

# Finding New Physics without learning about it: Anomaly Detection as a tool for Searches at Colliders

M. Crispim Romão<sup>1</sup>, N. F. Castro<sup>1,2</sup>, and R. Pedro<sup>1</sup>

<sup>1</sup>LIP, Av. Professor Gama Pinto 2, 1649-003 Lisboa, Portugal

<sup>2</sup>Departamento de Física, Escola de Ciências, Universidade do Minho, 4710-057 Braga, Portugal

June 11, 2020

## 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. 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 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 analysis. 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 [2] 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 other 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 avoid any dependence on BSM signals.

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 [3–11]. In ref. [3], an unsupervised bump hunting approach using CWoLa [12] is proposed, while in ref. [4], 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. [5] employs Neural Networks to compare the distribution of two samples and derive statistical tests to evaluate if any New Physics is present. In refs. [6–8] three different AE produce distributions of reconstruction errors to be used as anomaly scores, whereas [9] conjugates an AE with a Linear Outlier Factor. More recently, in [10, 11], novel non-ML approaches using density estimates are employed.

In this paper, we present three new unsupervised ML models for AD in the context of HEP collisions, in addition to an AE. In order to test their sensitivity to different BSM signals, the signals considered in [2] are used as benchmarks to access the performance of the proposed approach by comparing it with supervised DNN classifiers trained on the same benchmarks. In this way, we compare the performance of the AD methods to supervised DNNs. As such, we provide for the first time a comparison of 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) [13] and Isolation Forest (iForest) [14] 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) [15]. 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 peaks 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 counter part, 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 not 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.* section 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 [16] at leading order with a collision centre-of-mass energy of 13 TeV. Pythia 8.2 [17] was employed to simulate the parton shower and hadronisation, with the CMS CUETP8M1 [18] underlying event tuning and the NNPDF 2.3 [19] parton distribution functions. The detection of the collision products was accomplished with a multipurpose detector simulator, Delphes 3 [20]. 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- $\kappa_t$  algorithm [21] 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 [2]. For this reason, we studied the same BSM signals, namely the pair production of vector-like  $T$ -quarks (either produced via SM gluons [22] or BSM heavy gluons [23]) and  $tZ$  production through a flavour changing neutral current (FCNC) vertex [24]. We preselected events broadly compatible with the signal topologies commonly considered by the ATLAS and CMS experiments [25–28]: 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)<sup>1</sup>. 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

---

<sup>1</sup>The transverse plane is defined with respect to the proton colliding beams.

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  $t\bar{t}$ :  $p_T^{top} < 100$  GeV,  $p_T^{top} \in [100, 250]$  GeV,  $p_T^{top} > 250$  GeV;
- The scalar sum of the  $p_T$  of the hard-scatter outgoing particles for  $Z$ +jets:  $H_T < 250$  GeV,  $H_T \in [250, 500]$  GeV,  $H_T > 500$  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 split into the jet flavour as  $Zjj$  and  $Zbb$ , and 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. 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.

The SM cocktail used to train the AD methods is composed of the SM samples, each normalized to the expected yield after selection using the generation cross-section at leading order, computed with MADGRAPH5, and matched to a target luminosity of  $150 \text{ fb}^{-1}$ . 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:

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

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 [29–32], 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.

Table 1: Hyperparameter configuration for the Autoencoder.

Hyperparameter	Value
Encoder Layers	1
Encoder Units	128
Latent Space Dimension	16
Decoder Layers	1
Decoder Units	128
Activation	LeakyReLU
Optimiser	Adam

## 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 [33], but we changed the code as to take sample normalisation weights into account when computing the histograms. For the iForest, we based our implementation on the Scikit-Learn [34] through the `pyod` wrapper [33].

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 [34].

### 4.2 Deep methods

We implemented the AE in TensorFlow 2.1 [35]. The effect of the hyperparameters' choice was tested and we found the combination in table 1 to yield satisfactory reconstruction of the data with minimal overfitting.

Following [15], initially, we used the trained encoder of the AE as the DNN for the Deep SVDD. This procedure has seemingly two advantages. On the one hand, this would define the architecture of the Deep SVDD, which can be a practical challenge during model selection as different output dimensions will lead to distributions of distances in different

Table 2: Hyperparameter configuration for the supervised classifiers.

