CWoLa Hunting:
Extending the Bump Hunt with Machine Learning

Jack 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, MD 20742, USA*

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

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

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

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

E-mail: jhc296@umd.edu, khowe@fnal.gov, bpnachman@lbl.gov

ABSTRACT: The oldest and most robust technique to search for new particles is to look for 'bumps' in invariant mass spectra over smoothly falling backgrounds. This is a powerful technique, but only uses one-dimensional information. One can restrict the phase space to enhance a potential signal, but such tagging techniques require a signal hypothesis and training a classifier in simulation and applying it on data. We present a new method for using all of the available information (with machine learning) without using any prior knowledge about potential signals. Given the lack of new physics signals at the Large Hadron Collider (LHC), such model independent 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.

Contents

1 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 have a non-trivial internal structure 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

---

[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.

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 the LHC experiments and their predecesors, but only utilize simple objects and rely heavily on simulation for background estimation [23–32].

The tagging techniques used to isolate different jet types have increased in sophistication with the advent of modern machine learning classifiers [33–56]. 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 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 [38] which renders classifiers sub-optimal when training on simulation and testing on data. This is already apparent for existing multivariate classifiers where post-hoc mis-modeling corrections can be large [57–65]. Adversarial approaches can be used to mitigate potential mis-modeling effects during training, but at the cost of algorithmic performance [66]. The optimal solution is to train classifiers directly on data.

We propose a new method that combines resonance searches with recently proposed techniques for learning directly from data [67–69]. Simply stated, the new algorithm trains a fully supervised classifier to distinguish a signal region from a mass sideband. A bump hunt is then performed on the mass distribution after applying a threshold on the classifier output. This is Classification Without Labels (CWoLa) [68] 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. The key difference with respect to Ref. [68, 69] 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.

This paper is organized as follows. Section 2 formally introduces the CWoLa hunting approach and briefly discusses how auxiliary information can be useful for bump hunting. Then, Sec. 3 uses a simplified example to show how a neural network can be used to identify new physics structures from pseudodata. Finally, a realistic example based on a hadronic resonance search is presented in Sec. 4. Conclusions and future outlook are presented in Sec. 5.

2   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 X be a random variable that represents the invariant mass. The distribution of $X|$background is smooth while $X|$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: $M_1 = \{(X,Y)||X - m_0| < \delta\}$ and $M_2 = \{(X,Y)|\delta < |X - m_0| < \epsilon\}$ for $\epsilon > \delta$. 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, which is optimal for distinguishing signal from background by Ref. [68], and then performing a usual bump hunt on X 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.

The goal is for Y to incorporate as much local and global information as possible. 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 $\Pr(Y = 1|\text{signal}) = q$. 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 BSM, $N_s > 0$. Figure 2 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 sub-optimal - 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.
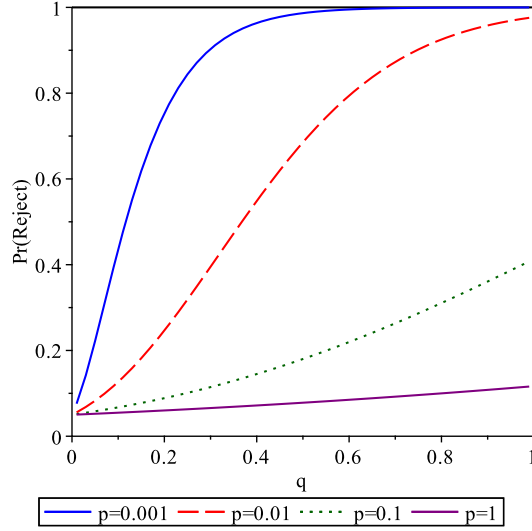
Figure 1. The probability to reject the SM as a function of q for fixed values of p as indicated in the legend when only considering events with $Y = 1$. The expected background is fixed at 1000 and the expected BSM is fixed at 20. When $q = 0$, there is no signal and therefore the rejection probability 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.

3   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 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 [70] $p_s(Y)/p_b(Y)$; 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.
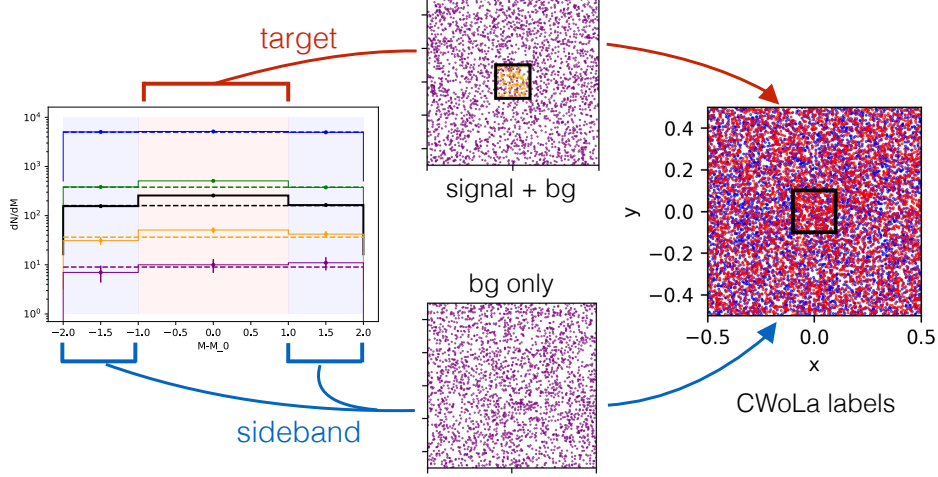
Figure 2. An illustration of the CWoLa procedure for the two-dimensional example presented in Sec. 3. The left plot shows the mass distribution for the three mass bins, which is uniform for the background. The blue line is the total number of events and the other lines represent thresholds on various 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).

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 KERAS [71] with a TENSORFLOW [72] 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 [73] 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
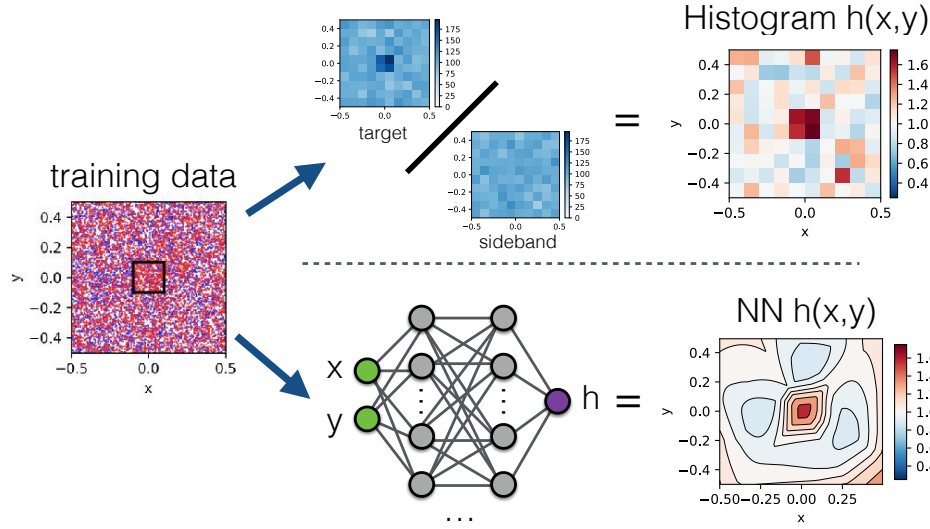
Figure 3. The CWoLa-labeled data can be used to construct an estimate for the optimal classifier $h(x, y) = \frac{p_b(x,y) + p_s(x,y)}{p_b(x,y)}$. The top path shows an estimate constructed by histogramming the observed training events in the (x, y) plane. The bottom path shows an estimate constructed by using a neural network trained as described in the text, which can be efficiently generalized to higher dimensional distributions.

improvement. This simple scheme is robust against enhancing statistical fluctuations but reduces the number of events used for the final search by a factor of three. 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.

