UC Berkeley
UC Berkeley Previously Published Works

Title
Comparing weak- and unsupervised methods for resonant anomaly detection

Permalink
https://escholarship.org/uc/item/7bz866xk

Journal
European Physical Journal C, 81(7)

ISSN
1434-6044

Authors
Collins, Jack H
Martín-Ramiro, Pablo
Nachman, Benjamin
et al.

Publication Date
2021-07-01

DOI
10.1140/epjc/s10052-021-09389-x

Peer reviewed

THE EUROPEAN
PHYSICAL JOURNAL C

Comparing weak- and unsupervised methods for resonant anomaly detection

Jack H. Collins[1,a], **Pablo Martín-Ramiro**[2,3,b], **Benjamin Nachman**[3,4,c], **David Shih**[5,d]

[1] SLAC National Accelerator Laboratory, Stanford University, Stanford, CA 94309, USA
[2] Instituto de Física Teórica, IFT-UAM/CSIC, Universidad Autónoma de Madrid, 28049 Madrid, Spain
[3] Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[4] Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA
[5] NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854, USA

Abstract Anomaly detection techniques are growing in importance at the Large Hadron Collider (LHC), motivated by the increasing need to search for new physics in a model-agnostic way. In this work, we provide a detailed comparative study between a well-studied unsupervised method called the autoencoder (AE) and a weakly-supervised approach based on the Classification Without Labels (CWoLa) technique. We examine the ability of the two methods to identify a new physics signal at different cross sections in a fully hadronic resonance search. By construction, the AE classification performance is independent of the amount of injected signal. In contrast, the CWoLa performance improves with increasing signal abundance. When integrating these approaches with a complete background estimate, we find that the two methods have complementary sensitivity. In particular, CWoLa is effective at finding diverse and moderately rare signals while the AE can provide sensitivity to very rare signals, but only with certain topologies. We therefore demonstrate that both techniques are complementary and can be used together for anomaly detection at the LHC.

Contents

[a] e-mail: jcollins@slac.stanford.edu
[b] e-mail: pmartin.ramiro@predoc.uam.es (corresponding author)
[c] e-mail: bpnachman@lbl.gov
[d] e-mail: shih@physics.rutgers.edu

1 Introduction

The LHC has the potential to address many of the most fundamental questions in physics. Despite all the searches for physics beyond the Standard Model (BSM) conducted by ATLAS [1,2] and CMS [3–5], no significant evidence of new physics has been found so far. These searches are designed to target specific new physics signals that would be produced by particular, well-motivated theoretical models. However, it is not feasible to perform a dedicated analysis for every possible topology and therefore some potential signals may be missed. This motivates the introduction of new methods that are less reliant on model assumptions and that are sensitive to a broad spectrum of new physics signatures.

A variety of machine-learning assisted anomaly detection techniques have been proposed that span the spectrum from completely supervised to completely unsupervised[1] [23–63] (see Refs. [60,64] for an overview). Two promising approaches are CWoLa Hunting [24,25] and deep autoencoders (AE) [27–32]:

- CWoLa Hunting is a weakly-supervised anomaly detection technique that uses the idea of Classification Without Labels (CWoLa) [65] and trains a classifier to distinguish two statistical mixed samples (typically a signal

---

[1] Citation block taken from the Living Review [6]. Background model dependent, non-machine learning models have also been studied experimentally - see Refs. [7–22].

region and a sideband region when used to search for new physics [24,25]) with different amounts of (potential) signal. The output of this classifier can then be used to select signal-like events. This method has already been tested in a real search by the ATLAS collaboration [44].

- Autoencoders are the basis for a fully-unsupervised anomaly detection technique that has been widely explored and used in many real-world scenarios. A deep autoencoder is a neural network that learns to compress data into a small latent representation and then reconstruct the original input from the compressed version. The AE can be trained directly on a background-rich sample to learn the features of background events and reconstruct them well. By contrast, it will struggle to reconstruct anomalous (e.g. signal) events. The reconstruction loss, defined by some chosen distance measure between the original and reconstructed event, can then be used as a classification score that selects anomalous events.

To date, there has not been a direct and detailed comparison between these two methods.[2] The goal of this paper will be to provide such a comparison, describe the strengths and weaknesses of the two approaches, and highlight their areas of complementarity.

We will focus on the new physics scenario where a signal is localized in one known dimension of phase space (in this case, the dijet invariant mass) on top of a smooth background. While CWoLa Hunting explicitly requires a setup like this to generate mixed samples, AEs technically do not, as they can function as anomaly detectors in a fully unsupervised setting. However, even for AEs one generally needs to assume something about the signal and the background in order to enable robust, data-driven background estimation.

In this scenario, both models can be trained to exploit the information in the substructure of the two jets to gain discriminating power between the signal and background events. CWoLa Hunting, being able to take advantage of the weak labels, should excel in the limit of moderately high signal rate in the sample because it is able to take advantage of learnt features of the signal. It should fail however in the limit of no signal. On the other hand, an unsupervised approach like the AE is fully agnostic to the specific features of the signal, and thus should be robust in the limit of low signal statistics. While the behaviour of these strategies in the high and low signal statistics limits can be understood on general grounds, it is the intermediate regime in which the two strategies might have a 'cross-over' in performance that is of most interest for

realistic searches. It is therefore worth analyzing in detail for some case studies the nature of this crossover and the degree of complementary of the strategies.

In this work, we provide a detailed comparative analysis of the performance of CWoLa Hunting and AEs at anomaly detection on a fully hadronic resonance search. After evaluating the ability of both methods to identify the signal events for different cross sections, we test whether they are able to increase the significance of the signal region excess. Here we emphasize the importance of going beyond the AUC metric and consider more meaningful performance metrics such as the Significance Improvement Characteristic (SIC). Furthermore, a realistic fit procedure based on ATLAS and CMS hadronic diboson searches is implemented. We will confirm the general behavior of AE and CWoLa Hunting approaches at large and small signal strengths described in the previous paragraph, and we will demonstrate quantitatively the existence of a cross-over region in a part of parameter space that could be of practical relevance. We conclude that the approaches have complementary sensitivity to different amounts or types of signals.

This paper is organized as follows. In Sect. 2, we describe the resonant hadronic new physics signal that we consider and the simulation details for the generated events. In Sect. 3, we introduce the details of CWoLa Hunting and the AE and explain how they can be successfully implemented in this type of new physics searches. We present results for the two models in Sect. 4 and discuss their performance at anomaly detection. Finally, the conclusions are presented in Sect. 5.

2 Simulation

In order to investigate the performance of CWoLa Hunting and AEs in a generic hadronic resonance search, we consider a benchmark new physics signal $pp \rightarrow Z' \rightarrow XY$, with $X \rightarrow jjj$ and $Y \rightarrow jjj$. There is currently no dedicated search to this event topology. The mass of the new heavy particle is set to $m_{Z'} = 3.5$ TeV, and we consider two scenarios for the masses of the new lighter particles: m_X, $m_Y = 500$ GeV and m_X, $m_Y = 300$ GeV. These signals typically produce a pair of large-radius jets J with invariant mass $m_{JJ} \simeq 3.5$ TeV, with masses of $m_J = 500, 300$ GeV and a three-prong substructure. These signals are generated in the LHC Olympics framework [60].

For both signal models, we generated 10^4 events. One million QCD dijet events serve as the background and are from the LHC Olympics [60] dataset. All the events were produced and showered using Pythia 8.219 [66] and the detector simulation was performed using Delphes 3.4.1 [67], with no pileup or multiparton interactions included. All jets are clustered with FastJet 3.3.2 [68] using the anti-k_t algorithm [69] with radius parameter $R = 1$.

---

[2] Recently, the authors of the Tag N' Train method [47] also made comparisons between these approaches with the aim of combining them. Our study has the orthogonal goal of directly comparing the two approaches in detail as distinct methods to understand their complementarity.

We require events to have at least one large-radius jet with $p_T > 1.2$ TeV and pseudo-rapidity $|\eta| < 2.5$. The two hardest jets are selected as the candidate dijet and a set of substructure variables are calculated for these two jets as shown in Fig. 1. In particular, the N-subjettiness variables τ_i^β were first proposed in Refs. [70,71] and probe the extent to which a jet has N subjets. All N-subjettiness variables are computed using FastJet 3.3.2 and angular exponent $\beta = 1$ unless otherwise specified in the superscript. The observable n_{trk} denotes the number of constituents in a given jet. Jets are ordered by mass in descending order.

3 Machine learning setup

In this section, we describe the machine learning setup and the strategies that we follow to train CWoLa Hunting and the AE approaches.

3.1 Classification without labels (CWoLa)

The strategy closely follows Refs. [24,25]. To begin, we use a set of high-level observables computed from the two leading jets. In particular, we consider the following set of input features for each jet:

$$Y_i = \left\{ m_J, \; \sqrt{\tau_1^{(2)}}/\tau_1^{(1)}, \; \tau_{21}, \; \tau_{32}, \; \tau_{43}, \; n_{\text{trk}} \right\}. \tag{3.1}$$

A reduced set of input features is shown in Fig. 1. Importantly, the correlation between this set of input features and m_{JJ} is minimal and not sufficient to sculpt artificial bumps in the absence of signal, as we will demonstrate in Sect. 4.

We select all of the events in the range $m_{JJ} \in [2800, 5200]$ GeV and split them uniformly in $\log(m_{JJ})$ in 30 bins. After selecting this range, 537304 background events remain in our sample. In order to test for a signal hypothesis with mass $m_{JJ} = m_{\text{peak}}$, where m_{peak} is the mean mass of the injected signal, we build a signal region and a sideband region. The former contains all of the events in the four bins centered around m_{peak}, while the latter is built using the three bins below and above the signal region. By doing this, we obtain a signal region in the range $m_{JJ} \in (3371, 3661)$ GeV with a width of 290 GeV, and a lower and upper sidebands that are 202 GeV and 234 GeV wide, respectively. The size of the signal region window depends on the signal width[3] and can be scanned for optimal performance. In Fig. 2, we show the binned distribution of a fraction of signal and all background events, with a signal-to-background ratio of $S/B = 6 \times 10^{-3}$

and a naive expected significance $S/\sqrt{B} = 1.8\sigma$ in the signal region. Note that if a signal is present in data, the signal region will have a larger density of signal events than the mass sidebands, which are mainly populated by background events by construction. In a real search the location of the mass peak of any potential signal would be unknown, and thus the mass hypothesis must be scanned, as described in Ref. [25].

After defining the signal and sideband regions, a CWoLa classifier is trained to distinguish the events of the signal region from the events of the sideband using the set of twelve input features that describe the jet substructure of each event, presented in Eq. (3.1). In this way, the CWoLa classifier will ideally learn the signal features that are useful to discriminate between both regions. It is important to remark that the classifier performance should be very poor when no signal is present in the signal region, but if a signal is present with anomalous jet substructure then the classifier should learn the information that is useful to distinguish the signal and sideband regions.

In this work, the classifiers that we use are fully connected neural networks with four hidden layers. The first layer has 64 nodes and a leaky Rectified Linear Unit (ReLU) [72] activation [73] (with an inactive gradient of 0.1), and the second through fourth layers have 32, 16 and 4 nodes respectively, with Exponential Linear Unit (ELU) activation [74]. The output layer has a sigmoid activation. The first three hidden layers are followed by dropout layers with a 20% dropout rate [75]. We use the binary cross-entropy loss function and the Adam optimizer [76] with learning rate of 0.001 and learning rate decay of 5×10^{-4}, batch size of 20480 and first and second moment decay rates of 0.8 and 0.99, respectively. The training data is reweighted such that the low and high sidebands have equal total weight, 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 reweighting procedure ensures that the two sideband regions have the same contribution to the training process in spite of their different event rates, and results in a classifier output peaked around 0.5 in the absence of any signal. All classifiers are implemented and trained using Keras [77] with TensorFlow [78] backend.

