






Finding new physics without learning about it: anomaly detection as a tool for searches at colliders

M. Crispim Romão¹ , N. F. Castro^{1,2,a} , R. Pedro¹ 

¹ LIP, Av. Professor Gama Pinto 2, 1649-003 Lisbon, Portugal

² Departamento de Física, Escola de Ciências, Universidade do Minho, 4710-057 Braga, Portugal

Received: 23 June 2020 / Accepted: 24 December 2020 / Published online: 15 January 2021

© The Author(s) 2021

Abstract In this paper we propose a new strategy, based on anomaly detection methods, to search for new physics phenomena at colliders independently of the details of such new events. For this purpose, machine learning techniques are trained using Standard Model events, with the corresponding outputs being sensitive to physics beyond it. We explore three novel AD methods in HEP: Isolation Forest, Histogram-Based Outlier Detection, and Deep Support Vector Data Description; alongside the most customary Autoencoder. In order to evaluate the sensitivity of the proposed approach, predictions from specific new physics models are considered and compared to those achieved when using fully supervised deep neural networks. A comparison between shallow and deep anomaly detection techniques is also presented. Our results demonstrate the potential of semi-supervised anomaly detection techniques to extensively explore the present and future hadron colliders' data.

1 Introduction

While the Standard Model of particle physics (SM) has been extremely successful in describing the experimental data accumulated so far, a significant number of open questions remains [1] and thus the search for new phenomena is a key aspect of the physics programme of present and future colliders. Given the practical difficulty of performing dedicated searches for all possible models and event topologies, inclusive searches and model-independent approaches are popular strategies to find a compromise between sensitivity and model independence of the experimental analyses. In fact, generic model-unspecific searches were conducted in the past by the D0 [2,3], CDF [4,5] and H1 [6,7] experiments at the Tevatron and HERA, respectively, and are also performed nowadays by the ATLAS [8] and CMS [9] Collaborations

of the Large Hadron Collider. Nonetheless, there is always the concern that a possible signal beyond the SM (BSM) is missed simply because the adopted strategy is not sensitive to it. In a previous work [10] we demonstrated that a possible direction to improve the sensitivity to BSM events without depending too much on the details of the considered signals is the supervised training of deep neural networks (DNN) since the performance of these networks does not significantly degrade when they are applied to another signal than the one used for training, as long as these signals are not very different from a topological point of view. A step forward in this direction is the use of anomaly detection (AD) methods, where only SM events are used in the training of the machine learning algorithm, allowing to isolate any BSM signal without knowing their details, avoiding any prior dependence and bias on the new physics that we are trying to discover.

The AD approach relies on identifying abnormal events in a data sample consisting, in the majority or completely, of normal events belonging to the same class. The problem is usually addressed by unsupervised learning with classical shallow algorithms running to identify the outlier events. In deep learning, Artificial Neural Networks such as Autoencoders (AE) have found their use as anomaly detectors since the error on the reconstruction of the inputs given by a model trained exclusively on normal events can be interpreted as an anomaly score. A known drawback of typical shallow methods, such as One-Class Support Vector Machines (OC-SVM), is the failure for high-dimensional data with many entries. This leads to a need for substantial feature engineering and dimensionality reduction before their application. On the other hand, the deep learning architecture of the AE family deals well with high-dimensional data and performs in anomaly detection despite not being trained specifically for discerning outlier events in the data.

The potential to isolate any unexpected signal from the SM prediction, commonly referred to as background, has motivated a growing interest for AD in HEP. For exam-

^a e-mail: Nuno.Castro@cern.ch (corresponding author)

ple, in Ref. [11], an unsupervised bump hunting approach using CWoLa [12] is proposed, while in Ref. [13], a Machine Learning (ML) model based on k-Nearest Neighbours is used to estimate event densities and assess how likely a new event is. Ref. [14] employs Neural Networks to compare the distribution of two samples and derive statistical tests to evaluate if any new physics is present. In Refs. [15–17] three different AE produce distributions of reconstruction errors to be used as anomaly scores, whereas [18] conjugates an AE with a linear outlier factor. More recently, in [19,20], novel non-ML approaches using density estimates are employed. On top of these examples, we also refer to the growing literature on the application of unsupervised or weakly supervised methods used to further understand the data generated at colliders [21–29].

The search for outlier events using anomaly detection techniques has a vast potential in the search for new phenomena in colliders, both at trigger (i.e. online) and analysis (i.e. offline) levels. Both applications have particular challenges and require dedicated efforts, namely in terms of the background modeling, event rates and statistical interpretation of the results. In this paper, we present three new unsupervised ML models for AD in the context of the offline analysis of HEP collisions, in addition to an AE, contributing to the path towards the use of such techniques by the experimental collaborations. In order to test their sensitivity to different BSM signals, the signals considered in [10] are used as benchmarks to assess the performance of the proposed approach by comparing it with supervised DNN classifiers trained on the same signals. In this way, we compare the performance of the AD methods to supervised DNNs. As such, we further contribute to the ongoing effort – see for example [15,29] – to systematically compare different unsupervised AD methods in searches for new physics.

2 Methods for anomaly detection

We use shallow and deep learning techniques trained on a data sample of Standard Model simulated events and test the ability of each model to identify new physics events with benchmark signals unseen during the training phase. Histogram-Based Outlier Detection (HBOS) [30] and Isolation Forest (iForest) [31] are the shallow models explored. These methods are guided to isolate instances of the data in the tails of the feature distributions and, unlike OC-SVM, are fast and scalable to high-dimensional data with many instances. As a deep model, we analyse the recently proposed Deep Support Vector Data Description (Deep SVDD) [32]. Contrary to an AE, the Deep SVDD is designed for outlier discovery. AEs, popularly used in AD tasks, are also explored.

2.1 Histogram-based outlier detection

In HBOS, a histogram is computed for each input feature and an anomaly score is derived based on how populated the bins where an instance falls on are. In the training phase, the predicted SM yields are used to construct the bins. On the test phase, the score of a new instance is computed as follows. For each of its features, we see in what bin of the histogram its value falls on, and assign an associated score of $\log_2(\text{Hist})$, with Hist being the density of the histogram where the instance value of the feature is, i.e. the height of the bin that contains that value. The total anomaly score is the sum across all features.

2.2 Isolation forest

The iForest algorithm randomly selects an input feature and a split value within the feature boundaries to recursively partition the data. The idea is that outliers are easier to isolate than normal instances of the data and the number of data splits can be used as a base for an anomaly score. In the training phase, the iForest model learns the feature boundaries from the training sample and on the test phase each event is isolated and its outlyingness is obtained.

2.3 Deep autoencoder

Deep AE is a deep architecture that learns to compress (encode) and then decompress (decode) data through a bottleneck intermediate layer that has a smaller dimensionality than the data. The AE is trained by minimising the reconstruction error, i.e. how different a decoded instance is from the original, through the training objective:

$$\min_{\mathcal{W}} \frac{1}{n} \sum_i ||\text{AE}(\mathbf{x}_i, \mathcal{W}) - \mathbf{x}_i||^2, \quad (1)$$

where \mathcal{W} are the weights of the AE, \mathbf{x}_i the feature vector of the i th event and n the total number of events. Since uncommon events will, in principle, be harder to reconstruct than more common ones, the reconstruction error can then be used as an anomaly score.

2.4 Deep support vector data description

The Deep SVDD architecture is designed in analogy to its shallow counterpart, the support vector data description, which in turn is closely related to OC-SVM. In SVDD, the data is mapped into an abstract feature space and, during training, we minimise the mean distance of data points to the centre of the data distribution in this space. In the deep version, this is implemented as follows. We initialise a DNN and calculate the average position of its outputs given the

training set. This will give us the centre of the distribution of the data in the space defined by the last layer of the DNN. Training is then performed as to minimise the distance of all points of the training set to this centre and can be expressed through the training objective:

$$\min_{\mathcal{W}} \frac{1}{n} \sum_i ||\text{DNN}(\mathbf{x}_i, \mathcal{W}) - \mathbf{c}||^2, \quad (2)$$

where \mathcal{W} are the weights of the DNN, \mathbf{c} the centre of the distribution in the output space, \mathbf{x}_i the feature vector of the i th event. In order to prevent pathological behaviours arising from trivial solutions associated with collapses of the whole distribution to \mathbf{c} , the DNN must have non-saturated activation functions, it must not have bias terms, and \mathbf{c} can be neither the origin of the output space nor a learnable parameter. The anomaly score of an event in a Deep SVDD is then deduced from how far from the centre, \mathbf{c} , the event lies.

