

Sensitivity to new physics phenomena in anomaly detection: A study of untunable hyperparameters

Fernando Abreu de Souza^{1,*}, Maura Barros^{1,†}, Nuno F. Castro^{1,‡}, Miguel Crispim Romão^{2,1,§},
Céu Neiva^{3,1,||} and Rute Pedro^{4,¶}

¹*LIP—Laboratório de Instrumentação e Física Experimental de Partículas, Departamento de Física,
Escola de Ciências, Universidade do Minho, 4701-057 Braga, Portugal*

²*Institute for Particle Physics Phenomenology, Durham University, Durham DH1 3LE, United Kingdom*

³*CeNTI—Centre for Nanotechnology and Advanced Materials, R. Fernando Mesquita 2785,
4760-034 Vila Nova de Famalicão, Portugal*

⁴*LIP—Laboratório de Instrumentação e Física Experimental de Partículas,
Avenida Professor Gama Pinto 2, 1649-003 Lisbon, Portugal*



(Received 28 May 2025; accepted 2 January 2026; published 4 February 2026)

The search for physics beyond the Standard Model (BSM) at collider experiments requires model-independent strategies to avoid missing possible discoveries of unexpected signals. Anomaly detection (AD) techniques offer a promising approach by identifying deviations from the Standard Model (SM) and have been extensively studied. The sensitivity of these methods to untunable hyperparameters has not been systematically compared, however. This study addresses it by investigating four semisupervised AD methods—AutoEncoders, Deep Support Vector Data Description, Histogram-based Outlier Score, and Isolation Forest—trained on simulated SM background events. In this paper, we study the sensitivity of these methods to BSM benchmark signals as a function of these untunable hyperparameters. Such a study is complemented by a proposal of a nonparametric permutation test using signal-agnostic statistics, which can provide a robust statistical assessment.

DOI: [10.1103/j2m3-wkkp](https://doi.org/10.1103/j2m3-wkkp)

I. INTRODUCTION

The Standard Model (SM) of particle physics has been remarkably successful in describing a wide range of experimental results. However, it fails to account for several observed phenomena [1], motivating ongoing searches for clues beyond the Standard Model (BSM). At the LHC, a broad program of BSM searches is conducted, typically based on specific signal models and event topologies. Ideally, however, these searches should be as general as possible to avoid missing possible discoveries of unexpected signals.

To address this, several efforts have focused on model-independent search strategies [2–9]. Nevertheless, such approaches do not guarantee sensitivity to all possible

BSM scenarios. Ensuring that search strategies remain broadly sensitive is therefore essential. One promising approach is the use of anomaly detection (AD) techniques, which aim to identify deviations from the expected distributions in datasets assumed to consist mostly of “normal” (i.e., SM-like) events. Community efforts such as the LHC Olympics 2020 Anomaly Detection Challenge [10] and the Dark Machines Collaboration [11] have explored a wide range of unsupervised and weakly supervised machine learning techniques for AD in collider data. These techniques have since been adopted by the ATLAS and CMS Collaborations for applications ranging from BSM searches [12–14] to jet substructure [15] and data quality monitoring [16]. Broadly speaking, AD approaches in HEP fall into two categories: outlier-detection methods, which aim to identify individual events that deviate from the bulk of the background, and density-based methods, which search for localized overdensities in feature space that may indicate new physics. While this work focuses on outlier detection, several efforts have successfully applied density-estimation techniques to BSM searches [17–20].

One of the early proposals of outlier-detection methods by some of the authors, presented in [21], used shallow and deep AD methods to search for new phenomena and compare the obtained limits with a dedicated supervised deep neural network (DNN). These models have been used

*Contact author: abreuocha@lip.pt

†Contact author: maura.barros@cern.ch

‡Contact author: nuno.castro@fisica.uminho.pt

§Contact author: miguel.romao@durham.ac.uk

||Contact author: ceu.neiva11@icloud.com

¶Contact author: rute.pedro@cern.ch

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

to enhance the sensitivity to new physics signals [22–24], study relations between different features [25], identify jets and anomalies within jet substructure [26–30], triggering events [31–33], and modeling and simulation [34,35].

While AutoEncoders (AEs) remain the most commonly used AD method, others have been explored. Histogram-based Outlier Scores (HBOS) have been used as a novelty detector to improve the exploration capabilities of artificial intelligence-guided BSM parameter space scans [36–38]. Deep Support Vector Data Description (Deep-SVDD) has also been adapted for collider data [39], where a supervised classifier is reformulated into an unsupervised anomaly detector. Weakly supervised anomaly detection methods have been applied to BSM searches as well [40,41]. In parallel, recent studies have investigated contrastive learning [42], topology-aware architectures [43], and symmetry-informed representations [44] to improve the structure and interpret the latent space. Despite these developments, a systematic comparison of different AD models across diverse signal scenarios and key untunable parameters remains missing.

At the same time, statistical interpretability remains a key challenge in model-independent searches. Methods that rely on anomaly scores frequently lack well-defined significance measures, making it difficult to assess discovery potential or compare across models. Recent efforts have focused on improving the statistical interpretation of AD results in the context of signal-agnostic searches [45,46].

This work presents a comparative study of four anomaly detection methods—AE, Deep-SVDD, HBOS, and Isolation Forest (iForest). While the overall objective is to distinguish anomalous signal events from background as in a classical binary classification problem, the models are trained exclusively on SM background simulation, thus single class labels, making the whole methodology semi-supervised. Their performance is investigated as a function of key untunable parameters. To complement performance metrics such as receiver operating characteristic (ROC) curves, we explore the $M\Delta$ and Cramér’s statistics, which allow to compare two sample distributions in a signal-agnostic way. Moreover, these are employed in a statistical permutation test for hypothesis testing, evaluating the sensitivity of each semisupervised method studied. This non-parametric test provides a robust way to quantify deviations from the SM without relying on signal-specific assumptions, reinforcing the generality of our methodology.

The paper is organized as follows. In Sec. II, we introduce the dataset of simulated events, comprising six different benchmark BSM signals and a SM background, all sharing a common final state with two leptons, one bottom jet, and large H_T . In Sec. III, we describe the semisupervised AD methods employed in this study, including both shallow and deep approaches. These methods are used to construct discriminants for generic new physics searches. In Sec. IV, we assess the sensitivity of each method to different BSM signals as a function of key

untunable hyperparameters, using the area under the ROC (ROC AUC) metric. In Sec. V, we introduce a statistical framework using permutation tests to evaluate the significance of observed deviations. Finally, in Sec. VI, we conclude.

II. DATASET SIMULATION DETAILS

The dataset used consists of simulated proton-proton collisions at a center-of-mass energy of 13 TeV. Events were generated at leading order using MadGraph5 [47], with Pythia8 [48] to simulate the parton shower and hadronization. The detection of the collision products was accomplished with the Delphes 3 [49] parametrized response using the default configuration, matching the CMS detector parameters. Jets and large-radius jets are reconstructed using the anti- κ_t algorithm [50] with a radius parameter of $R = 0.5$ and 0.8 , respectively.

A diverse set of BSM signals was simulated to benchmark the anomaly detection sensitivity study in different scenarios, from new resonances to new interactions only inducing small deviations from SM predictions. These signals, used solely for evaluation and not for defining or training the methods, include:

- (i) Heavy vectorlike quarks (HQ): Pair production of heavy vectorlike T quarks, with T masses $m_T = \{1.0, 1.4\}$ TeV [51];
- (ii) Flavor changing neutral current (FCNC): tZ production through a FCNC vertex [52];
- (iii) Randall-Sundrum (RS): Production of an RS radion R that decays into a pair of Z bosons and has a mass of $m_R = 4$ TeV [53];
- (iv) Two-Higgs-doublet model (2HDM): Top quark pair production in association with a heavy Higgs boson H' in the 2HDM framework, with $m_{H'} = 400$ GeV, inspired by the signal used in [54];
- (v) Left-right symmetric model (LRSB): Production of the right-handed counterpart of the W boson, W_R , decaying into a right-handed heavy neutrino, N_R , and a charged lepton, with masses $m_{W_R} = 6.5$ TeV and $m_{N_R} = 1.5$ TeV, inspired on the signal used in [55].

The chosen signals have a common final state consisting of two leptons, one bottom jet, and large H_T (> 500 GeV).¹ This motivates the definition of a broad phase space to focus our AD study. The SM processes constituting the background to these signals are Z + jets, top pair, and diboson production. To ensure a very good statistical representation, these processes were generated in samples of the kinematic space using event generation filters at parton level, as detailed in [56].