We implement a nested cross-validation procedure with five k-folds and therefore all data are used for training, validation and testing. We standardize all the input features from the training and validation sets using training information, and those from the test set using training and validation information. The full dataset is divided randomly, bin by bin, in five event samples of identical size. We set one of the samples aside for testing and perform four rounds of training and validation with the other four, using one of the subsets for validation each time. For each round, we train ten neural networks for 700 epochs on the same training and validation

---

[3] This is dominated by detector effects; for models with a non-trivial off-shell width, this may not be optimal.
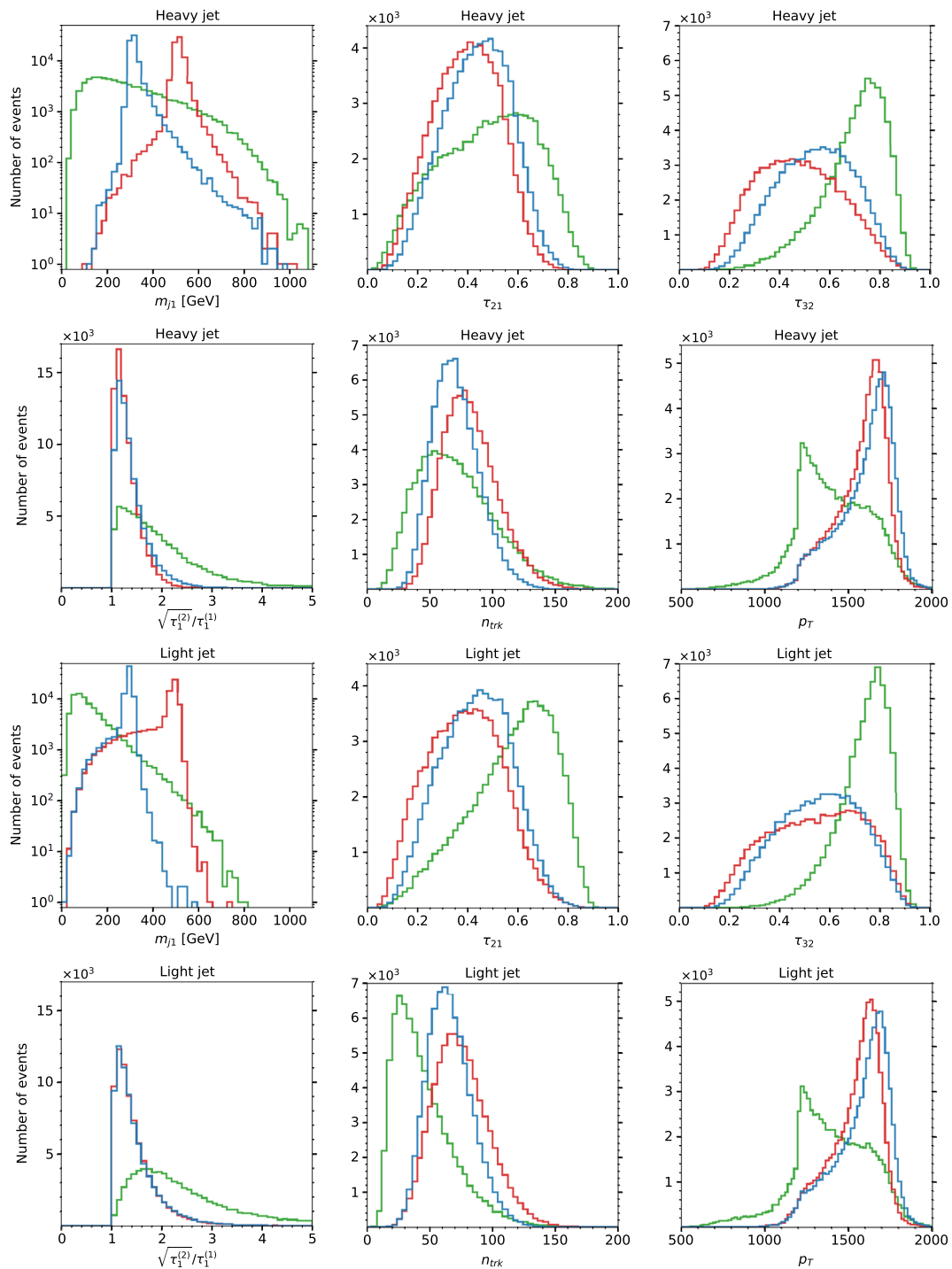
Fig. 1 A reduced set of the input features that we use for training the models are shown for Jet 1 (first and second rows) and Jet 2 (third and fourth rows) for the signals with $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV (red) and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV (blue), and the background (green). We plot the same number of signal and background events for visualization purpose

data, using a different initialization each time. We measure the performance of each classifier on validation data using the metric ϵ_{val}, defined as the true positive rate for the correct classification of signal region events, evaluated at a threshold

with a false positive rate $f = 1\%$ for incorrectly classifying events from the sideband region. Only the best out of the ten models is saved. We use an early stopping criterion to stop training if the validation performance has not improved for
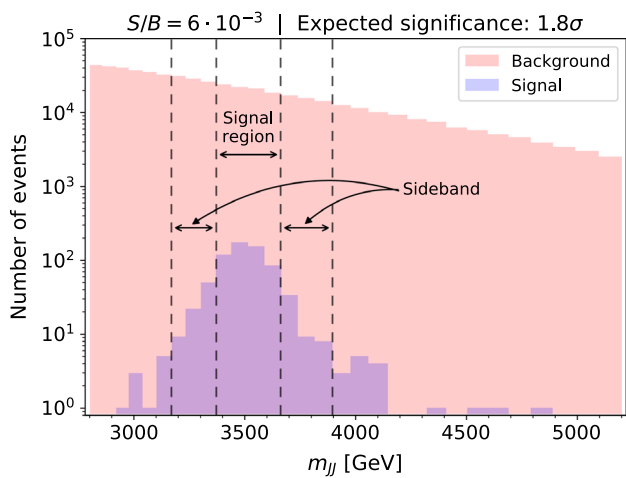
Fig. 2 Distribution of a fraction of signal and all background events on the m_{JJ} plane. Events are divided in 30 bins and a signal region and a sideband region are defined, as described in the text in Sect. 3.1. The amount of signal that has been injected corresponds to $S/B = 6 \times 10^{-3}$ and $S/\sqrt{B} = 1.8\sigma$ in the signal region

300 epochs. At the end of the four rounds, we use the mean of the outputs of the four selected models to build an ensemble model which is more robust on average than any individual model. This ensemble model is used to classify the events in the test set, and the $x\%$ most signal-like events are selected by applying a cut on the classifier output. This procedure is repeated for all five choices of test set, and the selected most signal-like events from each are combined into a signal-like sample. If a signal is present in data and CWoLa Hunting is able to find it, it will show as a bump in the signal region of the signal-like sample on the m_{JJ} plane, and standard bump-hunting techniques can be used to locate the excess.

It is worth mentioning that using an averaged model ensemble is important to reduce any potential overfitting. The cross-validation procedure ensures that even if an individual classifier learns any statistical fluctuations in the training data, each model will tend to overfit different regions of the phase space. As a result, the models will disagree in regions where overfitting has occurred, but will tend to agree in any region where a consistent excess is found.

3.2 Autoencoder

In this subsection we describe the strategy followed for the AE implementation. In the first place, we take the two leading jets in each event, ordered by mass, and consider the following set of input features for each jet:

$$Y_i = \{m_J, \ \tau_{21}, \ \tau_{32}, \ n_{\text{trk}}, \ p_T\} \ . \tag{3.2}$$

After analyzing different sets of input features, we found that the collection of 10 features presented in Eq. (3.2) led to

optimal performance. All the input features are standardized for the analysis.

Unlike the CWoLa method, the AE is trained on all the available background events in the full m_{JJ} range. The AE only requires a signal region and a background region for the purposes of background estimation through sideband interpolation. For the anomaly score itself (the reconstruction error), the AE is completely agnostic as to the m_{JJ} range of the signal.

In this work, the AE that we consider is a fully connected neural network with five hidden layers. The AE has an input layer with 10 nodes. The encoder has two hidden layers of 512 nodes, and is followed by a bottleneck layer with 2 nodes and linear activation. The decoder has two hidden layers of 512 nodes, and is followed by an output layer with 10 nodes and linear activation. All of the hidden layers have ReLU activation, and the first hidden layer in the encoder is followed by a Batch Normalization layer. We use the Minimum Squared Error (MSE) loss function and the Adam optimizer with learning rate of 10^{-4}, first and second moment decay rates of 0.9 and 0.999, respectively, and a mini-batch size of 128. In Appendix C we describe our quasi-unsupervised model-selection procedure. We use Pytorch [79] for implementing and training the AE.

In order to achieve a satisfactory generalization power, we decided to build an AE ensemble. For this purpose, we train fifty different models (i.e. the ensemble components) with random initialization on random subsamples of 50000 background events. Each model is trained for only 1 epoch. It is important to note that the training sample size and number of training epochs had a significant impact in the AE performance. When these are too large, the AE learns too much information and losses both generalization power and its ability to discriminate between signal and background events. Note that if there is a sufficient overlap between the distributions of signal and background events, then learning more about the background will not necessarily help to find the signal. For this reason, our training strategy gives the AE more generalization power and makes it more robust against overfitting.

The autoencoder ensemble is evaluated on the full dataset. The final MSE reconstruction loss of an event is obtained by computing the mean over the fifty different ensemble components. The optimal anomaly score is derived from the SIC curve as described in Appendix C. The results presented in this paper are for an AE trained on $S = 0$. We have verified that including relevant amounts of signal S do not significantly change the results. Therefore, for the sake of computational efficiency, we choose to present the AE trained with $S = 0$ everywhere.

4 Results

4.1 Signal benchmarks

Now we are ready to test the performance of CWoLa Hunting and the AE for different amounts of injected signal. Importantly, we will quantify the performance of CWoLa Hunting and the AE not using the full m_{JJ} range, but using a narrower slice $m_{JJ} \in (3371, 3661)$ GeV, the signal region defined in Sect. 3.1. This way, all performance gains from the two methods will be measured relative to the naive significance obtained from a simple dijet resonance bump hunt.

We define a set of eight benchmarks with a different number of injected signal events. For this purpose, to the current sample of 537304 background events in the range $m_{JJ} \in [2800, 5200]$ GeV, we add from 175 to 730 signal events. This results in a set of benchmarks distributed over the range $S/B \in [1.5 \times 10^{-3}, 7 \times 10^{-3}]$ in the signal region, corresponding to an expected naive significance in the range $S/\sqrt{B} \in [0.4, 2.1]$. To test the consistency of both models when no signal is present in data, we add a final benchmark with no signal events which allows us to evaluate any possible biases. For each S/B benchmark, the performance of CWoLa Hunting is evaluated across ten independent runs to reduce the statistical error using a random subset of signal events each time. After exploring a large range of cross sections, we decided to examine this range in S/B because it is sufficient to observe an intersection in the performance of the two methods. The observed trends continue beyond the limits presented here.

4.2 Supervised metrics

The performance of CWoLa Hunting and the AE in the signal region for different S/B ratios as measured by the Area Under the Curve (AUC) metric is presented in Fig. 3 for the two signal hypotheses considered in this work. Even though only a small fraction of signal events is used for training, the AUC metric is computed using all the available signal to reduce any potential overfitting. The results in both cases show that CWoLa Hunting achieves excellent discrimination power between signal and background events in the large S/B region, reaching AUC scores above 0.90 and approaching the 0.98 score from the fully supervised case. As the number of signal events in the signal region decreases, the amount of information that is available to distinguish the signal and sideband regions in the training phase becomes more limited. As a result, learning the signal features becomes more challenging and performance drops in testing. When the S/B ratio in the signal region is close to zero, the signal and sideband regions become nearly identical and the classifier should not be able to discriminate between both regions. For the benchmark with no signal events, the AUC scores

are only 0.43 and 0.59 for the signals with larger and smaller jet masses, respectively.[4] It is interesting to note that, in the absence of signal, the AUC should converge to 0.5. However, we will see that the presence of background events (from a statistical fluctuation) with a feature distribution that partially overlaps with the one from signal events, located in a region of the phase space with low statistics, allows the classifier to learn some information that turns out to be useful to discriminate between signal and background. Importantly, this does not imply that the information learnt by the classifier will be useful for enhancing the signal excess, as we discuss in detail below. By contrast, the AE performance is solid and stable across the whole S/B range. The reason is that, once the AE learns to reconstruct background events, its performance is independent of the number of signal events used for training as long as the contamination ratio is not too large. In our analysis, this ratio is always below 0.1% so the AE is trained on the full sample of background events with $S = 0$ for computational efficiency. Interestingly, the AUC curves from CWoLa Hunting and the AE cross at $S/B \sim 3 \times 10^{-3}$.

The most standard way of measuring the performance of a given model is through the Receiver Operating Characteristic (ROC) curve, and the area under this curve, the AUC metric. These two metrics are useful to compare the overall performance of different models in many classification tasks. However, the goal of a resonant anomaly detection search is to find a localized signal over a large background. For this purpose, the most important variables to consider are the signal-to-background ratio (S/B) and the naive expected significance ($S/\sqrt{B}$). With this in mind, we will consider the Significance Improvement Characteristic (SIC) [80] to measure the performance of CWoLa Hunting and the AE at enhancing the significance of the signal excess. The SIC metric measures the significance improvement after applying a cut in the classifier output. In particular, any given cut will keep a fraction ϵ_S of signal events and a fraction ϵ_B of background events, which are defined as the signal and background efficiencies of the cut. The significance improvement for this cut is thus given by SIC $= \epsilon_S/\sqrt{\epsilon_B}$.