2.5 Supervised classifier

In addition, we trained a supervised classifier, based on deep neural networks, for each benchmark signal (*c.f.* Sect. 3). This will provide us with a baseline with which to compare the AD algorithms performance.

3 Simulated datasets

We tested the different AD methods in the context of collider searches and our dataset is composed of simulated proton-proton collision events. The samples were generated with MADGRAPH5_MCATNLO 2.6.5 [33] at leading order with a collision centre-of-mass energy of 13 TeV. Pythia 8.2 [34] was employed to simulate the parton shower and hadronisation, with the CMS CUETP8M1 [35] underlying event tuning and the NNPDF 2.3 [36] parton distribution functions. The detection of the collision products was accomplished with a multipurpose detector simulator, Delphes 3 [37]. The configuration of Delphes was kept to the default, matching the parameters of the CMS detector. Jets and large-radius jets are reconstructed using the anti- κ_r algorithm [38] with a radius parameter of $R = 0.5$ and 0.8 , respectively.

One of our goals is to compare the AD performance to the one obtained with dedicated supervised deep learning, which we explored previously [10]. For this reason, we studied the same BSM signals, namely the pair production of vector-like T -quarks (either produced via SM gluons [39] or BSM heavy gluons [40]) and tZ production through a flavour changing neutral current (FCNC) vertex [41]. In total, seven benchmark signals were generated: $T\bar{T}$ with $m_T = 1.0, 1.2, 1.4$ TeV produced via SM gluon or a massive 3 TeV gluon, and tZ FCNC production.

We preselected events broadly compatible with the signal topologies commonly considered by the ATLAS and CMS experiments [42–45]: at least two final state leptons (i.e. electrons or muons), at least one b -tagged jet, and large scalar sum of transverse momentum (p_T) of all reconstructed particles in the event ($H_T > 500$ GeV).¹ The most important SM processes compatible with the event selection topology are Z +jets, top pair ($t\bar{t}$) production and dibosons (WW , WZ and ZZ). The generation of each of these processes was sampled in kinematic regions to ensure a good statistical representation across the entire phase space, and especially in the tails of the distributions, where anomalous events are particularly expected. This sampling employed event generation filters at parton level according to:

- The top/anti-top p_T (p_T^{top}) for $t\bar{t}$: $p_T^{\text{top}} < 100$ GeV, $p_T^{\text{top}} \in [100, 250]$ GeV, $p_T^{\text{top}} > 250$ GeV;
- The scalar sum of the p_T of the hard-scatter outgoing particles for Z +jets: $S_T < 250$ GeV, $S_T \in [250, 500]$ GeV, $S_T > 500$ GeV;
- W/Z p_T ($p_T^{W/Z}$) for dibosons: $p_T^{W/Z} < 250$ GeV, $p_T^{W/Z} \in [250, 500]$ GeV, $p_T^{W/Z} > 500$ GeV.

In order to ensure a reasonable statistics across the relevant phase space, the Z +jets simulation was further split into the jet flavour as Zjj and Zbb . Over 18 M events were simulated: 500 k per signal sample, 8 M for Z +jets, 3 M for $t\bar{t}$ and 1.5 M per diboson sample.

Furthermore, the generated events were also hadronised with Herwig 7 [46,47], employing NNPDF 2.3 [36] parton distribution functions, in order to produce an alternative set of samples to test the robustness of AD techniques against uncertainties on the parton shower and hadronisation modelling.

The SM cocktail used to train the AD methods is composed of the SM simulated samples, each normalised to the expected yield after selection using the generation cross-section at leading order, computed with MADGRAPH5, and matched to a target luminosity of 150 fb^{-1} . This normalisation is parsed as a form of event weights to the AD method. The data features correspond to basic information constituted of the four-momenta of the reconstructed particles as provided by the Delphes simulation:

- (η, ϕ, p_T, m) of the 5 leading jets and large-radius jets;
- (η, ϕ, p_T) of the 2 leading electrons and muons;
- multiplicity of jets, large-radius jets, electrons and muons;
- (E_T, ϕ) of the missing transverse energy.

¹ The transverse plane is defined with respect to the proton colliding beams.

Some of these features manifest an accumulation of density at the origin. This happens for objects that might not have been reconstructed, such as sub-leading large-radius jets or flavour-explicit leptons. This will produce density functions for these features, which are not continuous and can hinder the performance of deep learning models. In light of Universal Approximation Theorems for neural networks [48–51], we know that neural networks are only guaranteed to approximate any *continuous* function when given enough capacity, i.e. enough width and/or units. Therefore, it is only reasonable to assume that when the features are described by non-continuous densities, a neural network will have to learn a non-continuous function during training that will be difficult to learn as it is not guaranteed that it can be approximated. Consequently, we prepared the data with a second set of features that aims to mitigate this. This second set of features, which we refer to as *sanitised*, retains only the events with one large-radius jet while dropping the features of all sub-leading large-radius jets. In addition, we keep only the two leading leptons regardless of the flavour, dropping the remainder.

4 Implementation details and training

The data were split into train, validation and test sets with equal proportions to guarantee similar statistical representativity at each stage. When hyperparameters were tuned, the metrics used to help choosing the best configuration were computed on the validation set. A statistically independent test set was used to evaluate the performance of the AD methods in isolating BSM signals.

4.1 Shallow methods

We implemented the HBOS algorithm based on the `pyod` Python toolkit [52], but we changed the code to take sample normalisation weights into account when computing the histograms. For the iForest, we based our implementation on the Scikit-Learn [53] through the `pyod` wrapper [52].

For both the HBOS and the iForest implementations the data was preprocessed by a standardisation step, which sets all the features means to 0 and their standard deviation to unity, followed by a principal component rotation, where we retained the full dimensionality of the feature space. The purpose of this rotation is to remove linear correlations between the features, an assumption that is required by these methods. The preprocessing steps were implemented with Scikit-Learn [53].

4.2 Deep methods

We implemented the deep models in TensorFlow 2.3 [54]. In order to find the best hyperparameters for the deep architec-

tures, we implemented a bayesian hyperparameter optimisation step using the Python package `optuna` [55]. The hyperparameter optimisation step made use of the `optuna` built-in tree-structured parzen estimator [56] to suggest new hyperparameter combinations, over a loop of increasing number of maximum epochs to improve search efficiency. In addition, manually discovered promising hyperparameter combinations were added to the evaluation queue.

A crucial hyperparameter to be fixed before the hyperparameter optimisation loop is that of the dimensionality of the latent space of the AE and the embedding space of the Deep SVDD. The reason to fix it is twofold: on the one hand an AE hyperparameter optimisation step will always prefer a large latent space, which will fix it to the highest value possible during search; on the other hand, it is difficult to compare distributions of distances on different dimensions, making model comparison and selection for the Deep SVDD challenging. In [32] the second problem was circumvented by reusing the AE encoder as a pre-trained Deep SVDD. In this work, we let the Deep SVDD to be trained from scratch, but fixed the embedding dimension to be the same as of the latent space of the AE. We did not use the encoder of a trained AE as we observed that this led to instabilities during training and difficulties in reproducing the same results. Instead, we fixed the embedding dimension and optimised the remainder hyperparameters of the Deep SVDD using the same Bayesian search. The latent space dimension of the AE and the embedding space dimension of the Deep SVDD was set to 16, as it is roughly a quarter of the input dimensionality.

For the Deep SVDD, the vector \mathbf{c} , which represents the centre of the distribution of the data in the embedding space, was calculated as follows. First, we defined the model and initialise all its learning parameters. Next, and before any optimiser step, we forward pass the whole training set through this network and calculate the weighted average of the outputs. This will then be the centre of mass of the distribution in the embedding space and therefore sets \mathbf{c} .