Hyperparameter	Value
Hidden Layers	2
Hidden Layers Units	128
Batch Normalisation	True
Activation	LeakyReLU
Optimiser	Adam

dimensional spaces that are not easily comparable. Secondly, this would provide a pre-trained encoder, which in principle would lead to faster and more stable training. However, during our tests, we found that using pre-trained encoders led to very different results, highlighting the sensitivity of this method to the initial conditions set by the computation of  $\mathbf{c}$ . Instead, we used the hyperparameter configuration for the encoder but trained the Deep SVDD from scratch. We found that this led to far more stable training of the Deep SVDD and reproducible outcomes. **In addition, we observed that Batch Normalisation also leads to trivial collapses to the centre, as the output mean of a Batch Normalisation layer is a learnable parameter that behaves as an effective bias term.**

Since we forwent utilising the trained encoder in the Deep SVDD, we let the AE train with biases, while they were not allowed in the Deep SVDD. This improved the performance of the AE without worsening overfit. Both the AD and the Deep SVDD 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  $10^{-6}$  to  $2 \times 10^{-2}$ . 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 of 0.995 at the end of each epoch, exponentially decreasing it. 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. Training was stopped if no improvement of the loss on the validation set was observed for 200 epochs for the AE, and 300 for the Deep SVDD.<sup>2</sup> The AE was trained using mini-batches of size 4096, while the Deep SVDD was trained in mini-batches of size 1024.

For both the AD 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  $[0, 1]$ .

For each of the considered benchmark signals we also trained a DNN to perform a binary classification between signal and SM background prediction. The hyperparameters used were the same for all cases and are shown in table 2. Each classifier was trained with

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

the same learning rate cycle as the one used with the AE and Deep SVDD, on mini-batches of size 4096. The training was interrupted if no improvement in the validation set was observed for 200 epochs.

### 4.3 Feature impact on reconstructions

For both feature sets, full and sanitised, the AE trained with no sign of overfitting to the training data. However, 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  $\eta$  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.

Furthermore, as seen in the  $\eta$  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  $\eta$  and  $\phi$  variables were always difficult to reconstruct in our working methodology, and even increasing the capacity of the AE did not completely solve the issue. However, systematically study how one can improve the reconstruction of features using a deterministic AE is outside the scope of this work and as such we defer such concerns to a 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 and validation samples and both feature sets. We notice that for the full features set both the deep AD models, and more importantly the Deep SVDD, manifest a more pronounced difference between the training and validation sample distributions. This is a somehow surprising result, as in both cases we used the same architecture and there are slightly less events in the sanitised features samples.

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 distributions. 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. In this figure, we notice how the shallow methods are highly related between each other for both feature sets. In



