Extending the search for new resonances with machine learning

Jack H. Collins,[1,2,*] Kiel Howe,[3,†] and Benjamin Nachman[4,5,‡]

[1]*Maryland Center for Fundamental Physics, Department of Physics, University of Maryland, College Park, Maryland 20742, USA*

[2]*Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA*

[3]*Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA*

[4]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

[5]*Simons Institute for the Theory of Computing, University of California, Berkeley, Berkeley, California 94720, USA*

(Received 30 May 2018; revised manuscript received 21 November 2018; published 28 January 2019)

The oldest and most robust technique to search for new particles is to look for "bumps" in invariant mass spectra over smoothly falling backgrounds. We present a new extension of the bump hunt that naturally benefits from modern machine learning algorithms while remaining model agnostic. This approach is based on the classification without labels (CWoLa) method where the invariant mass is used to create two potentially mixed samples, one with little or no signal and one with a potential resonance. Additional features that are uncorrelated with the invariant mass can be used for training the classifier. Given the lack of new physics signals at the Large Hadron Collider (LHC), such model-agnostic approaches are critical for ensuring full coverage to fully exploit the rich datasets from the LHC experiments. In addition to illustrating how the new method works in simple test cases, we demonstrate the power of the extended bump hunt on a realistic all-hadronic resonance search in a channel that would not be covered with existing techniques.

DOI: 10.1103/PhysRevD.99.014038

I. INTRODUCTION

Searching for new resonances as bumps in the invariant mass spectrum of the new particle decay products is one of the oldest and most robust techniques in particle physics, from the ρ meson discovery [1] and earlier up through the recent Higgs boson discovery [2,3]. This technique is very powerful because sharp structures in invariant mass spectra are not common in background processes, which tend to produce smooth distributions. As a result, the background can be estimated directly from data by fitting a shape in a region away from the resonance (sideband) and then extrapolating to the signal region. It is often the case that the potential resonance mass is not known *a priori* and a technique like the BumpHunter [4] is used to scan the invariant mass distribution for a resonance. In some cases, the objects used to construct the invariant mass (e.g., jet substructure) and their surroundings (e.g., presence of additional forward jets) have properties that can be used to increase the signal purity. Both ATLAS and CMS[1] have conducted extensive searches for resonances decaying into jets originating from generic quarks and gluons [5–7], from boosted W [8,9], Z [10], Z' [11–13] or Higgs bosons [14–16], from b-quarks [17], as well as from boosted top quarks [18,19]. There is some overlapping sensitivity in these searches, but in general the sensitivity is greatly diminished away from the target process (see e.g., [20–22] for examples). It is not feasible to perform a dedicated analysis for every possible topology and so some signals may be missed. Global searches for new physics have been performed by the LHC experiments and their predecessors, but only utilize simple objects and rely heavily on simulation for background estimation [23–34].

The tagging techniques used to isolate different jet types have increased in sophistication with the advent of modern machine learning classifiers [35–58]. These new algorithms can use all of the available information to achieve optimal classification performance and could significantly improve the power of hadronic resonance searches. Deep learning

*jhc296@umd.edu
†khowe@fnal.gov
‡bpnachman@lbl.gov

[1]The references here are the run 2 results; run 1 results can be found within the cited papers. The techniques described in this paper also apply to leptonic or photonic final states, but jets are used as a prototypical example due to their inherent complex structure.

techniques are able to outperform traditional methods by exploiting subtle correlations in the radiation pattern inside jets. These correlations are not well modeled in general [40] which renders classifiers suboptimal when training on simulation and testing on data. This is already apparent for existing multivariate classifiers where *post hoc* mismodeling corrections can be large [59–67]. Ideally, one would learn directly from data (if possible) and/or combine with other approaches to mitigate potential mismodeling effects during training (e.g., with adversaries [68]).

We propose a new method that combines resonance searches with recently proposed techniques for learning directly from data [69–72]. Simply stated, the new algorithm trains a fully supervised classifier to distinguish a signal region from a mass sideband using auxiliary observables which are decorrelated from the resonance variable under the background-only hypothesis. A bump hunt is then performed on the mass distribution after applying a threshold on the classifier output. This is classification without labels (CWoLa) [70] where the two mixed samples are the signal region and sideband and the signal is a potential new resonance and the background is the Standard Model continuum. The algorithm naturally inherits the property of CWoLa that it is fully based on data and thus is insensitive to simulation mis-modeling.[2] The key difference with respect to Refs. [70,71] is that the signal process need not be known *a priori*. Therefore, we can become sensitive to new signatures for which we did not think to construct dedicated searches.

In addition to CWoLa, the extended bump hunt shares some features with the sPlot technique [73]. Our proposed extension to the bump hunt makes use of auxiliary features to enhance the presence of signal events over background events in a target distribution, where the signal is expected to be resonant. Similarly, sPlot provides a procedure for using auxiliary features ("discriminating variables" in the language of Ref. [73]) to extract the distribution of signal and background events in a target distribution ("control variable" in Ref. [73]). In both cases, the auxiliary features must be uncorrelated with the target feature. One main difference between the methods is that the extended bump hunt uses machine learning to identify regions of phase space that are signal-like. A second key distinction between methods is that sPlot takes the distribution of the auxiliary features as input, whereas this information is not required for the extended bump hunt.

This paper is organized as follows. Section II formally introduces the CWoLa hunting approach and briefly discusses how auxiliary information can be useful for bump hunting. Then, Sec. III uses a simplified example to show how a neural network can be used to identify new physics structures from pseudodata. A complete procedure for applying the CWoLa hunting approach is given in Sec. IV. Finally, a realistic example based on a hadronic resonance search is presented in Sec. V. Conclusions and the future outlook are presented in Sec. VI.

[2]The algorithm also inherits the assumptions of the CWoLa method. In this context, the main assumption will be that the signal region and the sideband region can only be distinguished with the mass. More details on this are in the next sections.

II. BUMP HUNTING USING CLASSIFICATION WITHOUT LABELS

In a typical resonance search, events have at least two objects whose four-vectors are used to construct an invariant mass spectrum. The structure of these objects as well as other information in the event may be useful for distinguishing signal from background even though there may be no other resonance structures. Let m_{res} be a random variable that represents the invariant mass. The distribution of m_{res} given background is smooth while m_{res} given signal is expected to be localized near some m_0. Let Y be another random variable that represents all other information available in the events of interest. Define two sets of events:

$$M_1 = \{(m_{\text{res}}, Y) | |m_{\text{res}} - m_0| < \delta\} \quad \text{(the signal region)} \tag{2.1}$$

$$M_2 = \{(m_{\text{res}}, Y) | \delta < |m_{\text{res}} - m_0| < \epsilon\} \quad \text{(the sideband region)}, \tag{2.2}$$

where $\epsilon > \delta$. The value of δ is chosen such that M_1 should have much more signal than M_2 and the value of ϵ is chosen such that the distribution of Y is nearly the same between M_1 and M_2. CWoLa hunting entails training a classifier to distinguish M_1 from M_2 using Y and then performing a usual bump hunt on m_{res} after placing a threshold on the classifier output. This procedure is then repeated for all mass hypotheses m_0. Note that nothing is assumed about the distribution of Y other than that it should be nearly the same for M_1 and M_2 under the background-only hypothesis.

Ideally, Y incorporates as much information as possible about the properties of the objects used to construct the invariant mass and their surroundings. The subsequent sections will show how this can be achieved with neural networks. To build intuition for the power of auxiliary information, the rest of this section provides analytic scaling results for a simplified bump hunt with the most basic case: $Y \in \{0, 1\}$.

Suppose that we have two mass bins M_1 and M_2 and the number of expected events in each mass bin is N_b. Further suppose that the signal is in at most one of the M_i (not required in general) and the expected number of signal events is N_s. A version of the bump hunt would be to compare the number of events in M_1 and M_2 to see if they are significantly different. As a Bernoulli random variable, Y is uniquely specified by $\Pr(Y = 1)$. Define $\Pr(Y = 1|\text{background}) = p$ and $\Pr(Y = 1|\text{signal}) = q$.

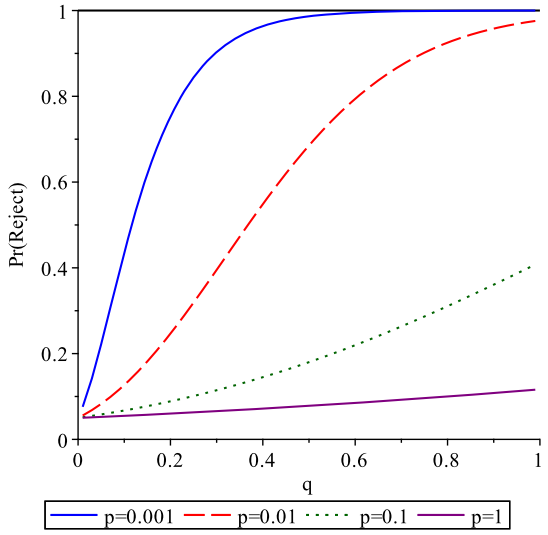

FIG. 1. The probability to reject the SM as a function of q for fixed values of p as indicated in the legend when only considering events with $Y = 1$. The expected background is fixed at 1000 and the expected BSM is fixed at 20. When $q = 0$, there is no signal and therefore the rejection probability is 5%, by construction. When $q = p = 1$, Y is not useful, but the probability to reject is above 5% simply because there is an excess of events inclusively.

The purpose of CWoLa hunting is to incorporate the information about Y into the bump hunt. By only considering events with $Y = 1$, the significance of the signal scales as $qN_s/\sqrt{N_b p}$. Therefore, the information about Y is useful when $q > \sqrt{p}$.

More quantitatively, suppose that we declare discovery of new physics when the number of events with $Y = 1$ in M_1 exceeds the number of events with $Y = 1$ in M_2 by some amount. Under the background-only case, for $N_b \gg 1$, the difference between the number of events in M_1 and M_2 with $Y = 1$ is approximately normally distributed with mean 0 and variance $2N_b p$. If we want the probability for a false positive to be less than 5%, then the threshold value is simply $\sqrt{2N_b p} \times \Phi^{-1}(0.95)$, where Φ is the cumulative distribution function of a standard normal distribution. Ideally, we would like to reject the SM often when there is beyond the SM (BSM), $N_s > 0$. Figure 1 shows the probability to reject the SM for a one-bin search using $N_b = 1000$ and $N_s = 20$ for different values of p as a function of q. The case $p = q = 1$ corresponds to the standard search that does not gain from having additional information. However, away from this case, there can be a significant gain from using Y, especially when p is small and q is close to 1. In the case where Y is a *truth bit*, i.e., $p = 1 - q = 0$, the SM is rejected as long as a single BSM event is observed. By construction, when $q \to 0$ (for $p > 0$), the rejection probability is 0.05. Note that when $q < p$, only considering events with $Y = 1$ is suboptimal—this is a feature that is corrected in the full CWoLa hunting approach.

While the model used here is simple, it captures the key promise of CWoLa hunting that will be expanded upon in more detail in the next sections. In particular, the main questions to address are: how to find Y and how to use the information about Y once it is identified.

III. ILLUSTRATIVE EXAMPLE: LEARNING TO FIND AUXILIARY INFORMATION

This section shows how to identify the useful attributes of the auxiliary information using a neural network. The example used here is closer to a realistic case, but is still simplified for illustration. Let the auxiliary information $Y = (x, y)$ be two-dimensional and assume that Y and the invariant mass are independent given the process (signal or background). This auxiliary information will become the jet substructure observables in the next section. For simplicity, for each process Y is considered to be uniformly distributed on a square of side length ℓ centered at the origin. The background has $\ell = 1$ $(-0.5 < x < 0.5, -0.5 < y < 0.5)$ and the signal follows $\ell = w$ $(-w/2 < x < w/2, -w/2 < y < w/2)$. Similarly to the full case, suppose that there are three bins of mass for a given mass hypothesis: a signal region $m_0 \pm \Delta$ and mass sidebands $(m_0 - 2\Delta, m_0 - \Delta)$, $(m_0 + \Delta, m_0 + 2\Delta)$. As in the last section, the signal is assumed to only be present in one bin (the signal region) with N_s expected events. There are N_b expected background events in the signal region and $N_b/2$ expected events in each of the mass sidebands.