All deep models were trained with a custom cosine-cyclical learning rate with warmup. The warmup phase was set to a 25 epochs period, where the learning rate linearly increased from an initial value, `Initial LR`, to its maximum value, `Max LR`, both to be optimised during the `optuna` loop. The cycle was set with a period of 50 epochs, during each period the learning rate oscillates between the maximum learning rate down to an order of magnitude lower. During the cycle phase, the maximum learning rate was multiplied by a factor, `gamma`, at the end of each epoch, exponentially decreasing it, which was optimised during the bayesian optimisation loop. We found this type of learning rate to significantly improve the converge speed of both AE and Deep SVDD, as well as to improve the training stability in terms of reproducibility of the final outcome. The training was stopped if no improvement of the loss on the validation

Table 1 Hyperparameter search spaces. The sampling for Initial LR and Max LR was performed logarithmically. For the AE the number of layers corresponds to both the number of encoder and decoder layers

Hyperparameter	Possible values
Number of Layers	[1, 5]
Number of Units	[32, 256]
Initial LR	$[10^{-8}, 10^{-3}]$
Max LR	$[10^{-3}, 10^{-1}]$
Gamma	[0.95, 0.999] in steps of 0.001
Weight Decay	{0, 10^{-9} , 10^{-8} , 10^{-7} , 10^{-6} , 10^{-5} , 10^{-4} , 10^{-3} }
Clipnorm	{None, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0}

Table 2 Best hyperparameter configurations for both deep AD models on both feature sets

Hyperparameter	AE		Deep SVDD	
	Full features	Sanitised features	Full features	Sanitised features
Number of layers	5	3	2	1
Number of Units	171	93	47	128
Initial LR	7×10^{-6}	4.487459×10^{-7}	5.834093×10^{-8}	10^{-6}
Max LR	0.023328	0.063960	0.005186	0.02
Gamma	0.998	0.992	0.971	0.995
Weight Decay	10^{-6}	0.0	10^{-9}	10^{-8}
Clipnorm	0.10	100.0	100.0	None

set was observed for 200 epochs for the AE, 300 for the Deep SVDD, and 100 for the supervised classifiers, after which the weights of the best epoch were kept, persisting the best models at every stage.² The AE was trained using mini-batches of size 4096, while the Deep SVDD and the supervised classifiers were trained in mini-batches of size 1024. All hidden layers activation functions were set to LeakyReLU.

In addition, all models were trained with the Adam optimiser [57], through the weight-decay wrapper provided by Tensorflow-Addons in order to implement weight-decay regularisation compatible with Adam [58]. The value of the weight-decay was optimised during the hyperparameter search loop. Furthermore, since the Deep-SVDD cannot have non-homogeneous learnable parameters, i.e. biases, we implemented a non-trainable Batch Normalisation or otherwise the learnable mean would effectively behave as a bias term and lead to trivial collapse solutions. Since preventing trivial solutions requires not using saturating activation functions and learnable batch normalisation layers, one would expect only shallower networks to be successfully trained in order to avoid vanishing and exploding gradients. To mitigate this, we allowed for the gradients to be norm-clipped to a value to be optimised.

The hyperparameter optimisation loop details can be found in Table 1. The best combinations were chosen by minimising validation loss, and the final configurations for

the AD models for both feature sets can be seen in Table 2. We do not present the best hyperparameters for the supervised classifiers for brevity.

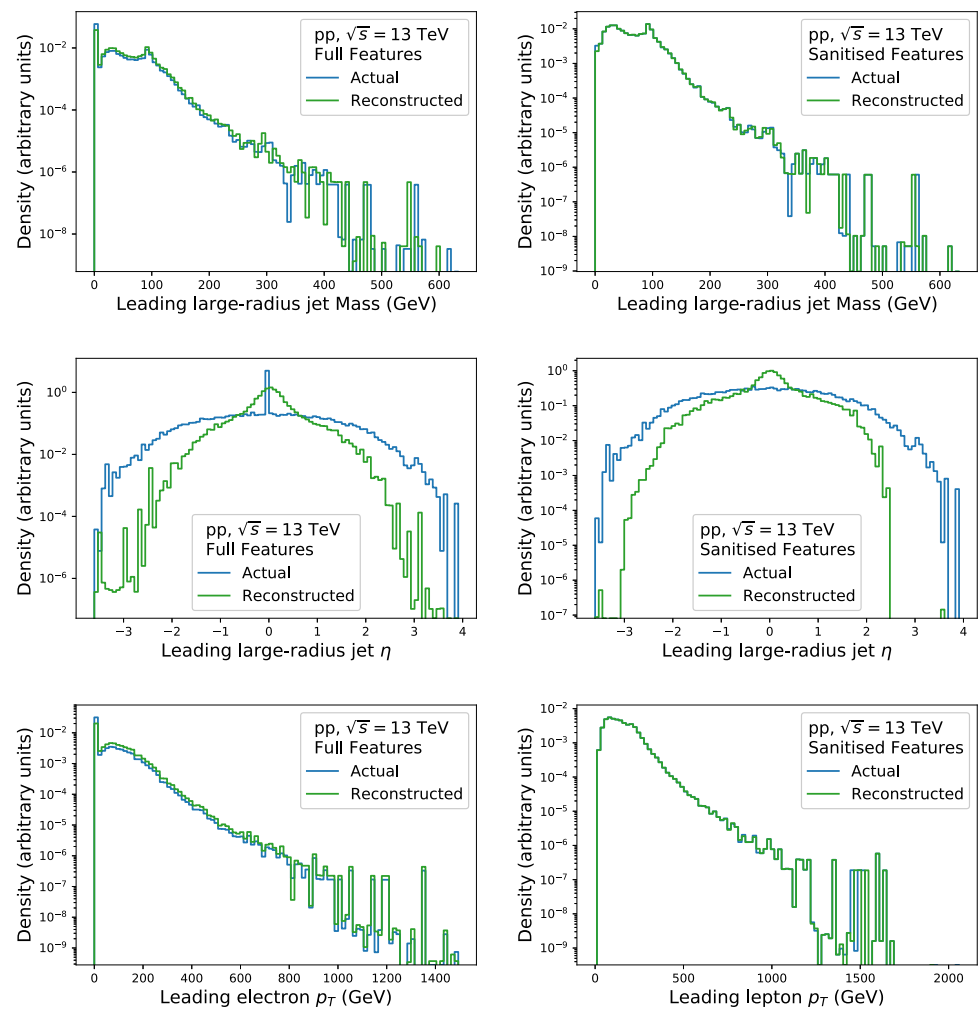
For both the AE and the Deep SVDD methods, the anomaly score was derived from the loss, i.e. Eqs. 1 and 2, by taking the base 10 logarithm of the values and scaling them as to fall in the interval [0, 1]. It is also important to reiterate that the signals samples were not used at any stage of both hyperparameters selection and AD model training.

4.3 Feature impact on reconstructions

When using the full feature set, which includes events with missing reconstructed objects, we observed that for features with pronounced accumulations at the origin, the reconstruction was degraded. In Fig. 1 we highlight this behaviour for three different features. For the mass of the leading large-radius jet, we notice how the accumulation in zero impacts the reconstruction of the rest of the spectrum. In the second case, concerning the η of the leading large-radius jet, we notice that for the case with zero accumulation the AE struggles to reconstruct values away from the mean, i.e. the origin. Removing the events without a leading large-radius jet has mitigated this problem. Finally, a similar behaviour as that of the large-radius jet is observed for the leading electron in the third case. Retaining only the two reconstructed leptons required at event pre-selection level provides a better result for the sanitised feature set.

² We allowed a larger patience for the Deep SVDD early stop criteria as we observed the loss to oscillate significantly at early stages.

Fig. 1 Distribution of some of the real input features and their reconstruction by the Autoencoder on the validation set. Left: Using all features set. Right: Using sanitised features set