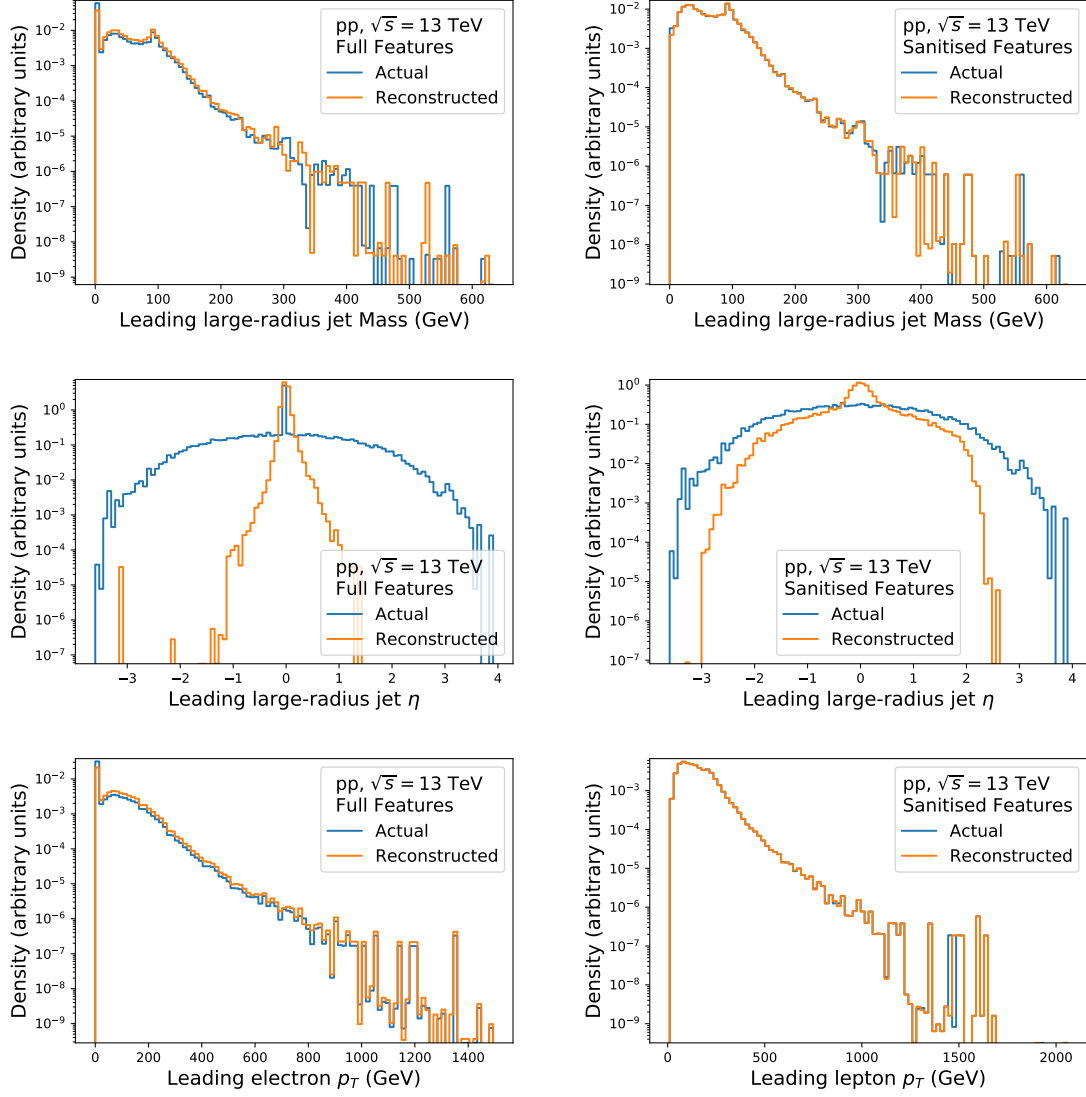


Figure 1: Distribution of some of the real input features and their reconstruction by the Autoencoder on the validation set. Left: Using all features set. Right: Using sanitised features set.