The model setup described above and used for the rest of this section is depicted in Fig. 2. The numerical examples presented below use $N_b = 10,000$, $N_s = 300$, and $w = 0.2$. Without using Y, these values correspond to $N_s/\sqrt{N_b} = 3\sigma$. The ideal tagger (one that is optimal by the Neyman-Pearson lemma [74]) should reject all events outside of the square in the (x, y) plane centered at zero with side length w. For the N_s and N_b used here, the expected significance of the ideal tagger is 15σ. The goal of this section is to show that without using any truth information, the CWoLa approach can recover much of the discriminating power from a neural network trained in the (x, y) plane. Note that optimal classifier is simply given by thresholding the likelihood ratio [74] $p_s(Y)/p_b(Y)$; in this two-dimensional case it is possible to provide an accurate approximation to this classifier without neural networks. However, these approximations often do not scale well with the dimensionality and will thus be less useful for the realistic example presented in the next section. This is illustrated in the context of the CWoLa hunting in Fig. 3.

To perform CWoLa hunting, a neural network is trained on (x, y) values to distinguish events in the mass sidebands from the signal region. Due to the simple nature of the example, it is also possible to easily visualize what the network is learning. A fully connected feed-forward network is trained using the PYTHON deep learning library

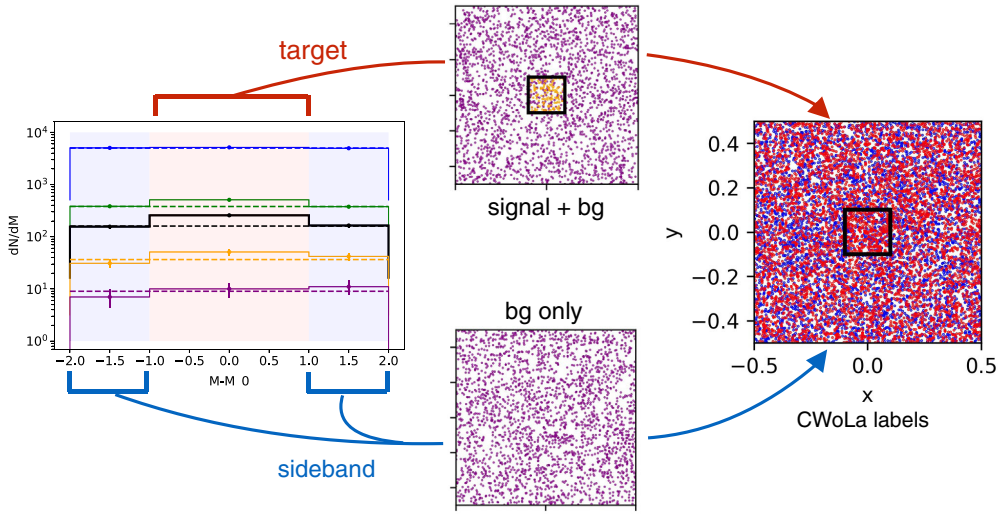

FIG. 2. An illustration of the CWoLa procedure for the simple two-dimensional uniform example presented in Sec. III. The left plot shows the mass distribution for the three mass bins, which is uniform for the background. The blue line is the total number of events and the other lines represent thresholds on various neural networks described in the text leading up to Fig. 5. The center plots show the (x, y) distribution for the events in each mass bin with truth labels (purple for background and yellow for signal). The black square is the true signal region for this example model, with signal distributed uniformly inside. The right plot shows the combined distribution in the (x, y) plane with CWoLa labels that can be used to train a classifier even without any truth-level information (red for target window, blue sideband).

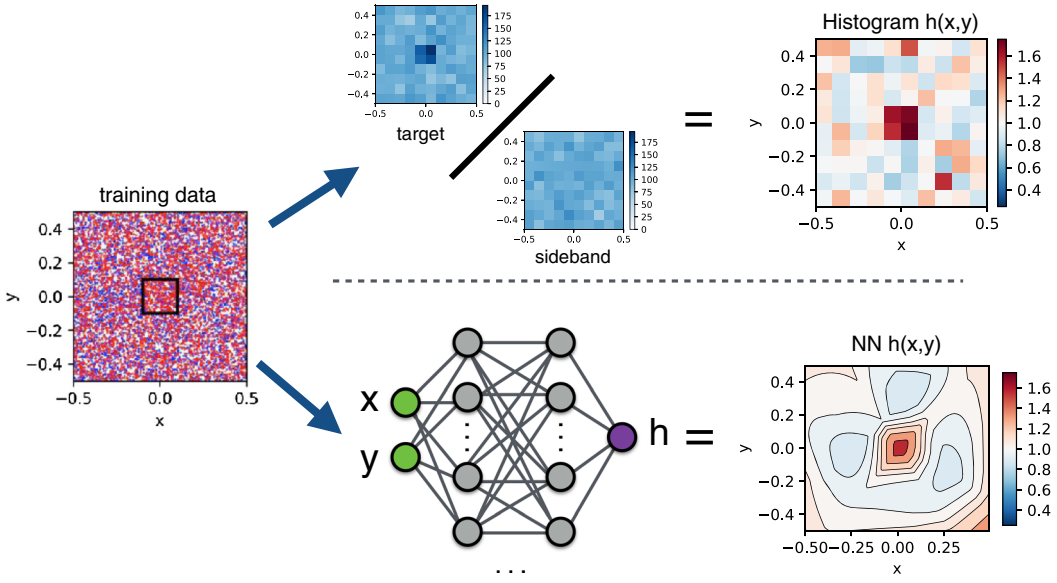

FIG. 3. The CWoLa-labeled data can be used to construct an estimate for the optimal classifier $h(x, y) = \frac{p_b(x,y)+p_s(x,y)}{p_b(x,y)}$. The top path shows an estimate constructed by histogramming the observed training events in the (x, y) plane. The bottom path shows an estimate constructed by using a neural network trained as described in the text, which can be efficiently generalized to higher dimensional distributions. The optimal classifier would be 1 outside of the small box centered at the origin and 1.75 inside the box.

KERAS [75] with a TENSORFLOW [76] backend. The network has three hidden layers with (256, 256, 64) nodes. The network was trained with the categorical cross-entropy loss function using the ADAM algorithm [77] with a learning rate of 0.003 and a batch size of 1024. The data are split into three equal sets, one used for training, one for validation, and one for testing. The training is terminated based on the efficiency of the signal region cut on the validation data at a fixed false-positive rate of 2% for the sideband data. If it fails to improve for 60 epochs, the training is halted and the network reverts to the last epoch for which there was a training improvement. This simple scheme is robust against enhancing statistical fluctuations but reduces the number of events used for the final search

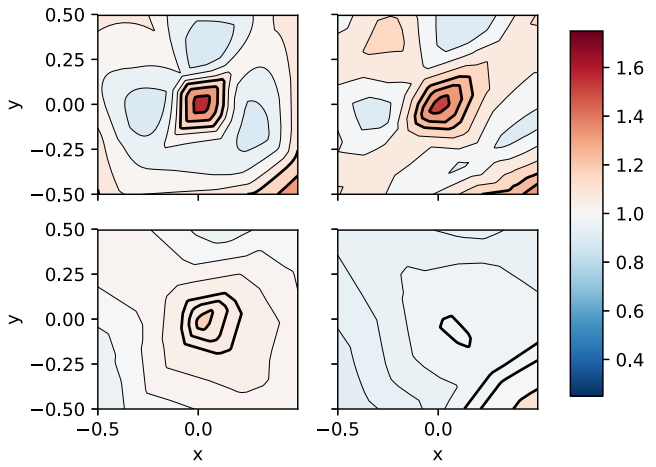

FIG. 4. The classifier $h(x, y)$ constructed from four independent training runs on the same example two-dimensional model dataset described in the text. The thick contours represent the cuts that would reduce the events in the test data target window by a factor of $\epsilon_{\text{test}} = 10\%, 5\%, 1\%$.

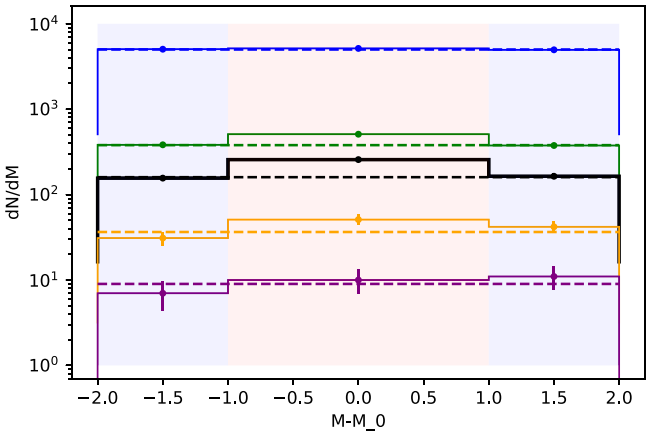

FIG. 5. The mass distribution after various threshold on the neural network classifier. The flat background fit from the sideband regions are the dashed lines, and the statistical uncertainty for each bin is shown by the error bars. The top histogram is the model before any threshold in the (x, y) plane, and from top to bottom respectively the histogram is given for efficiency thresholds of 10%, 5%, 1%, 0.2%. The significance is $\mathcal{S} = 3\sigma, 9.4\sigma, 10.8\sigma$, and 3.4σ for respectively no threshold, 10%, 5%, and 1%. The 0.2% threshold reduces the signal to no statistical significance.

by a factor of 3 as only the classifier output on the test set is used for the bump hunt. In the physical example described later, a more complicated scheme maximizes the statistical power of the available data.

Visualizations of the neural network trained as described above are presented in Fig. 4. In the top two examples, the network finds the signal region and correctly estimates the magnitude of the likelihood ratio. In both these cases, the network also overtrains on a (real) fluctuation in the training data, despite the validation procedure. Such regions will tend to decrease the effectiveness of the classifier, since a given cut threshold will admit more background in the test data. In the bottom left example of Fig. 4, the network finds a function approximately monotonic to $h(x, y)$ but with different normalization—while the cost function would have preferred to optimize this network to reach $h(x, y)$, the validation procedure cut off the optimization when the correct shape to isolate the signal region had been found. Due to the nature of the cuts, there is no performance loss for this network, since crucially it has found the correct shape near the signal region. The last network fails to converge to the signal region, and instead focuses its attention on the fluctuation in the training data. The variation in the network performance illustrates the importance of training multiple classifiers and using schemes to mitigate the impact of statistical fluctuations in the training dataset.

Figure 5 shows the mass distribution in the three bins after applying successfully tighter threshold on the neural network output. Since Y is not a truth bit, the data are reduced in both the signal region and the mass sidebands. For each threshold, the background expectation $\hat{n}_b$ assuming a uniform distribution is estimated by fitting a straight line to the mass sidebands. Then, the significance is estimated from the number of observed events in the signal region, n_o, via $\mathcal{S} \approx (n_o - \hat{n}_b)/\sqrt{\hat{n}_b}$. Of the threshold presented, the maximum significance corresponds to the 5% efficiency with $\mathcal{S} \approx 10.8\sigma$. Even though the ideal significance is 15σ, for the particular pseudodataset shown in Fig. 5, the ideal classifier significance is 13.9σ.