In order to find the localized signal over the large background, which we presented in Fig. 2, we will use the SIC metric to find the optimal cut in the classifiers output that leads to the maximal enhancement in $S/\sqrt{B}$ in the signal region. The SIC curves for CWoLa Hunting and the AE are shown in Fig. 4. The SIC curves are calculated using all the available signal and background events in the signal region. For CWoLa Hunting, the results show that the shape and the location of the peak of the SIC curve depend on the amount of injected signal used during training. In order to find the signal efficiency that leads to a maximal overall significance improvement for all S/B benchmarks, we analyze how the

---

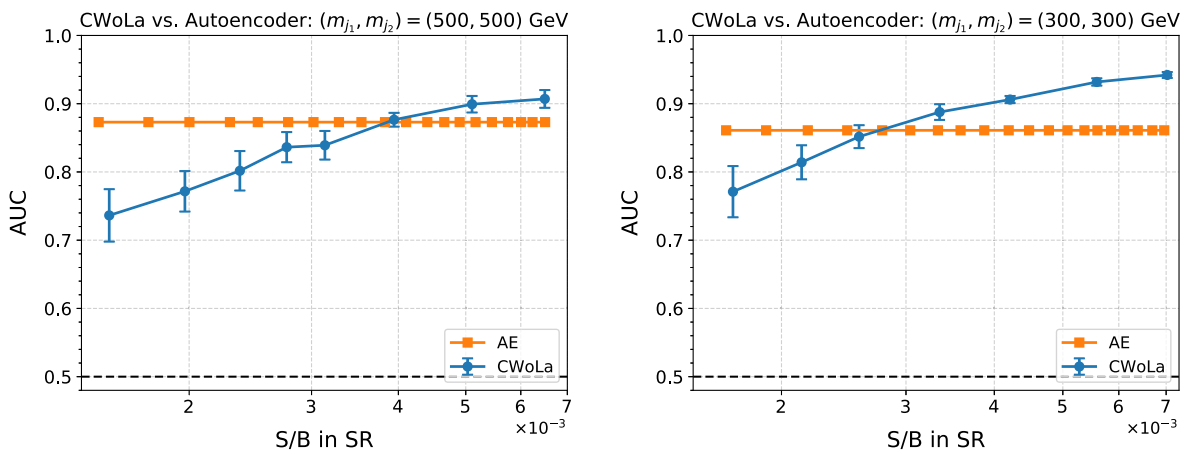[4] For visualization purpose, this benchmark is not shown in the plot.

Fig. 3 Performance of CWoLa Hunting (blue) and the AE (orange) as measured by the AUC metric on the signal with $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV (left plot) and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV (right plot). The error bars denote the standard deviation on the AUC metric from statistical uncertainties
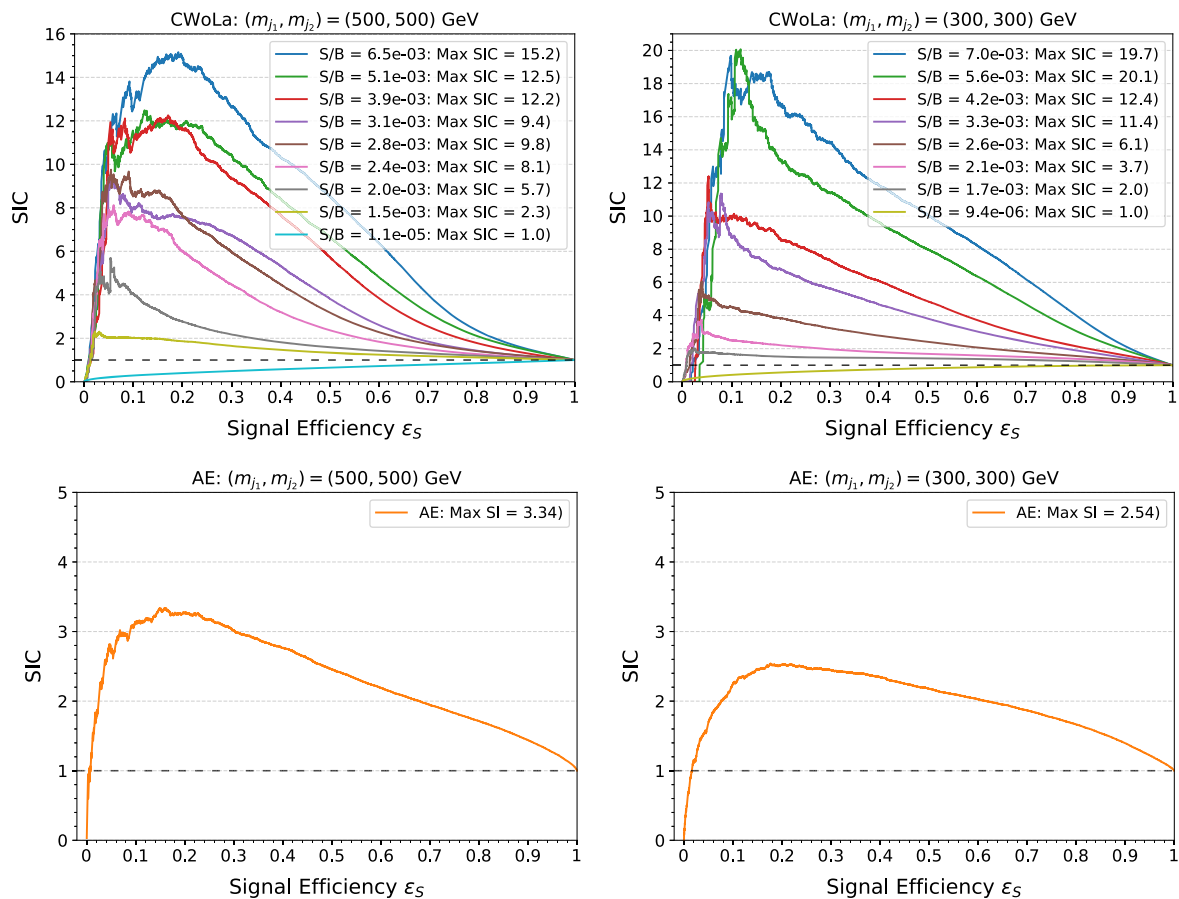

Fig. 4 The SIC curves for CWoLa Hunting (top row) and the AE (bottom row) are shown for the signals with $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV in the left and right plots, respectively. For CWoLa Hunting, a SIC curve is shown for each of the classifiers that were trained on mixed samples with different amounts of injected signal
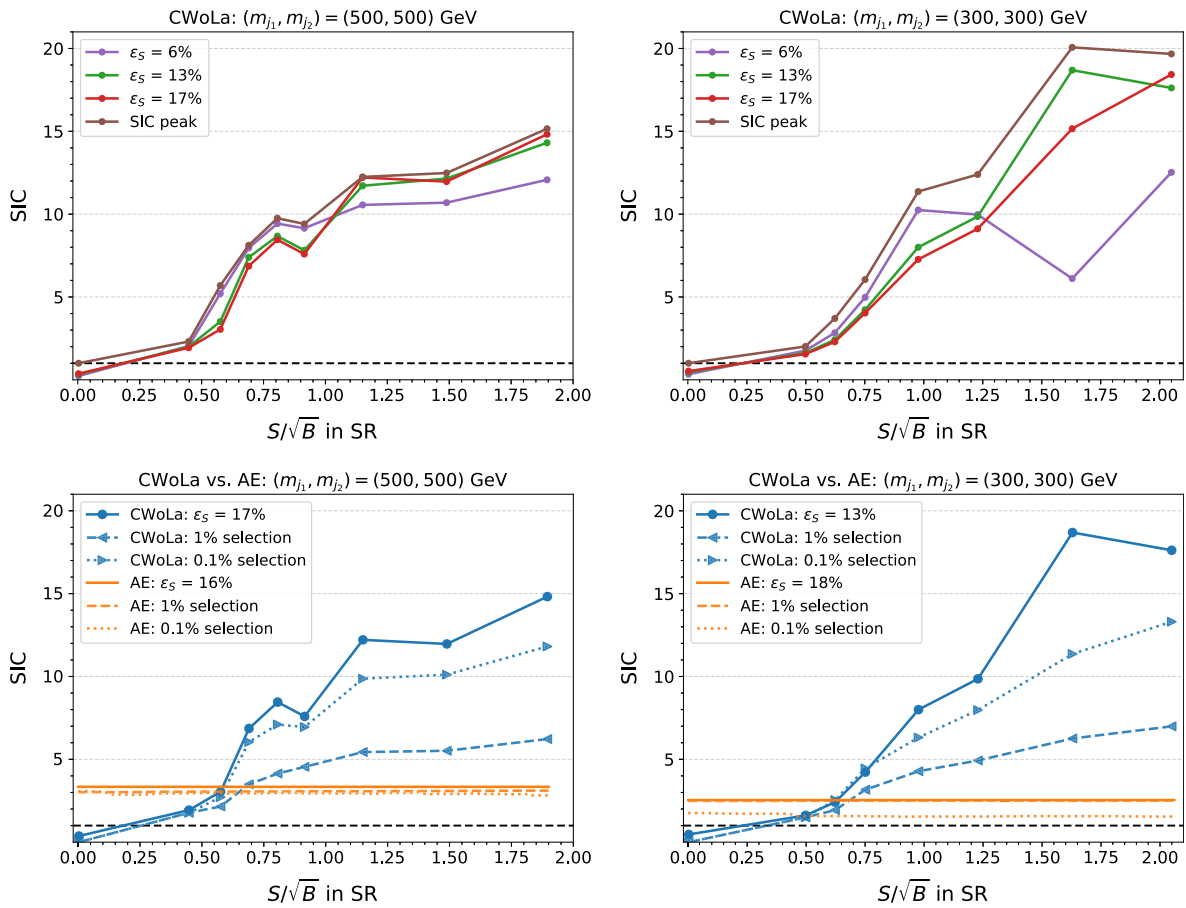
Fig. 5 Top row: The SIC value as a function of $S/\sqrt{B}$ for a set of fixed signal efficiencies is shown for the signals with $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV (left plot) and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV (right plot). The $\epsilon_S = 17\%$ and $\epsilon_S = 13\%$ signal efficiencies, respectively, maximize the overall significance improvement for all S/B bench-marks. Bottom row: The signal efficiencies are chosen such that the SIC values are maximized for CWoLa Hunting and the AE. The SIC values associated to the 1% and 0.1% are also shown for comparison. These values are calculated using only the fraction of signal that defines each S/B benchmark

SIC value changes as a function of $S/\sqrt{B}$ for a set of fixed signal efficiencies in the top row of Fig. 5. We find that the signal efficiencies that yield the maximum overall significance improvement for CWoLa Hunting are $\epsilon_S = 17\%$ and $\epsilon_S = 13\%$ for the high and low jet mass signals, respectively. For the AE, the optimal signal efficiencies are $\epsilon_S = 16\%$ and $\epsilon_S = 18\%$, respectively. Now we will use these optimal signal efficiencies to set an anomaly score threshold that maximizes the significant improvement in the signal region for each model. In practice, model independence would prevent picking a particular value and so we will later compare these optimized values with fixed values at round logarithmically spaced efficiencies.

4.3 Sideband fit and p-values

After evaluating the quality of the two methods at identifying the signal events among the background, we compare

how they perform at increasing the significance of the signal region excess. For this purpose, we performed a parametrized fit to the m_{JJ} distribution in the sideband region. We then interpolate the fitted background distribution into the signal region and evaluate the p-value of the signal region excess.

For the CWoLa method, we used the following 4-parameter function to fit the background:

$$\frac{d\sigma}{dm_{JJ}} = \frac{p_0(1-x)^{p_1}}{x^{p_2+p_3\ln(x)}}, \qquad (4.1)$$

where $x = m_{JJ}/\sqrt{s}$. We use the previous function to estimate the background in the range $m_{JJ} \in [2800, 5200]$ GeV. This function has been previously used by both ATLAS [81] and CMS [82] collaborations in hadronic heavy resonance searches.