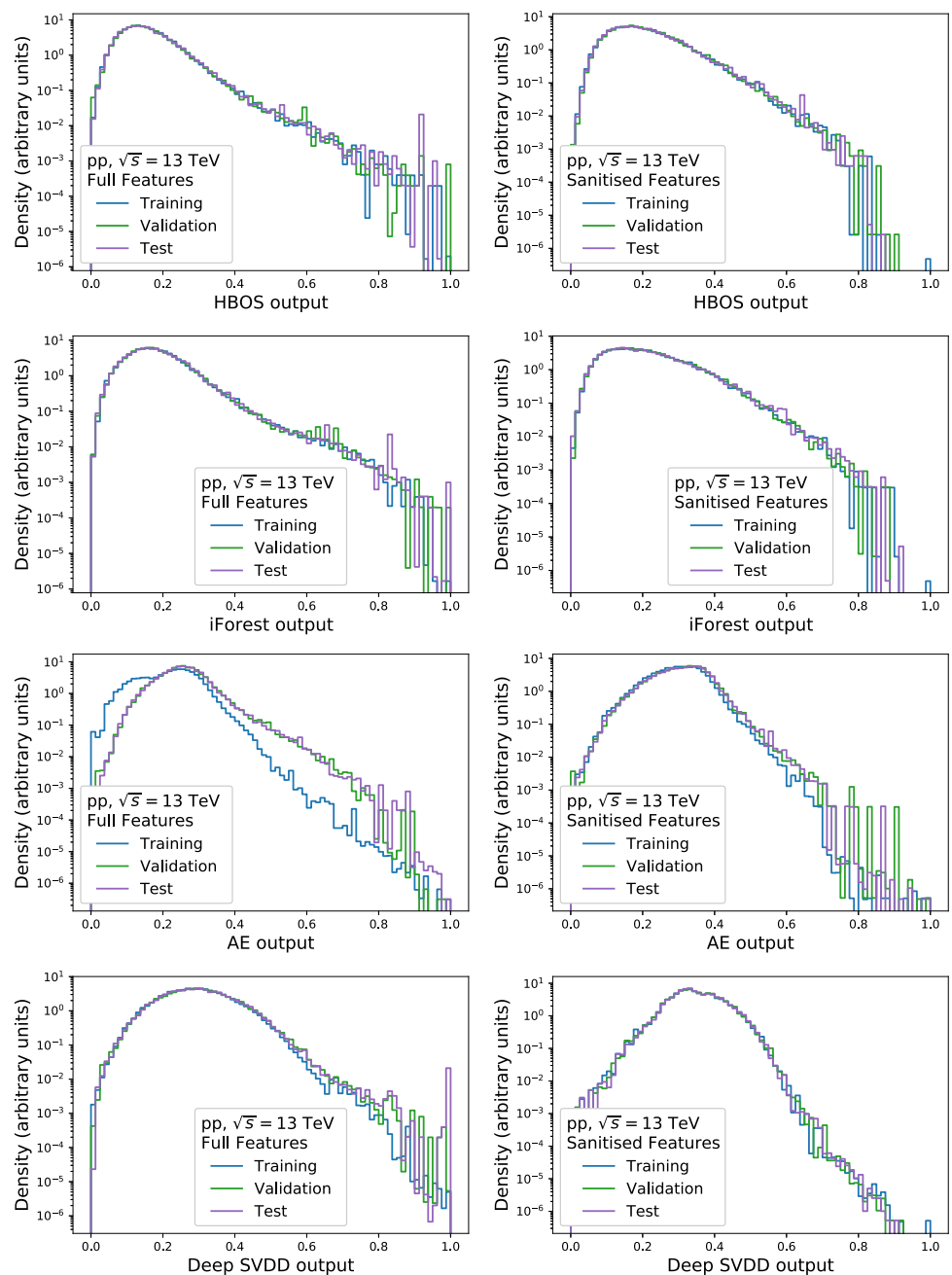
Furthermore, as seen in the η of the same large-radius jet distributions, removing the excess density around zero did not completely solve the reconstruction challenges of this variable. Indeed, we noticed that η and ϕ variables were always difficult to reconstruct in our working methodology, even after the hyperparameter optimisation step. We also note that transforming the 4-momenta variables to the cartesian coordinates did not resolve this issue. This problem highlights the challenges that DNN encounter when presented with inputs which would be better represented in varying length, such as recurrent neural networks and graph neural networks, which have been finding their way into HEP applications [59,60]. However, systematically study the best data representation and corresponding neural network architecture in order to provide optimal reconstruction of features using a deterministic AE is beyond the scope of this work and as such we defer such concerns to future work.

4.4 Anomaly scores for training and validation samples

The anomaly score distributions for each of the four AD methods are presented in Fig. 2 for the training, validation, and test samples and both feature sets. We notice that for the full features set both the deep AD models manifest a more pronounced difference between the training and validation sample distributions, with a significant difference in the AE case. However, we also observe that the validation sample follows same distribution as the test sample, where the upper limits on signal strength, i.e. the physical application, will be calculated. This provides some confidence that, even though these methods are overfitting to the training data, the observed behaviour for the validation set is expected to carry to the test set.

In Fig. 3 we show the distributions of four example features for the 10% most anomalous events under each AD method score – i.e. the events whose score lies in the 10% outlier quantile calculated on the validation distribution shown in Fig. 2, using the sanitised feature set. The figure shows that the AD algorithms are capturing the tails of distribu-

Fig. 2 Anomaly score for the different AD methods (HBOS, iForest, Autoencoder, Deep SVDD) for training, validation, and test data. The distributions are normalised to the unit area. Left: Using all features set. Right: Using sanitised features set



tions. However, we can see from the Jet Multiplicity distribution that the Deep SVDD seems to be capturing different events than the remaining AD methods, manifesting that the anomaly/outlyingness of an event can be very much dependent of the type of AD algorithm.

In Figs. 4 and 5 we present the distributions and the scatter plots of the anomaly scores for each process of the SM cocktail used in the AD model training. The correlation trends are similar across the individual SM processes. We notice how the shallow methods are highly related between each other for both feature sets over the validation sample. In contrast, both deep models show looser relation between their predic-

tions and the shallow predictions, and amongst themselves, for both feature sets. More interestingly, we notice how the Deep SVDD and the AE have a small correlation in the sanitised set. Again, these results point to the fact that different AD algorithms will be capturing different anomalous events.

5 Comparison of the AD methods for benchmark signals

In this section, we assess the performance of the trained AD models to discriminate signals from new physics, not present

Fig. 3 Distribution of some of the input features for the full validation set and for the 10% outlier quantile according to the anomaly score for the different AD methods using sanitised features. All distributions are normalised to the unit area

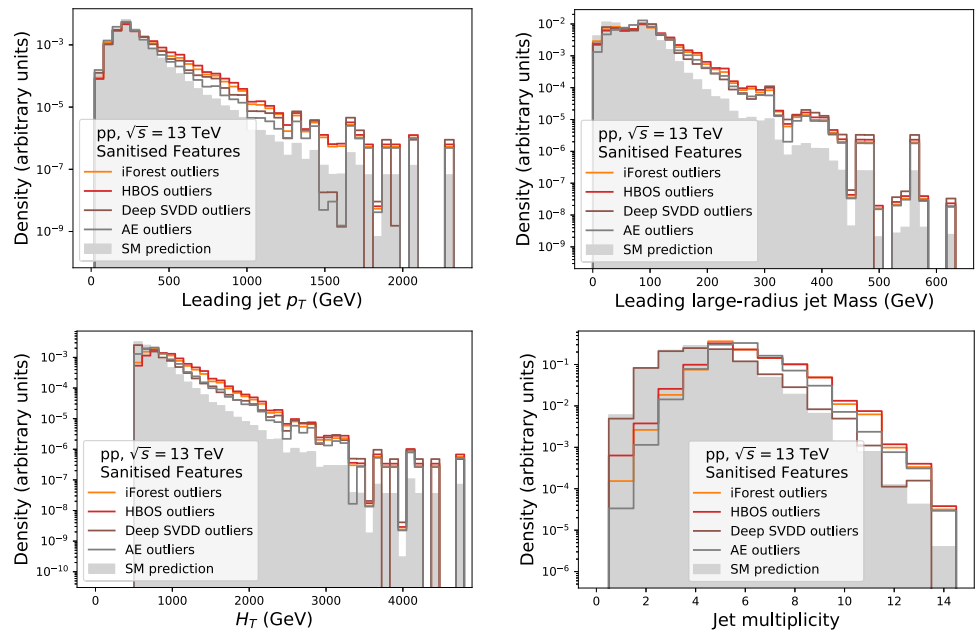


Fig. 4 Two-dimensional distribution of the anomaly scores for the different AD methods per SM process – $t\bar{t}$, Z+jets and diboson – using all features set. Diagonal: Distribution of the anomaly score per SM process