We can study the behavior of our neural network (NN) classifiers by looking at the significance generated by ensembles of models trained on signals of different strength, as shown in Fig. 6. The top histogram shows the significance for an ensemble of models trained on the example signal (blue) and on a control dataset with no signal (green). The control ensemble appears to be normally distributed around $s/\sqrt{b} = 0$, while the example signal ensemble is approximately normally distributed around 12σ (compared to 13.9σ for the ideal cut), along with a small $O(5\%)$ population of networks that fail to find the signal. The middle histogram shows the effect of decreasing the size of the signal region w_s while modifying N_s to maintain an expected significance of 15σ with ideal cuts. When w_s is decreased, the training procedure appears to have a harder time picking up the signal, possibly due to our choice of an operating point of 2% false-positive rate for the sideband validation. For $w_s = 0.1$ (green), about 50% of the networks effectively find the signal. For $w_s = 0.05$ (red), only about 5% find the signal. The bottom plot shows the effect of increasing w_s while keeping N_s fixed, so that the strength of the signal decreases. When the size of the signal region is doubled to $w_s = 0.4$ (green), giving an expected significance of 7.5σ, the network performs similarly to the $w_s = 0.2$ example (blue). When the signal distribution is identical to the background distribution

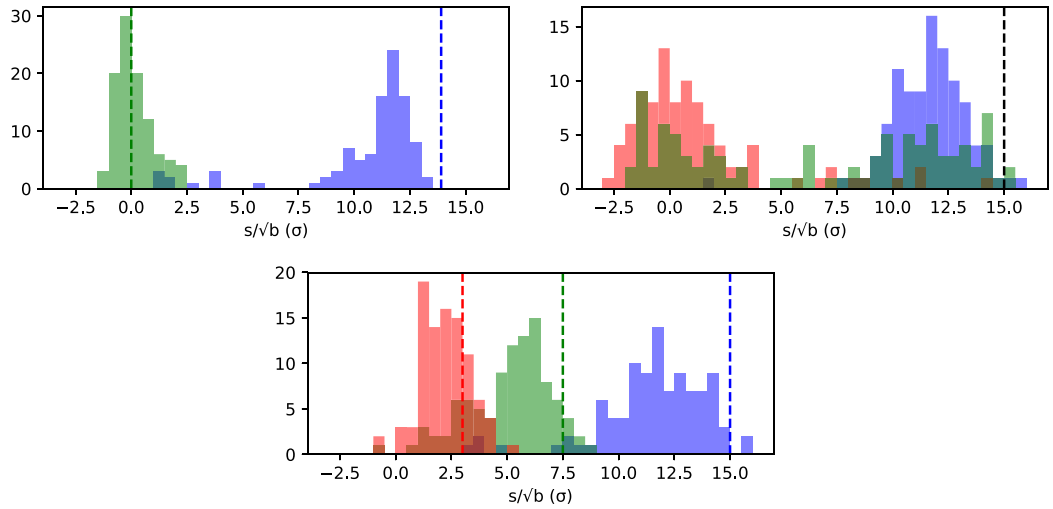

FIG. 6. Top left: Histogram of significance at a test threshold of 6% for 100 NN trained on the example toy model data (blue), and 100 NN trained on a control dataset with no signal present (green). The dashed green line gives the expected significance of 13.9σ for the example dataset with ideal cuts. Top right: Histogram of significance for ensembles with expected signal strength of 15σ with ideal cuts. The blue is $(N_b = 10000, N_s = 300, w_s = 0.2)$ at a test threshold of 6%, the green is $(N_B = 10000, N_s = 150, w_s = 0.1)$ at a test threshold of 1.5%, and the red is $(N_B = 10000, N_s = 75, w_s = 0.05)$ at a test threshold of 0.4%. For each ensemble, 100 independent instances of the dataset are generated and one NN is trained on each dataset. Bottom: Histogram of significance for ensembles with $(N_b = 10000, N_s = 300)$ and varying w_s. Blue is $w_s = 0.2$ at a test threshold of 6%, green is $w_s = 0.4$ at a test threshold of 24%, and red is $w_s = 1.0$ [for which the background and signal distribution in (x, y) are identical] at a test threshold of 50%. For each ensemble, 100 independent instances of the dataset are generated and one NN is trained on each dataset. The dashed line gives the expected significance for each ensemble.

($w_s = 1.0$, red), there is on average a small decrease in performance compared to simply not using a classifier.

IV. FULL METHOD

The previous sections use key elements of the full extended bump hunt but do not include all components, including the full background estimation and statistical analysis. This section gives a concrete prescription for applying the CWoLa hunting method in practice, which will be used in an explicit example in Sec. V. The setup is as in the previous sections: there is feature m_{res} where the signal is expected to be resonant and then a set of other features Y that are uncorrelated with m_{res}, but potentially useful for distinguishing signal from background. It is important to state that while a detailed model of Y is not required to perform the CWoLa hunting procedure that is described in the rest of the section, a limited model of Y is required to ensure the correlations with m_{res} are minimal. Such a model could come from simulation, from theory, or directly from a sufficiently signal-devoid data sample.

While in the presence of signal, the CWoLa hunting method would ideally learn systematic correlations between m_{res} and Y, instead, it may focus on statistical fluctuations in the background distributions. A naive application of CWoLa directly on the data may produce bumps in m_{res} by seeking local statistical excesses in the background distribution. This corresponds to a large look-elsewhere effect over the space of observables Y—the classifier may search this entire space and find the selection with the largest statistical fluctuation. In Sec. III, we took the approach of splitting the dataset into training, validation and test samples which eliminates this affect, since the statistical fluctuations in the three samples will be uncorrelated. However, applying this approach in practice would reduce the effective luminosity available for the search and thus degrade sensitivity. We therefore apply a cross-validation technique which allows all data to be used for testing while ensuring that event subsamples are never selected using a classifier that was trained on them. We split the events randomly, bin-by-bin, into five event samples of equal size. The first sample is set aside, and the first classifier is trained on the signal- and sideband-region events of the remaining four samples. This classifier may learn the statistical fluctuations in these event samples, but those will be uncorrelated with the fluctuations of the first sample. Applying the classifier to the set-aside event sample will then correspond to only one statistical test, eliminating the look-elsewhere effect. By repeating this procedure 5 times, each time setting aside one k-fold for testing and four for training and validation, all the data can be used for the bump hunt by adding up the selected events from each k-fold.

The algorithm we used for this procedure is summarized in Algorithm 1, and illustrated in Fig. 7. For each set-aside test set, we perform four rounds of training and validation using the four remaining data subsets. In each round, we set aside one of the remaining subsets as validation data, and

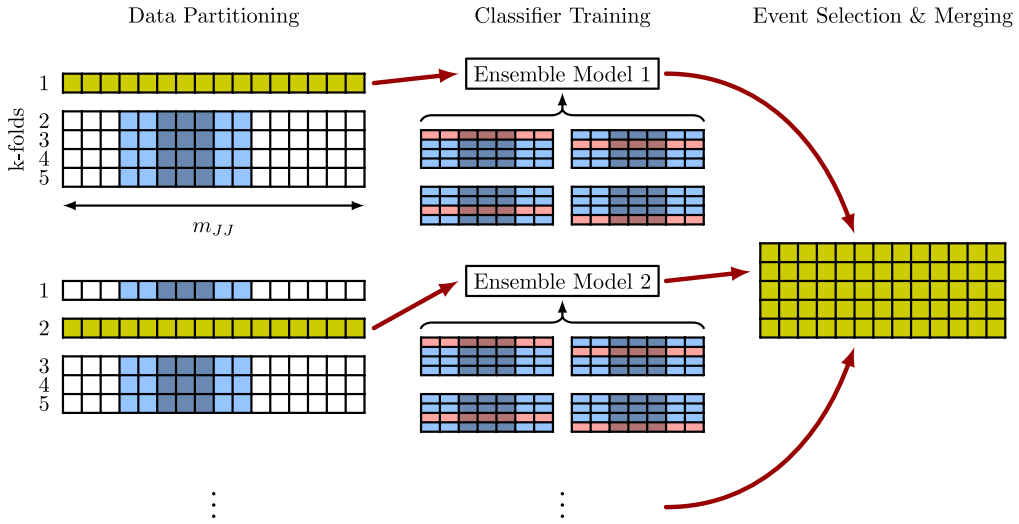

FIG. 7. Illustration of the nested cross-validation procedure. Left: The dataset is randomly partitioned bin-by-bin into five groups. Center: For each group $i \in \{1, 2, 3, 4, 5\}$ (the test set), an ensemble classifier is trained on the remaining groups $j \neq i$. There are four ways to split the four remaining groups into three for training and one for validation. For each of these four ways, many classifiers are trained and the one with best validation performance is selected. The ensemble classifier is then formed by the average of the four selected classifiers (one for each way to assign the training/validation split). Right: Data are selected from each test group using a threshold cut from their corresponding ensemble classifier. The selected events are then merged into a single m_{res} histogram.

the final three are used for training data. Only data falling in the signal and sideband regions are used for training and validation. The training and validation data are labeled as 0 or 1 if they fall in the sideband or signal regions, respectively. For each round, we train 20 NNs on the same training and validation data, using a different initialization each time. Each classifier is validated according to its performance as measured on validation data. Our performance metric ϵ_{val} is the true positive rate for correctly classifying a signal-region event as such, evaluated at a threshold with given false positive rate $s\%$ for incorrectly classifying a sideband region event as a signal region event. If a signal is present in the signal region and the classifier is able to find it, then it should be that $\epsilon_{\text{val}} > s\%$. On the other hand, if no signal is present then $\epsilon_{\text{val}} \simeq s\%$ is expected. Since we will be typically considering $\mathcal{O}(1\%)$-level signals we consider $s\% \sim 1\%$ in our test, and set $s\% = 0.5\%$ to generate our final results. For each of the twenty models, we end training if its performance

Algorithm 1. Nested cross-validation training and event selection procedure. See Appendix A for further details on the last two points.

Split dataset into five subsets stratified by m_{res} binning
for subset$_i$ *in subsets* **do**
  Set aside subset$_i$ as test data
  for subset$_j$ *in subsets*, $j \neq i$ **do**
    Validation data sideband = merge sideband bins of subset$_j$
    Validation data signal-region = merge signal-region bins of subset$_j$
    Training data sideband = merge sideband bins of remaining subsets
    Training data signal-region = merge signal-region bins of remaining subsets
    Assign signal-region data with label 1
    Assign sideband data with label 0
    Train twenty classifiers on training data, each with different random initialization
    model$_{i,j}$ = best of the twenty models, as measured by performance on validation data
  end
  model$_i = \sum_j$ model$_{i,j}/4$
  Select $Y\%$ most signal-like data points of subset$_i$, as determined by model$_i$. The threshold on the neural network to achieve $Y\%$ is determined using all other bins with large numbers of events and so the uncertainty on the value is negligible.
end
Merge selected events from each subset into new m_{res} histogram
Fit smooth background distribution to m_{res} distribution with the signal region masked
Evaluate p-value of signal region excess using fitted background distribution interpolated into the signal region.

has not improved in 300 epochs, and revert the model to a checkpoint saved at peak performance. We select the best of these 20 models, and discard the others. At the end of four rounds, the four selected models are averaged to form an ensemble model which is expected to be more robust than any individual model. The ensemble model is used to classify events in the test set, by selecting the $r\%$ most signal-like events. This procedure is repeated for all five choices of the test set, and the selected events from each are combined into a signal histogram in m_{res}. The presence of an identifiable signal will be indicated by a bump in the signal region, for which standard bump-hunting techniques can be used to perform a hypothesis test. The use of averaged ensemble models is important to reduce any performance degradation due to overfitting. Since each of the four models used to make each ensemble model has been trained on different training sets and with different random weight initialization, they will tend to overfit to different events. The models will therefore disagree in regions where overfitting has occurred, but will tend to agree in any region where a consistent excess is found.

Further technical details about the statistical methods can be found in Appendix A. Asymptotic formulas can be used to determine the local p-value of an excess, but such formulas must be validated using more computationally expensive methods for each application of CWoLa hunting, as is demonstrated in the Appendix A.

A. Interpreting the Results

The main result following the application of the method from Sec. IV is the local p-value. To determine the compatibility of the entire mass range with the no-resonance hypothesis, it is desirable to be able to compute a global p-value. In the result presented here, the mass bins were fixed ahead of time and were also nonoverlapping. Therefore, it is relatively simple to estimate a global p-value using e.g., a Bonferroni correction. However, this is not ideal (overconservative) when the mass bin width is scanned as part of the procedure. It is still possible to determine a global p-value, in the same spirit as the full BumpHunter statistic [4]. This would require a significant computational overhead as a large number of neural networks would need to be trained for each of many pseudoexperiments. An additional trials factor would be associated with scanning the threshold fraction on the neural network output. In the simplest approach, a small number of well-separated working points would be chosen, such as 10%, 1%, and 0.1%. These should be sufficiently different that the three local p-values could be treated as independent. However, a finer scan would require a proper assessment of the global p-value using pseudoexperiments. It may be possible to significantly reduce the computational cost by estimating the correlation between mass windows and threshold fractions in order to properly account for the look-elsewhere effect [78,79].