In total, 14.6 M events were simulated, split per process type as shown in Table I. The background processes

¹ H_T is the scalar sum of transverse momentum (p_T) of all reconstructed particles in the event.

TABLE I. Total number of simulated events for the background sample and for each of the signal models.

Sample	Simulated events
Background	12.5 M
HQ _{1.0TeV}	500 k
HQ _{1.4TeV}	500 k
FCNC	500 k
RS	150 k
2HDM	300 k
LRSM	150 k

composing the SM cocktail, used to train the AD methods, are normalized to the expected yield using the generation cross section for each process computed at leading order with `MadGraph5` and assuming a target luminosity of 150 fb^{-1} . The background and signal samples can be found in [57].

Each event is described by 32 features: the 4-momenta, in Cartesian basis, (p_x, p_y, p_z, m) of the two most energetic jets and of the most energetic large-R jet, (p_x, p_y, p_z) of the two most energetic leptons, (p_x, p_y) of the missing transverse momentum, the scalar sum of transverse momenta H_T , the multiplicities of muons, electrons, jets and large-R jets, a binary indicator for whether each of the two most energetic jets are b -tagged, and the large-R jet N-subjettiness variables, τ_N , with $N = [1, 5]$ [58]. Cartesian coordinates are used to remove ambiguities related to periodic variables, such as $\phi = \{0, 2\pi\}$,² which can lead to a topological obstruction in the dataset as the identification $\phi \sim \phi + 2\pi$ is not continuous or differentiable.

III. SEMISUPERVISED ANOMALY DETECTION: MODELS AND METHODOLOGY

A. Semisupervised learning methods

We use different shallow and deep learning algorithms to evaluate the sensitivity of semisupervised learning methods to new phenomena. The different algorithms were only trained on simulated SM events, i.e., the background, and their sensitivity to identify examples of potential new physics events was assessed using the previously mentioned benchmark signals. These were not used during training. Two shallow and two deep learning approaches are considered in this study, which were already discussed in [21]. The shallow methods considered, HBOS [59] and iForest [60], detect anomalies based on density or isolation principles, often requiring less computational power. The deep learning methods, Deep-SVDD [61] and AE, leverage DNN to learn more complex feature representation but, usually, require more computational power and training times.

²In collider physics experiments, the coordinate ϕ represents the azimuthal angle in the plane transverse to the colliding beam.

1. Autoencoder

An AE is a DNN trained to reconstruct data that is compressed in a bottleneck layer. It consists of an encoder, which compresses the input data into a lower-dimensional representation, the latent space bottleneck, and a decoder, which reconstructs the original data from this latent representation. The training process is done so that the reconstruction error is minimized, according to the loss function,

$$L = \mathbb{E}_{\mathbf{x} \sim \text{SM}}[||\text{AE}(\mathbf{x}, \mathcal{W}) - \mathbf{x}||^2], \quad (1)$$

where $\mathcal{W}$ are the learnable parameters of the AE, $\mathbf{x}$ is the feature vector of a SM event, and $\mathbb{E}_{\mathbf{x} \sim \text{SM}}$ denotes the expected value over SM events, which in this work is obtained through a weighted average over the training set. The reconstruction error quantifies the difference between the decoded and original data and it can be used as an anomaly score, as the AE is expected to learn the relations between different features from SM processes which might not hold to new physics processes.

2. Deep support vector data description

The Deep-SVDD also uses a DNN for the training. It maps the data into a hypersphere around its center-of-mass in a latent space, identifying anomalies as points lying far from the center. The training is done to minimize the distance of all points of the training set to this center, as expressed by the loss,

$$L = \mathbb{E}_{\mathbf{x} \sim \text{SM}}[||\text{Deep-SVDD}(\mathbf{x}, \mathcal{W}) - \mathbf{c}||^2], \quad (2)$$

where $\mathcal{W}$ are the Deep-SVDD trainable weights, $\mathbf{c}$ is the center-of-mass distribution in the output space, $\mathbf{x}$ is the feature vector of a SM event, and $\mathbb{E}_{\mathbf{x} \sim \text{SM}}$ denotes the expected value over SM events, which in this work is obtained through a weighted average over the training set. $\mathbf{c}$ is obtained by passing the whole dataset through the Deep-SVDD after initialization, but before any training takes place, and then taking the (weighted) average over all the embedding vectors, producing a center-of-mass of the training data in the target latent space. The loss minimization forces the network to find common patterns in data as to successfully embed it into the target space. The anomaly score in a Deep-SVDD is how far the event is from the center $\mathbf{c}$.

3. Isolation forest

The iForest algorithm is a tree-based AD method that isolates anomalies by successive random partitions of subsamples of the data. The partitions are represented as trees, which are grown until a maximum depth is achieved, or no further splitting is possible. As the partitions are random, the number of nodes of the tree that a data example needs to traverse until it reaches a leaf node is a

TABLE II. The untunable parameters of the different AD models.

Model	Untunable hyperparameter	Considered values
HBOS	Number of bins	[1, 31]
iForest	Number of estimators	[25, 200] in steps of 25
AE	Latent space dimension	[1, 31]
Deep-SVDD	Latent space dimension	[1, 31]

TABLE III. Hyperparameter search space for the AE.

Hyperparameter	Possible values
Number of layers	[1, 10]
Number of units	[8, 512] in steps of 8
Dropout rate	[0, 0.5] in steps of 0.1
Activation function	ReLU or LeakyReLU
Normalization	BatchNormalization, LayerNormalization, or None

measurement of how inlier it is. Conversely, outliers require fewer splits to be isolated, arriving at a leaf node after traversing fewer nodes. The discriminant is therefore the iForest score, which takes the form

$$\text{iForest}(\mathbf{x}) = 2^{-\frac{\mathbb{E}[h(\mathbf{x})]}{d}}, \quad (3)$$

where d is the average traversable path in a binary tree of the same depth, $h(\mathbf{x})$ is the number of nodes that a data example with a feature vector, $\mathbf{x}$, has to travel in a given tree, and $\mathbb{E}$ is the average over all the trees in the forest [60]. iForests were recently proposed as an anomaly detection model in the context of transient detection in microlensing data [62].

4. Histogram-based outlier score

HBOS is perhaps the most simple of all the semi-supervised models considered in this work. It computes a histogram for each feature while assuming feature independence. Unlike methods based on pairwise comparison or distance measure, HBOS estimates the probability density per histogram bin. To compute the anomaly score, the probability of each feature value falling into its respective bin is determined and a score of $\log_2(\text{HIST})$ is associated, where HIST is the density (height) of that bin. The total anomaly score is given by the sum across all features.

B. Training and tunable hyperparameters

The dataset was split equally into train, validation and test sets to guarantee statistical representativeness at each stage. Only SM events were used to train the semisupervised methods. To account for statistical fluctuations of the data and the stochastic character of the semisupervised methods, a collection of 10 independent models was trained

for each method, each training using a different subset of the training dataset and with a different model initialization, when applicable. The hyperparameters of some of the models were optimized, using performance metrics evaluated on the validation set. The final model evaluation, analysis, and statistical test were performed on the test set. All features were standardized setting their mean to zero and standard deviation to unity.

The choice of a model's hyperparameters can have a huge impact on its performance, making hyperparameter tuning a crucial step in optimizing the model's performance. For supervised tasks, the hyperparameters can be efficiently tuned by using techniques like grid search or Bayesian optimization. However, for semisupervised tasks, one might not have a sensible and meaningful metric to validate a choice of hyperparameters. In the work at hand, this is evident, as certain parameters are more difficult to fine-tune: the latent space size of AE and Deep-SVDD, the number of bins of HBOS, and the number of estimators in iForest. As our methodology is strictly semisupervised, where the AD models are only trained on SM processes, we lack a clear validation criterion on how well an AD model separates SM from BSM events. Consequently, these hyperparameters are typically selected at the beginning of the analysis, before further model optimization, based on the dataset properties and dimensionality. A key goal of this work is to provide a comparison of the sensitivity of the different methods to new phenomena for different values of these so-called untunable hyperparameters. We will focus on the values of these untunable hyperparameters presented in Table II, and the results are presented in Sec. IV.

The HBOS method was based on the PYOD Python toolkit [63], with sample normalization weights incorporated when computing the histograms. The iForest implementation was based on Scikit-learn [64]. For both shallow methods, a principal component analysis (PCA) was applied to

TABLE IV. Hyperparameter search space for the Deep-SVDD.