The key question now is to evaluate the performance of the CWoLa hunting approach. Figure 5 shows the mass distribution in the three bins after applying successfully tighter
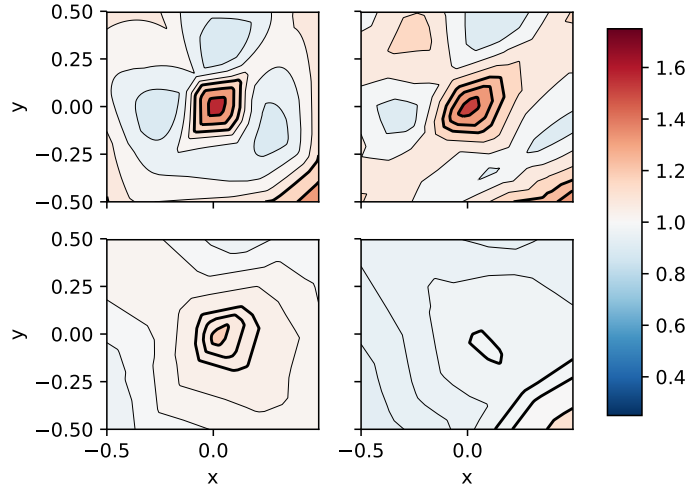
Figure 4. The classifier $h(x,y)$ constructed from four independent training runs on the same example two-dimensional model dataset described in the text. The thick contours represent the cuts that would reduce the events in the test data target window by a factor of $\epsilon_{\text{test}} = 10\%, 5\%, 1\%$.

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 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 a expected signifance of 7.5σ, the network performs similarly to the $w_s = 0.2$ example (blue).
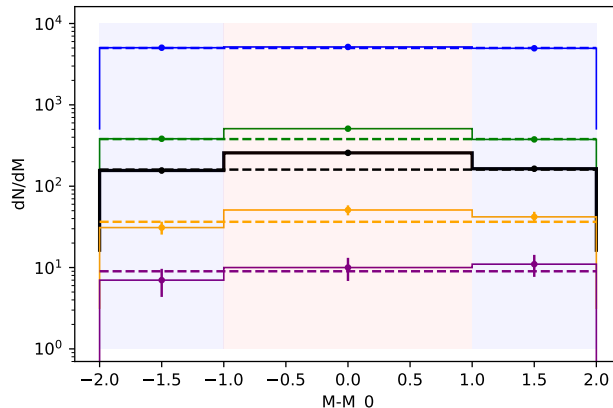
Figure 5. The mass distribution after various threshold on the neural network classifier. The flat background fit from the sideband regions are the dashed lines, and the statistical uncertainty for each bin is shown by the error bars. The top histogram is the model before any threshold in the (x, y) 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.

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

4 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. 1, 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 [33], 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[2] W/Z- [43, 74], H- [75, 76], top- [43, 74, 77], b- [44, 78], and quark-jet taggers [61, 79]. Simulated events with per-instance labels are used for training and then these classifiers are deployed in data and the performance is calibrated using special control samples with a relatively high purity. The fact that calibrations are required is evidence that these classifiers are sub-optimal when applied to data. This challenge 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 non-generic-jet

---

[2]These are the latest $\sqrt{s} = 13$ TeV results - see references within to find the complete history.
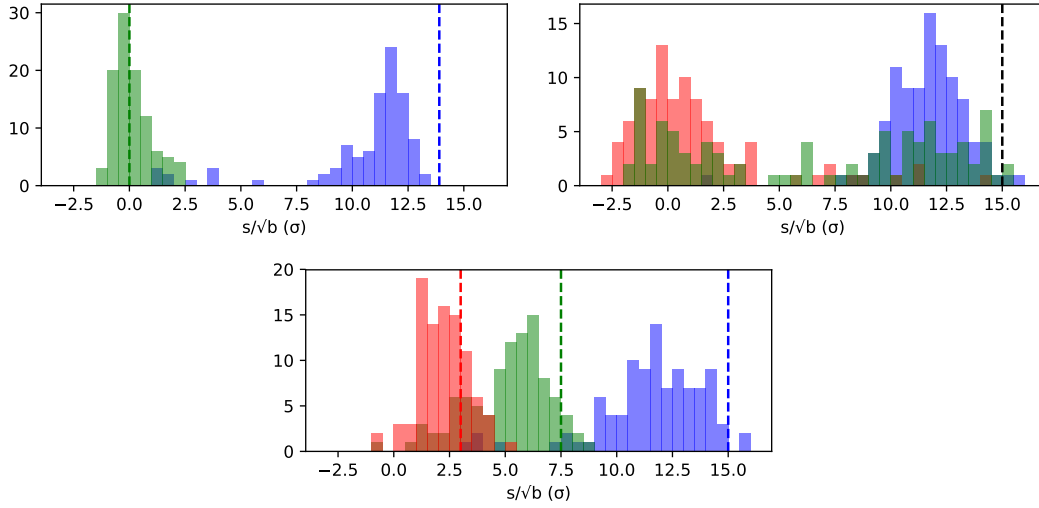
Figure 6. **Top-left:** histogram of significance at a test theshold of 6% for 100 NN trained on the example toy model data (blue), and 100 NN trained on a control dataset with no signal present (green). The dashed green line gives the expected significance of 13.9σ for the example dataset with ideal 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.

structure could find these new intermediate particles and thus also find the heavy resonance. This was the approach taken in [80], but that method is fully supervised and so suffers the usual theory prior bias and potential sources of mismodelling. Here we will illustrate how the CWoLa hunting approach could be used instead to find such a signal. The next section (Sec. 4.1) describes the benchmark model in more detail, as well as the simulation details for both signal and background.

4.1 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, for example, in the warped extra dimensional construction of [22, 81, 82]. The typical opening angle between the two W bosons resulting from the X decay is given by $\Delta R(W, W) \simeq 4 \, m_X/m_{W'}$ 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 [83] v2.5.5 to generate 10^4 signal events, using a model file implementing the tensor couplings of [82] and selecting only the fully hadronic decays of the three W bosons. The events are showered using Pythia 8.226 [84], and are passed through the fast detector simulator Delphes 3.4.1 [85]. Jets are clustered from energy-flow tracks and towers using the FastJet [86] implementation of the anti-k_t algorithm [87] 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 [88] 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 Chromo-dynamic (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 behaviour 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.

