called SA-CWOLA, makes use of physics encoded in simulations to automatically decorrelate classification features from m_{jj}. SA-CWOLA is shown to recover most of the performance of CWOLA from the ideal case where dependencies are ignored in the training. Both the SALAD and SA-CWOLA methods are able to exploit the physical priors of simulations without relying on simulations for background predictions. Thus, in these methods background-only simulations provide a critical tool for mitigating the sensitivity of the classifiers on dependencies between the resonant feature and the classifier features.

These weakly supervised methods are particularly promising, but they are not the only recently-proposed machine-learning-based anomaly detection methods. In particular, unsupervised methods also have great potential. The Anomaly Detection with Density Estimation (ANODE)

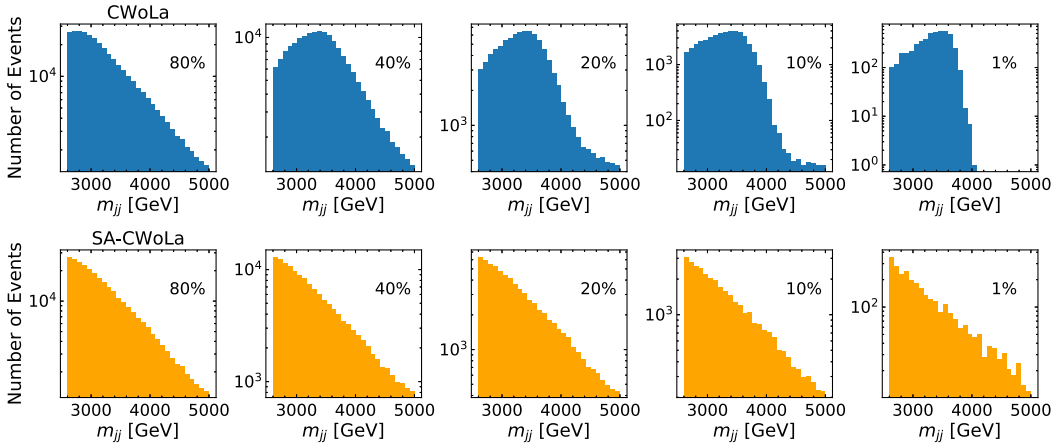

FIG. 9. Histograms of m_{jj} for CWOLA (top) and SA-CWOLA (bottom) for various thresholds on the classifiers in the background-only case.

[28] does not use simulation at all and has been shown to be relatively robust to dependencies between the resonant feature and the classifier features. Additionally, autoencoder (AE) methods have been combined with explicit decorrelation to build in robustness to such dependencies [18]. Since AEs do not use the presence of a signal to obtain signal sensitivity, it is expected that decorrelation is not useful for achieving signal sensitivity, only for estimating the background. Preliminary studies in Ref. [35] (e.g., Fig. 2, top right) indicate that the corresponding AE performance for signals similar to the ones probed in this paper are worse than the weakly supervised techniques. It will be interesting and important to perform a detailed and robust comparison between weakly supervised and unsupervised methods in the future.

Each of these unsupervised and semisupervised methods have advantages and weaknesses and it is likely that multiple approaches will be required to achieve broad sensitivity to BSM physics. Therefore, it is critical to study the sensitivity of each technique to dependencies and propose modifications where possible to build robustness. This paper is an important step in the decorrelation program for automated anomaly detection with machine learning. Tools like the ones proposed here may empower higher-dimensional versions of the existing ATLAS search [15] as well as other related searches by other experiments in the near future.

CODE AND DATA

The code for this paper can be found at https://github.com/bnachman/DCTRHunting and the simulated data are available from the LHC Olympics [66].

ACKNOWLEDGMENTS

B. N. would like to thank Jack Collins for useful discussions and Jesse Thaler for helpful feedback on the manuscript. This work was supported by the Department of Energy, Office of Science under contract No. DE-AC02-05CH11231. K. B. was supported in part by NSF PHY REU Grant No. 1949923. L. L. P. was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI). B. N. would like to thank NVIDIA for providing Volta GPUs for neural network training.

APPENDIX: ADDITIONAL LHC OLYMPICS PERFORMANCE PLOTS

Here we present plots to give more information regarding the classification performance discussed in Sec. IV.