For the AE, we find that this function does not fit well the distribution of surviving events on m_{JJ} after applying a cut on the reconstruction error. Instead, we found that a
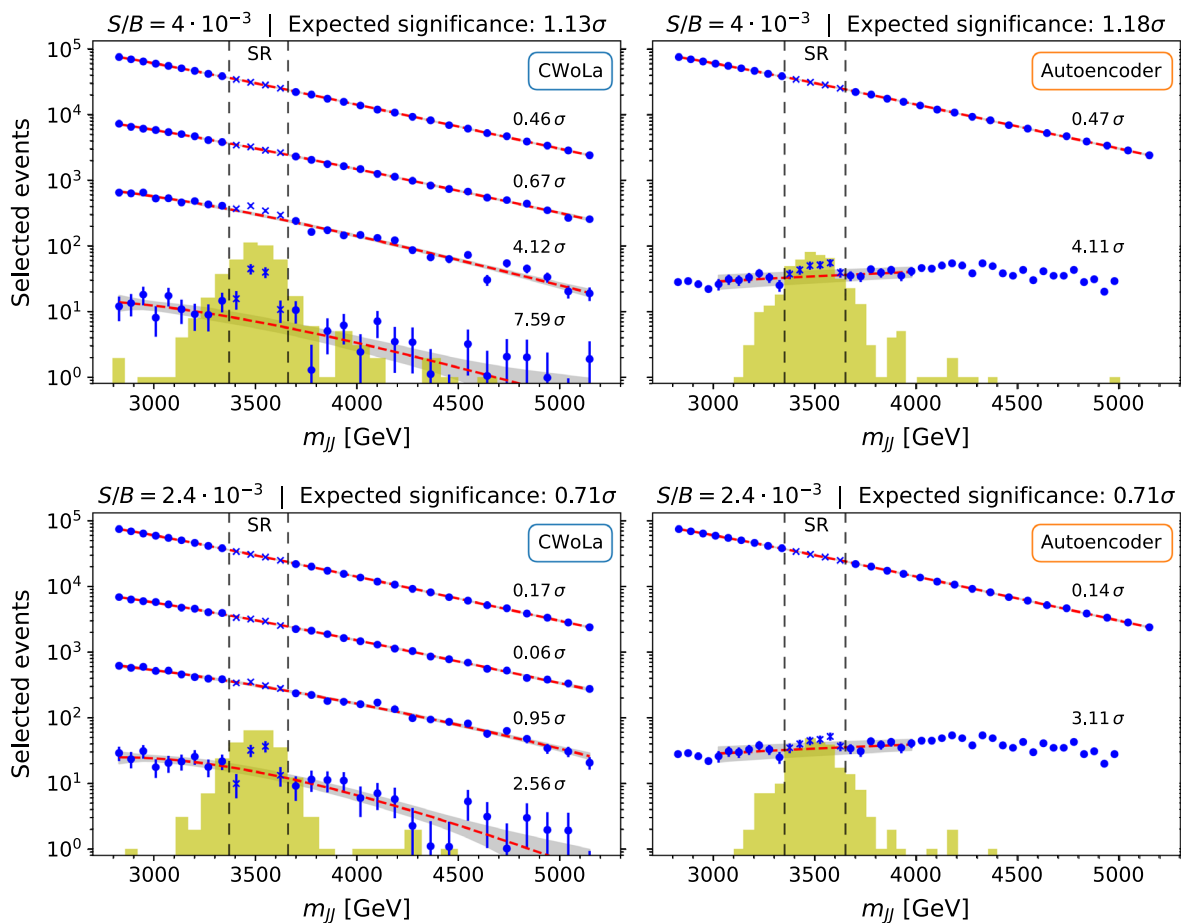
Fig. 6 Significance of the signal region excess after applying different cuts using the classifier output for CWoLa Hunting (left plots) and the AE (right plots), for one of the runs corresponding to the benchmarks with $S/B \simeq 4 \times 10^{-3}$ (top row) and $S/B \simeq 2.4 \times 10^{-3}$ (bottom row) on the signal with $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV. For CWoLa, we show the 100%, 10%, 1%, 0.04% most signal-like events. For the AE, we show the 100% and 0.6% event selections. In both cases, the smallest cut corresponds to the optimal cut according to the SIC curve. The blue crosses denote the event selection in each signal region bin, while the blue circles represent the event selection in each bin outside of the signal region. The dashed red lines indicate the fit to the events outside of the signal region, the grey band indicates the fit uncertainty and the injected signal is represented by the green histogram

simple linear fit (on a narrower sideband region) is able to describe the background distribution on the sideband with good accuracy and it is sensitive to an excess on the signal region for the cuts that we considered. For the cut based on the SIC curve and the 1% cut, the fit is implemented on the range $m_{JJ} \in [3000, 4000]$ GeV. For the 0.1% cut, we need to extend this range to $m_{JJ} \in [2800, 4400]$ GeV. This range extension produces a better fit χ^2 in the sideband and mitigates a small bias in the predicted signal at $S = 0$.[5]

The validity of sideband interpolation relies on the assumption that the m_{JJ} distribution for background events surviving a cut can still be well modelled by the chosen functional forms. This is likely to be the case so long as the selec-

tion efficiency of the tagger on background events is smooth and monotonic in m_{JJ}, and most simply[6] if it is constant in m_{JJ} (which would require signal features uncorrelated with m_{JJ}).

In Fig. 6, we show the fit results for CWoLa Hunting and the AE for one of the runs corresponding to the benchmarks with $S/B \simeq 4 \times 10^{-3}$ and $S/B \simeq 2.4 \times 10^{-3}$ on the signal with $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV. After applying different cuts using the classifiers outputs, the significance of signal region bump is significantly increased. For the benchmark with more injected signal, CWoLa Hunting yields a substantial significance increase of up to 7.6σ, while the AE is able to increase the bump significance by up to 4.1σ. When the amount of injected signal is reduced, the results show that CWoLa Hunting becomes weaker and it rises the excess

---

[5] As we tighten the cut, we will see that the fraction of events that survive at the lower end of the m_{JJ} distribution is significantly smaller than for higher invariant masses. This extends the linear behaviour to the range $m_{JJ} \in [2800, 4400]$ GeV for the 0.1% cut.

[6] Complete decorrelation is sufficient, but not necessary to prevent bump-sculpting [83].
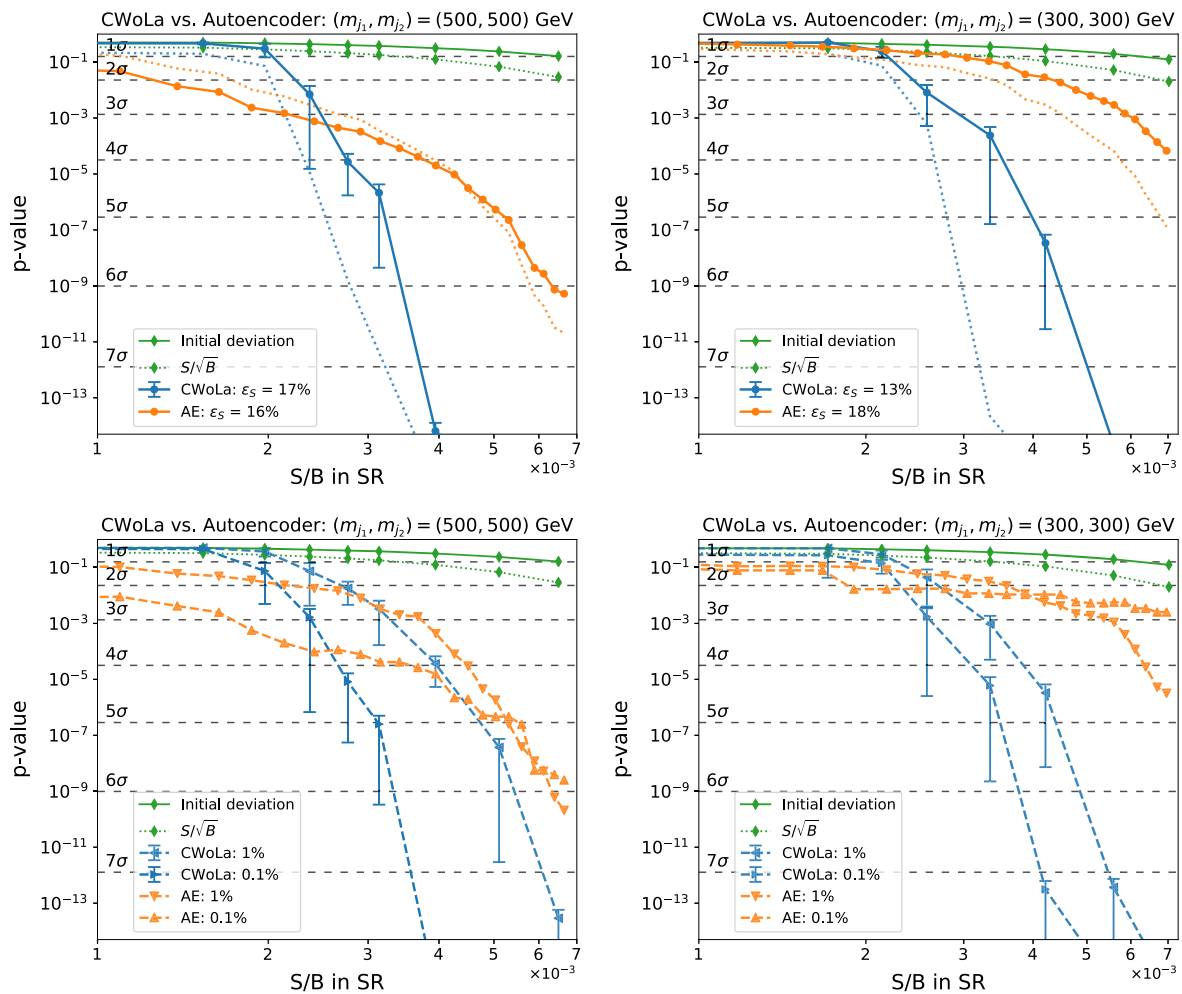
Fig. 7 The significance of the signal region excess after applying different cuts for CWoLa Hunting and the AE, for the signals with $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV, is shown in the left and right plots, respectively. The plots in the top row show the cuts that maximize the overall significance improvement for all benchmarks according to the SIC curve, while the bottom row plots show results for fixed, predetermined cuts. The best cuts for CWoLa Hunting (blue) correspond to the 17% and 13% signal efficiencies for the signals with high and low jet masses, respectively. For the AE (orange), the best cuts correspond to the 16% and 18% signal efficiencies, respectively. The dotted lines denote the naive expected significance, $S/\sqrt{B}$. The round cuts from the bottom plots show the 1% and 0.1% event selections for CWoLa Hunting and the AE. The initial significance of the bump (100% selection) is shown in green

significance up to only 2.6σ. However, in this case the AE performs better than CWoLa Hunting, increasing the bump significance up to 3.1σ. This is an important finding because it suggests that CWoLa Hunting and the AE may be complementary techniques depending on the cross section. Note that the event distribution from the AE is clearly shaped due to some correlations between the input features and m_{JJ}. In particular, since the jet p_T is very correlated with m_{JJ}. However, the average jet p_T scales monotonically (and roughly linearly) with m_{JJ}, which means that no artificial bumps are created and the distribution post-selection is still well modelled by the chosen fit function. Finally, note that the fit to the raw distribution (i.e. no cut applied) is lower than the naive expected significance $S/\sqrt{B}$ due to a downward fluctuation

in the number of background events in the signal region, as discussed in Appendix A.