4.2   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 can not 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 is that the classifier should be very poor 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 $\tau_1^{(2)}$ [89]. 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 [80, 90–92], 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

$$\text{For each jet:} \quad Y_i = \left(m_J, \ \sqrt{\tau_1^{(2)}}/\tau_1^{(1)}, \ \tau_{21}, \ \tau_{32}, \ \tau_{43}, \ n_{\mathrm{trk}} \right), \tag{4.1}$$

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 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.

Not only is it possible that a classifier may learn systematic correlations between m_{JJ} and Y, it may also learn statistical fluctuations in the background distributions. A naive application of CWoLa directly on the data may produce bumps in m_{JJ} 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. 3, 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 to LHC data 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 five 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

---
Algorithm 1: Nested cross-validation training and event selection procedure.

Split dataset into 5 subsets stratified by m_{JJ} 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 = merge signal bins of subset$_j$

        Training data sideband = merge sideband bins of remaining subsets

        Training data signal = merge signal bins of remaining subsets

        Assign signal region data with label 1

        Assign sideband region 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$_i = \sum_j (\text{model}_{i,j})/4$

    Select $Y\%$ most signal-like data points of subset$_i$, as determined by model$_i$

end

Sum selected events from each subset

Perform statistical test on the selected events for an excess in the signal region

---

the four remaining data subsets. In each round, we set aside one of the remaining subsets as validation data, and 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 labelled as 0 or 1 if they fall in the sideband or signal regions, respectively. For each round, we train twenty 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 $X\%$ 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 $\epsilon_{\text{val}} > X\%$. On the other hand, if no signal is present then $\epsilon_{\text{val}} \simeq X\%$ is expected. Since we will be typically considering $\mathcal{O}(1\%)$-level signals, we will take $X\% \sim 1\%$ in our tests. For each of the twenty models, we end training if its performance has not improved in 300 epochs, and revert the model to a checkpoint saved at peak performance. We select the best of these twenty models, and discard the others. At the end of four rounds, the four selected models are 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 $Y\%$ most signal-like events. This procedure is repeated for all five choices of test set, and the selected events from
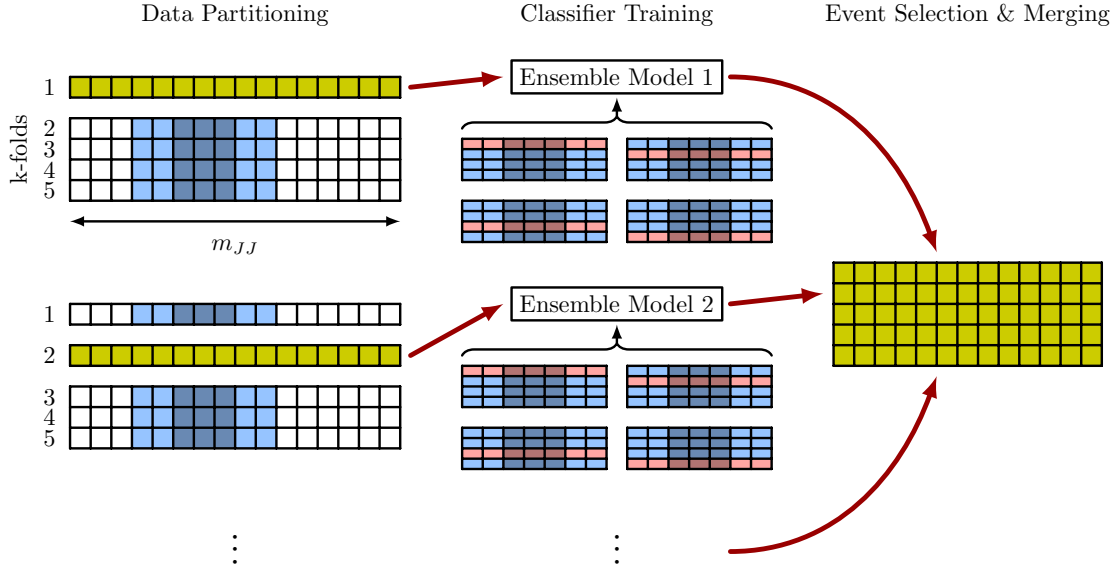
Figure 7. Illustration of the nested cross-validation procedure. **Left:** the dataset is randomly partitioned bin-by-bin into five groups. **Center:** for each group, an ensemble classifier is trained on the remaining groups. For each of the four possible combinations of these four groups into three training groups and one validation group, a set of invidual classifiers are trained and the one with best validation performance is selected. The ensemble classifier is formed by the average of the four selected individual classifiers. **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_{JJ} histogram.

each are combined into a signal histogram in m_{JJ}. 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.

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 (ReLU) activation (using an inactive gradient of 0.1), and second through fourth layers of 32, 16, 4 nodes respectively with Exponential Linear Unit (ELU) activation [93]. The output node has a sigmoid activation. The first three hidden layers are regulated with dropout layers with 20% dropout rate [94]. The neural networks are trained to minimize binary cross-entropy loss using the Adam optimizer with learning rate of 0.001, batch size of 20000, 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
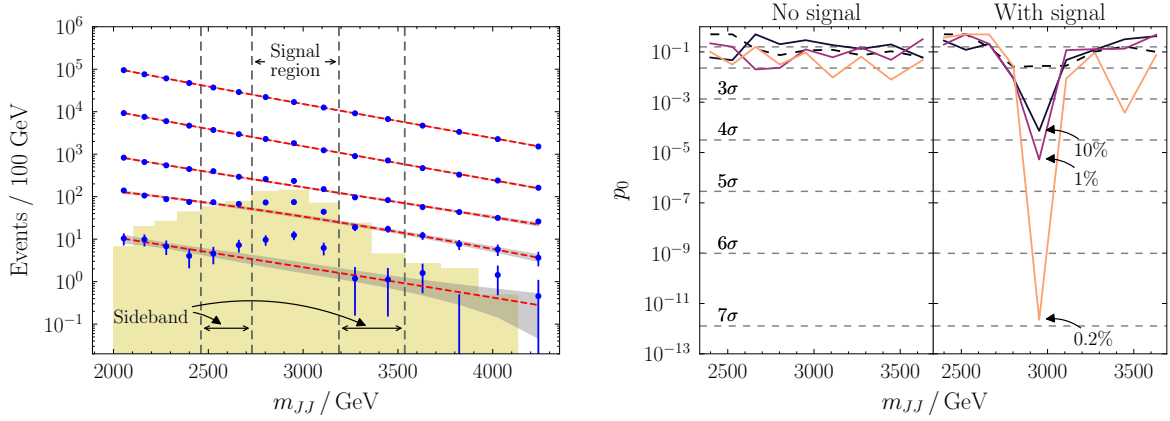
Figure 8. **Left:** m_{JJ} distribution of dijet events (including injected signal, indicated by the filled histogram) before and after applying jet substructure cuts using the NN classifier output for the $m_{JJ} \simeq 3$ TeV mass hypothesis. The dashed red lines indicate the fit to the data points 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}.

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.