Hyperparameter	Possible values
Number of layers	[1, 10]
Number of units	[8, 512] in steps of 8
Dropout rate	0
Activation function	ReLU or LeakyReLU
β_1	[0.85, 0.95]
β_2	[0.99, 0.9999]
Weight decay	0.0, 1×10^{-9} , 1×10^{-8} , 1×10^{-7} , 1×10^{-6} , 1×10^{-5} , 1×10^{-4} or 1×10^{-3}

remove the linear correlations between the features.³ The PCA was implemented using Scikit-learn. The values considered are represented in Table II.

The deep models were implemented with TensorFlow 2.11 [65]. Several were trained to scan the AE and the Deep-SVDD latent space dimension across the range defined in Table II. For each value of the untunable hyperparameters, OPTUNA [66] was used to optimize the tunable hyperparameters, within the search space detailed in Tables III and IV.

The AE was optimized over 1000 epochs, with early stopping applied using patience of 50 epochs, and 100 OPTUNA trials to maximize the quality of the reconstruction as measured by the coefficient of determination,⁴ R^2 , as the training process can be seen as a supervised regression problem. Therefore, for a fixed dimension of the latent space, there is a meaningful metric to validate and compare different values of the remaining hyperparameters in Table III. Furthermore, to prevent features with higher nominal values from dominating the loss function, the AE was trained to reconstruct the standardized version of the features.

For the Deep-SVDD, for each value of the latent space dimension we optimized the rest of the hyperparameters in Table IV over 10,000 epochs, with early stopping applied using patience of 100 epochs, and 100 trials to minimize the average distance to the center. The best hyperparameter combination resulting from the optimization loop for each value of the untunable parameter was selected and used to train 10 models, each on an independent subsample of the training set. This allows us to estimate the statistical uncertainty as the standard deviation of the results obtained from the model collection.

³This is especially important for HBOS as it assumes feature independence. For the iForest, we found in the early stages of this work that it helps boost discrimination. The reason for this is that the partitions are along the feature axis, which can lead to an excessive coverage of the feature space if there are strong linear correlations between features. By removing the linear dependency of the features, the iForest will partition along the principal components of the dataset, increasing the sensitivity outside of these directions, which in turn makes the iForest more sensitive to new phenomena where correlations between features might differ.

⁴With $R^2 = 1 - \mathbb{E}[||x - \hat{x}||^2] / \mathbb{E}[||x - \mathbb{E}[x]||^2]$, where x are the feature vectors and $\hat{x}$ their reconstruction, with the expected value being computed using the weights. Intuitively, the coefficient of determination quantifies the ratio of variance that is being described by the regression.

IV. SENSITIVITY DEPENDENCY ON THE UNTUNABLE HYPERPARAMETER

In this section, we evaluate how the choice of untunable hyperparameters, presented in Table II, influences the sensitivity of each semisupervised method to new physics phenomena using benchmark signals. To quantify sensitivity, we employ the ROC AUC, a widely used and interpretable metric in similar analyses. However, it is important to note that ROC AUC requires true labels, making it a supervised discrimination metric. In Sec. VA, we will explore alternative metrics suitable for signal-agnostic studies and their application in developing a semisupervised statistical test for detecting new physics phenomena in a sample.

Figure 1 illustrates the dependence of the ROC AUC on untunable hyperparameters for each semisupervised method. The results indicate that deep methods (AE and Deep-SVDD) exhibit some variation in background-signal separation concerning latent space dimensionality. For instance, the AE initially shows a decrease in sensitivity for HQ signals, stabilizing at a latent space dimension of approximately $\gtrsim 15$. For the FCNC (2HDM) signal, sensitivity remains nearly stable, with a slight increase (decrease) observed beyond a latent space dimension of $\gtrsim 20$. Despite these variations, the AE sensitivity changes are generally within statistical uncertainty across most signals. Likewise, the Deep-SVDD does not exhibit a monotonic relationship between ROC AUC and latent space dimensionality, maintaining a largely flat profile. This suggests that, for both deep learning models, latent space dimensionality has a minimal impact on discriminative power across a broad range of signals. Similarly, shallow methods show negligible variation, as evidenced by the stable sensitivity of HBOS and iForest to their untunable hyperparameters. The dashed lines in Fig. 1 indicate, for each signal, the value of the ROC AUC for the feature that maximizes it, i.e., the feature with the highest discrimination power. This “best feature” serves as a reference for evaluating the AD methods; it shows the discrimination achievable with the simplest approach—an univariate analysis over a single feature—and thus approximately provides a discrimination benchmark among the individual features considered. In all the cases, we see that the semisupervised methods are at best as good as the most discriminating feature for each signal, and at worst only slightly worse than the best feature. This observation is

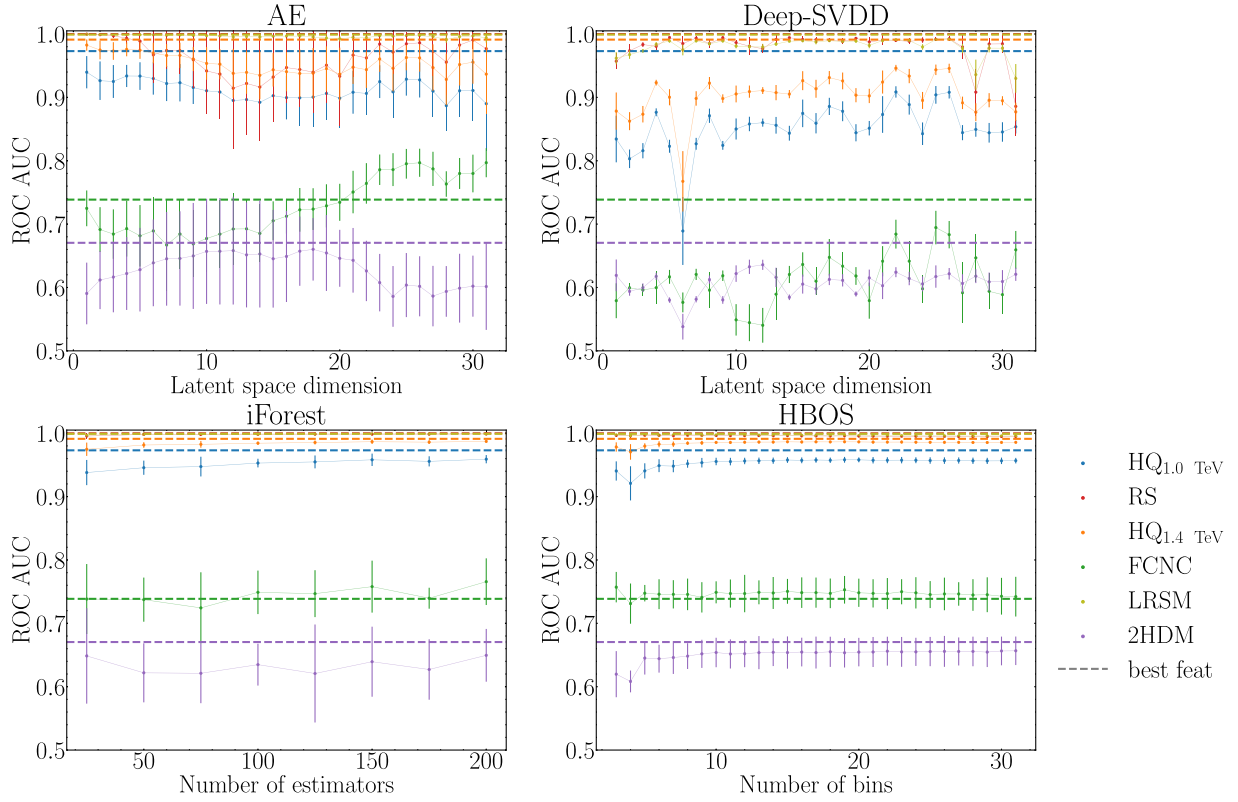


FIG. 1. ROC AUC for each signal and semisupervised method as a function of its untunable hyperparameter. The ROC AUC for the most discriminating feature for each signal is also displayed as a dashed line. For every point, a collection of 10 independent models was trained (using different training subsamples and seeds when applicable), and the uncertainty bars represent the standard deviation across these models.

reassuring, as it means that the semisupervised methods “conserve” the separation provided by the best feature without knowing which one it is as this identification is signal dependent.