In order to systematically study if CWoLa Hunting and the AE could be complementary techniques depending on the cross section, we analyze their performance at increasing the significance of the signal region excess for different S/B benchmarks and the two signal hypotheses in Fig. 7. The top two plots show the cuts on the classifier output that lead to the largest overall significance improvement according to the SIC curve. For CWoLa Hunting, we show the median p-values from the ten independent runs for every benchmark corresponding to the 17% (top left) and 13% (top right) signal efficiencies, which correspond to fractions of signal-like events between 0.04% and 1.7% depending on the bench-
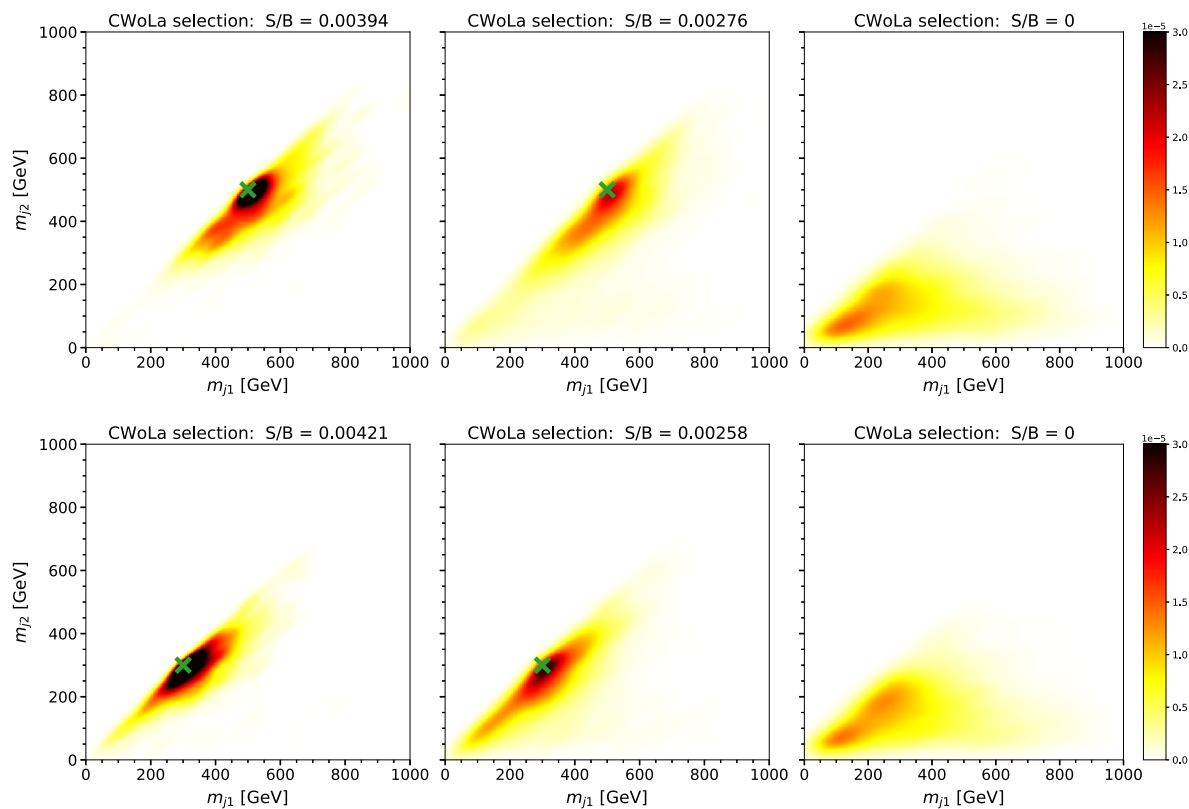
Fig. 8 Density of events on the (m_{j_1}, m_{j_2}) plane for the most signal-like events selected by CWoLa Hunting for the signal hypothesis $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV (top row) and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV (bottom row). From left to right, we show results for three benchmarks with $S/B \simeq 4 \times 10^{-3}, 2.8 \times 10^{-3}, 0$. The location of the injected signal is indicated by a green cross. Note that the upper right plot shows a small statistical fluctuation that disappears when averaging over a larger number of simulations

mark. The error bars represent the Median Absolute Deviation. Note that the fit result does not always agree with the naive expected significance, $S/\sqrt{B}$, due to the high uncertainties among the ten independent classifiers and the small fractions of events considered in some cases. For the AE, we show the p-values associated to the 16% (top left) and 18% (top right) signal efficiencies, which correspond to the 0.36% and 0.63% most signal-like events, respectively.

Importantly, there are other cuts that enhance the significance of the signal region excess, as shown in the bottom plots of Fig. 7. In a real experimental search, with no previous knowledge about any potential new physics signal, the two models would be able to find the signal for fixed round cuts of 1% and 0.1%. For the AE, these cuts are applied in the signal region to derive an anomaly score above which all the events in the full m_{JJ} range are selected.

The statistical analysis demonstrates two things. First, CWoLa Hunting is able to increase the significance of the signal region excess up to $3\sigma - 8\sigma$ for S/B ratios above $\sim 3 \times 10^{-3}$ for both signal hypotheses, even when the original fit shows no deviation from the background-only hypothesis. By contrast, the AE shows a superior performance below

this range for the signal with $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV, boosting the significance of the excess up to $2\sigma - 3\sigma$ in the low S/B region where CWoLa Hunting is not sensitive to the signal. Importantly, there is again a crossing point in the performance of the two methods as measured by their ability to increase the significance of the excess. Therefore, our results show that the two methods are complementary for less-than-supervised anomaly detection. Second, it is clear that the AE is not able to increase the bump excess for the signal with $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV below $S/B \sim 3 \times 10^{-3}$, even when it reaches a fairly solid AUC score, as shown in Fig. 3. This means that even though the AE is able to classify a sizeable fraction of signal events correctly, there is a significant fraction of background events that yield a larger reconstruction error than the signal events. In other words, the AE does not consider the signal events as sufficiently anomalous and finds more difficult to reconstruct part of the background instead. Therefore, cutting on the reconstruction error does not result in a larger fraction of signal in the selected events. By construction, this is the main limitation of the AE: it focuses its attention in anything that seems
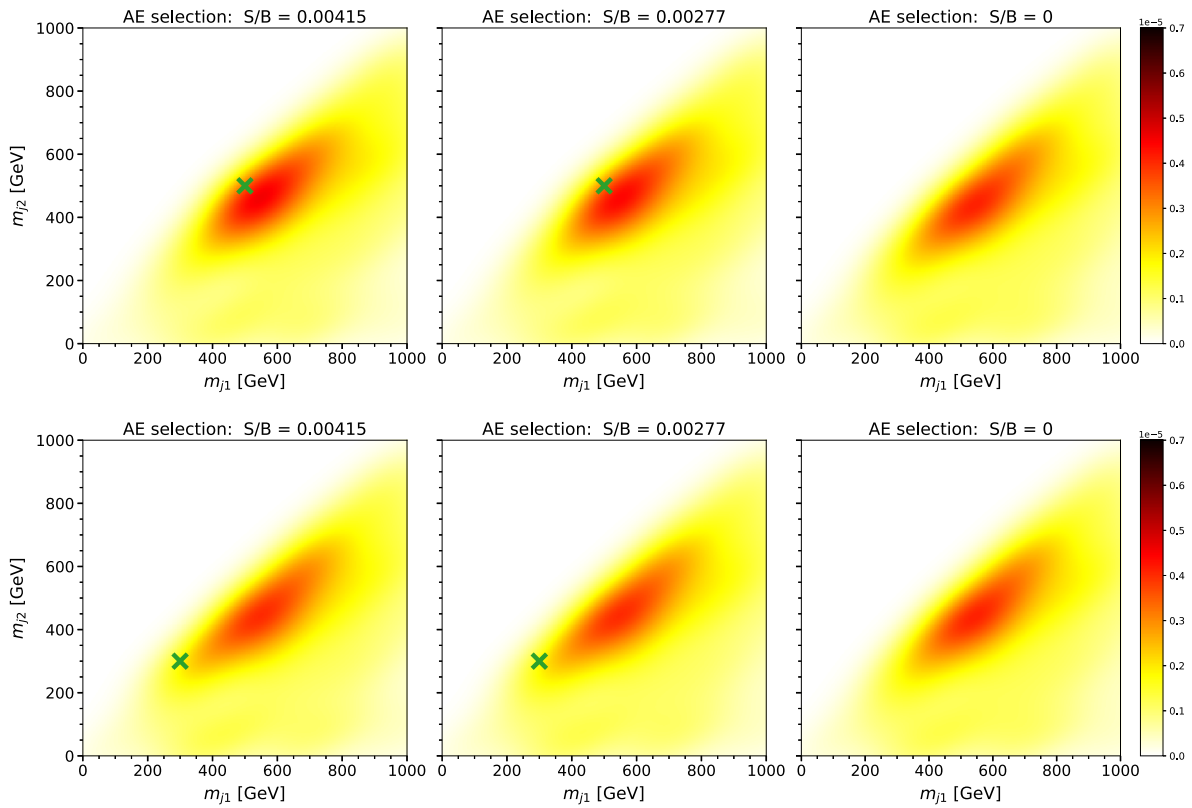
Fig. 9 Density of events on the (m_{j_1}, m_{j_2}) plane for the most signal-like events selected by the AE for the signal hypothesis $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV (top row) and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV (bottom row). From left to right, we show results for three benchmarks with $S/B \simeq 4 \times 10^{-3}, 2.8 \times 10^{-3}, 0$. The location of the injected signal is indicated by a green cross

anomalous, whether it is an exciting new physics signal or something that we consider less exotic.

Finally, it is important to analyze the performance of CWoLa Hunting and the AE when training on no signal. For consistency, both models should not sculpt any bumps on the m_{JJ} distribution when no signal is present on data. For CWoLa Hunting, the expected significance at $S/B = 0$ is 0σ for all cuts. For the AE, we find that the excess significance at $S/B = 0$ is 0.89σ, 0.56σ and 1.06σ for the SIC-based, 1% and 0.1% cuts, respectively. We checked that this is caused by the shaping of the m_{JJ} distribution and the small statistical fluctuations that appear for such tight cuts. We remark that this effect is not produced by the signal.

4.4 What did the machine learn?

In order to illustrate this point, we can examine what the classifiers have learnt by looking at the properties of the events which have been classified as signal-like for three benchmarks with $S/B \simeq 4 \times 10^{-3}, 2.8 \times 10^{-3}, 0$. In Figs. 8 and 9 we show the density of events on the (m_{j_1}, m_{j_2}) plane for the most signal-like events selected by CWoLa Hunting and the

AE, respectively. The cuts applied in each case correspond to the 0.1% cut. For CWoLa Hunting, it is clear that the classifier is able to locate the signal for the two mass hypotheses. In addition, note that the upper and lower right plots show a small statistical fluctuation that is produced by the different fractions of signal-like events represented in each plot, which disappears when averaging over a larger number of simulations.

The AE similarly identifies the high mass signal point, but fails to identify the low mass one. This can be most easily understood by observing the selection efficiency as a function of the two jet masses for the trained AE, shown in Fig. 10. In the left plot, we show the total number of events on the (m_{j_1}, m_{j_2}) plane. In the middle and right plots, we show the selection efficiencies for the 1% and 0.1% cuts. These results illustrate that the AE has learnt to treat high mass jets as anomalous (since these are rare in the training sample), and so the $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV signal is more easily reconstructed than high mass QCD events. In other words, high mass QCD events are regarded as more anomalous than signal events, and a sufficiently high selection cut on the AE reconstruction error will eliminate the signal. We remark again that this is one of the main limitations of the
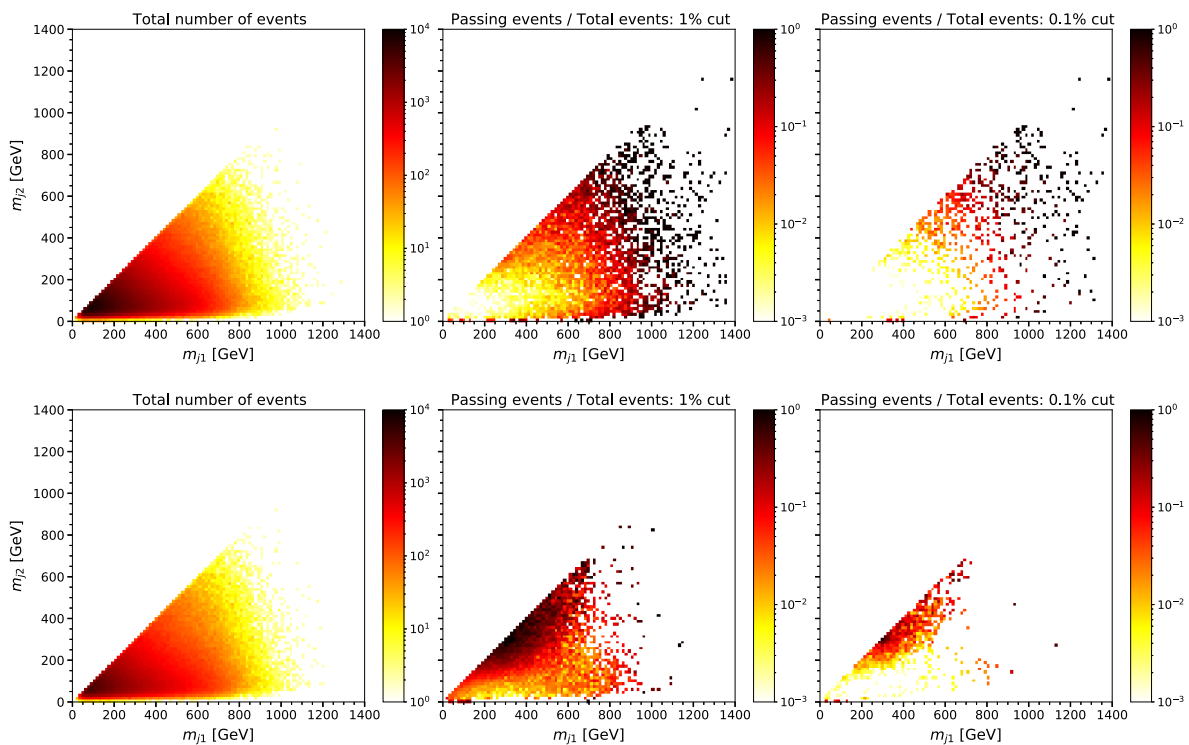
Fig. 10 The total density of events on the (m_{j_1}, m_{j_2}) plane is plotted on the left. The 1% and 0.1% selection efficiencies for the AE and CWoLa are plotted on the middle and right images, respectively. The top row shows results for the AE, while the bottom row shows results for CWoLa. The selection efficiency in a given bin is defined as the number of events passing the x% cut divided by the total number of events in that bin. These results correspond to the signal with $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV and $S/B \simeq 4 \times 10^{-3}$

AE. Therefore, it is crucial to find the cut that maximizes the fraction of signal within the most anomalous events. As shown in Fig. 13 in Appendix B, that cut corresponds to the anomaly score that maximizes the SIC curve in the signal region. In contrast, the bottom row of Fig. 10 shows that CWoLa is able to learn the signal features.