One final remark is about how one would use CWoLa hunting to set limits. In the form described above, the CWoLa hunting approach is designed to find new signals in data without any model assumptions. However, it is also possible to recast the lack of an excess as setting limits on particular BSM models. Given a simulated sample for a particular model, it would be possible to set limits on this model by mixing the simulation with the data and training a series of classifiers as above and running toy experiments, re-estimating the background each time. This is similar to the usual bump hunt, except that there is more computational overhead because the background distribution is determined in part by the neural networks, and the distribution in expected signal efficiencies cannot be determined except by these toy experiments.[3] In the absence of an excess, it is also possible to directly recast the results by taking the classifier trained on data with no significant signal. However, without a real excess, the classifier will have nothing to learn. Such a classifier will likely not be useful for any particular signal model. Therefore, while it is technically possible to do a standard reinterpretation of the results, the most powerful limit setting requires access to the data to retrain the neural networks for an injected signal.

V. PHYSICAL EXAMPLE

This section uses a dijet resonance search at the LHC to show the potential of CWoLa hunting in a realistic setting. As discussed in Sec. I, both ATLAS and CMS have a broad program targeting resonance decays into a variety of SM particles. Due to significance advances in jet substructure-based tagging [35], searches involving hadronic decays of the SM particles can be just as if not more powerful than their leptonic counterparts. The usual strategy for these searches is to develop dedicated single-jet classifiers, including[4] W/Z- [45,81], H- [82,83], top- [45,81,84], b- [46,85], and quark-jet taggers [63,86]. Simulated events with per-instance labels are used for training and then these classifiers are deployed in data. However, the best classifier in data may not be the best classifier in simulation. This problem is alleviated when learning directly from data.

Learning directly from data has another advantage—the decay products of a new heavy resonance may themselves be beyond the SM. If the massive resonance decays into new light states such as BSM Higgs bosons or dark sector particles that decay hadronically, then no dedicated SM tagger will be optimal [20,21]. A tagger trained to directly find nongeneric-jet structure could find these new intermediate particles and thus also find the heavy resonance.

[3]This complicates the legacy utility of the results, but it would be possible to tweak procedures like those advocated by RECAST [80] in which neutral networks would be automatically trained for a new signal model.

[4]These are the latest $\sqrt{s} = 13$ TeV results—see references within to find the complete history.

This was the approach taken in [87], but that method is fully supervised and so suffers the usual theory prior bias and potential sources of mismodeling. Here we will illustrate how the CWoLa hunting approach could be used instead to find such a signal. The next section (Sec. V A) describes the benchmark model in more detail, as well as the simulation details for both signal and background.

A. Signal and background simulation

For a benchmark signal, we consider the process $pp \to W' \to WX$, $X \to WW$, where W' and X are a new vector and scalar particle respectively. This process is predicted, e.g., in the warped extra dimensional construction of [22,88,89]. The typical opening angle between the two W bosons resulting from the X decay is given by $\Delta R(W,W) \simeq 4m_X/m_{W'}$ for $2m_W \ll m_X \ll m_{W'}$, and so the X particle will give rise to a single large-radius jet in the hadronic channel when $m_X \lesssim m_{W'}/4$. Taking the mass choices $m_{W'} = 3$ TeV and $m_X = 400$ GeV, the signal in the fully hadronic channel is a pair of large-radius jets J with $m_{JJ} \simeq 3$ TeV, one of which has a jet mass $m_J \simeq 80$ GeV and a two-pronged substructure, and the other has mass $m_J \simeq 400$ GeV with a four-prong substructure which often is arranged as a pair of two-pronged subjets.

Events are generated with MADGRAPH5_aMC@NLO [90] v2.5.5 to generate 10^4 signal events, using a model file implementing the tensor couplings of [89] and selecting only the fully hadronic decays of the three W bosons. The events are showered using PYTHIA 8.226 [91], and are passed through the fast detector simulator DELPHES 3.4.1 [92]. Jets are clustered from energy-flow tracks and towers using the FASTJET [93] implementation of the anti-k_t algorithm [94] with radius parameter $\Delta R = 1.2$. We require events to have at least two ungroomed large-radius jets with $p_T > 400$ GeV and $|\eta| < 2.5$. The selected jets are groomed using the soft drop algorithm [95] in grooming mode, with $\beta = 0.5$ and $z_{\text{cut}} = 0.02$. The two hardest groomed jets are selected as a dijet candidate, and a suite of substructure variables are recorded for these two jets. With the same simulation setup, 4.45×10^6 quantum chromodynamic (QCD) dijet events are generated with parton level cuts $p_{T,j} > 300$ GeV, $|\eta_j| < 2.5$, $m_{jj} > 1400$ GeV.

In order to study the behavior of the CWoLa hunting procedure both in the presence and absence of a signal, we produce samples both with and without an injected signal. The events are binned uniformly in $\log(m_{JJ})$, with 15 bins in the range 2001 GeV $< m_{JJ} <$ 4350 GeV.

B. Training a Classifier

In order to test for a signal with mass hypothesis $m_{JJ} \simeq m_{\text{res}}$, we construct a "signal region" consisting of all the events in the three bins centered around m_{res}. We also construct a low- and a high-mass sideband consisting of the events in the two bins below and above the signal region, respectively. The mass hypothesis will be scanned over the range 2278 GeV $\leq m_{\text{res}} \leq$ 3823 GeV, to avoid the first and last bins that cannot have a reliable background fit without constraints on both sides of the signal region. The signal region width is motivated by the width of the m_{JJ} peak for the benchmark signal process described earlier. Because all particles in the process are very narrow, this width corresponds to the resolution allowed by the jet reconstruction and detector smearing effects and will be relevant for other narrow signal processes also. For processes giving rise to wider bumps, the width of the signal hypothesis could be scanned over just as we scan over the mass hypothesis. We will then train a classifier to distinguish the events in the signal region from those in the sideband on the basis of their substructure. The objective in constructing the training framework is that the classifier should be very poor (equal efficiency in signal region and sideband for any threshold) in the case that no signal is present in the signal region, but if a signal is present with unusual jet substructure then the classifier should be able to locate the signal and provide discrimination power between signal and SM dijet events.

The background is estimated by fitting the regions outside of the signal region to a smoothly falling distribution. In practice, this requires that the auxiliary information Y is nearly independent of m_{JJ}; otherwise, the distribution could be sculpted. To illustrate the problem, consider a classifier trained to distinguish the sideband and signal regions using the observables m_J and the N-subjettiness variable $\tau_1^{(2)}$ [96]. The ratio $m_J/\sqrt{\tau_1^{(2)}}$ is approximately the jet p_{T}, which is highly correlated with m_{JJ} for the background. While it is often possible to find ways to decorrelate substructure observables [87,97–99], we take a simpler approach and instead select a basis of substructure variables which have no strong correlations with m_{JJ}. We will use the following set of 12 observables which does not provide learnable correlations with m_{JJ} sufficient to create signal-like bumps in our simulated background dijet event samples, as we shall demonstrate later in this section:

$$\text{For each jet: } Y_i = \left(m_J, \sqrt{\tau_1^{(2)}}/\tau_1^{(1)}, \tau_{21}, \tau_{32}, \tau_{43}, n_{\text{trk}} \right), \tag{5.1}$$

where $\tau_{MN} = \tau_M^{(1)}/\tau_N^{(1)}$. The full training uses $Y = (Y_1, Y_2)$. All ratios of N-subjettiness variables are chosen to be invariant under longitudinal boosts, so that the classifier cannot learn p_T from m_J and the other observables. The two jets are ordered by jet mass, so that the first six observables Y_1 correspond to the heavier jet while the last six Y_2 correspond to the lighter jet. We find that while the bulk of the m_J distribution in our simulated background dijet samples do not vary strongly over the sampled range of m_{JJ}, the high mass tails of the heavy and light jet mass

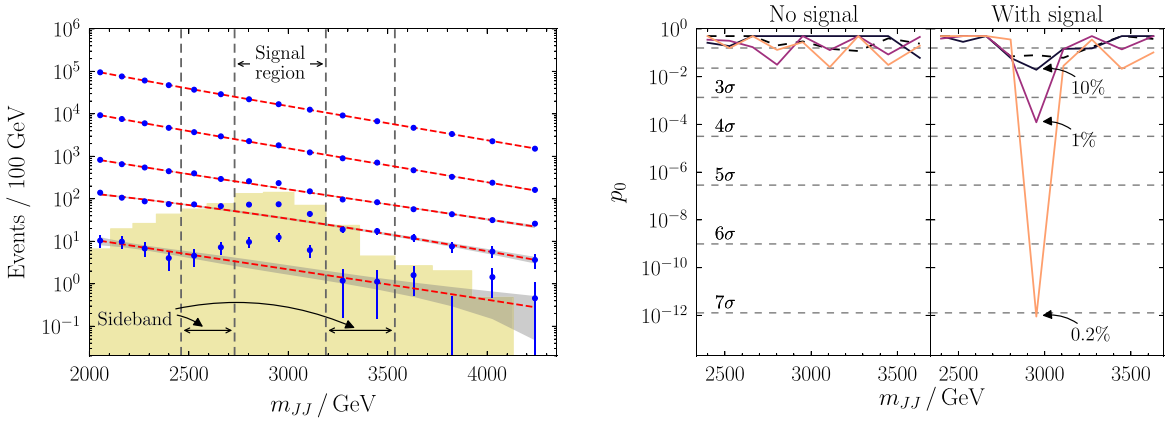

FIG. 8. Left: m_{JJ} distribution of dijet events (including injected signal, indicated by the filled histogram) before and after applying jet substructure cuts using the NN classifier output for the $m_{JJ} \simeq 3$ TeV mass hypothesis. The dashed red lines indicate the fit to the data points outside of the signal region, with the gray bands representing the fit uncertainties. The top dataset is the raw dijet distribution with no cut applied, while the subsequent datasets have cuts applied at thresholds with efficiency of 10^{-1}, 10^{-2}, 2×10^{-3}, and 2×10^{-4}. Right: Local p_0-values for a range of signal mass hypotheses in the case that no signal has been injected (left), and in the case that a 3 TeV resonance signal has been injected (right). The dashed lines correspond to the case where no substructure cut is applied, and the various solid lines correspond to cuts on the classifier output with efficiencies of 10^{-1}, 10^{-2}, and 2×10^{-3}.

distributions are sensitive to m_{JJ}. In lieu of a sophisticated decorrelation procedure, we simply reject outlier events which have $m_{J,A} > 500$ GeV and $m_{J,B} > 300$ GeV, where the subscripts A and B refer to the heavier and lighter jet respectively.

In our study, the classifiers used are dense neural networks built and trained using KERAS with a TENSORFLOW backend. We use four hidden layers consisting of a first layer of 64 nodes with a leaky rectified linear unit activation (using an inactive gradient of 0.1), and second through fourth layers of 32, 16, 4 nodes respectively with exponential linear unit activation [100]. The output node has a sigmoid activation. The first three hidden layers are regulated with dropout layers with 20% dropout rate [101]. The neural networks are trained to minimize binary cross-entropy loss using the ADAM optimizer with learning rate of 0.001, batch size of 20 000, first and second moment decay rates of 0.8 and 0.99, respectively, and learning rate decay of 5×10^{-4}. The training data is reweighted such that the low sideband has equal total weight to the high sideband, the signal region has the same total weight as the sum of the sidebands, and the sum of all events weights in the training data is equal to the total number of training events. This ensures that the NN output will be peaked around 0.5 in the absence of any signal, and ensures that low and high sideband regions contribute equally to the training in spite of their disparity in event rates.