4.3   Results

In the first study, we use a sample of 553388 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 $S/\sqrt{B} = 1.8$. Labelling 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.
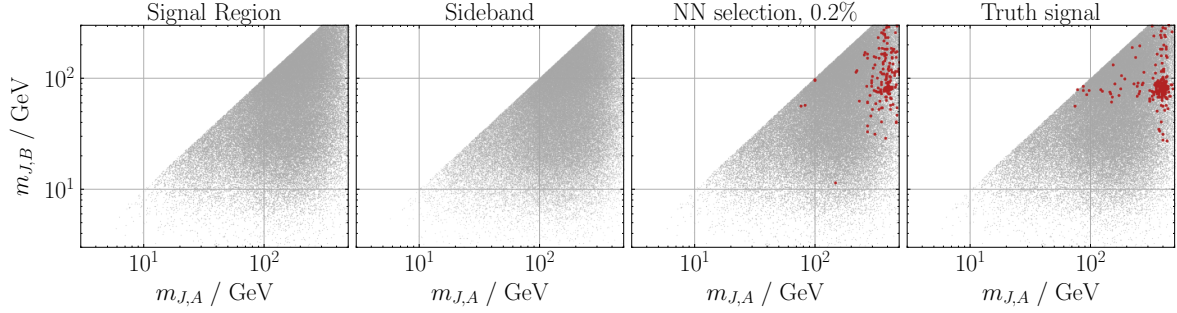
Figure 9. Events projected onto the 2D plane of the two jet masses. The classifiers are trained to discriminate events in the signal region (left plot) from those in the sideband (second plot). The third plot shows in red the 0.2% most signal-like events determined by the classifier trained in this way. The rightmost plot shows in red the truth-level signal events.

In Fig. 8 left, we plot the 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 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. The profile likelihood ratio is used as the test statistic, with the background fit parameters treated as nuisance, with pre-profile uncertainties taken from the background fit itself. The significance is estimated using asymptotic formulae describing the properties of the profile likelihood ratio statistic [95]. The right plot of Fig. 8 shows the signal significance for each signal mass hypothesis, in the case 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 $m_{\rm res} = 3$ 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.

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 observables substructure variables as measured on the same jet. The first column are all events in the signal region. The second column are 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. 4.3, we keep the number of background events fixed but vary the size of the signal, and plot truth-label ROC curves for each example. This allows us to directly asses 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 labelled 1 to 32 indicate the significance 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 labelled 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 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 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
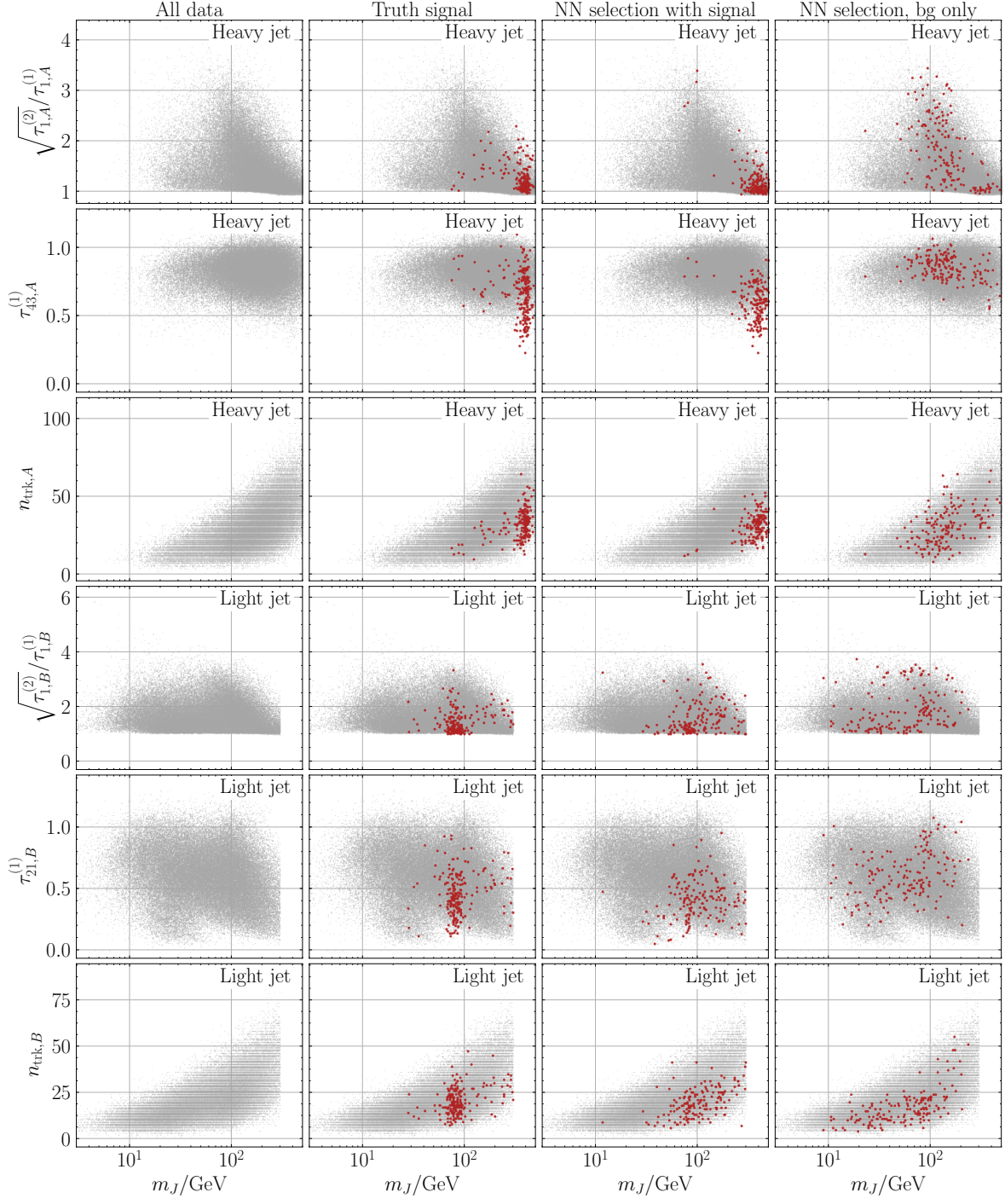
Figure 10. 2D projections of the 12D feature-space of the signal region dataset. **First column:** all signal region events. **Second column:** truth-level simulated signal events highlighted in red. **Third column:** 0.2% most signal-like events selected by the classifier desribed in section X. **Fourth column:** 0.2% most signal-like events selected by a classifier trained on the same sample but with true-signal events removed.
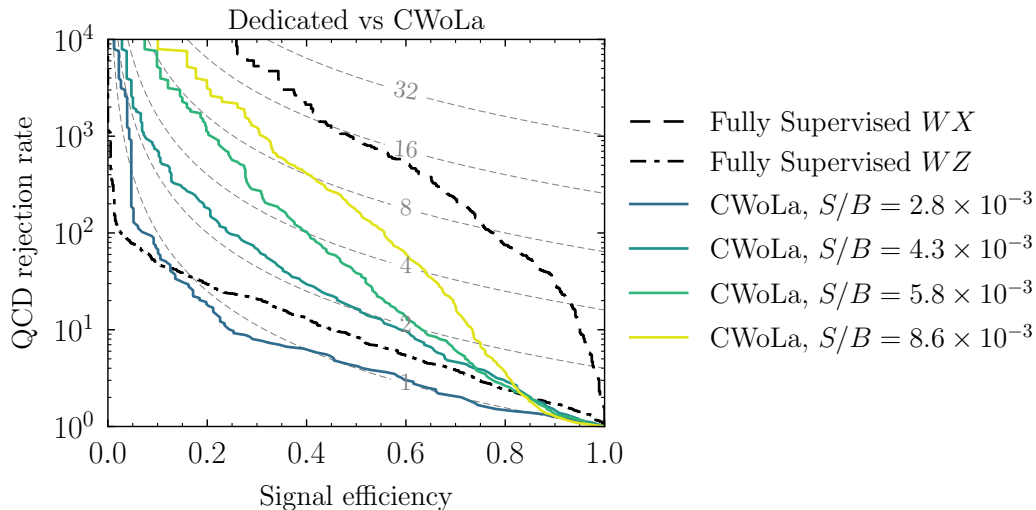
Figure 11. Truth-label ROC curves for taggers trained using CWoLa with varying number of signal events, compared to those for a dedicated tagger trained on pure signal and background samples (solid black) and one trained to discriminate W and Z jets from QCD (dashed black). The CWoLa examples have $B = 81341$ in the signal region and $S = (230, 352, 472, 697)$.