5 Conclusions

In this article, we have compared weakly-supervised and unsupervised anomaly detection methods, using Classification without Labels (CWoLa) Hunting and deep autoencoders (AE) as representative of the two classes. The key difference between these two methods is that the weak labels of CWoLa Hunting allow it to utilize the specific features of the signal overdensity, making it ideal in the limit of large signal rate, while the unsupervised AE does not rely on any information about the signal and is therefore robust to small signal rates.

We have quantitatively explored this complimentarity in a concrete case study of a search for anomalous events in fully hadronic dijet resonance searches, using as the target a physics model of a heavy resonance decaying into a pair of

three-prong jets. CWoLa Hunting was able to dramatically raise the significance of the signal in our benchmark points in order to breach 5σ discovery, but only if a sizeable fraction of signal is present ($S/B \gtrsim 4 \times 10^{-3}$). The AE maintained classification performance at low signal rates and had the potential to raise the significance of of one of our benchmark signals to the level of 3σ in a region where CWoLa Hunting lacked sensitivity.

Crucially, our results demonstrate that CWoLa Hunting is effective at finding diverse and moderately rare signals and the AE can provide sensitivity to rare signals, but only with certain topologies. Therefore, both techniques are complementary and can be used together for anomaly detection. A variety of unsupervised, weakly supervised, and semi-supervised anomaly detection approaches have been recently proposed (see e.g. Ref. [60]), including variations of the methods we have studied. It will be important to explore the universality of our conclusions across a range of models for anomaly detection at the LHC and beyond.

Data Availability Statement This manuscript has associated data in a data repository. [Authors' comment: The code to reproduce the results presented in this manuscript is publicly available in the following Github repository: https://github.com/pmramiro/CWoLa-and-AE-for-anomaly-detection.]

Appendix A: Background fit

In this appendix, we briefly describe the details about the fit procedure and discuss results from the fit to the background events. In order to evaluate the significance of any potential excess in the signal region, the total number of predicted signal region events is calculated by summing the individual predictions from each signal region bin. The systematic uncertainty of the fit in the signal region prediction is estimated by propagating the uncertainties in the fit parameters. We test the validity of the fit using a Kolmogorov–Smirnov test.

In Fig. 11 we show the fit to the background distribution using the 4-parameter function presented in Eq. (4.1). First, the Kolmogorov–Smirnov test yields a p-value of 0.99, which means that the fit describes the background distribution well outside of the signal region. In addition, the fit
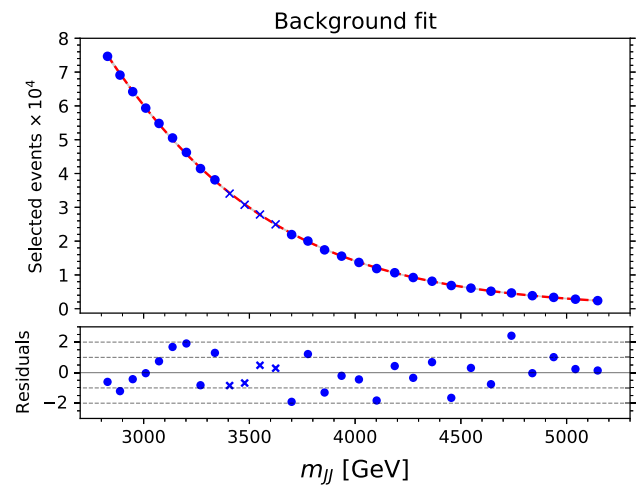
Fig. 11 Fit to the background distribution of dijet events and residuals from the fit. The signal region events are indicated by blue crosses

result produces a p-value of 0.5. However, the residuals indicate that the number of predicted events in the signal region is overestimated due to a local negative fluctuation of size $n = 123$ events.[7] As a result, the fit will always underestimate the excess significance when a signal is injected in the signal region. For example, if we introduce a number n of signal events in the signal region, the fit prediction will match the number of observed events and therefore the excess significance will be exactly zero, even when a signal has been injected.

Appendix B: Density of events for the optimal cut

See Figs. 12 and 13.

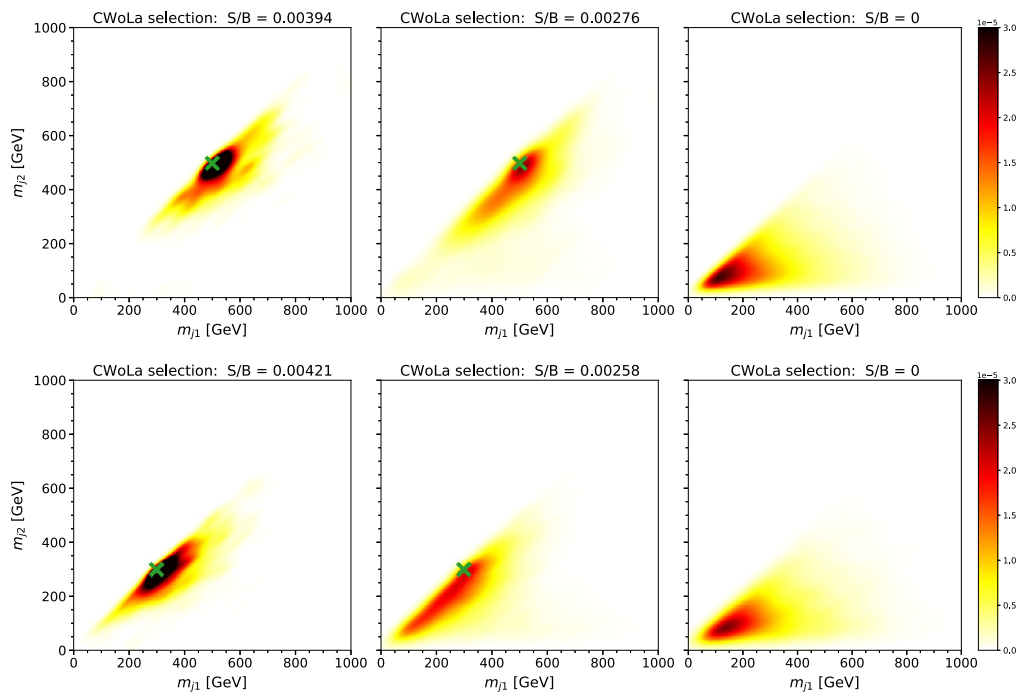[7] This has been validated as a fluctuation with an independent sample.

Fig. 12 Density of events on the (m_{j_1}, m_{j_2}) plane for the most signal-like events selected by CWoLa for the signal hypothesis $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV (top row) and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV (bottom row). The optimal cut is derived from the signal efficiency that maximizes the SIC curve. From left to right, we show results for three benchmarks with $S/B \simeq 4 \times 10^{-3}, 2.8 \times 10^{-3}, 0$. The location of the injected signal is indicated by a green cross. Note that the upper right plot shows a small statistical fluctuation that disappears when averaging over a larger number of simulations
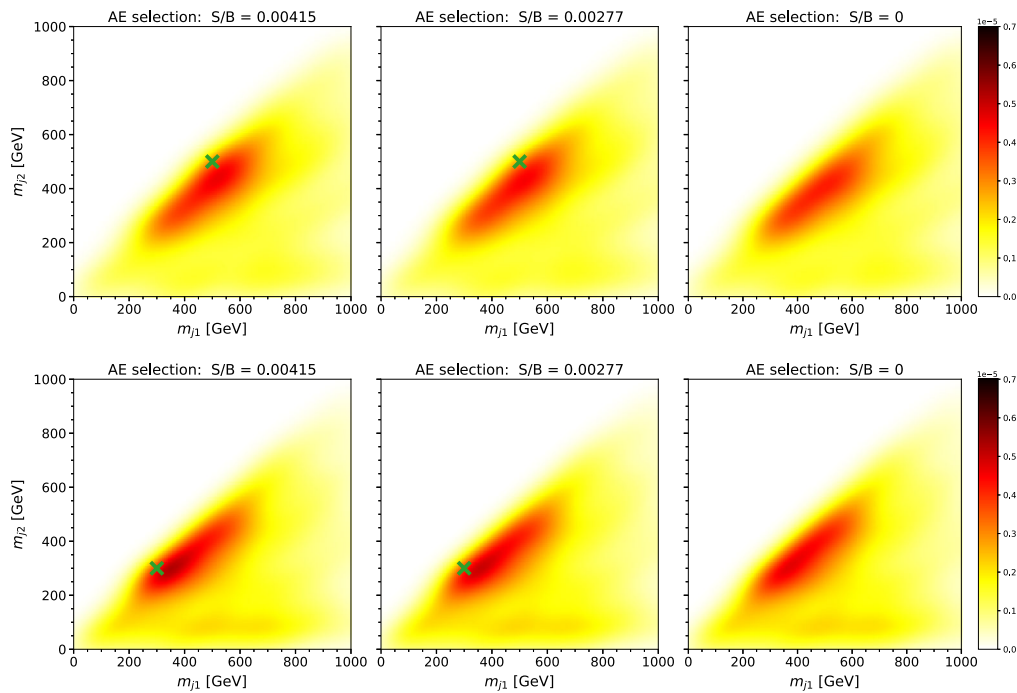
Fig. 13 Density of events on the (m_{j_1}, m_{j_2}) plane for the most signal-like events selected by the AE for the signal hypothesis $(m_{j_1}, m_{j_2}) = (500, 500)$ GeV (top row) and $(m_{j_1}, m_{j_2}) = (300, 300)$ GeV (bottom row). The optimal cut is derived from the signal efficiency that maximizes the SIC curve. From left to right, we show results for three benchmarks with $S/B \simeq 4 \times 10^{-3}, 2.8 \times 10^{-3}, 0$. The location of the injected signal is indicated by a green cross

Appendix C: Autoencoder model selection

Here we will motivate the selection of the AE model used in the main body of the paper. In general, the challenge or central tension of AE model selection for anomaly detection is to choose a model that strikes a good balance between compression and expressivity, between describing the bulk of the data just well enough (i.e. with the right size latent space) without swallowing up all the anomalies as well. Here we will put forth some guidelines that could be used to find this balance in an unsupervised way. While a complete study is well beyond the scope of this work, the two signals provide some evidence for the usefulness of these guidelines.

To begin, it is useful to consider the AE as consisting of three components:

1. Choice of input features.
2. Latent space dimension.
3. Rest of the architecture.

Our philosophy is that item 1 defines the type of anomaly we are interested in, and so cannot be chosen in a fully unsupervised way. In this paper, we chose the input features to be $(m_J, p_T, \tau_{21}, \tau_{32}, n_{trk})$ because we observed they did well in finding the 3-prong qqq signals. In contrast, item 2 and item 3 can be optimized to some extent independent of the anomaly (i.e. just from considerations of the background).

Our main handle for model selection will be the concept of FVU: *fraction of variance unexplained*. This is a commonly used statistical measure of how well a regression task is performing at describing the data. Let the input data be $\vec{x}_i, i = 1, \ldots, N$ and the (vector-valued) regression function being

$$\vec{y}_i = f(x_i). \tag{C.1}$$

Let the data to be described be $\vec{Y}_i$. (So, for an AE, $\vec{x}_i = \vec{Y}_i$.) Then the FVU F is

$$F = \frac{\frac{1}{N}\sum_{i=1}^{N}(\vec{Y}_i - \vec{y}_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(\vec{Y}_i - \langle\vec{Y}\rangle)^2}, \tag{C.2}$$

i.e. it is the MSE of the regression divided by the sample variance of the data. In the following, we will be working with features standardized to zero mean and unit variance, in which case the denominator (the sample variance) is just n, the number of input features, and F becomes

$$F = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\sum_{a=1}^{n}(Y_{ia} - y_{ia})^2, \tag{C.3}$$

i.e. it is the MSE of the regression normalized to the number of input features.