C. Results

We use a sample of 553 388 QCD dijet events with dijet invariant mass $m_{JJ} > 2001$ GeV, corresponding to a luminosity of 4.4 fb. We consider two cases: first, a background-only sample; and second, a sample in which a signal has been injected with $m_{JJ} \simeq 3000$ GeV, with 877 events in the range $m_{JJ} > 2001$ GeV. In the signal region 2730 GeV $< m_{JJ} <$ 3189 GeV, consisting of the three bins centered around 3000 GeV, there are 81341 background events and 522 signal events, corresponding to $S/B = 6.4 \times 10^{-3}$ and $S/\sqrt{B} = 1.8$. Labeling the bins 1 to 15, we perform the procedure outlined previously to search for signals in the background-only and background-plus-signal datasets in signal regions defined around bins 4–12. This leaves room to define a signal region three bins wide, surrounded by a low and high sideband each two bins wide.

In Fig. 8 (left), we plot the background-plus-signal dataset which survive cuts at varying thresholds using the output of the classifier trained on the signal bin centered around 3 TeV. The topmost distribution corresponds to the inclusive dijet mass distribution, while the subsequent datasets have thresholds applied on the neural network with overall efficiencies of 10%, 2%, and 0.5%, respectively. A clear bump develops at the stronger thresholds, indicating the presence of a 3 TeV resonance. The automated procedure used to determine the significance is explained in detail in Appendix A. In brief, we estimate the background in the signal region by performing a fit of a smooth three-parameter function to the event rates in all the bins besides those in the signal region. We perform a simple counting experiment in the signal region, using the profile likelihood ratio as the test statistic, with the background fit parameters treated as nuisance, with preprofile uncertainties taken from the background fit itself. The significance is estimated using asymptotic formulas describing the properties of the profile likelihood ratio statistic [102]. Figure 8 shows the signal significance for each signal mass hypothesis, in the case

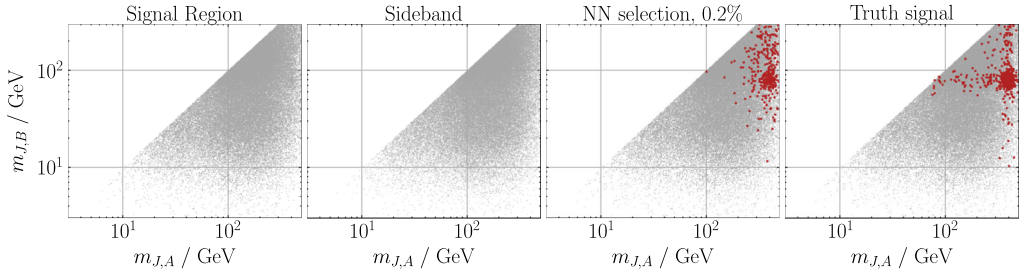

FIG. 9. Events projected onto the 2D plane of the two jet masses. The classifiers are trained to discriminate events in the signal region (left plot) from those in the sideband (second plot). The third plot shows in red the 0.2% most signal-like events determined by the classifier trained in this way. The rightmost plot shows in red the truth-level signal events.

that no signal is present (left), and in the case that the signal is present (right). We see that when no signal is present, no significant bump is created by our procedure. When a signal is present with $m_{\text{res}} = 3$ TeV, there is a significant bump which forms at this signal hypothesis, reaching 7σ at 0.2% efficiency. In Appendix B, we show the m_{JJ} distributions for each scan point used for the calculation of these p-values.

The fact that there is no significant bump in the left plot of Fig. 8 is an important method closure test. When deploying the CWoLa hunting approach in practice, we advocate to test the method in simulation in order to validate that there are no bump-catalyzing correlations in the selected classification features. A residual concern may be that there are correlations in the data which are not present in simulation. Residual correlations may come in two forms: process and kinematic. Process correlations occur when Y depends on the production channel (e.g., $pp \to qq$ or $pp \to gg$) and m_{JJ} also depends on the production channel; kinematic correlations are the case when m_{JJ} is correlated with Y given the process. Residual process correlations do not cause bumps because the m_{JJ} distribution of each process type (aside from signal) is smoothly falling. Thus, even if the classifier can exactly pick out one process, no bumps will be artificially sculpted. Residual kinematic correlations could cause artificially bumps in the m_{JJ} distribution. Physically, kinematic correlations occur because Y_i is correlated with $p_{\text{T,i}}$. One way to show in data that residual kinematic correlations are negligible is to use a *mixed sample* in which pairs of jets from different events are combined. As long as the potential signal fraction is small, this mixed sample will have no resonance peak. While the features chosen in this section were designed to be uncorrelated with m_{JJ} and not sculpt bumps, it may be possible to utilize correlated features in a modified CWoLa hunting procedure that includes systematic uncertainties for strong residual correlations. We leave studies of this possibility to future work.

We can investigate what the classifier has learnt by looking at the properties of events which have been classified as signal-like. In the first (second) plot of Fig. 9, events in the signal (sideband) region have been plotted on the plane of the jet masses of the heavier jet ($m_{J,A}$) and the lighter jet ($m_{J,B}$). After being trained to discriminate the events of the signal region from those of the sideband, the 0.2% most signal-like events as determined by the classifier are plotted in red in the third plot of Fig. 9, overlaid on top of the remaining events in gray. The classifier has selected a population of events with $m_{JA} \simeq 400$ GeV and $m_{JB} \simeq 80$ GeV, consistent with the injected signal. The final plot of the figure shows in red the truth-level signal events, overlaid on top of the truth-level background in grey.

Figure 10 shows some further 2D projections of the data. In each case, the x-axis is the jet mass of the heavier or the lighter jet in the top three or bottom three rows, respectively, while the y-axes correspond to observable substructure variables as measured on the same jet. The first column is all events in the signal region. The second column is truth-level signal events in red overlaid on truth-level background in gray. The third column is the 0.2% most signal-like events as determined by the classifier trained on this data. The fourth column shows the 0.2% most signal-like events as determined by a classifier trained on the data sample with no signal data, only background. We see that the tagger trained when signal is present has found a cluster of events with a 400 GeV jet with small $\tau_{43}^{(1)}$ and small n_{trk}; and an 80 GeV jet with relatively small $\sqrt{\tau_1^{(2)}}/\tau_1^{(1)}$, small $\tau_{21}^{(1)}$, and small n_{trk}. On the other hand, the events selected by the classifier trained on the background-only sample show no obvious clustering or pattern, and perhaps represent artifacts of statistical fluctuations in the training data.

The ability of the CWoLa approach to discriminate signal from background depends on the number of signal and background events in the signal and sideband regions. In Fig. 11, we keep the number of background events fixed but vary the size of the signal, and plot truth-label receiver operating characteristic (ROC) curves for each example. This allows us to directly assess the performance of the taggers for the signal. For varying thresholds, the x-axis corresponds to the efficiency on true signal events in the signal region, ϵ_S, while the y-axis represents the inverse efficiency on true QCD events in the signal region, ϵ_B. The gray dashed lines labeled 1 to 32 indicate the significance

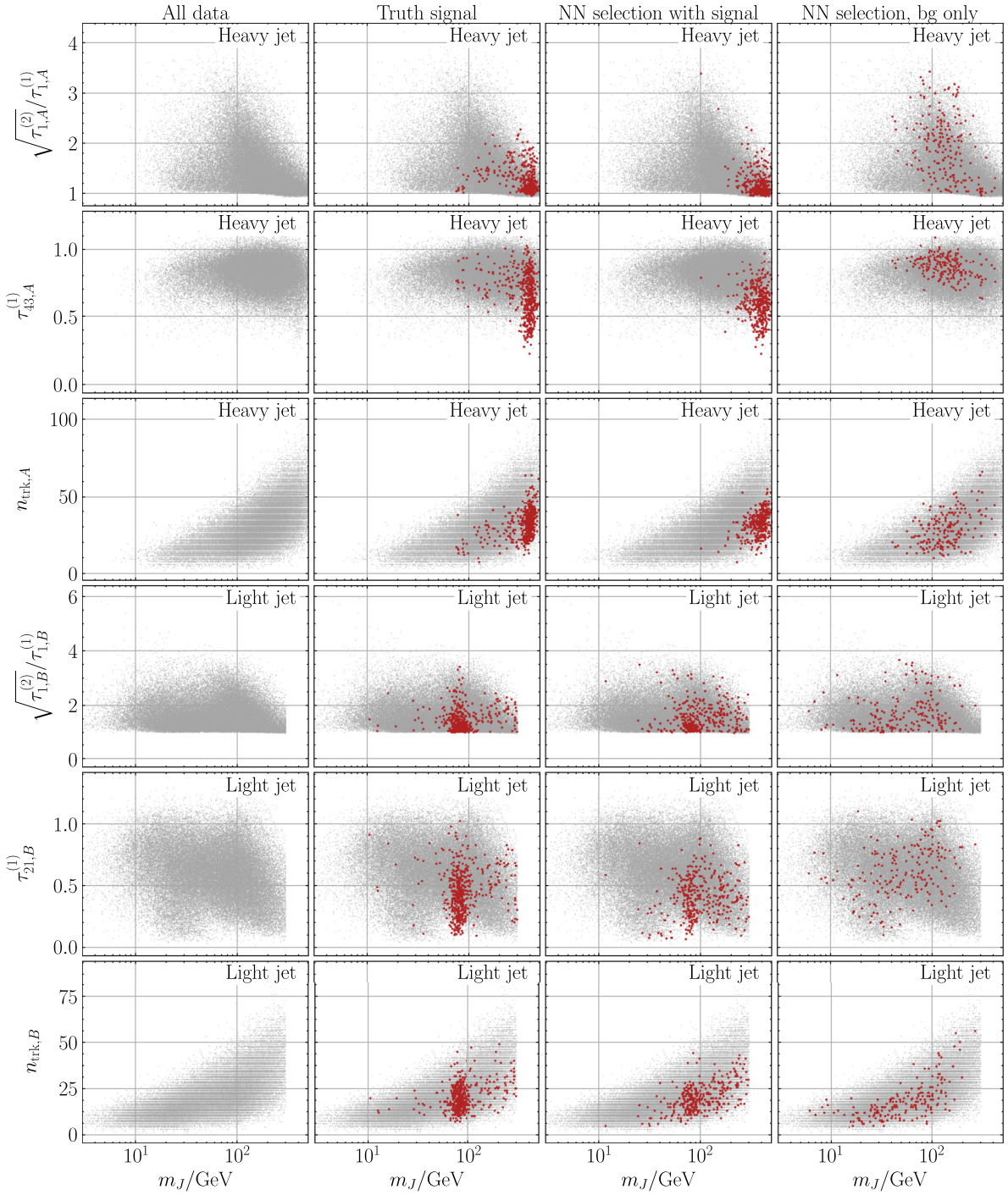

FIG. 10. 2D projections of the 12D feature-space of the signal region dataset. First column: All signal region events. Second column: Truth-level simulated signal events are highlighted in red. Third column: 0.2% most signal-like events selected by the classifier described in Sec. V are highlighted in red. Fourth column: Highlighted in red are the 0.2% most signal-like events selected by a classifier trained on the same sample but with true-signal events removed.

improvement, $\epsilon_S/\sqrt{\epsilon_B}$, which quantifies the gain in statistical significance compared to the raw m_{JJ} distribution with no cuts applied. In solid black we show the performance of a dedicated tagger trained with labeled signal and background events using a fully supervised approach. This gives a measure of the maximum achievable performance for this signal using the selected variables. A true dedicated tagger which could be used in a realistic dedicated search would be unlikely to reach this performance, since this would require careful calibration over 12 substructure variables with only simulated data available for the signal. While the CWoLa-based taggers do not reach the

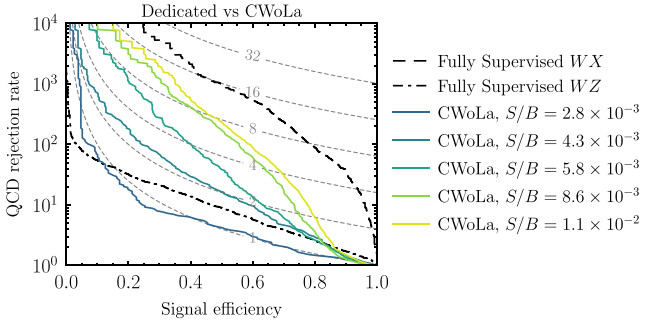

FIG. 11. Truth-label ROC curves for taggers trained using CWoLa with varying number of WX signal events, compared to those for a dedicated tagger trained on pure WX signal and background samples (dashed black) and one trained to discriminate W and Z jets from QCD (dot-dashed black). The CWoLa examples have $B = 81341$ in the signal region and $S = (230, 352, 472, 697, 927)$.

supervised performance in these examples, we find that performance does gradually improve with increasing statistics.