Figure 10 shows the performances of fully supervised classifiers trained on each input variable individually (indicated by the blue and orange lines) and on the combined set of inputs (indicated by the solid green line). Additionally, the right plot shows the performance when one variable is removed at a time. As expected, all combined variables provide significantly more information than any individual variable. Notable on this plot is that the features which provide information on the leading jet, $\tau_{2,1}^{J_2}$ and $m^{J_2} - m^{J_1}$, prove to be significantly better classifiers than those related to the subleading jet.

Figure 11 shows the mean and standard deviation of the models discussed in Sec. IV, over the course of 20 models. Particularly notable is the small variation on the SA-CWOLA method.

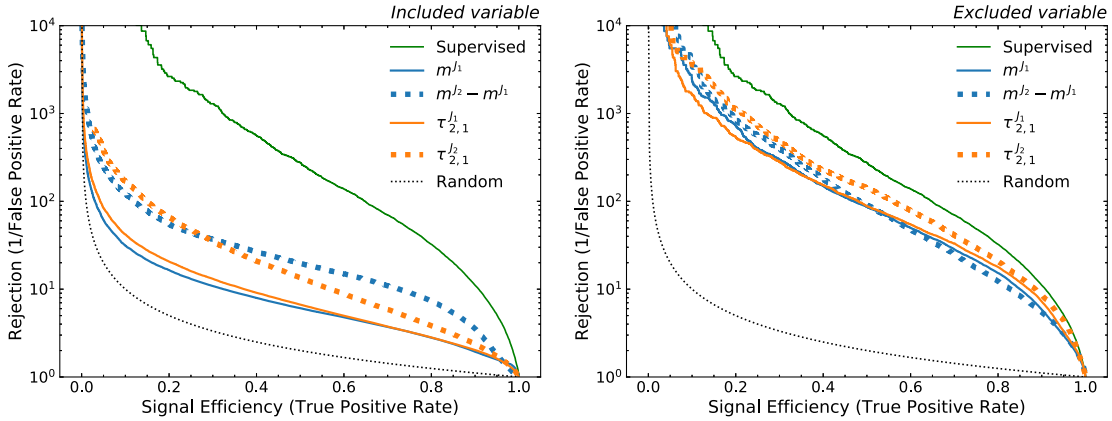

FIG. 10. ROC curves for fully supervised classifiers trained on the LHC Olympics dataset. Models trained only on the variable are shown at left, and models trained excluding the indicated variable are shown at right.