The results for AE imply that its discrimination power is largely independent of latent space dimensionality. However, this finding challenges the prevailing intuition that improved background reconstruction should enhance sensitivity to new

physics phenomena. Given that the AE is trained to model nonlinear relationships in SM events and detect deviations as potential indicators of new physics phenomena, one might expect a stronger correlation between reconstruction quality and sensitivity. Conversely, an AE that poorly reconstructs SM events should, at least in principle, be less effective in detecting anomalies. To investigate this, we now analyse the interplay between sensitivity and AE reconstruction quality.

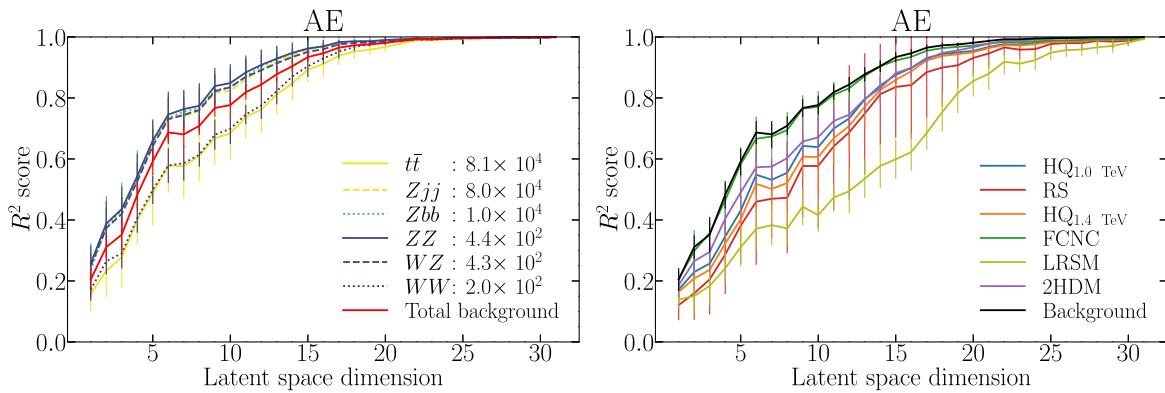


FIG. 2. R^2 score as a function of the latent space dimensionality of the AE. (Left) AE R^2 reconstruction score for each background process across different latent space dimensions. (Right) AE R^2 reconstruction score for each signal across different latent space dimensions. The uncertainty bars are the standard deviation obtained from the collection of 10 independent models.

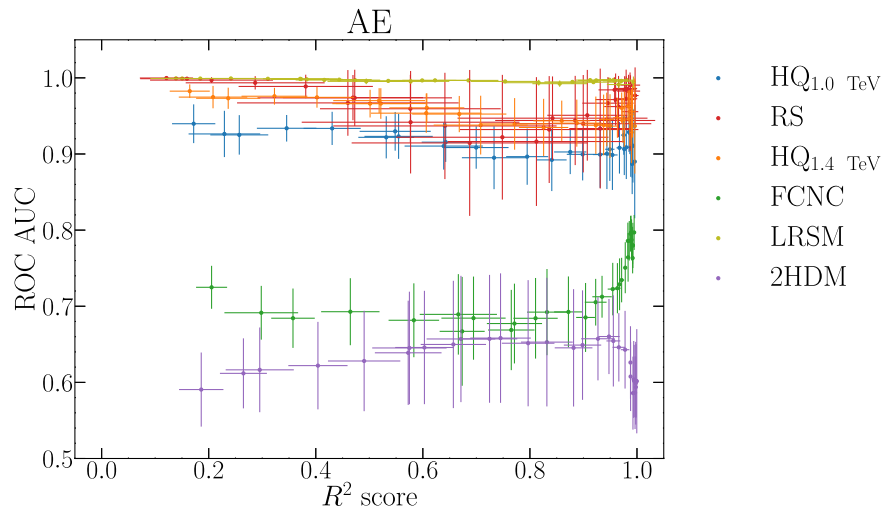


FIG. 3. ROC AUC as a function of AE reconstruction score R^2 for each signal. The uncertainty bars are the standard deviation obtained from the collection of 10 independent models.

In principle, the latent space dimension affects the reconstruction quality of both background and signal data. Ideally, AE models should capture the underlying physics of the SM while distinctly representing new BSM signals. However, as shown in Fig. 2, a higher latent space dimension does not necessarily lead to improved background-signal separation. The reconstruction quality of the AE, quantified by the R^2 score, increases monotonically with latent space dimension for both background (left pane) and signal (right pane). This is of course expected, as a higher latent space dimensionality allows for a more complete description of the training data as less information will be lost by the encoding bottleneck. However, if higher reconstruction quality correlated with improved sensitivity, we would expect a corresponding increase in ROC AUC with latent space dimension, yet this contradicts the findings in Fig. 1. Nevertheless, we observe that, for a fixed latent space dimension, the relative reconstruction quality of different signals aligns with their ranking in terms of AE sensitivity in Fig. 1.

Figure 3 further supports this observation by plotting ROC AUC against the reconstructed R^2 score for each signal. The results reveal no monotonic correlation between ROC AUC and R^2 . More critically, we find that discrimination power does not converge to 0.5 (random classification) as $R^2 \rightarrow 0$, nor does it approach 1.0 (perfect classification) as $R^2 \rightarrow 1$.

These observations lead to a key conclusion: while for a fixed latent space dimension, discrimination power depends on the relative difference in reconstruction error between signal and background, it does not depend on the absolute reconstruction quality of the background, which increases monotonically with the latent space dimension. This suggests that background-signal separation remains largely invariant across different latent space dimensions.

Additionally, one might assume that different background processes could have optimal AE latent space dimensions, resulting in noticeable differences in their reconstruction quality. However, the left pane of Fig. 2 does not support this hypothesis for the training dataset used in this analysis. Instead, reconstruction quality differences across background processes remain small up to a latent space dimension of approximately 20, beyond which the AE achieves $R^2 \sim 1$, effectively learning how to reconstruct all background processes. This lack of variability is likely due to the stringent selection criteria applied in this analysis, which include requiring two leptons, one bottom jet, and a large H_T . These cuts yield highly similar events, diminishing potential differences that might otherwise lead to distinct optimal embedding dimensions.⁵

Finally, while this discussion has focused on AE, where R^2 provides an interpretable semisupervised metric, the same intuition applies to Deep-SVDD. Since the Deep-SVDD latent space serves as an embedding space similar to the AE latent space, its sensitivity to new physics phenomena should also be largely invariant across different latent space dimensions, which is demonstrated in Fig. 1.

V. SIGNAL AGNOSTIC STATISTICAL TEST WITH PERMUTATIONS

In this section, we introduce a statistical test based on sample permutations and signal-agnostic test statistics to evaluate the sensitivity of semisupervised models to new phenomena. In particular, we study their dependence on untunable hyperparameters.

⁵See, for example, [43] on a discussion on the topology and geometry of the AE latent spaces for different physical processes.

A. Test statistics

In Sec. IV, we assessed the discrimination dependence of the considered semisupervised methods to untunable hyperparameters using one of the most common metrics in machine learning classification problems, the ROC AUC. While widely used and easy to interpret, the ROC AUC is inherently a supervised metric that requires true labels, making it unsuitable as a test statistic for unexpected new phenomena. Additionally, the ROC AUC has several limitations, such as insensitivity to symmetric but distinct distributions,⁶ an implicit assumption that the class of interest has a higher median than the negative class, and a tendency to saturate quickly, especially for signals with large values in the tails of the background distribution.

To address these issues, we consider two alternative label-free test statistics; $M\Delta$ and Cramér's test (Cr), which we define below.

$M\Delta$: $M\Delta$ is inspired by the Kolmogorov-Smirnov test for comparing two sample distributions [67,68]. Given two samples, A and B , $M\Delta$ for a univariate quantity x is defined as

$$M\Delta_{A,B} = \max_x |\text{eCDF}_A(x) - \text{eCDF}_B(x)|, \quad (4)$$

where $\text{eCDF}_i(x)$, $i = A, B$ are the empirical cumulative distribution functions (eCDFs) of x in samples A, B . $M\Delta$ thus represents the maximum difference between two eCDFs.⁷ This statistic has several advantages; it is insensitive to the ordering of distributions (i.e., $M\Delta_{AB} = M\Delta_{BA}$), it can detect differences in symmetric eCDFs, and it produces values in the range $[0, 1]$, making it easy to interpret and compare.