We also show in the dashed black curve the performance of the W/Z tagger in identifying this signal for which the tagger is not designed. This tagger is trained on a sample of $pp \to W' \to WZ$ events in the fully hadronic channel. In this case, the tagger is trained on the individual W/Z jets themselves rather than the dijet event, as is typical in the current ATLAS and CMS searches. In producing the ROC curve, dijet events are considered to pass the tagging requirement if both large-radius jets pass a threshold cut on the output of the W/Z-tagger. We see that for $\epsilon_B \sim 10^{-4}$, which is a typical background rejection rate for recent hadronic diboson searches, the signal rate is negligible since the X-jet rarely passes the cuts. This illustrates that CWoLa hunting may find unexpected signals which are not targeted by existing dedicated searches is S/B if high enough. If S/B is too low, then the CWoLa hunting approach is not able to identify the signal and it underperforms compared with the search that is targeting a different signal model.

The datasets and code used for the case study can be found at Refs. [103,104].

VI. CONCLUSIONS

We have presented a new anomaly detection technique for finding BSM physics signals directly from data. The central assumption is that the signal is localized as a bump in one variable in which the background is smooth, and that other features are available for additional discrimination power. This allows us to identify potential signal-enhanced and signal-depleted event samples with almost identical background characteristics on which a classifier can be trained using the classification without labels approach. In the case that a distinctive signal is present, the trained classifier output becomes an effective discriminant between signal events and background events, while in the case that no signal is present the classifier output shows no clear pattern. An event selection based on a threshold cut on the classifier output produces a smooth distribution if no signal is present and produces a bump if a signal is present, and so standard bump hunting techniques can be used on the selected distribution.

The prototypical example used here is the dijet resonance search in which the dijet mass is the one-dimensional feature where the signal is localized. Related quantities could also be used, such as the single jet mass for boosted resonance searches [11–13] or the average mass of pair produced objects [105–110]. Jet substructure information was used to augment information from just the dijet mass and a CWoLa classifier was trained using a deep neural network to discriminate signal region events from sideband events based on their substructure distributions. Additional local information such as the number of leptons inside the jets, the number of displaced vertices, etc. could be used in the future to ensure sensitivity to a wide variety of models. Furthermore, event-level information such as the number of jets or the magnitude of the missing transverse momentum could be added to an extended CWoLa hunt.

The CWoLa hunting strategy is generalizable beyond this single case study. To summarize, the essential requirements are:

(1) There is one bump-variable m_{res} in which the background forms a smooth distribution, for which there is a background model such a parametric function, and a signal can be expected to be localized as a bump. This was the variable m_{JJ} in the dijet case study.
(2) There are additional features Y in the events which may potentially provide discriminating power between signal and background, but the detailed topology of the signal in these variables is not known in advance. This was the set of substructure variables in the dijet study.
(3) The background distribution in Y should not have strong correlations with m_{res} over the resonance width of the signal. In the case that such correlations exist, it may be possible to find a transformation of the variables that removes these correlations before being fed into the classifier, or alternatively to train the classifier in such a way that penalizes shaping of the m_{res} distribution outside of the signal region. Closure tests in simulation or with mixed samples in data can be used to confirm that Y is not strongly correlated with m_{res}.

By harnessing the power of modern machine learning, CWoLa hunting and other weakly supervised strategies may provide the key to uncovering BSM physics lurking in the unique datasets already collected by the LHC experiments.

ACKNOWLEDGMENTS

We appreciate helpful discussions with and useful feedback on the manuscript from Timothy Cohen, Aviv Cukierman, Patrick Fox, Jack Kearney, Zhen Liu, Eric Metodiev, Brian Nord, Bryan Ostdiek, Matt Schwartz, and Jesse Thaler. We would also like to thank Peizhi Du for providing the UFO file for the benchmark signal model used in Sec. V. The work of J. H. C. is supported by NSF under Grant No. PHY-1620074 and by the Maryland Center for Fundamental Physics (MCFP). The work of B. N. is supported by the DOE under Contract No. DE-AC02-05CH11231. This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

APPENDIX A: STATISTICAL ANALYSIS

The significance of the bump is evaluated in the following way. First, after selecting the signal-like events in each cross-validation sample using its corresponding classifier, we merge the selected events from the k samples into a single selected dataset. This dataset is binned in m_{JJ}, and we estimate the background by fitting a smooth, parametric function to the dataset with the signal region masked out. We use the following three-parameter function (also used in the ATLAS [9] and CMS [111] searches for fully hadronic diboson resonances[5]):

$$\frac{dN}{dm_{JJ}} = p_0 \frac{(1 - m_{JJ}/\sqrt{s})^{p_1}}{(m_{JJ}/\sqrt{s})^{p_2}}, \tag{A1}$$

which is fitted using a least-squares fit. The number of events in the signal region is predicted by summing the predictions in each of the three signal region bins. The systematic uncertainty in this fit is estimated by propagating linearly the uncertainties on the fit parameters onto an uncertainty in the signal region prediction. The fits and fit uncertainties are indicated in the left plot in Fig. 8 by the red dashed lines and gray bands. We tested the goodness of fit of this functional form in background-only simulations using Kolmogorov-Smirnov tests, and in any real search we would advocate similar tests in simulation. In the case that simulation is not completely reliable, it is possible to define data validation regions using nonsignal selections in order to provide a cross-check of the fit function, as is done in e.g., Ref. [9]. In the case of CWoLa hunting, this would entail selecting events in nonsignal windows of the classifier output. For example, if using a 1% selection for the signal search, one could use other percentile windows of the NN output to define nonsignal selections with similar statistics which should be well fitted by the fit function under both the null and alternate hypotheses.

Since the shape of the signal in the signal region is *a priori* unknown we base our hypothesis test on the total number of events in the signal region. We form the profile likelihood ratio,

$$\lambda_0 = \frac{\mathcal{L}(\mu = 0, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\theta})}, \tag{A2}$$

where μ indicates the signal rate and θ is the nuisance parameter associated with the systematic uncertainties on the background prediction. In the numerator, $\hat{\hat{\theta}}$ represents the best fit value for the nuisance parameter in the background-only hypothesis $\mu = 0$, while in the denominator $\hat{\mu}$ and $\hat{\theta}$ represent the combined best fit for μ and θ. The likelihood is formed from a product of a Poisson factor for the number of events in the signal region, and a Gaussian constraint for the background nuisance parameter,

$$\mathcal{L}(\mu, \theta) = \text{Poiss}(n|b + \theta + \mu) e^{-\theta^2/(2\sigma^2)}, \tag{A3}$$

where n is the observed number of events in the signal region, b is the number of background events predicted by the sideband fit, θ is the nuisance parameter associated with the systematic uncertainty for the background prediction, and σ is the uncertainty on that nuisance parameter.

Our test statistic is

$$q_0 = \begin{cases} -2\log(\lambda_0), & \hat{\mu} > 0, \\ 0, & \hat{\mu} \leq 0. \end{cases} \tag{A4}$$

Using asymptotic formulas [102] gives a significance $Z = \sqrt{q_0}$ and $p_0 = 1 - \Phi(Z)$, where Φ is the cumulative distribution function of the normal distribution.

The null hypothesis is that the dijet invariant mass distribution after selection by the classifiers is well described by the smooth functional form of Eq. (A1). This requires that prior to any classification, the spectrum is smooth (already assumed by ATLAS and CMS) and that the classifiers are not able to generate localized features in the mass distribution following the CWoLa hunting procedure. In order to use the asymptotic formulas from Ref. [102], the bin counts in the selected, merged datasets must be Poissonian. The rest of the section investigates the validity of these approximations.

Let $f(x)$ represent the function described by Eq. (A1) prior to any classification and consider a dataset with $N_i^{(\text{uncut})}$ events in m_{JJ} bin i (bin center $m_{\text{res},i}$) with $N_i^{(\text{uncut})} \sim \text{Poiss}(f(m_{\text{res},i}))$. Let Y be a set of auxiliary

[5]More complex procedures for fitting the background such as Gaussian processes are also possible [112] but their use is beyond our scope.

observables whose probability distributions are independent of m_{res}. The goal is to demonstrate that the p-values reported from the statistical procedure described above are accurate. To begin, the dataset with $N_i^{(\text{uncut})}$ events is partitioned into k samples with equal probability for an event to be assigned to one of the samples. Next, a classifier is trained to discriminate signal region events from sideband events using all subsamples except the jth. The classifier is then used to select a fraction ϵ of events in the held out jth sample, using all other bins to determine ϵ.[6] This means that

$$n_{j;i}^{(\text{cut})} \sim \text{Poiss}\left(\frac{\epsilon}{k} f(m_{\text{res},i})\right). \tag{A5}$$

In the case that the cross-validated selected event rates in the k samples are uncorrelated, then it would follow that after merging these datasets the total selected event rate distribution would be given by

$$N_i^{(\text{cut})} = \sum_j n_{j;i}^{(\text{cut})} \sim \text{Poiss}(\epsilon f(m_{\text{res},i})). \tag{A6}$$

However, because the events in one sample are used to train a classifier applied to the other samples, it cannot necessarily be assumed that the event rates are uncorrelated between samples. If strong correlations are expected between selected samples then in order to calculate reliable p-values the test statistic would need to be calibrated by running many toy experiments on either new simulated event samples or on bootstrapped samples, with the NNs trained fresh each time. Since this is a computationally expensive procedure, it is preferable if a simpler alternative is available.

In order to check that the simpler approach (assuming no correlations between cross-validated samples) is valid, we have performed an empirical test of this effect in the following way. We generated 10^3 toy datasets with binned event counts drawn from Poisson distributions with means determined by the distribution of Eq. (A1), with parameters obtained by a fit to the uncut dijet dataset used in Sec. V. Each event has 12 auxiliary variables, as in Sec. V, but with these variables drawn from a random uniform distribution in the range [0, 1]. NNs were trained using a cross-validation procedure exactly as described in Sec. V, except for the following modifications that were required to reduce the computational time required. We used fourfold cross validation (rather than fivefold), trained only four NNs per iteration from which the best was selected (rather than 20), and the NNs were trained with a patience of 100 epochs of no improvement in validation performance before stopping (instead of 300 epochs). The trained NNs were used to select the 1% most "signal-like" events for each toy. For each toy we then calculated the test statistic for rejection of the null hypothesis, and the distribution of these test statistics is shown by the black markers in Fig. 12.

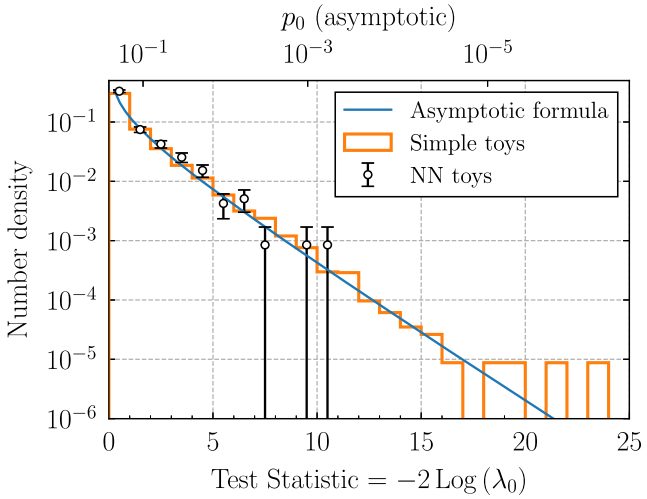

FIG. 12. Toy test statistic distributions. Black data points: 10^3 toys with NN training and cross-validation procedure; error bars represent $\sqrt{N}$ Poisson uncertainty. Orange histogram: 10^5 toys with random selection (no NN training or cross validation). Blue: Asymptotic formula.

Additionally, we generated 10^5 toy datasets with m_{res} drawn in the same way. Instead of training NNs to select events, we randomly selected 1% of events. For each toy we then calculated the test statistic for rejection of the null hypothesis, and the distribution of these test statistics is shown by the orange histogram in Fig. 12. Finally, we show the expected asymptotic distribution with the blue line.