the cuts. This illustrates that CWoLa hunting may find unexpected signals which are not targeted by existing dedicated searches.

5   Conclusions

We have presented a new anomaly detection technique for finding BSM physics signals directly from data. The only information required about the signal is that there is a known, one-dimensional feature space over which it can be localized. 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 [96–101]. A Classification Without Labels approach is used to train fully supervised classifiers to distinguish mass sidebands from the signal region in order to use all of the available information. In our all-hadronic diboson resonance search example, jet substructure information was used to augment information from just the dijet mass and the CWoLa classifier was trained using a deep neural network. 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. By harnessing the power of modern machine learning, CWoLa hunting may provide the key to uncovering BSM physics lurking in the unique datasets already collected by the LHC experiments.

Acknowledgments

A   Statistical Analysis

The significance of the bump is evaluated in the following way. First, we estimate the background by masking out the signal region and performing a least squares fit to the remaining data points using the following three-parameter function (also used in the ATLAS [9] and CMS [102] searches for fully hadronic diboson resonances):

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

The systematic uncertainty in this fit is estimated by propagating linearly the uncertainties on the fit parameters onto uncertainties in the binned signal rate predictions. The fits and fit uncertainties are indicated in the left plot in Fig. 8 by the red dashed lines and gray bands.

Using the data points in the signal region, we then form the profile likelihood ratio

$$\lambda_0 = \frac{\mathcal{L}(\mu = 0, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\theta})} \tag{A.2}$$

where μ indicates the signal rate and θ are nuisance parameters associated with the systematic uncertainties on the background prediction and with the signal shape. In the numerator, $\hat{\hat{\theta}}$ represents the best fit value for the nuisance parameters for fixes signal rate $\mu = 0$, while in the denominator $\hat{\mu}$ and $\hat{\theta}$ represent the combined best fit for μ and θ. The likelihoods are formed from product of a Poisson factor for each bin, and a Gaussian constraint for the background nuisance parameters

$$\mathcal{L}(\mu, \theta) = \left(\prod^{i} \text{Poiss}(n_i | b_i + \theta_i^{(\text{bg})} + s_i) \right) \exp\left(-\frac{1}{2} \vec{\theta}^{(\text{bg})} \cdot \Sigma^{-1} \cdot \vec{\theta}^{(\text{bg})} \right) \tag{A.3}$$

where n_i is the observed number of events in bin i, b_i is the number of events predicted by the sideband fit, $\theta_i^{(\text{bg})}$ is the nuisance parameter associated with the systematic uncertainty

for the background prediction, and Σ is the covariance matrix for these nuisance parameters. The parameter s_i is the signal hypothesis for bin i. The signal rate in each bin is determined by the overall rate μ, and two shape parameters determining the fraction of events fall into each of the three bins

$$(s_1, s_2, s_3) = (f_1\mu, f_2(1 - f_1)\mu, (1 - f_1 - f_2(1 - f_1))\mu) \,. \tag{A.4}$$

The shape parameters f_1 and f_2 are profiled in the fit, subject to the contraint $0 \leq f_1, f_2 \leq 1$.

Our test statistic is

$$q_0 = \begin{cases} -2\log(\lambda_0), & \hat{\mu} > 0, \\ 0, & \hat{\mu} \leq 0. \end{cases} \tag{A.5}$$

Using asymptotic formulae [95], gives a significance $Z = \sqrt{q_0}$ and $p_0 = 1 - \Phi(Z)$, where Φ is the cumulative distribution function of the normal distribution.
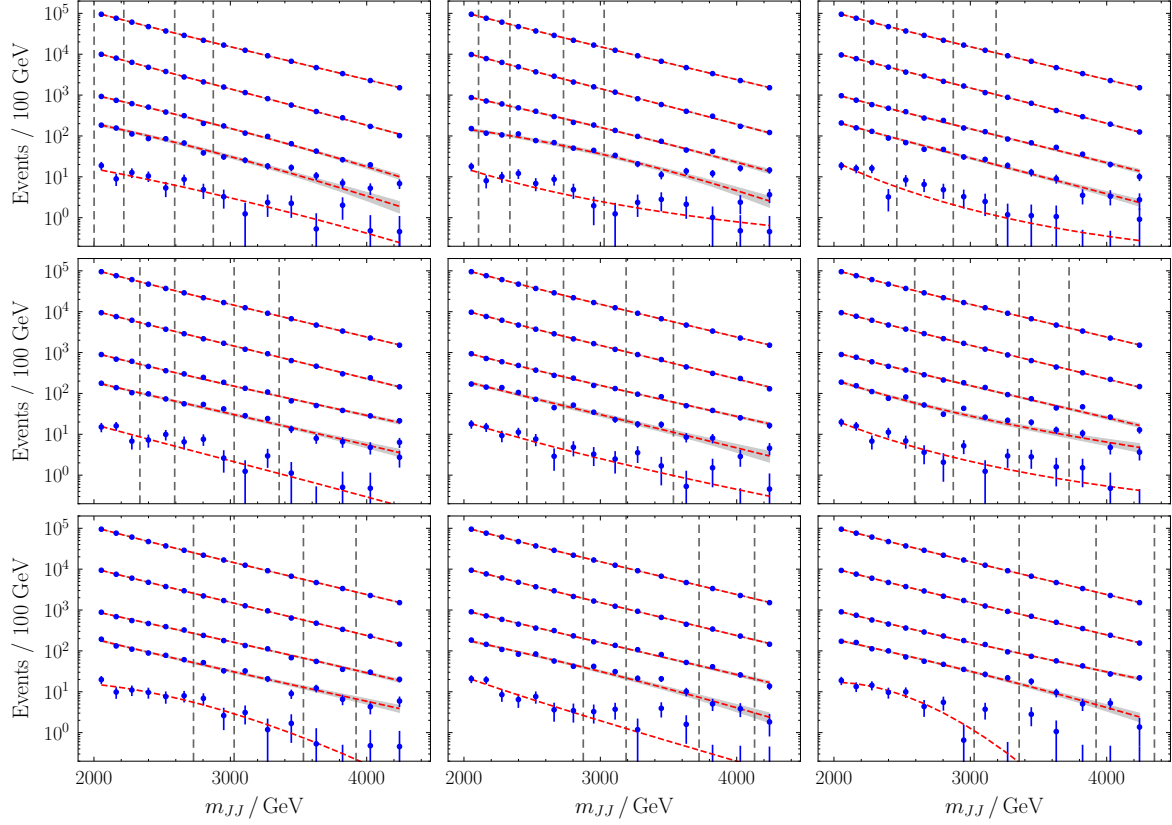
B Dijet Mass Scans