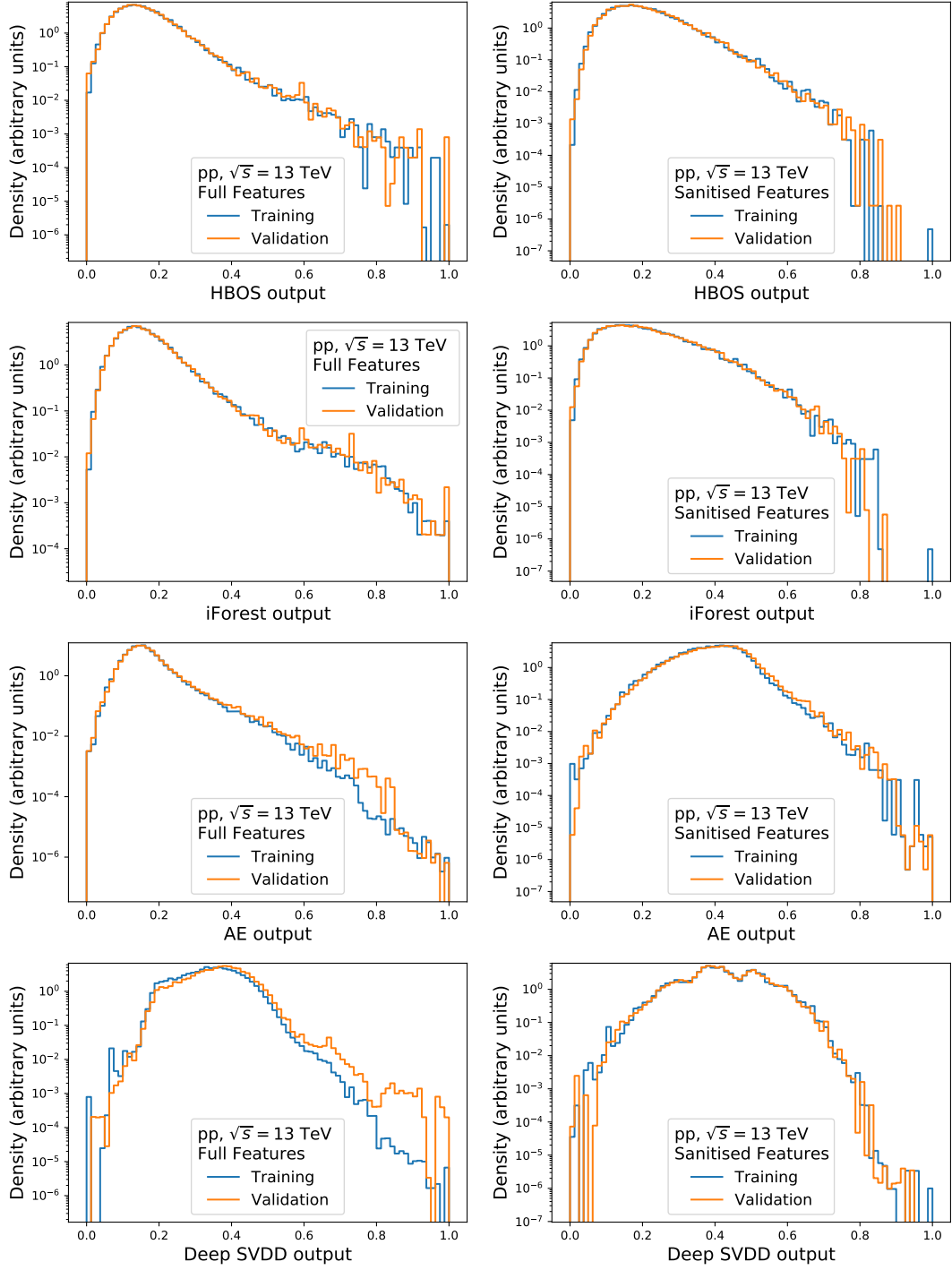


Figure 2: Anomaly score for the different AD methods (HBOS, iForest, Autoencoder, Deep SVDD) for training and validation data. The training and validation distributions are normalised to the unit area. Left: Using all features set. Right: Using sanitised features set.

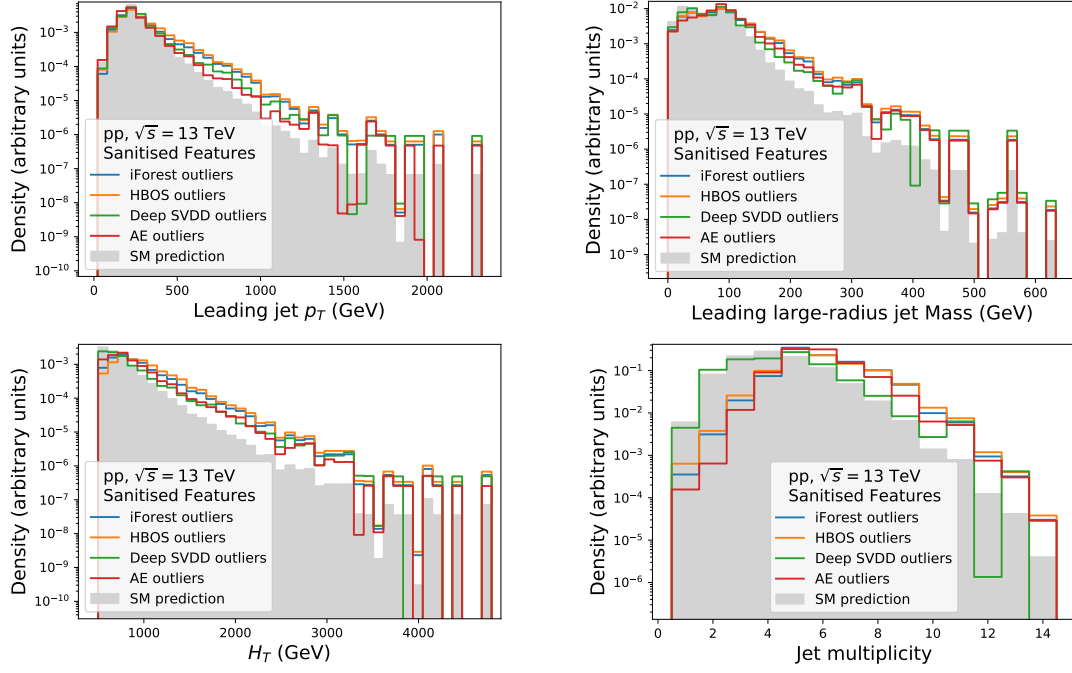


Figure 3: Distribution of some of the input features for the full validation set and for the 10% outlier quantile according to the anomaly score for the different AD methods using sanitised features. All distributions are normalised to the unit area.