Cr: Cr was introduced in [69], inspired by a one-sample goodness-of-fit test statistic proposed by Harald Cramér [70]. It is defined as the integral of the squared difference between two empirical cumulative distribution functions. Given two samples, A and B , Cr for a univariate quantity x is given by

$$Cr_{A,B} = \int_{-\infty}^{\infty} |\text{eCDF}_A(x) - \text{eCDF}_B(x)|^2 dx. \quad (5)$$

Since Cr is unbounded for unbounded x , it is more challenging to interpret and compare across distributions.⁸

⁶Since all momentum and missing energy components are symmetric around zero, the ROC AUC can be misleading. Even when background and signal distributions differ in shape, their symmetry can cause the ROC AUC to remain close to 0.5, failing to reflect the true discriminative power. This issue is particularly relevant when using the best feature ROC AUC as a baseline for comparison.

⁷In this work, the eCDFs are always computed with simulation weights to reflect the true physical distributions.

⁸One possible way to improve interpretability is to scale the two-sample distributions into a fixed interval. However, even with this adjustment, we found that the resulting values were not suitable for direct comparison across different discriminant samples, as the domain size varies significantly between different features and semisupervised methods.

Nevertheless, Cr is highly sensitive to differences in symmetric distributions. Its primary advantage, compared to the other test statistics considered, is its sensitivity to the tails of the discriminant distribution; a property that proves particularly useful when performing a two-sample permutation test, as discussed in the next section. We also notice that, although similar to the Cramér-von Mises and the 2-Wasserstein distances, the Cr test used in this work is distinct and has the benefit of being easier to compute than these other two.

B. Statistical test with permutations

The test statistics presented earlier quantify the discrimination power of semisupervised methods at background signal separation in the same way that ROC AUC can be used in supervised methods. However, a hypothesis testing methodology is necessary to claim evidence for new phenomena in observed data. Usually, this is derived from the ratio of likelihoods under the background-only and signal hypotheses. The question arises on how to perform such a hypothesis test in a signal-agnostic way when the signal hypothesis is unknown *a priori*.

In this work, we employ the two-sample test statistics presented above with a permutation test to achieve a hypothesis testing methodology for AD high-energy physics analyses. Permutation tests are used to assess whether two independent samples, X and Y , obtained through some underlying distributions, $X \sim p_X$ and $Y \sim p_Y$, are in fact generated by the same distribution. Therefore, the null hypothesis for the two-sample statistical test is that the two samples come from the same underlying distribution,

$$H_0: p_X = p_Y. \quad (6)$$

The choice of the permutation test is motivated by it being a nonparametric two-sample statistical test that can be combined with any suitable test statistic without the need for prior knowledge about the shape of distributions. This is exactly what is needed in signal-agnostic analyses where there is no knowledge about the nature of the new physics signal. The proposed methodology for the two-sample permutation test works is listed below, which is also presented in a diagrammatic form in Fig. 4:

- (1) Consider two samples, control $\mathcal{C}$ and analysis $\mathcal{A}$, where $\mathcal{C}$ can be thought of as the background simulation produced by the experiment and $\mathcal{A}$ as the recorded experimental data;
- (2) Compute the observed test: $t_{\text{obs}} = t(\mathcal{C}, \mathcal{A})$;
- (3) Randomly pool $\mathcal{C} \cup \mathcal{A}$ into two samples: $\mathcal{P}_1^i, \mathcal{P}_2^i$, where $i = 1, \dots, N$ is the permutation label;
- (4) Calculate $t(\mathcal{P}_1^i, \mathcal{P}_2^i)$, where t is the chosen test statistic;
- (5) Repeat steps 3 and 4 for N permutations to derive the distribution of the test statistic under the null hypothesis, $P(t|H_0)$;

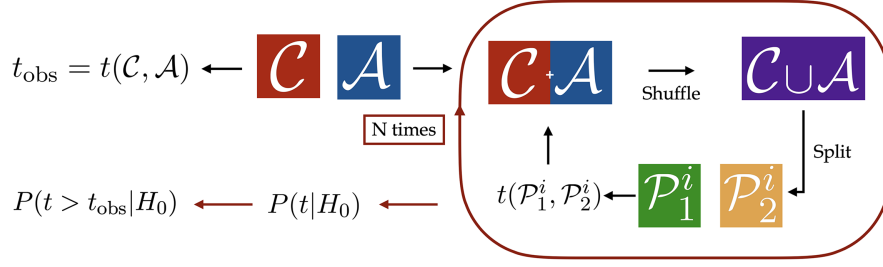
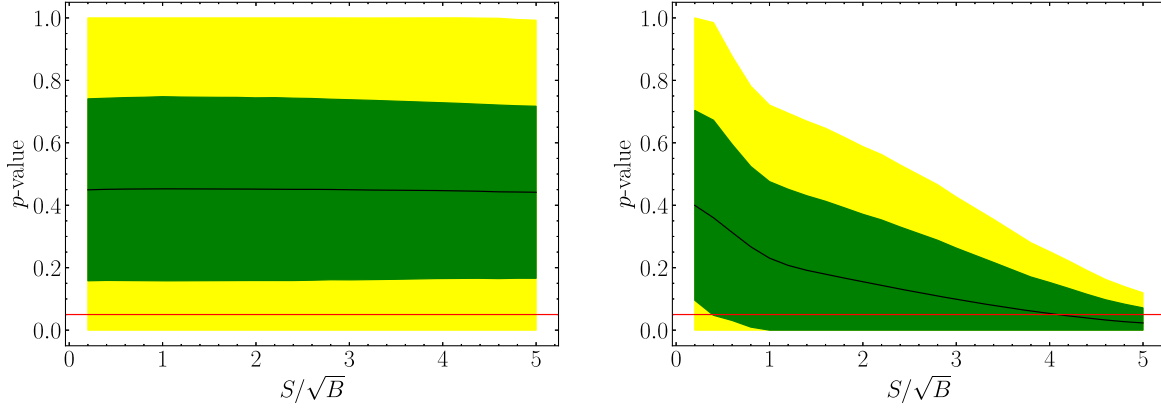


FIG. 4. Statistical test with permutations.


 FIG. 5. p -value as a function of $S/\sqrt{B}$ for the 2HDM signal (left) and the RS signal (right) using Cr as the statistic in the statistical test for the collection of AE with the latent space dimension of 31. The uncertainty bands correspond to 1 and 2 standard deviations of the p -values obtained from the collection of 10 independent models, while the red line indicates the p -value at 95% confidence level.

- (6) The one-sided p -value of the test is given as

$$p_{\text{value}} = P(t > t_{\text{obs}} | H_0) = \frac{\#(t > t_{\text{obs}})}{N}^9;$$

- (7) For $p_{\text{value}} \leq 0.05$, there is evidence to reject the null hypothesis, H_0 , with 95% confidence.

It is important to note that, strictly speaking, the null hypothesis is whether both samples, $\mathcal{C}$ and $\mathcal{A}$, are originated by the same distribution. However, if $\mathcal{C}$ is prepared in such way that it only includes SM events, $\mathcal{C} \sim p_{\text{SM}}$, the null hypothesis becomes equivalent to the statement that data is also solely composed of SM events, i.e., $\mathcal{A} \sim p_{\text{Data}} = p_{\text{SM}}$, and therefore we obtain a semisupervised statistical test on the presence of new phenomena in the data.¹⁰

As a proof of concept, we start by producing the control and analysis sets from an equal split of the test set. The control sample is composed of background events normalized to the expected yield B at 150 fb^{-1} , while the analysis

sample is composed of a mixture of background and signal events normalized to $S + B$. In this splitting step, we found that stratifying according to the SM processes instead of purely random splitting was crucial to guarantee a stable statistical test. This can be somewhat surprising as our dataset has very good statistics across all SM processes. However, random splitting does not preserve the correct relative yields of SM processes, which have significantly different cross sections. This imbalance has a profound impact on the outcome of the statistical test.

To assess how the Cr permutation test responds to signal proportion in terms of resulting sensitivity, we vary $S/\sqrt{B}$ of the analysis set. The resulting analysis distribution is then normalized to the total control yield, to base the test on the difference between the distributions' shape and complement a counting experiment. In Fig. 5 we show the p -values for rejecting the null hypothesis using AE discriminants in the presence of 2HDM and RS signals. Results show no discrimination for 2HDM even with signal contamination up to $S/\sqrt{B} \geq 5$. For the RS signal, the monotonic decrease in p -value with the increase of the signal significance is pronounced, contributing to validating our methodology. In this case, we would find evidence to reject H_0 , and therefore evidence supporting the presence of new phenomena, for $S/\sqrt{B} \gtrsim 4$. Furthermore, the

⁹Since we are considering statistical distances for our test statistics, the evidence against the null hypothesis will be embodied by large values of the test statistic, therefore the p -value is computed on the right-hand side of the $P(t | H_0)$.