In Figs. 12, 13 we plot the dijet invariant mass distributions before and after applying tagger cuts over the full range of the mass scan described in Sec. 4. The p-values calculated from the top four distributions in these plots are displayed in Fig. 8 (right).

References

[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* $\overline{p} + p \rightarrow 2\pi^+ + 2\pi^- + n\pi^0$, *Phys. Rev.* **126** (Jun, 1962) 1858–1863.

[2] **ATLAS** Collaboration, G. Aad et al., *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett.* **B716** (2012) 1–29, [arXiv:1207.7214].

[3] **CMS** Collaboration, S. Chatrchyan et al., *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys. Lett.* **B716** (2012) 30–61, [arXiv:1207.7235].

[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, CERN,Geneva, Switzerland 17-20 January 2011*, 2011. arXiv:1101.0390.

[5] **ATLAS** Collaboration, M. Aaboud et al., *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.* **D96** (2017), no. 5 052004, [arXiv:1703.09127].

[6] **CMS** Collaboration, A. M. Sirunyan et al., *Search for dijet resonances in proton?proton collisions at* $\sqrt{s} = 13$ *TeV and constraints on dark matter and other models*, *Phys. Lett.* **B769** (2017) 520–542, [arXiv:1611.03568]. [Erratum: Phys. Lett.B772,882(2017)].

[7] **CMS** Collaboration, V. Khachatryan et al., *Search for narrow resonances decaying to dijets in proton-proton collisions at* $\sqrt{(s)} = 13$ *TeV*, *Phys. Rev. Lett.* **116** (2016), no. 7 071801, [arXiv:1512.01224].

Figure 12. 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 $10^{-1}, 10^{-2}, 2 \times 10^{-3}$, 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 ommitted.
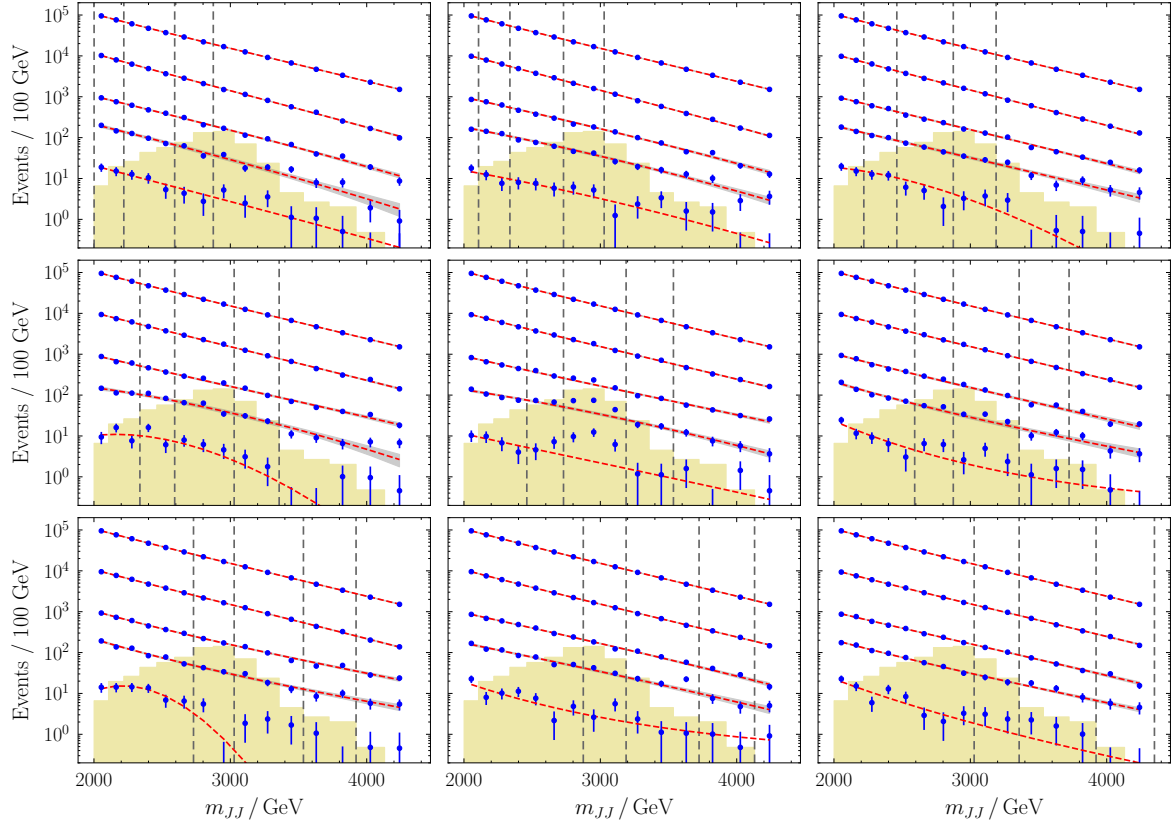
Figure 13. Dijet invariant mass distributions in the resonance mass scan in the case that the signal is present (indicated by the filled histogram). Otherwise, the plots are as described in the caption to Fig. 12.

[8] **CMS** Collaboration, A. M. Sirunyan et al., *Search for massive resonances decaying into WW, WZ, ZZ, qW, and qZ with dijet final states at sqrt(s) = 13 TeV*, arXiv:1708.05379.

[9] **ATLAS** Collaboration, M. Aaboud et al., *Search for diboson resonances with boson-tagged jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Lett. **B777** (2018) 91–113, [arXiv:1708.04445].

[10] **CMS** Collaboration, A. M. Sirunyan et al., *Search for high-mass $Z\gamma$ resonances in proton-proton collisions at $\sqrt{s} = 8$ and 13 TeV using jet substructure techniques*, Phys. Lett. **B772** (2017) 363–387, [arXiv:1612.09516].

[11] **CMS** Collaboration, A. M. Sirunyan et al., *Search for Low Mass Vector Resonances Decaying to Quark-Antiquark Pairs in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV*, Phys. Rev. Lett. **119** (2017), no. 11 111802, [arXiv:1705.10532].