The key feature of Fig. 12 is that the test statistic distribution for the NN toys shows no apparent deviation from that for the simple toys or from the asymptotic form. We therefore find no evidence of any distortion caused by correlations between cross-validation samples in this toy experiment.

Finally, it is worth remarking that the p-value computed with the above procedure is only local. If a local p-value is below some threshold, a follow-up, dedicated analysis using an orthogonal dataset should target the identified region of phase space with no trials factor penalty. One could also estimate a global p-value in a standard way using e.g., a Bonferroni correction. Other methods like the full BumpHunter statistic could be used [4] but that is not the standard practice in the current ATLAS and CMS diboson resonance searches.

[6]The number of events used to determine ϵ is sufficiently large that the uncertainty on the value of the NN used to achieve ϵ efficiency is negligible.

APPENDIX B: DIJET MASS SCANS

In Figs. 13 and 14 we plot the dijet invariant mass distributions before and after applying tagger cuts over the full range of the mass scan described in Sec. V. The p-values calculated from the top four distributions in these plots are displayed in Fig. 8 (right).

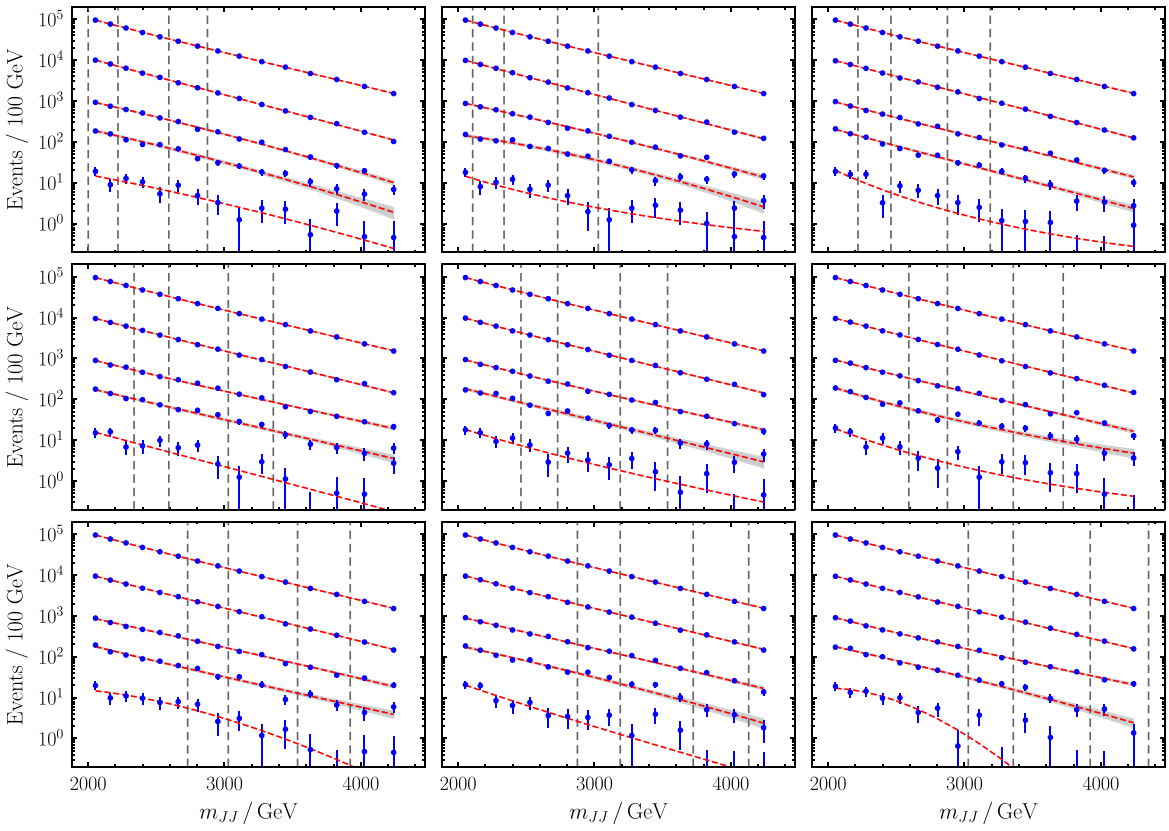

FIG. 13. Dijet invariant mass distributions in the resonance mass scan in the case that no signal is present. In each plot, the signal and sideband regions used for training the tagger are indicated by the vertical dashed lines. The top distribution in each plot is the original dijet mass distribution, and the subsequent distributions have had cuts applied at efficiencies of 10^{-1}, 10^{-2}, 2×10^{-3}, and 2×10^{-4} respectively. The dashed red curves are fits determined by weighted least squares, and the gray bands the corresponding systematic uncertainty in the fit. In the lowest distribution of each plot the fit is poor as the presence of low-count bins makes the least-squares fit inappropriate—the fit line is kept to guide the eye, but the fit uncertainty band is omitted.

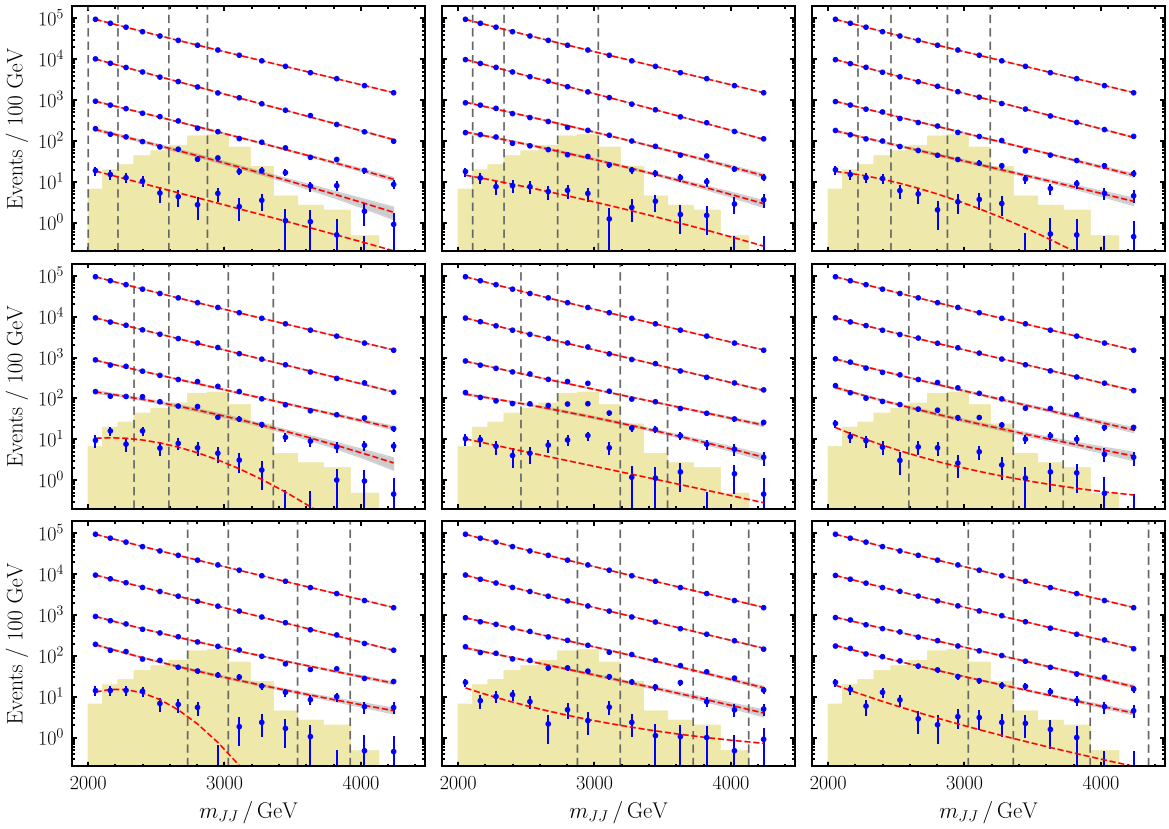

FIG. 14. Dijet invariant mass distributions in the resonance mass scan in the case that the signal is present (indicated by the filled histogram). Otherwise, the plots are as described in the caption to Fig. 13.

[1] J. Button, G. R. Kalbfleisch, G. R. Lynch, B. C. Maglić, A. H. Rosenfeld, and M. L. Stevenson, Pion-pion interaction in the reaction $\bar{p} + p \to 2\pi^+ + 2\pi^- + n\pi^0$, Phys. Rev. **126**, 1858 (1962).

[2] G. Aad *et al.* (ATLAS Collaboration), Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B **716**, 1 (2012).

[3] S. Chatrchyan *et al.* (CMS Collaboration), Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, Phys. Lett. B **716**, 30 (2012).

[4] G. Choudalakis, On hypothesis testing, trials factor, hypertests and the BumpHunter, in *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, 2011* (CERN, Geneva, Switzerland, 2011).

[5] M. Aaboud *et al.* (ATLAS Collaboration), Search for new phenomena in dijet events using 37 fb^{-1} of pp collision data collected at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Rev. D **96**, 052004 (2017).

[6] A. M. Sirunyan *et al.* (CMS Collaboration), Search for dijet resonances in proton-proton collisions at $\sqrt{s} =$ 13 TeV and constraints on dark matter and other models, Phys. Lett. B **769**, 520 (2017); Erratum, Phys. Lett. B **772**, 882(E) (2017).

[7] V. Khachatryan *et al.* (CMS Collaboration), Search for Narrow Resonances Decaying to Dijets in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. Lett. **116**, 071801 (2016).

[8] A. M. Sirunyan *et al.* (CMS Collaboration), Search for massive resonances decaying into WW, WZ, ZZ, qW, and qZ with dijet final states at $\sqrt{s} = 13$ TeV, Phys. Rev. D **97**, 072006 (2018).

[9] M. Aaboud *et al.* (ATLAS Collaboration), Search for diboson resonances with boson-tagged jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Lett. B **777**, 91 (2018).

[10] A. M. Sirunyan *et al.* (CMS Collaboration), Search for high-mass $Z\gamma$ resonances in proton-proton collisions at $\sqrt{s} = 8$ and 13 TeV using jet substructure techniques, Phys. Lett. B **772**, 363 (2017).

[11] A. M. Sirunyan *et al.* (CMS Collaboration), Search for Low Mass Vector Resonances Decaying to Quark-Antiquark Pairs in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. Lett. **119**, 111802 (2017).