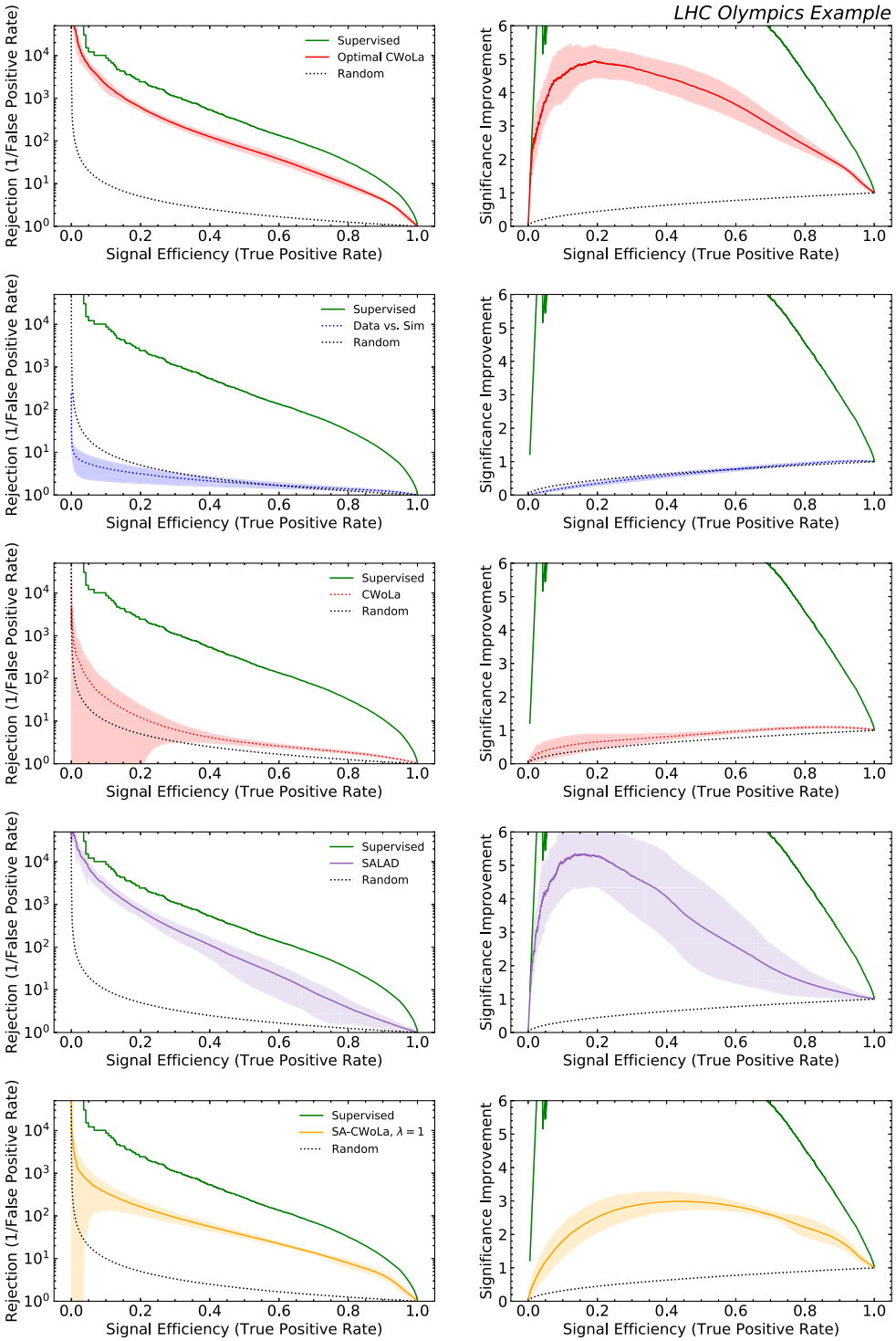

FIG. 11. Spread of models from Sec. IV, in mean and standard deviation.

[1] ATLAS Collaboration, Exotic physics searches (2018), https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults.

[2] ATLAS Collaboration, Supersymmetry searches (2018), https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults.

[3] ATLAS Collaboration, Higgs and Diboson Searches, 2019, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/HDBS.

[4] CMS Collaboration, Cms exotica public physics results, 2018, https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO.

[5] CMS Collaboration, Cms supersymmetry physics results, 2018, https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS.

[6] CMS Collaboration, Cms beyond-two-generations (b2g) public physics results, 2018, https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G.

[7] LHCb Collaboration, Publications of the QCD, Electroweak and Exotica Working Group, 2019, http://lhcbproject.web.cern.ch/lhcbproject/Publications/LHCbProjectPublic/Summary_QEE.html.

[8] A. J. Larkoski, I. Moult, and B. Nachman, Jet substructure at the large hadron collider: A review of recent advances in theory and machine learning, Phys. Rep. **841**, 1 (2020).

[9] D. Guest, K. Cranmer, and D. Whiteson, Deep learning and its application to LHC physics, Annu. Rev. Nucl. Part. Sci. **68**, 161 (2018).

[10] M. Abdughani, J. Ren, L. Wu, J. M. Yang, and J. Zhao, Supervised deep learning in high energy phenomenology: A mini review, Commun. Theor. Phys. **71**, 955 (2019).

[11] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao, and T. Wongjirad, Machine learning at the energy and intensity frontiers of particle physics, Nature (London) **560**, 41 (2018).

[12] R. T. D'Agnolo and A. Wulzer, Learning new physics from a machine, Phys. Rev. D **99**, 015014 (2019).

[13] J. H. Collins, K. Howe, and B. Nachman, Anomaly Detection for Resonant New Physics with Machine Learning, Phys. Rev. Lett. **121**, 241803 (2018).

[14] J. H. Collins, K. Howe, and B. Nachman, Extending the search for new resonances with machine learning, Phys. Rev. D **99**, 014038 (2019).