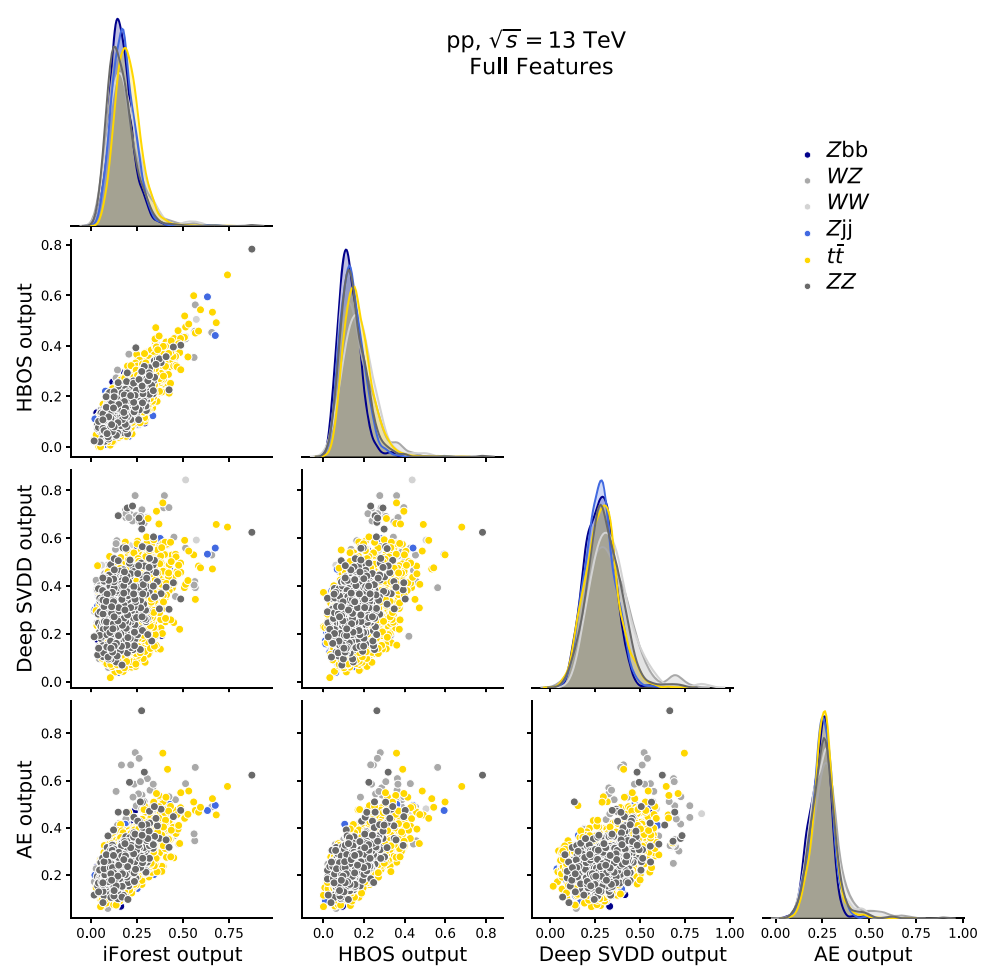
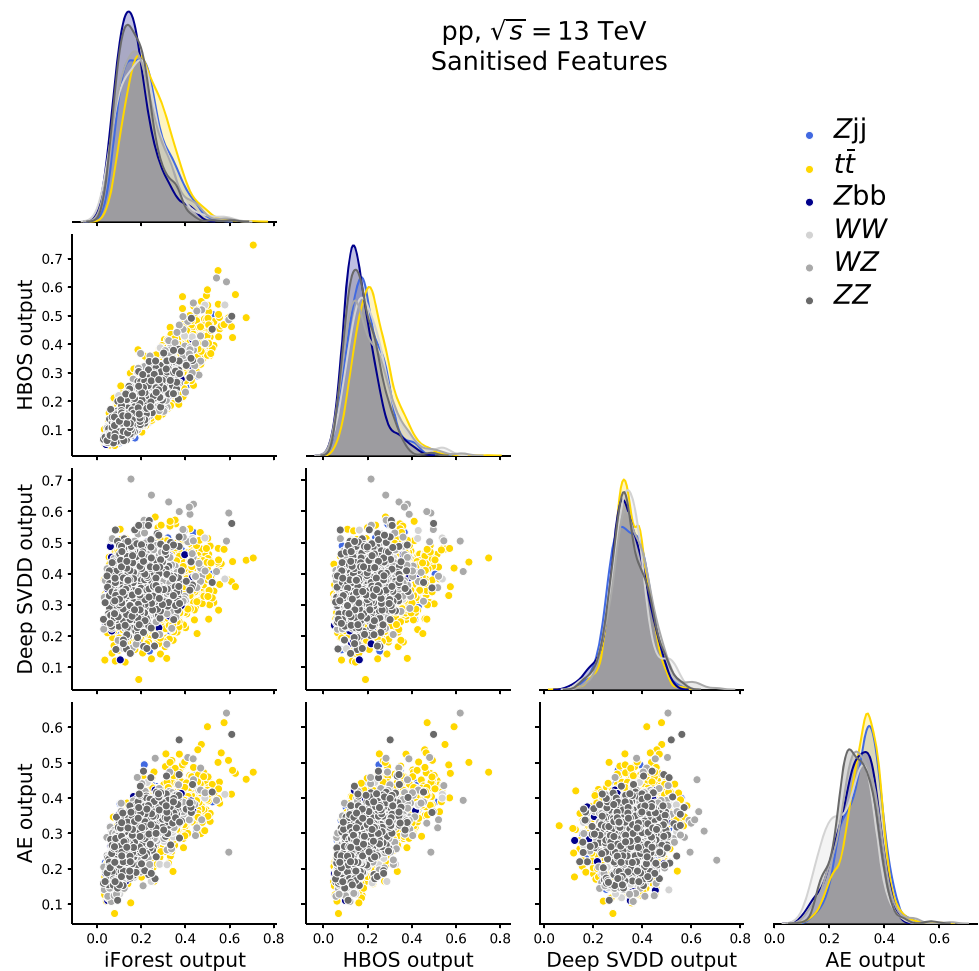


Fig. 5 Two-dimensional distribution of the anomaly scores for the different AD methods per SM process – $t\bar{t}$, Z +jets and diboson – using sanitised features set. Diagonal: Distribution of the anomaly score per SM process



in the SM cocktail used for their development. The performance metric is based on the 95% confidence level (CL) upper limit on the signal strength μ , defined as the ratio between the expected upper limit on the signal cross-section, normalised to the corresponding theory prediction, computed at leading order. Such limits were obtained by fitting the AD score distribution of the test data set and were computed using the CL_s method [61], as implemented in OpTHyLiC [62]. Poissonian statistical uncertainties on each bin of the distributions were included in the limit computation, assuming an integrated luminosity of 150 fb^{-1} .

5.1 Anomaly score distributions

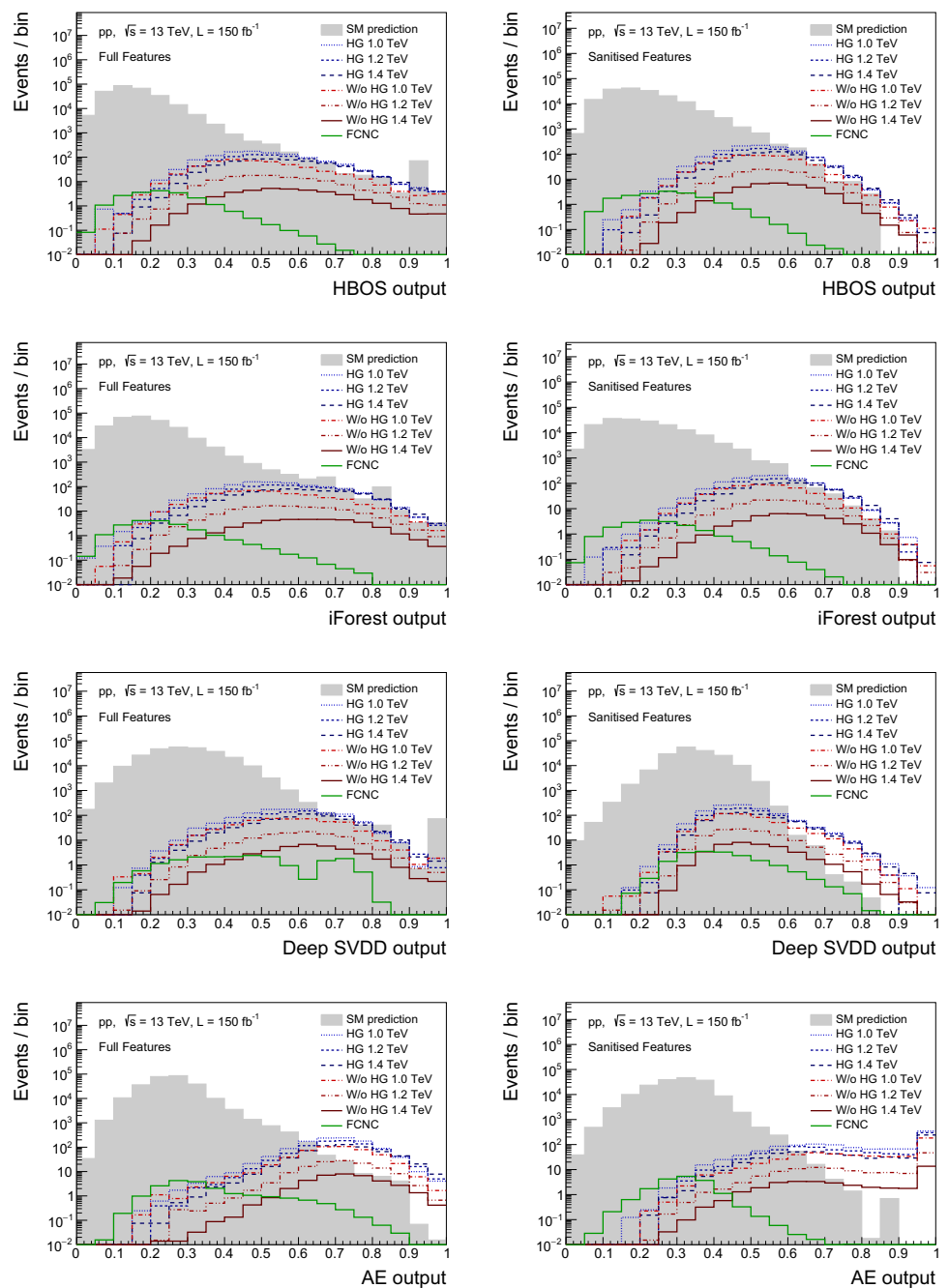
In Fig. 6 we present the output distributions of the four AD models trained on both feature sets, for the SM prediction and each benchmark signal. We observe that the shallow methods have similar behaviour for both feature sets, and in each of them, the FCNC signal follows a distribution that is very close to the one followed by the SM processes. In contrast, the vector-like T -quarks are being assigned on average higher anomaly scores.