¹⁰Of course, one could point out that any severe enough mismodeling can lead to a rejection of the null hypothesis, i.e., the SM. However, this is a common challenge in searches for new phenomena and not of this methodology in itself.

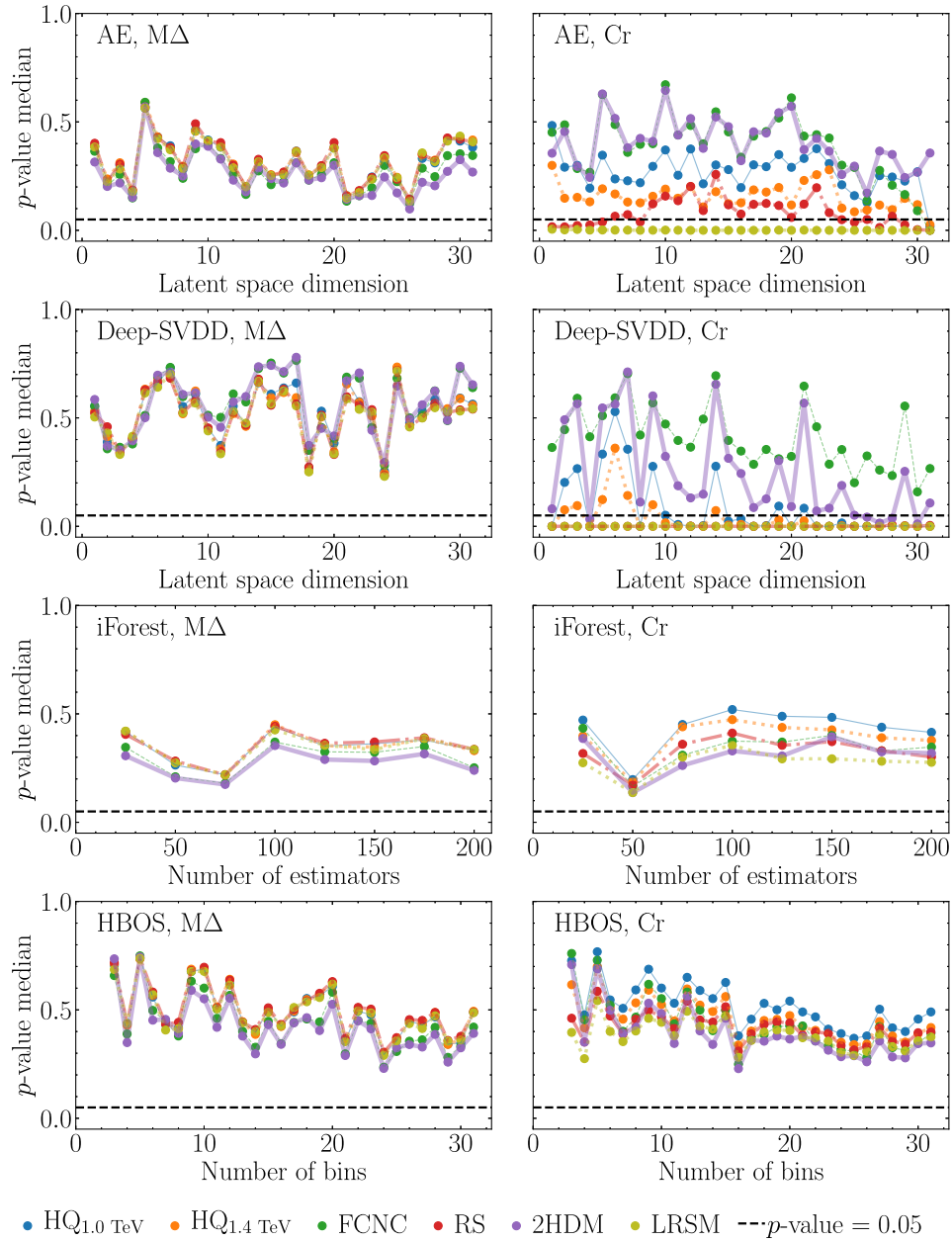


FIG. 6. Median p -value for each signal over the collection of 10 independent models using $M\Delta$ (left) and Cr (right) as a function of the untunable hyperparameter. The signal in the analysis sample is normalized such that $S/\sqrt{B} = 2$.

spread of the p -values over the collection of AE models is quite broad when there is no sensitivity, as t_{obs} is expected to be distributed similarly to $P(t|H_0)$ under H_0 , leading to naturally large fluctuations.

C. Results

In the previous section, we showed the potential of rejecting the SM-only hypothesis in the presence of new phenomena using the permutation test on a semisupervised discriminant. We now extend the proof of concept to all

signals under study, and use it as a metric for testing the dependency of semisupervised models introduced in Sec. III to the possible untunable hyperparameters, as presented in Sec. IV with the ROC AUC. The two test statistics introduced in Sec. V are employed, and the signals in the analysis sample are normalized such that $S/\sqrt{B} = 2$. This choice poses an interesting scenario wherein the signal significance of a counting experiment is noticeable for a signal expectation but otherwise interpreted as an insignificant fluctuation of the expected background. Therefore, we study whether the statistical test could provide evidence

for new phenomena in a regime where it could be easily overlooked, enabling a semisupervised search to obtain evidence to reject the SM-only hypothesis.

For convenience, we summarize here the evaluation procedure as follows:

- (1) For each semisupervised learning method (AE, Deep-SVDD, iForest, and HBOS), we trained a collection of 10 independent models on background-only (SM) data while varying the corresponding untunable hyperparameters;
- (2) For each trained semisupervised method in the previous step, we computed the anomaly score distributions for each benchmark signal;
- (3) For each anomaly score distribution, we apply the statistical test with permutations using two test statistics, $M\Delta$ and Cr, with the benchmarked signals injected with a normalization satisfying $S/\sqrt{B} = 2$.

The full results are presented in Fig. 6 for each signal type using the $M\Delta$ and Cr test statistics, showing the median p -value over the collection of 10 trained models as a function of the untunable hyperparameter.

The first observation standing out is the importance of the choice of test statistics, as $M\Delta$ fails to produce evidence for the presence of new phenomena in $\mathcal{A}$ (median p -value is always > 0.05) across all signals and AD methods irrespective of the value of the untunable hyperparameter. On the other hand, Cr shows to be more promising at capturing evidence for the rejection of the SM-only hypothesis, with sensitivity varying across signals, AD methods, and their configurations.

Secondly, the Cr-based results show significantly lower p -values for the deep learning methods. This is somehow surprising as the shallow methods consistently outperformed the Deep-SVDD in terms of ROC AUC, cf. Fig. 1. This reinforces the fact that the ROC AUC can be a misleading metric to measure semisupervised sensitivity, as argued earlier. The increased sensitivity of deep learning can be associated with the long-tailed nature of their outputs. Both shallow methods produce a bounded-value anomaly score: there is a finite number of trees in the iForest and a finite number of bins contributing to the HBOS score. Oppositely, the deep learning models produce unbounded outputs, as an event can be arbitrarily badly reconstructed by the AE or land arbitrarily far from the center-of-mass of the distribution in the Deep-SVDD latent space. The combination of these long-tailed distributions with the Cr test statistics, which integrates the differences of the eCDFs across the discriminant domain, explains the increased sensitivity observed. The same effect does not happen when using $M\Delta$, since this test statistic only captures the maximum difference of the eCDFs.

The third important feature is the impact that the untunable hyperparameters have on the sensitivity. Overall, the results show no clear dependence on this hyperparameter choice, with no notable trend or an

outstanding optimal value. However, the variation with the untunable hyperparameter is still significant indicating that its choice should not be overlooked. The observations suggest that an approach to the problem could be to aggregate results from AD models with different untunable hyperparameters. While in this work we took a visual approach by assessing the p -value variation over the different hyperparameter values in plots such as Fig. 6, other approaches that quantitatively aggregate different p -values, such as those presented in [45], could provide an alternative way of assessing sensitivity across all values of the hyperparameters.

Finally, the results show some sensitivity complementarity between the AE and the Deep-SVDD, with the Deep-SVDD presenting in general a higher sensitivity to both HQ signals. This further corroborates the hypothesis already pointed out in [21] that different AD methods capture different notions of anomaly and do not always agree on what anomalous events are.

VI. CONCLUSIONS

In this work, we investigate the influence of untunable hyperparameters of AD methods on the sensitivity of signal-agnostic searches for new physics phenomena for a selection of AD algorithms. A set of shallow (iForest and HBOS) and deep (AE and Deep-SVDD) semisupervised AD methods are trained, and their performance is assessed with a set of BSM signals.