[15] ATLAS Collaboration, Dijet Resonance Search with Weak Supervision Using 13 TeV pp Collisions in the ATLAS Detector, Phys. Rev. Lett. **125**, 131801 (2020).

[16] R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, Learning multivariate new physics, Eur. Phys. J. C **81**, 89 (2021).

[17] M. Farina, Y. Nakai, and D. Shih, Searching for new physics with deep autoencoders, Phys. Rev. D **101**, 075021 (2020).

[18] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or what?, SciPost Phys. **6**, 030 (2019).

[19] T. S. Roy and A. H. Vijay, A robust anomaly finder based on autoencoder, arXiv:1903.02032.

[20] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Variational autoencoders for new physics mining at the large hadron collider, J. High Energy Phys. 05 (2019) 036.

[21] A. Blance, M. Spannowsky, and P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches, J. High Energy Phys. 10 (2019) 047.

[22] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, Novelty detection meets collider physics, Phys. Rev. D **101**, 076015 (2020).

[23] A. De Simone and T. Jacques, Guiding new physics searches with unsupervised learning, Eur. Phys. J. C **79**, 289 (2019).

[24] A. Mullin, H. Pacey, M. Parker, M. White, and S. Williams, Does SUSY have friends? A new approach for LHC event analysis, J. High Energy Phys. 02 (2021) 160.

[25] G. M. A. Casa, Nonparametric semisupervised classification for signal detection in high energy physics, arXiv:1809.02977.

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

[27] A. Andreassen, B. Nachman, and D. Shih, Simulation assisted likelihood-free anomaly detection, Phys. Rev. D **101**, 095004 (2020).

[28] B. Nachman and D. Shih, Anomaly detection with density estimation, Phys. Rev. D **101**, 075042 (2020).

[29] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, A generic anti-QCD jet tagger, J. High Energy Phys. 11 (2017) 163.

[30] M. R. Crispim, N. Castro, R. Pedro, and T. Vale, Transferability of deep learning models in searches for new physics at colliders, Phys. Rev. D **101**, 035042 (2020).

[31] M. C. Romao, N. Castro, J. Milhano, R. Pedro, and T. Vale, Use of a generalized energy Mover's distance in the search for rare phenomena at colliders, Eur. Phys. J. C **81**, 192 (2021).

[32] O. Knapp, G. Dissertori, O. Cerri, T. Q. Nguyen, J.-R. Vlimant, and M. Pierini, Adversarially learned anomaly detection on CMS open data: Re-discovering the top quark, Eur. Phys. J. Plus **136**, 236 (2021).

[33] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc, Learning the latent structure of collider events, J. High Energy Phys. 10 (2020) 206.

[34] M. C. Romao, N. Castro, and R. Pedro, Finding new physics without learning about it: Anomaly detection as a tool for searches at colliders, Eur. Phys. J. C **81**, 27 (2021).

[35] O. Amram and C. M. Suarez, Tag N' train: A technique to train improved classifiers on unlabeled data, J. High Energy Phys. 01 (2021) 153.

[36] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, and T. Golling, Variational autoencoders for anomalous jet tagging, arXiv:2007.01850.

[37] C. K. Khosa and V. Sanz, Anomaly awareness, arXiv:2007.14462.

[38] P. Thaprasop, K. Zhou, J. Steinheimer, and C. Herold, Unsupervised outlier detection in heavy-ion collisions, Phys. Scr. **96**, 064003 (2021).

[39] J. Aguilar-Saavedra, F. Joaquim, and J. Seabra, Mass unspecific supervised tagging (MUST) for boosted jets, J. High Energy Phys. 03 (2021) 012.

[40] J. Aguilar-Saavedra and B. Zaldívar, Jet tagging made easy, Eur. Phys. J. C **80**, 530 (2020).

[41] G. Louppe, M. Kagan, and K. Cranmer, Learning to pivot with adversarial networks, Adv. Neural Inf. Process. Syst. **30**, 981 (2017).