[12] **CMS** Collaboration, A. M. Sirunyan et al., *Search for low mass vector resonances decaying into quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **01** (2018) 097, [arXiv:1710.00159].

[13] **ATLAS** Collaboration, M. Aaboud et al., *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*, arXiv:1801.08769.

[14] **CMS** Collaboration, A. M. Sirunyan et al., *Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair*, Phys. Rev. Lett. **120** (2018), no. 7 071802, [arXiv:1709.05543].

[15] **CMS** Collaboration, A. M. Sirunyan et al., *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*, arXiv:1710.04960.

[16] **ATLAS** Collaboration, M. Aaboud et al., *Search for heavy resonances decaying to a W or Z boson and a Higgs boson in the $q\bar{q}^{(')}b\bar{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Lett. **B774** (2017) 494–515, [arXiv:1707.06958].

[17] **ATLAS** Collaboration, M. Aaboud et al., *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. **B759** (2016) 229–246, [arXiv:1603.08791].

[18] **CMS** Collaboration, A. M. Sirunyan et al., *Search for $t\bar{t}$ resonances in highly boosted lepton+jets and fully hadronic final states in proton-proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **07** (2017) 001, [arXiv:1704.03366].

[19] **ATLAS** Collaboration, M. Aaboud et al., *Search for $W' \rightarrow tb$ decays in the hadronic final state using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, arXiv:1801.07893.

[20] J. A. Aguilar-Saavedra, *Running bumps from stealth bosons*, Eur. Phys. J. **C78** (2018), no. 3 206, [arXiv:1801.08129].

[21] J. A. Aguilar-Saavedra, *Stealth multiboson signals*, Eur. Phys. J. **C77** (2017), no. 10 703, [arXiv:1705.07885].

[22] 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*, To Appear.

[23] **ATLAS Collaboration** Collaboration, *A model independent general search for new phenomena with the ATLAS detector at $\sqrt{s} = 13$ TeV*, Tech. Rep. ATLAS-CONF-2017-001, CERN, Geneva, Jan, 2017.

[24] **CMS Collaboration** Collaboration, *Model Unspecific Search for New Physics in pp Collisions at sqrt(s) = 7 TeV*, Tech. Rep. CMS-PAS-EXO-10-021, CERN, Geneva, 2011.

[25] **ATLAS Collaboration** Collaboration, *A general search for new phenomena with the ATLAS detector in pp collisions at sort(s)=7 TeV.*, Tech. Rep. ATLAS-CONF-2012-107, CERN, Geneva, Aug, 2012.

[26] *A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV*, Tech. Rep. ATLAS-CONF-2014-006, CERN, Geneva, Mar, 2014.

[27] **H1** Collaboration, A. Aktas et al., *A General search for new phenomena in ep scattering at HERA*, *Phys. Lett.* **B602** (2004) 14–30, [`hep-ex/0408044`].

[28] **H1** Collaboration, F. D. Aaron et al., *A General Search for New Phenomena at HERA*, *Phys. Lett.* **B674** (2009) 257–268, [`arXiv:0901.0507`].

[29] **D0** Collaboration, B. Abbott et al., *Search for new physics in eμX data at DØ using Sherlock: A quasi model independent search strategy for new physics*, *Phys. Rev.* **D62** (2000) 092004, [`hep-ex/0006011`].

[30] **D0** Collaboration, V. M. Abazov et al., *A Quasi model independent search for new physics at large transverse momentum*, *Phys. Rev.* **D64** (2001) 012004, [`hep-ex/0011067`].

[31] **CDF** Collaboration, T. Aaltonen et al., *Model-Independent and Quasi-Model-Independent Search for New Physics at CDF*, *Phys. Rev.* **D78** (2008) 012002, [`arXiv:0712.1311`].

[32] **CDF** Collaboration, T. Aaltonen et al., *Global Search for New Physics with 2.0 fb^{-1} at CDF*, *Phys. Rev.* **D79** (2009) 011101, [`arXiv:0809.3781`].

[33] 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`].

[34] J. Cogan, M. Kagan, E. Strauss, and A. Schwarztman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, *JHEP* **02** (2015) 118, [`arXiv:1407.5675`].

[35] L. G. Almeida, M. Backović, M. Cliche, S. J. Lee, and M. Perelstein, *Playing Tag with ANN: Boosted Top Identification with Pattern Recognition*, *JHEP* **07** (2015) 086, [`arXiv:1501.05968`].

[36] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, *Jet-images - deep learning edition*, *JHEP* **07** (2016) 069, [`arXiv:1511.05190`].

[37] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, *Jet Substructure Classification in High-Energy Physics with Deep Neural Networks*, *Phys. Rev.* **D93** (2016), no. 9 094034, [`arXiv:1603.09349`].

[38] J. Barnard, E. N. Dawe, M. J. Dolan, and N. Rajcic, *Parton Shower Uncertainties in Jet Substructure Analyses with Deep Neural Networks*, *Phys. Rev.* **D95** (2017), no. 1 014018, [`arXiv:1609.00607`].

[39] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, *Deep-learning Top Taggers or The End of QCD?*, *JHEP* **05** (2017) 006, [`arXiv:1701.08784`].

[40] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, *Deep-learned Top Tagging with a Lorentz Layer*, arXiv:1707.08966.

[41] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, JHEP **01** (2017) 110, [arXiv:1612.01551].

[42] G. Louppe, K. Cho, C. Becot, and K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics*, arXiv:1702.00748.

[43] **ATLAS Collaboration** Collaboration, *Performance of Top Quark and W Boson Tagging in Run 2 with ATLAS*, Tech. Rep. ATLAS-CONF-2017-064, CERN, Geneva, Aug, 2017.

[44] **ATLAS Collaboration** Collaboration, *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*, Tech. Rep. ATL-PHYS-PUB-2017-013, CERN, Geneva, Jul, 2017.

[45] **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*, ATL-PHYS-PUB-2017-004 (2017).

[46] **CMS** Collaboration, *Heavy flavor identification at CMS with deep neural networks*, CMS-DP-2017-005 (Mar, 2017).

[47] **CMS** Collaboration, *CMS Phase 1 heavy flavour identification performance and developments*, CMS-DP-2017-013 (May, 2017).

[48] **ATLAS** Collaboration, *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*, ATL-PHYS-PUB-2017-003 (2017).

[49] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, *Jet Constituents for Deep Neural Network Based Top Quark Tagging*, arXiv:1704.02124.

[50] K. Datta and A. Larkoski, *How Much Information is in a Jet?*, JHEP **06** (2017) 073, [arXiv:1704.08249].

[51] K. Datta and A. J. Larkoski, *Novel Jet Observables from Machine Learning*, JHEP **03** (2018) 086, [arXiv:1710.01305].

[52] **ATLAS** Collaboration, *Quark versus Gluon Jet Tagging Using Jet Images with the ATLAS Detector*, ATL-PHYS-PUB-2017-017 (2017).

[53] **CMS** Collaboration, *New Developments for Jet Substructure Reconstruction in CMS*, CMS-DP-2017-027 (2017).

[54] K. Fraser and M. D. Schwartz, *Jet Charge and Machine Learning*, arXiv:1803.08066.