Our criteria for whether it is worth adding another latent space dimension to the AE is whether it substantially reduces the FVU. Here the measure of "substantially reduces" is whether it decreases the FVU by significantly more than $1/n$. A decrease of $1/n$ (or less) suggests that the AE is merely memorizing one of the input features via the extra latent space dimension. In that case, adding the latent space dimension should not help with anomaly detection. Meanwhile, a decrease in FVU of significantly more than $1/n$ suggests that the latent space dimension is learning something nontrivial about the inner workings of the data, capturing one of the underlying correlations. In this case adding the latent space dimension may help with anomaly detection performance.

We will demonstrate the effectiveness of this model selection criteria using the two signals considered in this paper, $Z'(3500) \rightarrow X(m)X(m)$, $X \rightarrow qqq$ events with $m = 500$ GeV and $m = 300$ GeV.

We scan over the size of the latent space and hidden layers, $n_{latent} = 1, 2, 3, 4, \ldots$ and $n_{hidden} = 128, 256, 512$, respectively. For each architecture and choice of input features we train 10 models with random initializations on a random subset of 50000 background jets.

For evaluation, we feed all 1M QCD events and all the signal events to the trained models. We compute the following metrics for each model: $\langle MSE\rangle_{bg}$, $\sigma(MSE)_{bg}$, $\max(SIC)$ where the SIC is computed by cutting on the MSE distribution. For all three metrics, we only compute them using the MSE distribution in a window (3300, 3700) GeV in m_{JJ}.

Shown in Fig. 14 is the FVU versus the number of latent dimensions, for 5 input features and different AE architectures. Each point represents the average MSE obtained from 10 independent trainings. We see that the FVU versus n_{latent} plot has a characteristic shape, with faster-than-
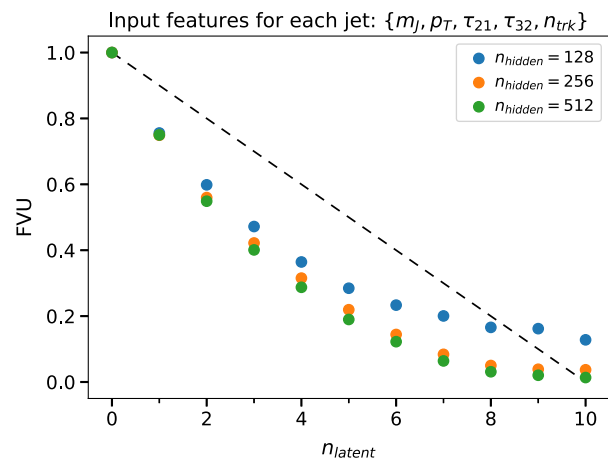

Fig. 14 FVU vs number of latent dimensions, for 10 input features and different AE architectures. The diagonal line is $1 - n_{latent}/n = 1 - n_{latent}/10$ indicating the nominal case of a latent dimension just memorizing one of the input features
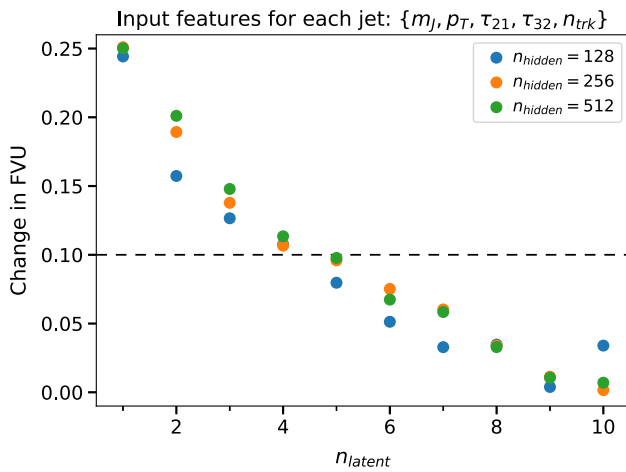
Fig. 15 Decrease in FVU from adding one more latent dimension vs number of latent dimensions, for 10 input features and different AE architectures. The horizontal line is $1/n = 1/10$ indicating the nominal case of a latent dimension just memorizing one of the input features

nominal decrease for small n_{latent} (the AE is learning non-trivial correlations in the data) and then leveling out for larger

n_{latent} (the AE is not learning as much and is just starting to memorize input features).

In Fig. 15 we show the decrease in FVU with each added latent dimension, versus the number of latent dimensions. From this we see that $n_{latent} = 1, 2, 3$ add useful information to the AE but beyond that the AE may not be learning anything useful.

We also see from these plots that the FVU decreases with more n_{hidden} as expected, although it seems to be levelling off by the time we get to $n_{hidden} = 512$. This makes sense – for fixed n_{latent} the bottleneck is fixed, so increasing n_{hidden} just increases the complexity of the correlations that the AE can learn, with no danger of becoming the identity map. This suggests that the best AE anomaly detector will be the largest n_{hidden} that we can take for fixed n_{latent}, although the gains may level off for n_{hidden} sufficiently large.

Now we examine the performance of the various AE models on anomaly detection of the 300 GeV and 500 GeV 3-prong signals. The max(SIC) versus n_{latent} is shown in Fig. 16. We see that there is decent performance on both signals for $n_{latent} = 2, 3, 4, 5$ with $n_{latent} = 2$ being especially
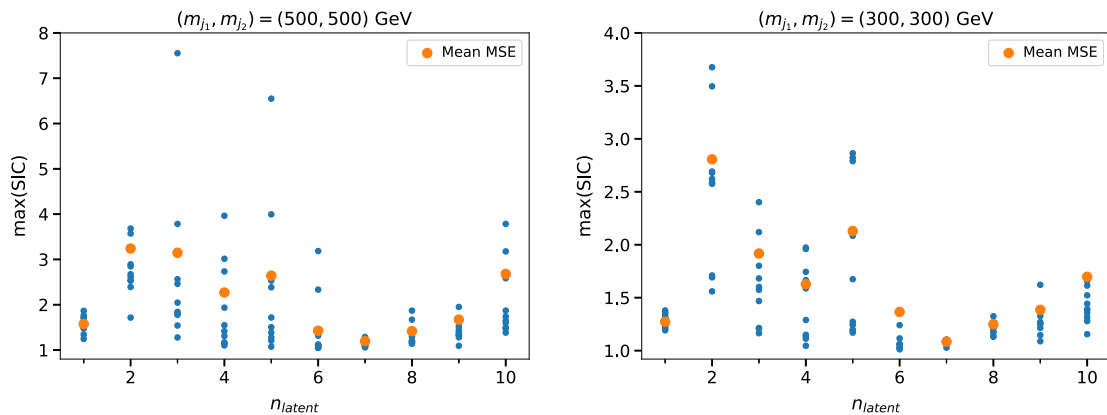

Fig. 16 max(SIC) vs. n_{latent} for the 500 GeV signal (left) and 300 GeV signal (right), 5 input features, and $n_{hidden} = 512$. The blue dots are the maxSICs for each of the 10 independent trainings, while the orange dot is the max(SIC) obtained from the average of the 10 MSE distributions
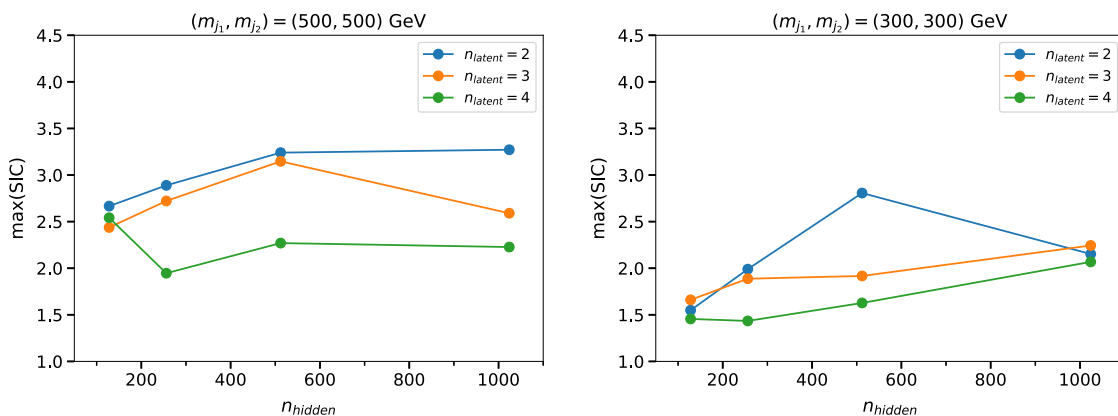

Fig. 17 max(SIC) of the averaged MSE distributions vs. n_{hidden} for the 500 GeV signal (left) and 300 GeV signal (right), 5 input features, and $n_{latent} = 2, 3, 4$

good for both.[8] This is roughly in line with the expectations from the FVU plots. Importantly, if we restricted to $n_{latent} = 2, 3$ which have the larger decreases in FVU, we would not miss out on a better anomaly detector.

Finally in Fig. 17, we show the max(SIC) for the MSE distributions averaged over 10 trainings vs n_{hidden}, for $n_{latent} = 2, 3, 4$. We see that generally the trend is rising or flat with increasing n_{hidden}, which is more or less consistent with expectations.

To summarize, we believe we have a fairly model-independent set of criteria for AE model selection, based on the FVU, which empirically works well on our two signals. Admittedly this is too small of a sample size to conclude that this method really works; it would be interesting to continue to study this in future work. Based on these criteria, we fix the AE model in this paper to have $n_{latent} = 2$ and $n_{hidden} = 512$.