For the deep models, we observe a significant difference in distribution shapes when we switch from the full feature set to the sanitised feature set. In particular, we notice how the Deep SVDD provides significant better capacity to isolate signal with the sanitised feature set. For the AE, the FCNC distribution becomes more similar to the SM background when using the sanitised feature set, as it happens to the shallow methods. In both cases, the anomaly score distributions for the signals have their mass shifted to the right, meaning that on average abnormal signals have higher anomaly scores than the SM events and that this behaviour is more noticeable in the deep models.

5.2 Expected upper limits

We fit the distributions presented in Fig. 6, to determine upper limits on the signal strength. In Table 3 we show the central values of the upper limit on μ and the associated statistical uncertainties. In Fig. 7 are presented the same central values but normalised to the first line, i.e. to the supervised DNN using the full feature set. We observe that the deep models, both AD and supervised, had significant per-

Fig. 6 Distribution of the AD discriminant for the SM prediction and each signal type: tZ production by FCNC, $T\bar{T}$ production via heavy gluon or without heavy gluon for $m_T = \{1.0, 1.2, 1.4\}$ TeV. The distributions are normalised to the generation cross-section and to an integrated luminosity of 150 fb^{-1} . Left: Using all features set. Right: Using sanitised features set



formance impact by switching the feature set. In particular, we noticed how the Deep SVDD significantly improved when using the sanitised features for all cases. Furthermore, the AE has a sensitivity similar to supervised DNN for signals with vector-like quarks. On the other hand, the shallow models retained the same discriminating power when changing the features.

Another relevant result that we observe is how, with sanitised features, the AE seems to focus more strongly on the out tails of the distributions and therefore provides upper limits that are competitive to those derived using a supervised

discriminant. On a different direction, the Deep SVDD produced similar discriminant power for all signals, including the FCNC, which is far more similar to the SM distribution than the signals with VLQ. This reinforces the idea that different AD algorithms are capturing outliers differently and might indicate, for instance, that although having worst performance when compared to AE, Deep SVDD might be interesting in searches for signals of new physics implying small deviations of the SM. A more detailed study of this behaviour, as well as of the propagation of systematic sources of uncertainties through these methods is left for a future study.

Table 3 95% CL upper limit on the signal strength μ of each benchmark signal for the different AD methods using the full feature set and the sanitised set and for a dedicated supervised DNN model trained

on the full feature set. The statistical uncertainties, including the effect from limited statistics in the simulated datasets, are also shown

Model	Benchmark signal						
	FCNC	HG			No HG		
		1.0 TeV	1.2 TeV	1.4 TeV	1.0 TeV	1.2 TeV	1.4 TeV
Full features							
Supervised DNN	7_{-2}^{+3}	$0.05_{-0.03}^{+0.06}$	$0.019_{-0.007}^{+0.010}$	$0.013_{-0.004}^{+0.008}$	$0.03_{-0.01}^{+0.02}$	$0.12_{-0.04}^{+0.06}$	$0.4_{-0.2}^{+0.2}$
H_T	100_{-20}^{+60}	$0.14_{-0.05}^{+0.07}$	$0.16_{-0.06}^{+0.08}$	$0.16_{-0.05}^{+0.08}$	$0.4_{-0.1}^{+0.3}$	$1.0_{-0.3}^{+0.5}$	$1.8_{-0.6}^{+0.9}$
Deep SVDD	10_{-1}^{+8}	$0.15_{-0.04}^{+0.07}$	$0.17_{-0.05}^{+0.08}$	$0.21_{-0.07}^{+0.09}$	$0.3_{-0.1}^{+0.2}$	$1.1_{-0.3}^{+0.5}$	$3.1_{-0.9}^{+1.4}$
AE	30_{-8}^{+20}	$0.029_{-0.009}^{+0.014}$	$0.03_{-0.01}^{+0.02}$	$0.04_{-0.01}^{+0.02}$	$0.06_{-0.02}^{+0.03}$	$0.21_{-0.07}^{+0.10}$	$0.6_{-0.2}^{+0.3}$
HBOS	100_{-20}^{+60}	$0.15_{-0.05}^{+0.07}$	$0.17_{-0.05}^{+0.08}$	$0.19_{-0.06}^{+0.09}$	$0.4_{-0.1}^{+0.1}$	$1.0_{-0.3}^{+0.5}$	$2.7_{-0.9}^{+1.3}$
iForest	100_{-3}^{+100}	$0.19_{-0.06}^{+0.10}$	$0.23_{-0.08}^{+0.12}$	$0.26_{-0.09}^{+0.14}$	$0.5_{-0.2}^{+0.2}$	$1.4_{-0.5}^{+0.7}$	4_{-2}^{+2}
Sanitised features							
Supervised DNN	6_{-2}^{+3}	$0.008_{-0.003}^{+0.004}$	$0.009_{-0.003}^{+0.005}$	$0.006_{-0.001}^{+0.003}$	$0.009_{-0.003}^{+0.005}$	$0.04_{-0.01}^{+0.03}$	$0.3_{-0.1}^{+0.2}$
H_T	100_{-30}^{+50}	$0.14_{-0.05}^{+0.07}$	$0.16_{-0.06}^{+0.08}$	$0.16_{-0.05}^{+0.08}$	$0.4_{-0.1}^{+0.3}$	$1.0_{-0.3}^{+0.5}$	$1.8_{-0.6}^{+0.9}$
Deep SVDD	10_{-2}^{+7}	$0.08_{-0.02}^{+0.04}$	$0.08_{-0.02}^{+0.04}$	$0.09_{-0.02}^{+0.04}$	$0.15_{-0.04}^{+0.06}$	$0.5_{-0.1}^{+0.2}$	$1.4_{-0.4}^{+0.6}$
AE	100_{-1}^{+100}	$0.0053_{-0.0005}^{+0.0006}$	$0.0068_{-0.0006}^{+0.0006}$	$0.0089_{-0.0008}^{+0.0007}$	$0.0104_{-0.0012}^{+0.0009}$	$0.042_{-0.004}^{+0.004}$	$0.15_{-0.01}^{+0.02}$
HBOS	100_{-20}^{+60}	$0.19_{-0.06}^{+0.11}$	$0.21_{-0.07}^{+0.13}$	$0.22_{-0.07}^{+0.14}$	$0.4_{-0.1}^{+0.2}$	$1.1_{-0.4}^{+0.6}$	$2.7_{-0.9}^{+1.7}$
iForest	100_{-5}^{+100}	$0.18_{-0.06}^{+0.09}$	$0.19_{-0.07}^{+0.09}$	$0.19_{-0.07}^{+0.09}$	$0.4_{-0.1}^{+0.3}$	$1.0_{-0.3}^{+0.6}$	$2.4_{-0.8}^{+1.2}$

For comparison, Table 3 also shows the limits for each signal type obtained by fitting the distribution of the scalar sum of transverse momentum (p_T) of all reconstructed particles in the event (H_T) as a simpler, but commonly used [42], alternative to the use of ML methods. While the shallow methods are always worst than a simple H_T fitting, the AE performs significantly better for all the benchmark signals, with the exception of the FCNC case, when the sanitized features are used.