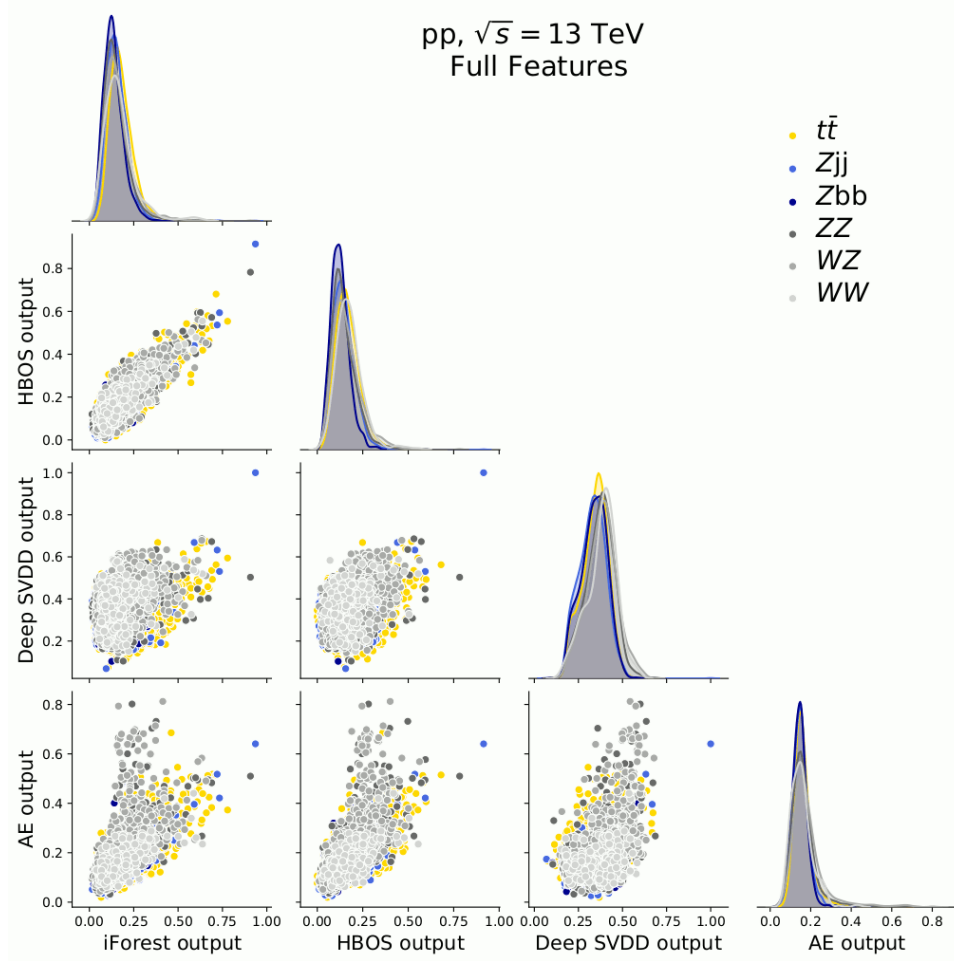


Figure 4: Two-dimensional distribution of the anomaly scores for the different AD methods per SM process -  $t\bar{t}$ ,  $Z$ +jets and diboson using all features set. Diagonal: Distribution of the anomaly score per SM process.

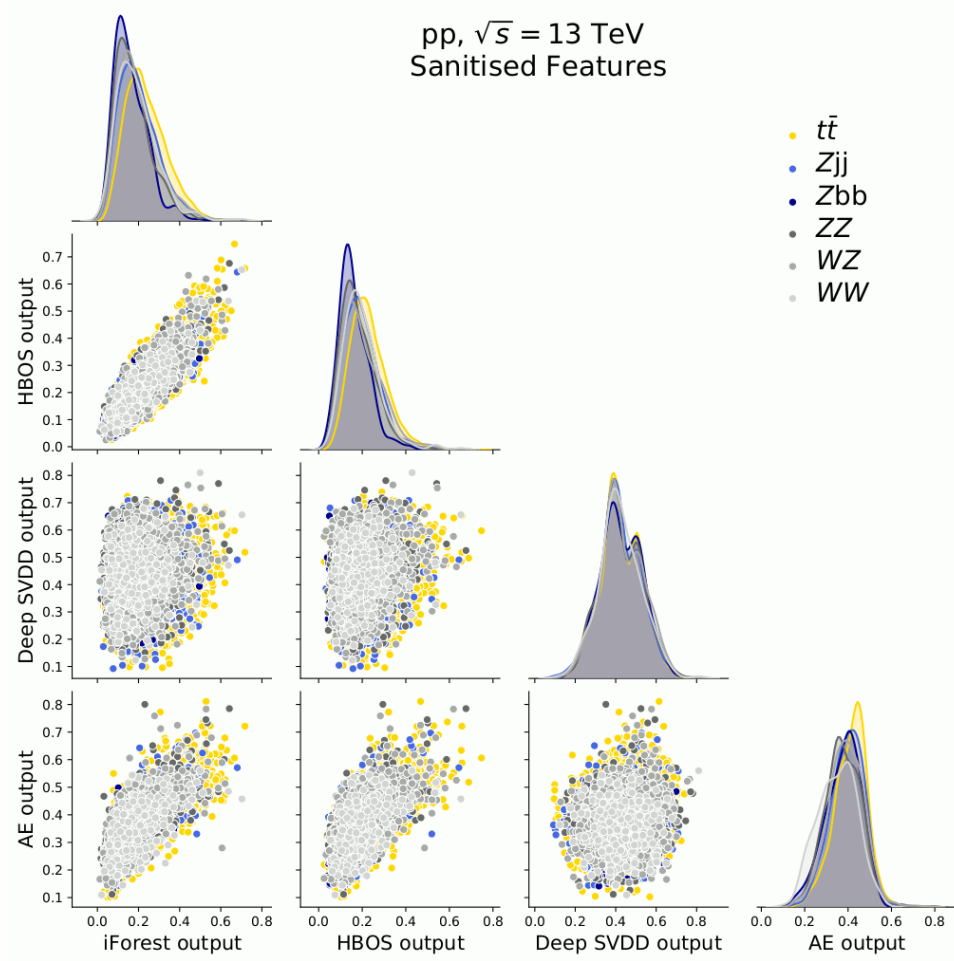


Figure 5: Two-dimensional distribution of the anomaly scores for the different AD methods per SM process -  $t\bar{t}$ ,  $Z$ +jets and diboson process -  $t\bar{t}$ ,  $Z$ +jets and diboson using sanitised features set. Diagonal: Distribution of the anomaly score per SM process.