[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] J. Stevens and M. Williams, uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers, J. Instrum. **8**, P12013 (2013).
[45] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Sgaard, Decorrelated jet substructure tagging using adversarial neural networks, Phys. Rev. D **96**, 074034 (2017).
[46] L. Bradshaw, R. K. Mishra, A. Mitridate, and B. Ostdiek, Mass agnostic jet taggers, SciPost Phys. **8**, 011 (2020).
[47] ATLAS Collaboration, Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS, Report No. ATL-PHYS-PUB-2018-014 (2018), ēhttp://cds.cern.ch/record/2630973.
[48] L.-G. Xia, QBDT, a new boosting decision tree method with systematical uncertainties into training for high energy physics, Nucl. Instrum. Methods Phys. Res., Sect. A **930**, 15 (2019).
[49] C. Englert, P. Galler, P. Harris, and M. Spannowsky, Machine Learning Uncertainties with Adversarial Neural Networks, Eur. Phys. J. C **79**, 4 (2019).
[50] 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).
[51] G. Kasieczka and D. Shih, DisCo Fever: Robust Networks Through Distance Correlation, Phys. Rev. Lett. **125**, 122001 (2020).
[52] G. Kasieczka, M. Schwartz, D. Shih, and B. Nachman, ABCDisCo: Automating the ABCD method with machine learning, Phys. Rev. D **103**, 035021 (2021).
[53] E. M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.
[54] F. Chollet, Keras, https://github.com/fchollet/keras, 2017.
[55] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, Tensorflow: A system for large-scale machine learning, in *OSDI* (2016), Vol. 16, pp. 265–283.
[56] D. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.
[57] A. Andreassen and B. Nachman, Neural networks for full phase-space reweighting and parameter tuning, Phys. Rev. D **101**, 091901 (2020).
[58] K. Cranmer, J. Pavez, and G. Louppe, Approximating likelihood ratios with calibrated discriminative classifiers, arXiv:1506.02169.
[59] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, Parameterized neural networks for high-energy physics, Eur. Phys. J. C **76**, 235 (2016).
[60] A. M. Sirunyan *et al.* (CMS Collaboration), Search for narrow and broad dijet resonances in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on dark matter mediators and other new particles, J. High Energy Phys. 08 (2018) 130.
[61] G. Aad *et al.* (ATLAS Collaboration), Search for new resonances in mass distributions of jet pairs using 139 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, J. High Energy Phys. 03 (2020) 145.
[62] M. Frate, K. Cranmer, S. Kalia, A. Vandenberg-Rodes, and D. Whiteson, Modeling smooth backgrounds and generic localized signals with Gaussian processes, arXiv:1709.05681.
[63] S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. **9**, 60 (1938).
[64] A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Am. Math. Soc. **54**, 426 (1943).
[65] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur. Phys. J. C **71**, 1554 (2011); , Erratum, Eur. Phys. J. C **73**, 2501 (2013).
[66] G. Kasieczka, B. Nachman, and D. Shih, R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge, 2019, https://doi.org/10.5281/zenodo.2629073.
[67] T. Sjöstrand, S. Mrenna, and P. Z. Skands, PYTHIA6.4 physics and manual, J. High Energy Phys. 05 (2006) 026.
[68] T. Sjostrand, S. Mrenna, and P. Z. Skands, A brief introduction to PYTHIA8.1, Comput. Phys. Commun. **178**, 852 (2008).
[69] M. Bahr *et al.*, Herwig++ physics and manual, Eur. Phys. J. C **58**, 639 (2008).
[70] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lematre, A. Mertens, and M. Selvaggi, DELPHES 3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.
[71] A. Mertens, New features in Delphes 3, J. Phys. Conf. Ser. **608**, 012045 (2015).
[72] M. Selvaggi, DELPHES 3: A modular framework for fast-simulation of generic collider experiments, J. Phys. Conf. Ser. **523**, 012033 (2014).
[73] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, Eur. Phys. J. C **72**, 1896 (2012).
[74] M. Cacciari and G. P. Salam, Dispelling the N^3 myth for the k_t jet-finder, Phys. Lett. B **641**, 57 (2006).
[75] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k(t) jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[76] J. Thaler and K. Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness, J. High Energy Phys. 02 (2012) 093.
[77] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.
[78] J. H. Collins, P. Martín-Ramiro, B. Nachman, and D. Shih, Comparing weak- and unsupervised methods for resonant anomaly detection, arXiv:2104.02092.