References

1. ATLAS Collaboration, Exotic physics searches (2018). https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults
2. ATLAS Collaboration, Supersymmetry searches (2018). https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults
3. CMS Collaboration, Cms exotica public physics results (2018). https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO
4. CMS Collaboration, Cms supersymmetry physics results (2018). https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS
5. CMS Collaboration, CMS beyond-two-generations (b2g) public physics results (2018). https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G
6. M. Feickert, B. Nachman, A Living review of machine learning for particle physics. arXiv:2102.02770 [hep-ph]
7. B. Knuteson, A quasi-model-independent search for new high p_T physics at D0. Ph.D. thesis, University of California at Berkeley. https://www-d0.fnal.gov/results/publications_talks/thesis/knuteson/thesis.ps (2000)
8. D0 Collaboration, B. Abbott et al., 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). https://doi.org/10.1103/PhysRevD.62.092004. arXiv:hep-ex/0006011
9. D0 Collaboration, V.M. Abazov et al., A quasi model independent search for new physics at large transverse momentum. Phys. Rev. D **64**, 012004 (2001). https://doi.org/10.1103/PhysRevD.64.012004. arXiv:hep-ex/0011067
10. D0 Collaboration, B. Abbott et al., A quasi-model-independent search for new high p_T physics at DØ. Phys. Rev. Lett. **86**, 3712–3717 (2001). https://doi.org/10.1103/PhysRevLett.86.3712. arXiv:hep-ex/0011071
11. H1 Collaboration, F.D. Aaron et al., A general search for new phenomena at HERA. Phys. Lett. B **674**, 257–268 (2009). https://doi.org/10.1016/j.physletb.2009.03.034. arXiv:0901.0507 [hep-ex]
12. H1 Collaboration, A. Aktas et al., A general search for new phenomena in ep scattering at HERA. Phys. Lett. B **602**, 14–30 (2004). https://doi.org/10.1016/j.physletb.2004.09.057. arXiv:hep-ex/0408044
13. K.S. Cranmer, Searching for new physics: contributions to LEP and the LHC. PhD thesis, Wisconsin University, Madison (2005). http://weblib.cern.ch/abstract?CERN-THESIS-2005-011
14. CDF Collaboration, T. Aaltonen et al., Model-independent and quasi-model-independent search for new physics at CDF. Phys. Rev. D **78**, 012002 (2008). https://doi.org/10.1103/PhysRevD.78.012002. arXiv:0712.1311 [hep-ex]
15. CDF Collaboration, T. Aaltonen et al., Model-independent global search for new high-p(T) physics at CDF. arXiv:0712.2534 [hep-ex]
16. CDF Collaboration, T. Aaltonen et al., Global search for new physics with 2.0 fb^{-1} at CDF. Phys. Rev. D **79**, 011101 (2009). https://doi.org/10.1103/PhysRevD.79.011101. arXiv:0809.3781 [hep-ex]
17. CMS Collaboration, C. Collaboration, MUSiC, a model unspecific search for new physics, in pp collisions at $\sqrt{s} = 8$ TeV
18. 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). http://cds.cern.ch/record/1360173
19. CMS Collaboration, MUSiC, a model unspecific search for new physics, in pp collisions at sqrt(s)=13 TeV
20. ATLAS Collaboration, M. Aaboud et al., A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment. Eur. Phys. J. C **79**, 120 (2019). https://doi.org/10.1140/epjc/s10052-019-6540-y. arXiv:1807.07447 [hep-ex]
21. ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV, ATLAS-CONF-2014-006 no. ATLAS-CONF-2014-006 (2014). https://cds.cern.ch/record/1666536
22. ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at sort(s)=7 TeV, ATLAS-CONF-2012-107 (2012). https://cds.cern.ch/record/1472686
23. R.T. D'Agnolo, A. Wulzer, Learning new physics from a machine. Phys. Rev. D **99**(1), 015014 (2019). https://doi.org/10.1103/PhysRevD.99.015014. arXiv:1806.02350 [hep-ph]
24. J.H. Collins, K. Howe, B. Nachman, Anomaly detection for resonant new physics with machine learning. Phys. Rev. Lett. **121**(24), 241803 (2018). https://doi.org/10.1103/PhysRevLett.121.241803. arXiv:1805.02664 [hep-ph]
25. J.H. Collins, K. Howe, B. Nachman, Extending the search for new resonances with machine learning. Phys. Rev. D **99**(1), 014038 (2019). https://doi.org/10.1103/PhysRevD.99.014038. arXiv:1902.02634 [hep-ph]
26. R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Learning multivariate new physics. arXiv:1912.12155 [hep-ph]
27. M. Farina, Y. Nakai, D. Shih, Searching for new physics with deep autoencoders. arXiv:1808.08992 [hep-ph]
28. T. Heimel, G. Kasieczka, T. Plehn, J.M. Thompson, QCD or what? SciPost Phys. **6**(3), 030 (2019). https://doi.org/10.21468/SciPostPhys.6.3.030. arXiv:1808.08979 [hep-ph]
29. T.S. Roy, A.H. Vijay, A robust anomaly finder based on autoencoder. arXiv:1903.02032 [hep-ph]
30. O. Cerri, T.Q. Nguyen, M. Pierini, M. Spiropulu, J.-R. Vlimant, Variational autoencoders for new physics mining at the large hadron collider. JHEP **05**, 036 (2019). https://doi.org/10.1007/JHEP05(2019)036. arXiv:1811.10276 [hep-ex]
31. A. Blance, M. Spannowsky, P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches. JHEP **10**, 047 (2019). https://doi.org/10.1007/JHEP10(2019)047. arXiv:1905.10384 [hep-ph]

---

[8] We also see a rise in performance for very large n_{latent} which is puzzling and mysterious.

32. B. Bortolato, B.M. Dillon, J.F. Kamenik, A. Smolkovič, Bump hunting in latent space. arXiv:2103.06595 [hep-ph]
33. J. Hajer, Y.-Y. Li, T. Liu, H. Wang, Novelty detection meets collider physics. arXiv:1807.10261 [hep-ph]
34. A. De Simone, T. Jacques, Guiding new physics searches with unsupervised learning. Eur. Phys. J. C **79**(4), 289 (2019). https://doi.org/10.1140/epjc/s10052-019-6787-3. arXiv:1807.06038 [hep-ph]
35. A. Mullin, H. Pacey, M. Parker, M. White, S. Williams, Does SUSY have friends? A new approach for LHC event analysis. arXiv:1912.10625 [hep-ph]
36. G.M. Alessandro Casa, Nonparametric semisupervised classification for signal detection in high energy physics. arXiv:1809.02977 [hep-ex]
37. B.M. Dillon, D.A. Faroughy, J.F. Kamenik, Uncovering latent jet substructure. Phys. Rev. D **100**(5), 056002 (2019). https://doi.org/10.1103/PhysRevD.100.056002. arXiv:1904.04200 [hep-ph]
38. A. Andreassen, B. Nachman, D. Shih, Simulation assisted likelihood-free anomaly detection. Phys. Rev. D **101**(9), 095004 (2020). https://doi.org/10.1103/PhysRevD.101.095004. arXiv:2001.05001 [hep-ph]
39. B. Nachman, D. Shih, Anomaly detection with density estimation. Phys. Rev. D **101**, 075042 (2020). https://doi.org/10.1103/PhysRevD.101.075042. arXiv:2001.04990 [hep-ph]
40. J.A. Aguilar-Saavedra, J.H. Collins, R.K. Mishra, A generic anti-QCD jet tagger. JHEP **11**, 163 (2017). https://doi.org/10.1007/JHEP11(2017)163. arXiv:1709.01087 [hep-ph]
41. M. Romão Crispim, N. Castro, R. Pedro, T. Vale, Transferability of deep learning models in searches for new physics at colliders. Phys. Rev. D **101**(3), 035042 (2020). https://doi.org/10.1103/PhysRevD.101.035042. arXiv:1912.04220 [hep-ph]
42. M.C. Romao, N. Castro, J. Milhano, R. Pedro, T. Vale, Use of a generalized energy mover's distance in the search for rare phenomena at colliders. arXiv:2004.09360 [hep-ph]
43. O. Knapp, G. Dissertori, O. Cerri, T.Q. Nguyen, J.-R. Vlimant, M. Pierini, Adversarially learned anomaly detection on CMS open data: re-discovering the top quark. arXiv:2005.01598 [hep-ex]
44. A. Collaboration, Dijet resonance search with weak supervision using 13 TeV pp collisions in the ATLAS detector. arXiv:2005.02983 [hep-ex]
45. B.M. Dillon, D.A. Faroughy, J.F. Kamenik, M. Szewc, Learning the latent structure of collider events. arXiv:2005.12319 [hep-ph]
46. M.C. Romao, N. Castro, R. Pedro, Finding new physics without learning about it: anomaly detection as a tool for searches at colliders. arXiv:2006.05432 [hep-ph]
47. O. Amram, C.M. Suarez, Tag N' Train: a technique to train improved classifiers on unlabeled data. arXiv:2002.12376 [hep-ph]
48. T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, T. Golling, Variational autoencoders for anomalous jet tagging. arXiv:2007.01850 [hep-ph]
49. C.K. Khosa, V. Sanz, Anomaly awareness. arXiv:2007.14462 [cs.LG]
50. P. Thaprasop, K. Zhou, J. Steinheimer, C. Herold, Unsupervised outlier detection in heavy-ion collisions. arXiv:2007.15830 [hep-ex]
51. S. Alexander, S. Gleyzer, H. Parul, P. Reddy, M.W. Toomey, E. Usai, R. Von Klar, Decoding dark matter substructure without supervision. arXiv:2008.12731 [astro-ph.CO]
52. J.A. Aguilar-Saavedra, F.R. Joaquim, J.F. Seabra, Mass unspecific supervised tagging (MUST) for boosted jets. arXiv:2008.12792 [hep-ph]
53. K. Benkendorfer, L.L. Pottier, B. Nachman, Simulation-assisted decorrelation for resonant anomaly detection. arXiv:2009.02205 [hep-ph]
54. A.A. Pol, V. Berger, G. Cerminara, C. Germain, M. Pierini, Anomaly detection with conditional variational autoencoders. arXiv:2010.05531 [cs.LG]
55. V. Mikuni, F. Canelli, Unsupervised clustering for collider physics. arXiv:2010.07106 [physics.data-an]
56. M. van Beekveld, S. Caron, L. Hendriks, P. Jackson, A. Leinweber, S. Otten, R. Patrick, R. Ruiz de Austri, M. Santoni, M. White, Combining outlier analysis algorithms to identify new physics at the LHC. arXiv:2010.07940 [hep-ph]
57. S.E. Park, D. Rankin, S.-M. Udrescu, M. Yunus, P. Harris, Quasi anomalous knowledge: searching for new physics with embedded knowledge. arXiv:2011.03550 [hep-ph]
58. D.A. Faroughy, Uncovering hidden patterns in collider events with Bayesian probabilistic models. arXiv:2012.08579 [hep-ph]
59. G. Stein, U. Seljak, B. Dai, Unsupervised in-distribution anomaly detection of new physics through conditional density estimation. arXiv:2012.11638 [cs.LG]
60. G. Kasieczka et al., The LHC Olympics 2020: a community challenge for anomaly detection in high energy physics. arXiv:2101.08320 [hep-ph]
61. P. Chakravarti, M. Kuusela, J. Lei, L. Wasserman, Model-independent detection of new physics signals using interpretable semi-supervised classifier tests. arXiv:2102.07679 [stat.AP]
62. J. Batson, C.G. Haaf, Y. Kahn, D.A. Roberts, Topological obstructions to autoencoding. arXiv:2102.08380 [hep-ph]
63. A. Blance, M. Spannowsky, Unsupervised event classification with graphs on classical and photonic quantum computers. arXiv:2103.03897 [hep-ph]
64. B. Nachman, Anomaly detection for physics analysis and less than supervised learning. arXiv:2010.14554 [hep-ph]
65. E.M. Metodiev, B. Nachman, J. Thaler, Classification without labels: learning from mixed samples in high energy physics. arXiv:1708.02949 [hep-ph]
66. T. Sjostrand, S. Mrenna, P.Z. Skands, A brief introduction to PYTHIA 8.1. Comput. Phys. Commun. **178**, 852–867 (2008). https://doi.org/10.1016/j.cpc.2008.01.036. arXiv:0710.3820 [hep-ph]
67. DELPHES 3 Collaboration, J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, M. Selvaggi, DELPHES 3, A modular framework for fast simulation of a generic collider experiment. JHEP **02**, 057 (2014). https://doi.org/10.1007/JHEP02(2014)057. arXiv:1307.6346 [hep-ex]
68. M. Cacciari, G.P. Salam, G. Soyez, FastJet user manual. Eur. Phys. J. C **72**, 1896 (2012). https://doi.org/10.1140/epjc/s10052-012-1896-2. arXiv:1111.6097 [hep-ph]
69. M. Cacciari, G.P. Salam, G. Soyez, The Anti-k(t) jet clustering algorithm. JHEP **04**, 063 (2008). https://doi.org/10.1088/1126-6708/2008/04/063. arXiv:0802.1189 [hep-ph]
70. J. Thaler, K. Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness. JHEP **02**, 093 (2012). https://doi.org/10.1007/JHEP02(2012)093. arXiv:1108.2701 [hep-ph]
71. J. Thaler, K. Van Tilburg, Identifying boosted objects with N-subjettiness. JHEP **03**, 015 (2011). https://doi.org/10.1007/JHEP03(2011)015. arXiv:1011.2268 [hep-ph]
72. A. Maas, A. Hannun, A. Ng, Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Conference on Machine Learning, Atlanta (2013)
73. V. Nair, G. Hinton, Rectified linear units improve restricted Boltzmann machines, vol. 27, pp. 807–814 (2010)
74. D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (ELUS) (2015)
75. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929–1958 (2014). http://jmlr.org/papers/v15/srivastava14a.html
76. D. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv:1412.6980 [cs]
77. F. Chollet, Keras. https://github.com/fchollet/keras (2017)

78. 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. OSDI **16**, 265–283 (2016)

79. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: an imperative style, high-performance deep learning library. In: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., pp. 8024–8035 (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

80. J. Gallicchio, J. Huth, M. Kagan, M.D. Schwartz, K. Black, B. Tweedie, Multivariate discrimination and the Higgs + W/Z search. JHEP **04**, 069 (2011). https://doi.org/10.1007/JHEP04(2011)069. arXiv:1010.3698 [hep-ph]

81. ATLAS Collaboration, G. Aad et al., Search for new resonances in mass distributions of jet pairs using 139 fb^{-1} of pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector. JHEP **03**, 145 (2020). https://doi.org/10.1007/JHEP03(2020)145. arXiv:1910.08447 [hep-ex]

82. C.M.S. Collaboration, A.M. Sirunyan et al., Search for narrow and broad dijet resonances in proton-proton collisions at $\sqrt{s}$ = 13 TeV and constraints on dark matter mediators and other new particles. JHEP **08**, 130 (2018). https://doi.org/10.1007/JHEP08(2018)130. arXiv:1806.00843 [hep-ex]

83. O. Kitouni, B. Nachman, C. Weisser, M. Williams, Enhancing searches for resonances with machine learning and moment decomposition. arXiv:2010.09745 [hep-ph]