The results show that these unsupervised AD algorithms are reasonably sensitive to new signals, with a maximum degradation relative to the supervised DNN of around an order of magnitude on the μ exclusion limits, for the worst cases, and no significant impact for the best ones. Interestingly, in previous work where DNN trained on different models were used to discriminate between the background and other signals [10], we observed similar trends when training deep neural networks on signals different from those used for the classification.

6 Robustness of the anomaly detection methods

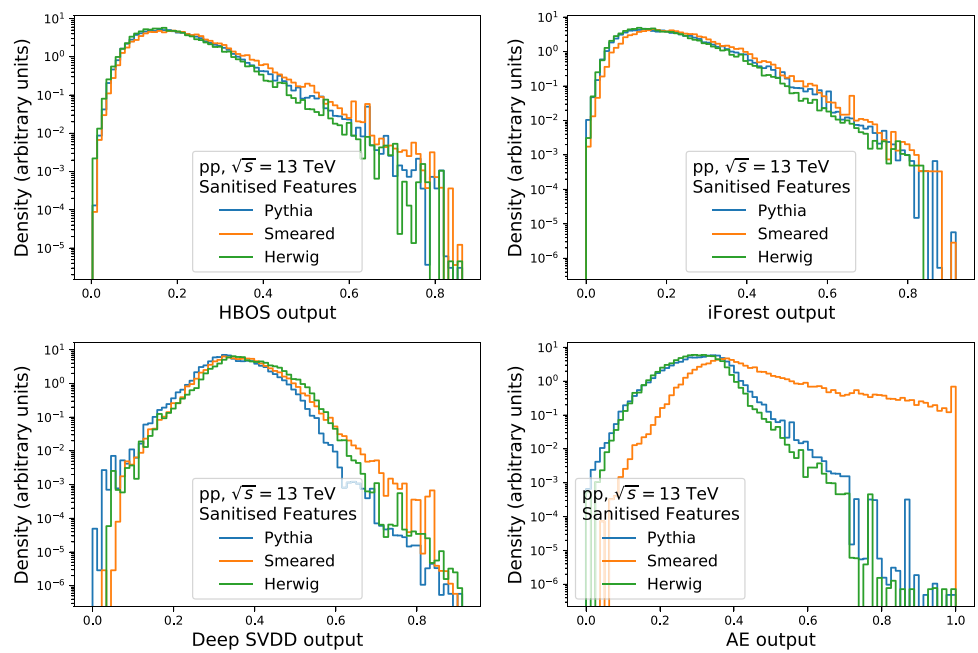
In order to study the robustness of the presented models against background mismodelling we performed two simple experiments. In the first experiment, we smeared the p_T of all objects with a Gaussian noise with standard deviation

of 0.1. For the second experiment, we switch the hadroniser from Phytia to Herwig, whilst maintaining everything else the same. The outputs of the AD models trained on the original Pythia sample with the sanitised features for both cases are presented in Fig. 8.

For the p_T smeared test, we observe that the mass of the output distribution of each AD model is shifted to the right, meaning that the new sample is deemed more anomalous than the original Pythia sample. More interestingly, we observe a considerable change in shape of the output distribution for the AE, suggesting that this method is specially sensitive to mismodelling of the p_T . On the other hand, the Deep SVDD seems more robust to this smear, although it pushes some background to the region where one would expect signal. Finally, we notice that the shallow methods, being simpler, are clearly more robust against p_T smearing. In addition, we derived the expected upper limits with the smeared p_T using the AD models trained on the original sample. We observed that the shallow methods produced values of μ compatible with those presented in Table 3 within the statistical uncertainty, while the results for deeper models got worse. The central values for μ for the Deep SVDD increased on average 2 to 3 times, while still maintaining exclusion power. For the AE, however, the limits worsen by two orders of magnitude, as one would expect from Fig. 8.

For the Herwig sample we notice, once again, that the output distributions for the shallow methods are considerably

Fig. 8 Anomaly score for the different AD methods (HBOS, iForest, Autoencoder, DeepSVDD) for the sanitised features on the test data of the original sample (Pythia), p_T smeared sample (Smeared), and Herwig sample (Herwig). The distributions are normalised to the unit area



		FCNC	HG 1.0 TeV	HG 1.2 TeV	HG 1.4 TeV	W/o HG 1.0 TeV	W/o HG 1.2 TeV	W/o HG 1.4 TeV	
Model	Supervised DNN	1	1	1	1	1	1	1	Full Features
	AE	5	0.6	2	3	2	2	2	
	Deep SVDD	2	3	9	16	13	9	9	
	HBOS	17	3	9	14	14	9	8	
	iForest	22	4	12	19	18	12	10	
	Supervised DNN	0.9	0.15	0.5	0.44	0.3	0.4	0.8	
AE	21	0.1	0.37	0.66	0.39	0.37	0.43		
Deep SVDD	2	2	5	7	6	4	4		
HBOS	17	4	12	17	15	9	8		
iForest	22	3	10	14	17	9	7		
		Signal							

Fig. 7 95% CL upper limits on μ normalised to the limit obtained for the supervised DNN model

less modified, while they differ from the expected Pythia output distributions for the deep methods. For the deep methods we observe different effects. While for the Deep SVDD the distribution moves to the right, for the AE it seems to move to the left. As before, we produced expected upper limits using the Herwig samples and compared them to the ones obtained using the original Pythia sample in Table 3. Just

like with the p_T smeared case, the shallow methods proved to be the more robust with μ values compatible with the ones derived with the Pythia sample. For the deep methods, the Deep SVDD produced limits around twice as large as for the Pythia sample, but still with smaller degradation than those obtained with the p_T smeared case. For the AE we observed a degradation as severe as with the p_T smeared case, with the limits worsening by two orders of magnitude.

These two tests suggest that the methods presented in this work can be sensitive to mismodelling, and point to the need for a thorough study of the impact of systematic uncertainties and how to mitigate the effect of such uncertainties on the sensitivity to new phenomena beyond the Standard Model. Such comprehensive study is outside the scope of the presented work.

7 Conclusions

In this work, we studied four distinct unsupervised AD algorithms, two shallow and two deep, which were trained on simulated SM events. The resulting trained models provided us with an anomaly score that was then used to perform upper bounds on seven benchmark signals covering three classes of new physics: FCNC interaction, SM gluon VLQ production, and heavy gluon VLQ production. Even though all algorithms eventually targeted events at the tails of the original SM distributions, they capture different events and are therefore learning different notions of *outlyingness*. This was clearly observed on how the Deep SVDD and the AE performed between VLQ and FCNC signals. Upper limits

on the signal strength were obtained by fitting the output distributions of each AD model using the CL_s method. We showed that the deep models outperform the shallow ones, and each deep model performed differently depending on the broader class of signals being tested. This result suggests that different AD algorithms are suitable to isolate different types of BSM physics and are complementary to each other in unsupervised generic searches for new physics.

Acknowledgements We thank Guilherme Milhano, Maria Ramos and Guilherme Guedes for the careful reading of the manuscript and for the useful discussions. We also thank Ana Peixoto and Tiago Vale for providing the MadGraph cards used for the simulation of the beyond the Standard Model samples. We acknowledge the support from FCT Portugal, Lisboa2020, Compete2020, Portugal2020 and FEDER under project PTDC/FIS-PAR/29147/2017. The computational part of this work was supported by INCD (funded by FCT and FEDER under the project 01/SAICT/2016 nr. 022153) and by the Minho Advanced Computing Center (MACC). The Titan Xp GPU card used for the training of the Deep Neural Networks developed for this project was kindly donated by the NVIDIA Corporation.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: The simulated data used in this work was obtained using publicly available software and all the required technical details to reproduce it are given in the paper. The obtained results should also be reproducible from the provided information. The authors are available to provide any information the readers might need.]

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. Funded by SCOAP³.