contrast, both deep models show looser relation between their predictions 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 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  $\mu$ , defined as the ratio between the expected upper limit on the signal cross-section, normalized 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 [36], as implemented in OpTHyLiC [37]. Poissonian statistical uncertainties on each bin of the distributions were included in the limit computation, assuming an integrated luminosity of  $150 \text{ fb}^{-1}$ .

### 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  $\mu$  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 performance 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

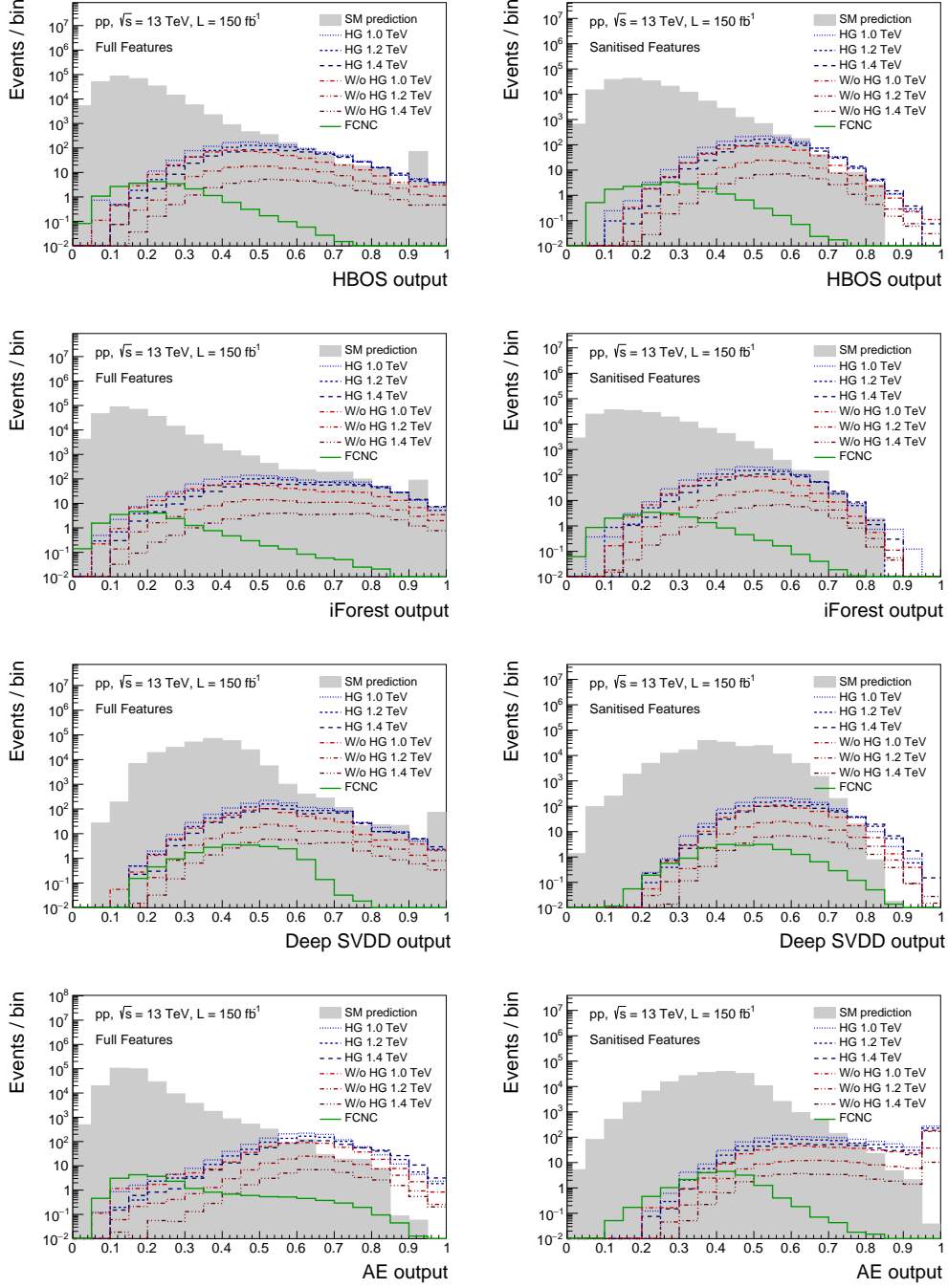


Figure 6: Distribution of the AD discriminant for the SM prediction and each signal type:  $t\bar{t}Z$  production by FCNC,  $T\bar{T}$  production via heavy gluon or without heavy gluon for  $m_T = \{1.0, 1.2, 1.4\}$  TeV. The distributions are normalised to the generation cross-section and to an integrated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

Table 3: 95% CL upper limit on the signal strength  $\mu$  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.