[12] A. M. Sirunyan *et al.* (CMS Collaboration), Search for low mass vector resonances decaying into quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 01 (2018) 097.
[13] M. Aaboud *et al.* (ATLAS Collaboration), Search for light resonances decaying to boosted quark pairs and produced in association with a photon or a jet in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Lett. B **788**, 316 (2019).
[14] A. M. Sirunyan *et al.* (CMS Collaboration), Inclusive Search for a Highly Boosted Higgs Boson Decaying to a Bottom Quark-Antiquark Pair, Phys. Rev. Lett. **120**, 071802 (2018).
[15] A. M. Sirunyan *et al.* (CMS Collaboration), Search for a massive resonance decaying to a pair of Higgs bosons in the four b quark final state in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Lett. B **781**, 244 (2018).
[16] M. Aaboud *et al.* (ATLAS Collaboration), Search for heavy resonances decaying to a W or Z boson and a Higgs boson in the $q\bar{q}^{(\prime)}b\bar{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Lett. B **774**, 494 (2017).
[17] M. Aaboud *et al.* (ATLAS Collaboration), Search for resonances in the mass distribution of jet pairs with one or two jets identified as b-jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Lett. B **759**, 229 (2016).
[18] A. M. Sirunyan *et al.* (CMS Collaboration), Search for $\mathrm{t\bar{t}}$ resonances in highly boosted lepton + jets and fully hadronic final states in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 07 (2017) 001.
[19] M. Aaboud *et al.* (ATLAS Collaboration), Search for $W' \to tb$ decays in the hadronic final state using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Lett. B **781**, 327 (2018).
[20] J. A. Aguilar-Saavedra, Running bumps from stealth bosons, Eur. Phys. J. C **78**, 206 (2018).
[21] J. A. Aguilar-Saavedra, Stealth multiboson signals, Eur. Phys. J. C **77**, 703 (2017).
[22] K. Agashe, J. H. Collins, P. Du, S. Hong, D. Kim, and R. K. Mishra, Detecting a boosted diboson resonance, J. High Energy Phys. 11 (2018) 027.
[23] ATLAS Collaboration, A model independent general search for new phenomena with the ATLAS detector at $\sqrt{s} = 13$ TeV, CERN Technical Report No. ATLAS-CONF-2017-001, 2017.
[24] CMS Collaboration, Model unspecific search for new physics in pp collisions at $\sqrt{s} = 7$ TeV, CERN Technical Report No. CMS-PAS-EXO-10-021, 2011.
[25] ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV, CERN Technical Report No. ATLAS-CONF-2012-107, 2012.
[26] ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV, CERN Technical Report No. ATLAS-CONF-2014-006, 2014.
[27] A. Aktas *et al.* (H1 Collaboration), A general search for new phenomena in ep scattering at HERA, Phys. Lett. B **602**, 14 (2004).
[28] F. D. Aaron *et al.* (H1 Collaboration), A general search for new phenomena at HERA, Phys. Lett. B **674**, 257 (2009).
[29] B. Abbott *et al.* (D0 Collaboration), Search for new physics in $e\mu X$ data at DØ using Sherlock: A quasi-model-independent search strategy for new physics, Phys. Rev. D **62**, 092004 (2000).
[30] V. M. Abazov *et al.* (D0 Collaboration), A quasi-model-independent search for new physics at large transverse momentum, Phys. Rev. D **64**, 012004 (2001).
[31] T. Aaltonen *et al.* (CDF Collaboration), Model-independent and quasi-model-independent search for new physics at CDF, Phys. Rev. D **78**, 012002 (2008).
[32] T. Aaltonen *et al.* (CDF Collaboration), Global search for new physics with 2.0 fb^{-1} at CDF, Phys. Rev. D **79**, 011101 (2009).
[33] B. Knuteson, Ph.D. thesis, University of California at Berkeley, 2000.
[34] B. Knuteson, Systematic analysis of high-energy collider data, Nucl. Instrum. Methods Phys. Res., Sect. A **534**, 7 (2004).
[35] 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, arXiv:1709.04464.
[36] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, Jet-images: Computer vision inspired techniques for jet tagging, J. High Energy Phys. 02 (2015) 118.
[37] 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.
[38] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-images—deep learning edition, J. High Energy Phys. 07 (2016) 069.
[39] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, Jet substructure classification in high-energy physics with deep neural networks, Phys. Rev. D **93**, 094034 (2016).
[40] J. Barnard, E. N. Dawe, M. J. Dolan, and N. Rajcic, Parton shower uncertainties in jet substructure analyses with deep neural networks, Phys. Rev. D **95**, 014018 (2017).
[41] 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.
[42] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, Deep-learned top tagging with a Lorentz layer, SciPost Phys. **5**, 028 (2018).
[43] 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.
[44] G. Louppe, K. Cho, C. Becot, and K. Cranmer, QCD-aware recursive neural networks for jet physics, arXiv:1702.00748.
[45] ATLAS Collaboration, Performance of top quark and W boson tagging in run 2 with ATLAS, CERN Technical Report No. ATLAS-CONF-2017-064, 2017.
[46] ATLAS Collaboration, Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017–18 LHC run, CERN Technical Report No. ATL-PHYS-PUB-2017-013, 2017.
[47] ATLAS Collaboration, Identification of hadronically-decaying W bosons and top quarks using high-level

features as input to boosted decision trees and deep neural networks in ATLAS at $\sqrt{s} = 13$ TeV, CERN Report No. ATL-PHYS-PUB-2017-004, 2017.
[48] CMS Collaboration, Heavy flavor identification at CMS with deep neural networks, CERN Report No. CMS-DP-2017-005, 2017.
[49] CMS Collaboration, CMS phase 1 heavy flavor identification performance and developments, CERN Report No. CMS-DP-2017-013, 2017.
[50] ATLAS Collaboration, Identification of jets containing b-hadrons with recurrent neural networks at the ATLAS experiment, CERN Report No. ATL-PHYS-PUB-2017-003, 2017.
[51] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, Jet constituents for deep neural network based top quark tagging, arXiv:1704.02124.
[52] K. Datta and A. Larkoski, How much information is in a jet?, J. High Energy Phys. 06 (2017) 073.
[53] K. Datta and A. J. Larkoski, Novel jet observables from machine learning, J. High Energy Phys. 03 (2018) 086.
[54] ATLAS Collaboration, Quark versus gluon jet tagging using jet images with the ATLAS Detector, CERN Report No. ATL-PHYS-PUB-2017-017, 2017.
[55] CMS Collaboration, New developments for jet substructure reconstruction in CMS, CERN Report No. CMS-DP-2017-027, 2017.
[56] K. Fraser and M. D. Schwartz, Jet charge and machine learning, J. High Energy Phys. 10 (2018) 093.
[57] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, JUNIPR: A framework for unsupervised machine learning in particle physics, arXiv:1804.09720.
[58] S. Macaluso and D. Shih, Pulling out all the tops with computer vision and deep learning, J. High Energy Phys. 10 (2018) 121.
[59] G. Aad *et al.* (ATLAS Collaboration), Performance of b-jet identification in the ATLAS experiment, J. Instrum. **11**, P05013 (2016).
[60] S. Chatrchyan *et al.* (CMS Collaboration), Identification of b-quark jets with the CMS experiment, J. Instrum. **8**, P04013 (2013).
[61] G. Aad *et al.* (ATLAS Collaboration), Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector, Eur. Phys. J. C **74**, 3023 (2014).
[62] CMS Collaboration, Performance of quark/gluon discrimination in 8 TeV pp data, CERN Technical Report No. CMS-PAS-JME-13-002, 2013.
[63] CMS Collaboration, Performance of quark/gluon discrimination in 13 TeV data, CERN Technical Reports No. CMS-DP-2016-070 and No. CERN-CMS-DP-2016-070, 2016.
[64] 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).
[65] V. Khachatryan *et al.* (CMS Collaboration), Identification techniques for highly boosted W bosons that decay into hadrons, J. High Energy Phys. 12 (2014) 017.
[66] G. Aad *et al.* (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.
[67] CMS Collaboration, Boosted top jet tagging at CMS, CERN Technical Report No. CMS-PAS-JME-13-007, 2014.
[68] G. Louppe, M. Kagan, and K. Cranmer, Learning to pivot with adversarial networks, arXiv:1611.01046.
[69] L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman, Weakly supervised classification in high energy physics, J. High Energy Phys. 05 (2017) 145.
[70] 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.
[71] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, Learning to classify from impure samples, Phys. Rev. D **98**, 011502 (2018).
[72] T. Cohen, M. Freytsis, and B. Ostdiek, (Machine) learning to do more with less, J. High Energy Phys. 02 (2018) 034.
[73] M. Pivk and F. R. Le Diberder, SPlot: A statistical tool to unfold data distributions, Nucl. Instrum. Methods Phys. Res., Sect. A **555**, 356 (2005).
[74] J. Neyman and E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, Phil. Trans. R. Soc. A **231**, 289 (1933).
[75] F. Chollet, Keras. https://github.com/fchollet/keras.
[76] 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 Vol. 16 (USENIX association, Savannah, GA, 2016), pp. 265.
[77] D. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.
[78] E. Gross and O. Vitells, Trial factors for the look elsewhere effect in high energy physics, Eur. Phys. J. C **70**, 525 (2010).
[79] O. Vitells and E. Gross, Estimating the significance of a signal in a multidimensional search, Astropart. Phys. **35**, 230 (2011).
[80] K. Cranmer and I. Yavin, RECAST: Extending the impact of existing analyses, J. High Energy Phys. 04 (2011) 038.
[81] CMS Collaboration, W/Z/Top tagging for run 2 in CMS, CERN Reports No. CMS-DP-2015-043 and No. CERN-CMS-DP-2015-043, 2015.
[82] CMS Collaboration, b-tagging in boosted topologies, CERN Reports No. CMS-DP-2015-038 and No. CERN-CMS-DP-2015-038, 2015.
[83] ATLAS Collaboration, Boosted Higgs ($\to b\bar{b}$) boson identification with the ATLAS detector at $\sqrt{s} = 13$ TeV, CERN Technical Report No. ATLAS-CONF-2016-039, 2016.
[84] CMS Collaboration, Top tagging with new approaches, CERN Technical Report No. CMS-PAS-JME-15-002, 2016.
[85] CMS Collaboration, Performance of heavy flavor identification algorithms in proton-proton collisions at 13 TeV at the CMS experiment, CERN Reports No. CMS-DP-2017-012 and No. CERN-CMS-DP-2017-012, 2017.
[86] ATLAS Collaboration, Quark versus gluon jet tagging using charged particle multiplicity with the ATLAS detector, CERN Technical Report No. ATL-PHYS-PUB-2017-009, 2017.
[87] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, A generic anti-QCD jet tagger, J. High Energy Phys. 11 (2017) 163.

[88] K. Agashe, P. Du, S. Hong, and R. Sundrum, Flavor universal resonances and warped gravity, J. High Energy Phys. 01 (2017) 016.
[89] K. Agashe, J. H. Collins, P. Du, S. Hong, D. Kim, and R. K. Mishra, Dedicated strategies for triboson signals from cascade decays of vector resonances, arXiv:1711.09920.
[90] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.
[91] T. Sjostrand, S. Mrenna, and P. Z. Skands, A brief introduction to PYTHIA 8.1, Comput. Phys. Commun. **178**, 852 (2008).
[92] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.
[93] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, Eur. Phys. J. C **72**, 1896 (2012).
[94] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k(t) jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[95] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, Soft drop, J. High Energy Phys. 05 (2014) 146.
[96] J. Thaler and K. Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness, J. High Energy Phys. 02 (2012) 093.
[97] 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.
[98] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Sogaard, Decorrelated jet substructure tagging using adversarial neural networks, Phys. Rev. D **96**, 074034 (2017).
[99] I. Moult, B. Nachman, and D. Neill, Convolved substructure: Analytically decorrelating jet substructure observables, J. High Energy Phys. 05 (2018) 002.
[100] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), arXiv:1511.07289.
[101] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).
[102] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Asymptotic formulas for likelihood-based tests of new physics, Eur. Phys. J. C **71**, 1554 (2011); Erratum **73**, 2501 (2013).
[103] B. N. J. Collins and K. Howe, CWoLa Hunting: Data sample, Mendeley Data (2018).
[104] B. N. J. Collins and K. Howe, CWoLa Hunting: Code.
[105] M. Aaboud *et al.* (ATLAS Collaboration), A search for pair-produced resonances in four-jet final states at $\sqrt{s} =$ 13 TeV with the ATLAS detector, Eur. Phys. J. C **78**, 250 (2018).
[106] CMS Collaboration, A search for light pair-produced resonances decaying into at least four quarks, CERN Report No. CMS-PAS-EXO-17-022, 2018.
[107] G. Aad *et al.* (ATLAS Collaboration), Search for pair-produced massive coloured scalars in four-jet final states with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV, Eur. Phys. J. C **73**, 2263 (2013).
[108] G. Aad *et al.* (ATLAS Collaboration), A search for top squarks with R-parity-violating decays to all-hadronic final states with the ATLAS detector in $\sqrt{s} = 8$ TeV proton-proton collisions, J. High Energy Phys. 06 (2016) 067.
[109] S. Chatrchyan *et al.* (CMS Collaboration), Search for Pair-Produced Dijet Resonances in Four-Jet Final States in pp Collisions at $\sqrt{s} = 7$ TeV, Phys. Rev. Lett. **110**, 141802 (2013).
[110] V. Khachatryan *et al.* (CMS Collaboration), Search for pair-produced resonances decaying to jet pairs in proton-proton collisions at $\sqrt{s} = 8$ TeV, Phys. Lett. B **747**, 98 (2015).
[111] A. M. Sirunyan *et al.* (CMS Collaboration), Search for massive resonances decaying into WW, WZ or ZZ bosons in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 03 (2017) 162.
[112] 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.