References

1. J. Ellis, Outstanding questions: physics beyond the Standard Model. *Philos. Trans. R. Soc. Lond. A* **370**, 818–830 (2012)
2. V.M. Abazov et al., A Quasi model independent search for new physics at large transverse momentum. *Phys. Rev. D* **64**, 012004 (2001)
3. D0 Collaboration, Quasi-model-independent search for new high p_T physics at d0. *Phys. Rev. Lett.* **86**(17), 3712–3717 (2001)
4. CDF Collaboration, Model-independent and quasi-model-independent search for new physics at cdf. *Phys. Rev. D* **78**(1), 012002 (2008)
5. CDF Collaboration, Global search for new physics with 2.0 fb⁻¹ at cdf. *Phys. Rev. D* **79**(1), 011101 (2009)
6. H1 Collaboration, A General search for new phenomena in ep scattering at HERA. *Phys. Lett. B* **602**, 14–30 (2004)
7. H1 Collaboration, A General Search for New Phenomena at HERA. *Phys. Lett. B* **674**, 257–268 (2009)
8. ATLAS Collaboration, 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**(2), 120 (2019)
9. CMS Collaboration, Music: a model unspecific search for new physics in proton–proton collisions at $\sqrt{s} = 13$ TeV (2020). [arXiv:2010.02984](https://arxiv.org/abs/2010.02984)
10. M. Rom ao Crispim, N.F. Castro, R. Pedro, T. Vale, Transferability of deep learning models in searches for new physics at colliders. *Phys. Rev. D* **101**(3), 035042 (2020)
11. J. Collins, K. Howe, B. Nachman, Anomaly detection for resonant new physics with machine learning. *Phys. Rev. Lett.* **121**(24), 241803 (2018)
12. E.M. Metodiev, B. Nachman, J. Thaler, Classification without labels: learning from mixed samples in high energy physics. *J. High Energy Phys.* **2017**(10), 174 (2017)
13. A. De Simone, T. Jacques, Guiding new physics searches with unsupervised learning. *Eur. Phys. J. C* **79**(4), 1–15 (2019)
14. R.T. D'Agno, A. Wulzer, Learning new physics from a machine. *Phys. Rev. D* **99**(1), (2019)
15. O. Cerri, T.Q. Nguyen, M. Pierini, M. Spiropulu, J.R. Vlimant, Variational autoencoders for new physics mining at the large hadron collider. *J. High Energy Phys.* **2019**(5), (2019)
16. M. Farina, Y. Nakai, D. Shih, Searching for new physics with deep autoencoders. *Phys. Rev. D* **101**(7), (2020)
17. A. Blance, M. Spannowsky, P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches. *J. High Energy Phys.* **2019**(10), (2019)
18. J. Hajer, Y. Li, T. Liu, H. Wang, Novelty detection meets collider physics. *Phys. Rev. D* **101**(7), (2020)
19. B. Nachman, D. Shih, Anomaly detection with density estimation. *Phys. Rev. D* **101**(7), (2020)
20. A. Andreassen, B. Nachman, D. Shih, Simulation assisted likelihood-free anomaly detection. *Phys. Rev. D* **101**(9), (2020)
21. J.A. Aguilar-Saavedra, J. Collins, R.K. Mishra, A generic anti-QCD jet tagger. *J. High Energy Phys.* **2017**(11), 163 (2017)
22. T. Heimel, G. Kasieczka, T. Plehn, J.M. Thompson, QCD or what. *Sci. Post Phys.* **6**(030), 1808–08979 (2019)
23. B.M. Dillon, D.A. Faroughy, J.F. Kamenik, Uncovering latent jet substructure. *Phys. Rev. D* **100**(5), 056002 (2019)
24. R.T. d'Agno, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Learning multivariate new physics (2019). [arXiv:1912.12155](https://arxiv.org/abs/1912.12155)
25. J.H. Collins, K. Howe, B. Nachman, Extending the bump hunt with machine learning (2019). [arXiv:1902.02634](https://arxiv.org/abs/1902.02634)
26. O. Amram, C.M. Suarez, Tag n' train: a technique to train improved classifiers on unlabeled data (2020). [arXiv:2002.12376](https://arxiv.org/abs/2002.12376)
27. B.M. Dillon, D.A. Faroughy, J.F. Kamenik, M. Szewc, Learning the latent structure of collider events (2020). [arXiv:2005.12319](https://arxiv.org/abs/2005.12319)
28. ATLAS Collaboration, G Aad, et al. Dijet resonance search with weak supervision using $\sqrt{s} = 13$ tev pp collisions in the atlas detector. *Phys. Rev. Lett.* **125**(13):131801 (2020). <https://doi.org/10.1103/PhysRevLett.125.131801>
29. 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 (2020). [arXiv:2005.01598](https://arxiv.org/abs/2005.01598)
30. M. Goldstein, A. Dengel, Histogram-based outlier score (hbos): a fast unsupervised anomaly detection algorithm (2012)
31. F.T. Liu, K. M. Ting, Z. Zhou, Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08* (IEEE Computer Society, 2008), pp. 413–422
32. L. Ruff et al. Deep one-class classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research* (Stockholmsmässan, Stockholm, 2018), pp. 4393–4402

33. J. Alwall et al., The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP* **07**, 079 (2014)
34. T. Sjöstrand et al., An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.* **191**, 159–177 (2015)
35. CMS Collaboration, Event generator tunes obtained from underlying event and multiparton scattering measurements. *Eur. Phys. J. C* **76**(3), 155 (2016)
36. R.D. Ball et al., Parton distributions with LHC data. *Nucl. Phys. B* **867**, 244–289 (2013)
37. J. de Favereau et al., DELPHES 3, a modular framework for fast simulation of a generic collider experiment. *JHEP* **02**, 057 (2014)
38. M. Cacciari, G.P. Salam, G. Soyez, The anti- k_t jet clustering algorithm. *JHEP* **04**, 063 (2008)
39. J.A. Aguilar-Saavedra, Identifying top partners at LHC. *JHEP* **11**, 030 (2009)
40. J.P. Araque, N.F. Castro, J. Santiago, Interpretation of Vector-like Quark Searches: heavy Gluons in Composite Higgs Models. *JHEP* **11**, 120 (2015)
41. G. Durieux, F. Maltoni, C. Zhang, Global approach to top-quark flavor-changing interactions. *Phys. Rev. D* **91**(7), 074017 (2015)
42. ATLAS Collaboration, Search for pair and single production of vectorlike quarks in final states with at least one z boson decaying into a pair of electrons or muons in pp collision data collected with the atlas detector at $\sqrt{s} = 13\text{TeV}$. *Phys Rev D* **98**, 112010 (2018)
43. CMS Collaboration, Search for vector-like quarks in events with two oppositely charged leptons and jets in proton-proton collisions at $\sqrt{s} = 13$ tev. *Eur. Phys. J. C* **79**(4), 364 (2019)
44. ATLAS collaboration, Search for flavour-changing neutral current top-quark decays $t \rightarrow qz$ in proton-proton collisions at $\sqrt{s} = 13$ tev with the atlas detector. *JHEP* **2018**(7), 176 (2018)
45. CMS Collaboration, Search for associated production of a Z boson with a single top quark and for tZ flavour-changing interactions in pp collisions at $\sqrt{s} = 8$ TeV. *JHEP* **07**, 003 (2017)
46. M. Bahr et al., Herwig++ Physics and Manual. *Eur. Phys. J. C* **58**, 639–707 (2008)
47. J. Bellm et al., Herwig 7.0/Herwig++ 3.0 release note. *Eur. Phys. J. C* **76**(4), 196 (2016)
48. K. Hornik, M. Stinchcombe, H. White et al., Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**(5), 359–366 (1989)
49. G. Cybenko, Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **2**(4), 303–314 (1989)
50. K. Hornik, Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**(2), 251–257 (1991)
51. Z. Lu, H. Pu, F. Wang, Z. Hu, L. Wang, The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pp. 6231–6239 (2017)
52. Y. Zhao, Z. Nasrullah, Z. Li, Pyod: a python toolbox for scalable outlier detection. *J. Mach. Learn. Res.* **20**(96), 1–7 (2019)
53. F. Pedregosa et al., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
54. M. Abadi et al., TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org (20150)
55. T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631 (2019)
56. J.S. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization. *Adv. Neural Inf. Process. Syst.*, 2546–2554 (2011)
57. D.P. Kingma, J.Ba, Adam: a method for stochastic optimization (2014). [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
58. I. Loshchilov, F. Hutter, Decoupled weight decay regularization (2017). [arXiv:1711.05101](https://arxiv.org/abs/1711.05101)
59. J. Shlomi, P. Battaglia, J.-R. Vlimant, Graph neural networks in particle physics (2020). [arXiv:2007.13681](https://arxiv.org/abs/2007.13681)
60. D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban, D. Whiteson, Jet flavor classification in high-energy physics with deep neural networks. *Phys. Rev. D* **94**(11), 112002 (2016)
61. A.L. Read, Presentation of search results: The CL(s) technique. *J. Phys. G* **28**, 2693–2704 (2002)
62. E. Busato, D. Calvet, T. Theveneaux-Pelzer, OpTHyLiC: an optimised tool for hybrid limits computation. *Comput. Phys. Commun.* **226**, 136–150 (2018)