Model	Benchmark Signal						
	FCNC	1.0 TeV	HG 1.2 TeV	1.4 TeV	1.0 TeV	No HG 1.2 TeV	1.4 TeV
Full features							
Supervised DNN	$6^{+3}_{-2}$	$0.011^{+0.007}_{-0.004}$	$0.015^{+0.008}_{-0.005}$	$0.016^{+0.009}_{-0.005}$	$0.03^{+0.02}_{-0.01}$	$0.08^{+0.04}_{-0.03}$	$0.20^{+0.12}_{-0.07}$
Deep SVDD	$60^{+30}_{-20}$	$0.29^{+0.14}_{-0.09}$	$0.32^{+0.15}_{-0.10}$	$0.4^{+0.2}_{-0.1}$	$0.8^{+0.4}_{-0.2}$	$1.9^{+0.9}_{-0.6}$	$5^{+2}_{-1}$
AE	$30^{+10}_{-10}$	$0.06^{+0.04}_{-0.02}$	$0.06^{+0.05}_{-0.02}$	$0.06^{+0.04}_{-0.02}$	$0.12^{+0.08}_{-0.04}$	$0.4^{+0.2}_{-0.1}$	$1.0^{+0.6}_{-0.3}$
HBOS	$100^{+40}_{-30}$	$0.15^{+0.07}_{-0.05}$	$0.17^{+0.08}_{-0.05}$	$0.19^{+0.09}_{-0.06}$	$0.4^{+0.2}_{-0.1}$	$1.0^{+0.5}_{-0.3}$	$2.7^{+1.2}_{-0.9}$
iForest	$200^{+60}_{-40}$	$0.22^{+0.11}_{-0.07}$	$0.26^{+0.13}_{-0.09}$	$0.3^{+0.2}_{-0.1}$	$0.6^{+0.3}_{-0.2}$	$1.6^{+0.8}_{-0.6}$	$4^{+2}_{-1}$
Sanitised features							
Supervised DNN	$6^{+3}_{-2}$	$0.0035^{+0.0022}_{-0.0009}$	$0.006^{+0.003}_{-0.002}$	$0.009^{+0.004}_{-0.003}$	$0.014^{+0.010}_{-0.005}$	$0.07^{+0.04}_{-0.03}$	$0.15^{+0.09}_{-0.05}$
Deep SVDD	$60^{+30}_{-20}$	$0.25^{+0.13}_{-0.08}$	$0.16^{+0.08}_{-0.04}$	$0.12^{+0.05}_{-0.03}$	$0.5^{+0.2}_{-0.1}$	$1.0^{+0.5}_{-0.3}$	$2.0^{+0.8}_{-0.5}$
AE	$160^{+60}_{-50}$	$0.0099^{+0.0009}_{-0.0007}$	$0.0122^{+0.0006}_{-0.0009}$	$0.0152^{+0.0009}_{-0.0007}$	$0.0165^{+0.0007}_{-0.0011}$	$0.073^{+0.004}_{-0.004}$	$0.27^{+0.02}_{-0.02}$
HBOS	$110^{+50}_{-30}$	$0.19^{+0.11}_{-0.06}$	$0.21^{+0.12}_{-0.07}$	$0.23^{+0.14}_{-0.08}$	$0.4^{+0.2}_{-0.1}$	$1.1^{+0.7}_{-0.4}$	$2.7^{+1.7}_{-0.9}$
iForest	$140^{+60}_{-40}$	$0.3^{+0.2}_{-0.1}$	$0.4^{+0.2}_{-0.1}$	$0.4^{+0.2}_{-0.1}$	$0.8^{+0.4}_{-0.3}$	$2.2^{+1.2}_{-0.7}$	$5^{+3}_{-2}$

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 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.

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  $\mu$  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 [2], we observed similar trends when training deep neural networks on signals different from those used for the classification.



Model	Supervised DNN	1	1	1	1	1	1	Full Features
	AE	4	6	4	4	4	5	
	Deep SVDD	10	28	22	23	25	25	
	HBOS	18	14	11	12	12	13	
	iForest	25	21	17	19	18	21	
	Supervised DNN	1	0.33	0.4	0.6	0.5	1	Sanitised Features
	AE	27	0.94	0.82	0.95	0.52	0.97	
	Deep SVDD	10	24	10	7	15	13	
	HBOS	18	18	14	14	13	14	
	iForest	23	31	25	25	27	29	
		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
		Signal						

Figure 7: 95% CL upper limits on  $\mu$  normalised to the limit obtained for the supervised DNN model.