Using the ROC AUC as a metric for signal sensitivity, we conclude that the AE discriminative power is not directly related to the background reconstruction quality. Moreover, no significant variation is observed with the choice of untunable hyperparameters for most BSM signals and across all AD methods studied. Under this metric, the shallow methods outperform the Deep-SVDD for all signals and across all hyperparameter values, while being competitive to the AE, within the uncertainty pertaining to the stochastic nature of the trainings. For all cases except the Deep-SVDD, the ROC AUC was mostly compatible with the one obtained by the most discriminating feature. Since the selection of the best feature is signal-dependent, this indicates that the AD methods probed preserve the discrimination of the best feature in a signal-independent way.

Two test statistics ($M\Delta$ and Cr) were used to assess the sensitivity towards new phenomena where a significance threshold of p -value ≤ 0.05 is associated with evidence for new phenomena, or, more precisely, for the rejection of the SM-only null hypothesis. For a signal injection of $S/\sqrt{B} = 2$, the deep methods tested achieve sensitivity for most signals when using Cr as the test statistic. For $M\Delta$, we see no sensitivity for any method under study and any of the signals, highlighting the importance of choosing the right test statistic for the discriminant domain. The limitations of the ROC AUC as a measure of signal-agnostic

discrimination provide the underlying reasons for the little relation observed between ROC AUC and the p -values. However, the same qualitative conclusion is extracted from p -values and ROC AUCs considering an unobserved correlation between the signal sensitivity and untunable hyperparameters. Furthermore, the p -value metric unveils a significant sensitivity variation across the tested values. We also notice sensitivity complementarity between the AE and the Deep-SVDD, where a combination of the two would, in principle, improve performance. This reinforces the “no free lunch theorem” [71], stating that no single machine learning model can outperform all others for every task, which in the context of AD means that discriminants do not equally agree on what an anomaly is.

Our work also suggests new research avenues to develop the methodology proposed. Namely, the choice of the test statistic is crucial to produce a semisupervised statistical test of the SM-only hypothesis that is sensitive to new phenomena. Additionally, given how different semisupervised AD models produce strikingly different sensitivities, it is likely that other AD discriminants would improve upon the ones presented here. Furthermore, while we only considered a simple permutation test, we leave to future work a more detailed comparison between the permutation test and other approaches to two-sample statistical testing.

In summary, in this work, we have shown that the untunable hyperparameter choice in the semisupervised

AD models studied highly impacts the sensitivity of searches for new phenomena, suggesting that strategies built on the aggregation of models should be explored to tackle the problem. Finally, we introduced a method to produce semisupervised statistical tests on the SM-only null hypothesis, paving the way for purely semisupervised searches for new phenomena, complementary to supervised searches in experiments.

ACKNOWLEDGMENTS

This work is supported in part by the Portuguese Fundação para a Ciência e Tecnologia (FCT) under No. LA/P/0016/2020 and No. UID/50007/2023. F.A.d.S. is supported by FCT under the research Grant No. UI/BD/153105/2022. M.B. is supported by FCT under the research Grant No. SFRH/BD/151006/2021. M.C.R. is supported by the STFC under Grant No. ST/T001011/1. R. P. is supported by FCT under Grant No. 2021.01023.CEECIND. We would like to acknowledge Centro Nacional de Computação Avançada (CNCA) under No. 2023.10635.CPCA.A1.

DATA AVAILABILITY

The data that support the findings of this article are openly available [57].

-
- [1] John Ellis, Outstanding questions: Physics beyond the standard model, *Phil. Trans. R. Soc. A* **370**, 818 (2012).
 - [2] DØ Collaboration, Quasi-model-independent search for new physics at large transverse momentum, *Phys. Rev. D* **64**, 012004 (2001).
 - [3] DØ Collaboration, Quasi-model-independent search for new high p_T physics at DØ, *Phys. Rev. Lett.* **86**, 3712 (2001).
 - [4] CDF Collaboration, Model-independent and quasi-model-independent search for new physics at CDF, *Phys. Rev. D* **78**, 012002 (2008).
 - [5] CDF Collaboration, Global search for new physics with 2.0/fb at CDF, *Phys. Rev. D* **79**, 011101 (2009).
 - [6] H1 Collaboration, A general search for new phenomena in ep scattering at HERA, *Phys. Lett. B* **602**, 14 (2004).
 - [7] H1 Collaboration, A general search for new phenomena at HERA, *Phys. Lett. B* **674**, 257 (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**, 120 (2019).
 - [9] CMS Collaboration, MUSiC: A model-unspecific search for new physics in proton–proton collisions at $\sqrt{s} = 13$ TeV, *Eur. Phys. J. C* **81**, 629 (2021).
 - [10] G. Kasieczka *et al.*, The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics, *Rep. Prog. Phys.* **84**, 124201 (2021).
 - [11] T. Aarrestad *et al.*, The dark machines anomaly score challenge: Benchmark data and model independent event classification for the large hadron collider, *SciPost Phys.* **12**, 043 (2022).
 - [12] ATLAS Collaboration, Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector, *Phys. Rev. D* **108**, 052009 (2023).
 - [13] ATLAS Collaboration, Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $\sqrt{s} = 13$ TeV with the ATLAS detector, *Phys. Rev. Lett.* **132**, 081801 (2024).
 - [14] ATLAS Collaboration, Weakly supervised anomaly detection for resonant new physics in the dijet final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, *Phys. Rev. D* **112**, 072009 (2025).
 - [15] CMS Collaboration, Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton

- collisions at $\sqrt{s} = 13$ TeV, *Rep. Prog. Phys.* **88**, 067802 (2025).
- [16] CMS Collaboration, Autoencoder-based anomaly detection system for online data quality monitoring of the CMS electromagnetic calorimeter, *Comput. Software Big Sci.* **8**, 11 (2024).
- [17] R. Verheyen, Event generation and density estimation with surjective normalizing flows, *SciPost Phys.* **13**, 047 (2022).
- [18] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, Classifying anomalies through outer density estimation, *Phys. Rev. D* **106**, 055006 (2022).
- [19] G. Stein, U. Seljak, and B. Dai, Unsupervised in-distribution anomaly detection of new physics through conditional density estimation, in *Proceedings of the 34th Conference on Neural Information Processing Systems* (2020), [arXiv:2012.11638](#).
- [20] B. Nachman and D. Shih, Anomaly detection with density estimation, *Phys. Rev. D* **101**, 075042 (2020).
- [21] M. Crispim Romão, N. F. Castro, and R. Pedro, Finding new physics without learning about it: Anomaly detection as a tool for searches at colliders, *Eur. Phys. J. C* **81**, 27 (2021).
- [22] M. Farina, Y. Nakai, and D. Shih, Searching for new physics with deep autoencoders, *Phys. Rev. D* **101**, 075021 (2020).
- [23] A. Banda, C. K. Khosa, and V. Sanz, Strengthening anomaly awareness, [arXiv:2504.11520](#).
- [24] S. V. Chekanov, W. Islam, R. Zhang, and N. Luongo, ADFilter—A web tool for new physics searches with autoencoder-based anomaly detection using deep unsupervised neural networks, *Information* **16**, 258 (2025).
- [25] M. Crispim Romão, J. G. Milhano, and M. van Leeuwen, Jet substructure observables for jet quenching in quark gluon plasma: A machine learning driven analysis, *SciPost Phys.* **16**, 015 (2024).
- [26] T. Heimel, G. Kasieczka, T. Plehn, and J. Thompson, QCD or what?, *SciPost Phys.* **6**, 030 (2019).
- [27] T. S. Roy and A. H. Vijay, A robust anomaly finder based on autoencoders, [arXiv:1903.02032](#).
- [28] T. Finke, M. Krämer, A. Morandini, A. Mück, and I. Oleksiyuk, Autoencoders for unsupervised anomaly detection in high energy physics, *J. High Energy Phys.* **06** (2021) 161.
- [29] L. Apolinário, N. F. Castro, M. Crispim Romão, J. G. Milhano, R. Pedro, and F. C. R. Peres, Deep Learning for the classification of quenched jets, *J. High Energy Phys.* **11** (2021) 219.
- [30] F. Canelli, A. de Cosa, L. L. Pottier, J. Niedziela, K. Pedro, and M. Pierini, Autoencoders for semivisible jet detection, *J. High Energy Phys.* **02** (2022) 074.
- [31] E. Govorkova, E. Puljak, T. Aarrestad, T. James, V. Loncar, M. Pierini, A. A. Pol, N. Ghielmetti, M. Graczyk, S. Summers, J. Ngadiuba, T. Q. Nguyen, J. Duarte, and Z. Wu, Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider, *Nat. Mach. Intell.* **4**, 154 (2022).
- [32] B. M. Dillon, L. Favaro, T. Plehn, P. Sorrenson, and M. Krämer, A normalized autoencoder for LHC triggers, *SciPost Phys. Core* **6**, 074 (2023).
- [33] B. Bhattacharjee, P. Konar, V. S. Ngairangbam, and P. Solanki, LLPNet: Graph autoencoder for triggering light long-lived particles at HL-LHC, [arXiv:2308.13611](#).
- [34] P. Ilten, T. Menzo, A. Youssef, and J. Zupan, Modeling hadronization using machine learning, *SciPost Phys.* **14**, 027 (2023).
- [35] M. Touranakou, N. Chernyavskaya, J. Duarte, D. Gunopulos, R. Kansal, B. Orzari, M. Pierini, T. Tomei, and J. Vlimant, Particle-based fast jet simulation at the LHC with variational autoencoders, *Mach. Learn.* **3**, 035003 (2022).
- [36] J. Crispim Romão and M. Crispim Romão, Combining evolutionary strategies and novelty detection to go beyond the alignment limit of the Z3 3HDM, *Phys. Rev. D* **109**, 095040 (2024).
- [37] F. A. de Souza, N. F. Castro, M. Crispim Romão, and W. Porod, Exploring scotogenic parameter spaces and mapping uncharted dark matter phenomenology with multi-objective search algorithms, *J. High Energy Phys.* **10** (2025) 116.
- [38] F. A. de Souza, R. Boto, M. Crispim Romão, P. N. Figueiredo, J. Crispim Romão, and J. P. Silva, Unearthing large pseudoscalar Yukawa couplings with Machine Learning, *J. High Energy Phys.* **07** (2025) 268.
- [39] S. Caron, J. E. García Navarro, M. M. Llacer, P. Moskvitina, M. Rovers, A. R. Jiménez, R. Ruiz de Austri, and Z. Zhang, Universal anomaly detection at the LHC: Transforming optimal classifiers and the DDD method, *Eur. Phys. J. C* **85**, 415 (2025).
- [40] C. L. Cheng, G. Singh, and B. Nachman, Incorporating physical priors into weakly-supervised anomaly detection, *Phys. Rev. Lett.* **135**, 021801 (2025).
- [41] L. Brennan, T. A. Vami, O. Amram, S. Sekhar, Y. Takahashi, L. Moureaux, M. Sommerhalder, P. Maksimovic, and T. Cai, Weakly supervised anomaly detection with event-level variables, *Phys. Rev. D* **112**, 055040 (2025).
- [42] K. Metzger, L. Xu, M. Sodini, T. K. Arrestad, K. Govorkova, G. Grosso, and P. Harris, Anomaly preserving contrastive neural embeddings for end-to-end model-independent searches at the LHC, *Phys. Rev. D* **112**, 072011 (2025).
- [43] V. S. Ngairangbam, B. Rozwoda, K. Sakurai, and M. Spannowsky, Enhancing anomaly detection with topology-aware autoencoders, *Mach. Learn. Sci. Tech.* **6**, 045051 (2025).
- [44] V. Sanz, Learning symmetries in datasets, [arXiv:2504.05174](#).
- [45] G. Grosso and M. Letizia, Multiple testing for signal-agnostic searches of new physics with machine learning, *Eur. Phys. J. C* **85**, 4 (2025).
- [46] G. Grosso, M. Letizia, M. Pierini, and A. Wulzer, Goodness of fit by Neyman-Pearson testing, *SciPost Phys.* **16**, 123 (2024).
- [47] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, *J. High Energy Phys.* **07** (2014) 079.
- [48] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to Pythia 8.2, *Comput. Phys. Commun.* **191**, 159 (2015).
- [49] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (Delphes 3 collabo-

- ration), A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [50] M. Cacciari, G.P. Salam, and G. Soyez, The anti- k_t jet clustering algorithm, *J. High Energy Phys.* **04** (2008) 063.
- [51] ATLAS Collaboration, Combination of the searches for pair-produced vectorlike partners of the third-generation quarks at $\sqrt{s} = 13$ TeV with the ATLAS detector, *Phys. Rev. Lett.* **121**, 211801 (2018).
- [52] G. Durieux, F. Maltoni, and C. Zhang, Global approach to top-quark flavor-changing interactions, *Phys. Rev. D* **91**, 074017 (2015).
- [53] L. Randall and R. Sundrum, Large mass hierarchy from a small extra dimension, *Phys. Rev. Lett.* **83**, 3370 (1999).
- [54] ATLAS Collaboration, Search for heavy Higgs bosons with flavour-violating couplings in multi-lepton plus b-jets final states in pp collisions at 13 TeV with the ATLAS detector, *J. High Energy Phys.* **12** (2023) 081.
- [55] ATLAS Collaboration, Search for heavy Majorana or Dirac neutrinos and right-handed W gauge bosons in final states with charged leptons and jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, *Eur. Phys. J. C* **83**, 1164 (2023).
- [56] M. Crispim Romao, N. F. Castro, and R. Pedro, Simulated pp collisions at 13 TeV with 2 leptons + 1 b jet final state and selected benchmark Beyond the Standard Model signals, [10.5281/zenodo.5126747](https://arxiv.org/abs/10.5281/zenodo.5126747) (2021).
- [57] F. Abreu de Souza, M. Barros, N. F. Castro, M. Crispim Romão, and R. Pedro, Simulated pp collisions at 13 TeV for Standard Model background and beyond Standard Model signals with 2 leptons, 1-bjet and high HT, [10.5281/zenodo.15423467](https://arxiv.org/abs/10.5281/zenodo.15423467) (2025).
- [58] Jesse Thaler and Ken Van Tilburg, Identifying boosted objects with N-subjettiness, *J. High Energy Phys.* **03** (2011) 015.
- [59] M. Goldstein and A. Dengel, Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm (2012), <https://www.goldiges.de/publications/HBOS-KI-2012.pdf>.
- [60] F. T. Liu, K. M. Ting, and Z. Zhou, Isolation forest, in *2008 Eighth IEEE International Conference on Data Mining* (2008), pp. 413–422, <https://ieeexplore.ieee.org/document/4781136>.
- [61] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, Deep one-class classification, in *Proceedings of the 35th International Conference on Machine Learning*, edited by Jennifer Dy and Andreas Krause, Volume 80 of Proceedings of Machine Learning Research (PMLR, 2018), pp. 4393–4402, <https://proceedings.mlr.press/v80/ruff18a.html>.
- [62] M. Crispim Romao, D. Croon, and D. Godines, Anomaly detection to identify transients in LSST time series data, *Mon. Not. R. Astron. Soc.* **351**, 357 (2025).
- [63] Y. Zhao, Z. Nasrullah, and Z. Li, PYOD: A Python toolbox for scalable outlier detection, *J. Mach. Learn. Res.* **20**, 1 (2019), [arXiv:1901.01588](https://arxiv.org/abs/1901.01588).
- [64] F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12**, 2825 (2011), [arXiv:1201.0490](https://arxiv.org/abs/1201.0490).
- [65] M. Abadi *et al.*, TensorFlow: Large-scale machine learning on heterogeneous distributed systems, [arXiv:1603.04467](https://arxiv.org/abs/1603.04467).
- [66] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, OPTUNA: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19* (Association for Computing Machinery, New York, NY, USA, 2019), pp. 2623–2631, [10.1145/3292500.3330701](https://doi.org/10.1145/3292500.3330701).
- [67] A. N. Kolmogorov-Smirnov, Sulla determinazione empirica di una legge di distribuzione (1933), Vol. **4**, pp. 83–91, <https://api.semanticscholar.org/CorpusID:222427298>.
- [68] N. V. Smirnov, Table for Estimating the Goodness of Fit of Empirical Distributions, *Ann. Math. Stat.* **19**, 279 (1948).
- [69] L. Baringhaus and C. Franz, On a new multivariate two-sample test, *J. Multivariate Anal.* **88**, 190 (2004).
- [70] H. Cramér, On the composition of elementary errors, *Scand. Actuarial J.* **1928**, 13 (1928).
- [71] D. Wolpert, The lack of *a priori* distinctions between learning algorithms, *Neural Comput.* **8**, 1341 (1996).