[55] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, *JUNIPR: a Framework for Unsupervised Machine Learning in Particle Physics*, arXiv:1804.09720.

[56] S. Macaluso and D. Shih, *Pulling Out All the Tops with Computer Vision and Deep Learning*, arXiv:1803.00107.

[57] **ATLAS** Collaboration, G. Aad et al., *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11** (2016), no. 04 P04008, [arXiv:1512.01094].

[58] **CMS** Collaboration, S. Chatrchyan et al., *Identification of b-quark jets with the CMS experiment*, JINST **8** (2013) P04013, [arXiv:1211.4462].

[59] **ATLAS** Collaboration, G. Aad et al., *Light-quark and gluon jet discrimination in pp collisions at* $\sqrt{s} = 7$ TeV *with the ATLAS detector*, *Eur. Phys. J.* **C74** (2014), no. 8 3023, [arXiv:1405.6583].

[60] **CMS** Collaboration, *Performance of quark/gluon discrimination in 8 TeV pp data*, Tech. Rep. CMS-PAS-JME-13-002, 2013.

[61] **CMS Collaboration** Collaboration, *Performance of quark/gluon discrimination in 13 TeV data*, .

[62] **ATLAS** Collaboration, G. Aad et al., *Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at* $\sqrt{s} = 8$ *TeV*, *Eur. Phys. J.* **C76** (2016), no. 3 154, [arXiv:1510.05821].

[63] **CMS** Collaboration, V. Khachatryan et al., *Identification techniques for highly boosted W bosons that decay into hadrons*, *JHEP* **12** (2014) 017, [arXiv:1410.4227].

[64] **ATLAS** Collaboration, G. Aad et al., *Identification of high transverse momentum top quarks in pp collisions at* $\sqrt{s} = 8$ *TeV with the ATLAS detector*, *JHEP* **06** (2016) 093, [arXiv:1603.03127].

[65] **CMS** Collaboration, *Boosted Top Jet Tagging at CMS*, Tech. Rep. CMS-PAS-JME-13-007, 2014.

[66] G. Louppe, M. Kagan, and K. Cranmer, *Learning to Pivot with Adversarial Networks*, arXiv:1611.01046.

[67] L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman, *Weakly Supervised Classification in High Energy Physics*, *JHEP* **05** (2017) 145, [arXiv:1702.00414].

[68] E. M. Metodiev, B. Nachman, and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174, [arXiv:1708.02949].

[69] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, *Learning to Classify from Impure Samples*, arXiv:1801.10158.

[70] and, *Ix. on the problem of the most efficient tests of statistical hypotheses*, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **231** (1933), no. 694-706 289–337, [http://rsta.royalsocietypublishing.org/content/231/694-706/289.full.pdf].

[71] F. Chollet, "Keras." https://github.com/fchollet/keras, 2017.

[72] 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, pp. 265–283, 2016.

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

[74] **CMS Collaboration** Collaboration, *W/Z/Top tagging for Run 2 in CMS*, .

[75] **CMS Collaboration** Collaboration, *b-tagging in boosted topologies*, .

[76] **ATLAS Collaboration** Collaboration, *Boosted Higgs* $(\to b\bar{b})$ *Boson Identification with the ATLAS Detector at* $\sqrt{s} = 13$ *TeV*, Tech. Rep. ATLAS-CONF-2016-039, CERN, Geneva, Aug, 2016.

[77] **CMS Collaboration** Collaboration, *Top Tagging with New Approaches*, Tech. Rep. CMS-PAS-JME-15-002, CERN, Geneva, 2016.

[78] **CMS Collaboration** Collaboration, *Performance of heavy flavour identification algorithms in proton-proton collisions at 13 TeV at the CMS experiment*, .

[79] **ATLAS Collaboration** Collaboration, *Quark versus Gluon Jet Tagging Using Charged Particle Multiplicity with the ATLAS Detector*, Tech. Rep. ATL-PHYS-PUB-2017-009, CERN, Geneva, May, 2017.

[80] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, *A generic anti-QCD jet tagger*, *JHEP* **11** (2017) 163, [arXiv:1709.01087].

[81] K. Agashe, P. Du, S. Hong, and R. Sundrum, *Flavor Universal Resonances and Warped Gravity*, *JHEP* **01** (2017) 016, [arXiv:1608.00526].

[82] 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.

[83] 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*, *JHEP* **07** (2014) 079, [arXiv:1405.0301].

[84] T. Sjostrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852–867, [arXiv:0710.3820].

[85] **DELPHES 3** Collaboration, J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057, [arXiv:1307.6346].

[86] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, *Eur. Phys. J.* **C72** (2012) 1896, [arXiv:1111.6097].

[87] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k(t) jet clustering algorithm*, *JHEP* **04** (2008) 063, [arXiv:0802.1189].

[88] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, *Soft Drop*, *JHEP* **05** (2014) 146, [arXiv:1402.2657].

[89] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, *JHEP* **02** (2012) 093, [arXiv:1108.2701].

[90] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, *Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure*, *JHEP* **05** (2016) 156, [arXiv:1603.00027].

[91] 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.* **D96** (2017), no. 7 074034, [arXiv:1703.03507].

[92] I. Moult, B. Nachman, and D. Neill, *Convolved Substructure: Analytically Decorrelating Jet Substructure Observables*, arXiv:1710.06859.

[93] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, *Fast and accurate deep network learning by exponential linear units (elus)*, *arXiv preprint arXiv:1511.07289* (2015).

[94] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Dropout: A simple way to prevent neural networks from overfitting*, *Journal of Machine Learning Research* **15** (2014) 1929–1958.

[95] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J.* **C71** (2011) 1554, [arXiv:1007.1727]. [Erratum: Eur. Phys. J.C73,2501(2013)].

[96] **ATLAS** Collaboration, M. Aaboud et al., *A search for pair-produced resonances in four-jet final states at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Eur. Phys. J.* **C78** (2018), no. 3 250, [arXiv:1710.07171].

[97] **CMS** Collaboration, C. Collaboration, *A search for light pair-produced resonances decaying into at least four quarks*, .

[98] **ATLAS** Collaboration, G. Aad et al., *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.* **C73** (2013), no. 1 2263, [arXiv:1210.4826].

[99] **ATLAS** Collaboration, G. Aad et al., *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*, *JHEP* **06** (2016) 067, [arXiv:1601.07453].

[100] **CMS** Collaboration, S. Chatrchyan et al., *Search for pair-produced dijet resonances in four-jet final states in pp collisions at $\sqrt{s}=7??TeV$*, *Phys. Rev. Lett.* **110** (2013), no. 14 141802, [arXiv:1302.0531].

[101] **CMS** Collaboration, V. Khachatryan et al., *Search for pair-produced resonances decaying to jet pairs in proton?proton collisions at $\sqrt{s} = 8$ TeV*, *Phys. Lett.* **B747** (2015) 98–119, [arXiv:1412.7706].

[102] **CMS** Collaboration, A. M. Sirunyan et al., *Search for massive resonances decaying into WW, WZ or ZZ bosons in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **03** (2017) 162, [arXiv:1612.09159].