## 6 Conclusions

In this work, we developed 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.

## 7 Acknowledgments

We thank Guilherme Milhano for the careful reading of the manuscript and for the useful discussions. The authors also 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.

## References

- [1] J. Ellis. Outstanding questions: Physics beyond the Standard Model. *Phil. Trans. Roy. Soc. Lond. A*, 370:818–830, 2012.
- [2] M. Romão Crispim, N.F. Castro, R. Pedro, and T. Vale. Transferability of Deep Learning Models in Searches for New Physics at Colliders. *Phys. Rev. D*, 101(3):035042, 2020.
- [3] J. Collins, K. Howe, and B. Nachman. Anomaly detection for resonant new physics with machine learning. *Physical Review Letters*, 121(24), Dec 2018.
- [4] A. De Simone and T. Jacques. Guiding new physics searches with unsupervised learning. *The European Physical Journal C*, 79(4), Mar 2019.
- [5] R. T. D’Agnolo and A. Wulzer. Learning new physics from a machine. *Physical Review D*, 99(1), Jan 2019.

- [6] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J. R. Vlimant. Variational autoencoders for new physics mining at the large hadron collider. *Journal of High Energy Physics*, 2019(5), May 2019.
- [7] A. Blance, M. Spannowsky, and P. Waite. Adversarially-trained autoencoders for robust unsupervised new physics searches. *Journal of High Energy Physics*, 2019(10), Oct 2019.
- [8] M. Farina, Y. Nakai, and D. Shih. Searching for new physics with deep autoencoders. *Physical Review D*, 101(7), Apr 2020.
- [9] J. Hajer, Y. Li, T. Liu, and H. Wang. Novelty detection meets collider physics. *Physical Review D*, 101(7), Apr 2020.
- [10] B. Nachman and D. Shih. Anomaly detection with density estimation. *Physical Review D*, 101(7), Apr 2020.
- [11] A. Andreassen, B. Nachman, and D. Shih. Simulation assisted likelihood-free anomaly detection. *Physical Review D*, 101(9), May 2020.
- [12] E. M. Metodiev, B. Nachman, and J. Thaler. Classification without labels: learning from mixed samples in high energy physics. *Journal of High Energy Physics*, 2017(10), Oct 2017.
- [13] M. Goldstein and A. Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. 2012.
- [14] F. T. Liu, K. M. Ting, and Z. Zhou. Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08*, pages 413–422, USA, 2008. IEEE Computer Society.
- [15] 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*, pages 4393–4402, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.
- [16] 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.
- [17] T. Sjöstrand et al. An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.
- [18] The CMS Collaboration. Event generator tunes obtained from underlying event and multiparton scattering measurements. *Eur. Phys. J.*, C76(3):155, 2016.
- [19] R. D. Ball et al. Parton distributions with LHC data. *Nucl. Phys.*, B867:244–289, 2013.

- [20] J. de Favereau et al. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 02:057, 2014.
- [21] M. Cacciari, G. P. Salam, and G. Soyez. The anti- $k_t$  jet clustering algorithm. *JHEP*, 04:063, 2008.
- [22] J. A. Aguilar-Saavedra. Identifying top partners at LHC. *JHEP*, 11:030, 2009.
- [23] J. P. Araque, N. F. Castro, and J. Santiago. Interpretation of Vector-like Quark Searches: Heavy Gluons in Composite Higgs Models. *JHEP*, 11:120, 2015.
- [24] G. Durieux, F. Maltoni, and C. Zhang. Global approach to top-quark flavor-changing interactions. *Phys. Rev.*, D91(7):074017, 2015.
- [25] The 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  $\sqrt{s} = 13$  TeV. *Phys. Rev. D*, 98:112010, 2018.
- [26] The 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. *The European Physical Journal C*, 79(4):364, 2019.
- [27] The 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.
- [28] The 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.
- [29] K. Hornik, M. Stinchcombe, H. White, et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [30] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.
- [31] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.
- [32] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pages 6231–6239, 2017.
- [33] Y. Zhao, Z. Nasrullah, and Z. Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019.

- [34] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [35] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [36] A. L. Read. Presentation of search results: The CL(s) technique. *J. Phys.*, G28:2693–2704, 2002. [,11(2002)].
- [37] E. Busato, D. Calvet, and T. Theveneaux-Pelzer. OpTHyLiC: an Optimised Tool for Hybrid Limits Computation. *Comput. Phys. Commun.*, 226:136–150, 2018.