Published for SISSA by Springer
Received: November 25, 2020
Revised: July 30, 2021
Accepted: July 31, 2021
Published: September 6, 2021

Combining outlier analysis algorithms to identify new physics at the LHC

Melissa van Beekveld,[a,b,c] **Sascha Caron,**[b,c] **Luc Hendriks,**[b] **Paul Jackson,**[d] **Adam Leinweber,**[d] **Sydney Otten,**[b,e] **Riley Patrick,**[d] **Roberto Ruiz de Austri,**[f] **Marco Santoni**[d] **and Martin White**[d]

[a] *Rudolf Peierls Centre for Theoretical Physics, Clarendon Laboratory, 20 Parks Road, Oxford OX1 3PU, U.K.*
[b] *High Energy Physics, IMAPP, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, Netherlands*
[c] *Nikhef, Science Park 105, 1098 XG Amsterdam, Netherlands*
[d] *ARC Centre of Excellence for Dark Matter Particle Physics, University of Adelaide, North Terrace, SA 5005, Australia*
[e] *Gravitation and Astroparticle Physics Amsterdam (GRAPPA), Science Park 904, 1098 XH Amsterdam, Netherlands*
[f] *Instituto de Física Corpuscular, IFIC-UV/CSIC, Valencia, Spain*

E-mail: mcbeekveld@gmail.com, scaron@nikhef.nl, luc.hendriks@gmail.com, p.jackson@adelaide.edu.au, adam.leinweber@student.adelaide.edu.au, sydney.otten@rwth-aachen.de, riley.patrick@adelaide.edu.au, rruiz@ific.uv.es, marco.01.santoni@gmail.com, martin.white@adelaide.edu.au

Abstract: The lack of evidence for new physics at the Large Hadron Collider so far has prompted the development of model-independent search techniques. In this study, we compare the anomaly scores of a variety of anomaly detection techniques: an isolation forest, a Gaussian mixture model, a static autoencoder, and a β-variational autoencoder (VAE), where we define the reconstruction loss of the latter as a weighted combination of regression and classification terms. We apply these algorithms to the 4-vectors of simulated LHC data, but also investigate the performance when the non-VAE algorithms are applied to the latent space variables created by the VAE. In addition, we assess the performance when the anomaly scores of these algorithms are combined in various ways. Using supersymmetric benchmark points, we find that the logical AND combination of the anomaly scores yielded from algorithms trained in the latent space of the VAE is the most effective discriminator of all methods tested.

Keywords: Phenomenological Models, Supersymmetry Phenomenology

ArXiv ePrint: 2010.07940

https://doi.org/10.1007/JHEP09(2021)024

JHEP09(2021)024

Contents

1 Introduction

The 2012 discovery of the Higgs boson [1, 2] puts us at an intriguing milestone in the history of particle physics: the next discovery in collider physics is likely to arise from a theory whose precise details will not be known in advance. Whilst a plethora of well-motivated extensions to the Standard Model (SM) of particle physics have been proposed, there is no unambiguous prediction of the phenomenology that we can expect to observe at the Large Hadron Collider (LHC). Many searches for theories of interest such as supersymmetry are currently heavily optimised on simplified models, an approach which has been shown by phenomenological studies to leave a significant bulk of models unprobed. To give an example, this is to be expected in the case of simplified models of supersymmetry, since analyses are only optimised on vanishingly-thin hyperplanes in the total space of sparticle mass differences and branching ratios. Models not on those planes are not guaranteed to be reached by the analyses, and indeed frequently remain unexcluded [3–7].

In recent years, a number of techniques have been developed for performing signal model-independent searches with collider data. The D0 collaboration at the Tevatron developed an unsupervised, multivariate signal detection algorithm named SLEUTH [8–11], the H1 Collaboration [12, 13] at HERA used a 1-dimensional signal detection algorithm, and the CDF Collaboration [14, 15] at the Tevatron also developed a 1-dimensional signal-detection algorithm. The BUMPHUNTER algorithm has been operated in a similar vein

JHEP09(2021)024

at the LHC and the Tevatron [16]. Other model-independent LHC searches have been performed by the ATLAS and CMS collaborations, or with their publicly available data [17–21]. A promising approach that uses neural networks to compare observations with a set of reference events (via the definition of a suitable test statistic) was presented in [22, 23]. The use of autoencoders in jet substructure applications has been outlined in [24, 25], whilst the use of autoencoders for general LHC searches was also explored in [26], with a particular emphasis on proposing new anomaly score metrics that can increase the likelihood that anomalous data will be identified. Unsupervised anomaly detection techniques were considered in [27], which compared the isolation forest algorithm to histogram-based outlier detection, an autoencoder and the Deep Support Vector Data Description algorithm [28], focusing on vector-like T-quark production and tZ production through a flavour-changing neutral-current vertex. Our method differs from this by applying anomaly detection techniques within the latent space of a variational autoencoder and by combining anomaly scores from different algorithms in various ways. Variational autoencoders have been used to detect anomalies using high level features as inputs in [29], and using adversarial neural networks in [30]. An early use of Gaussian mixture models to find anomalies in the context of Higgs boson searches is detailed in CDF [31]. Further model-independent or weakly-supervised techniques for LHC discovery applications have been proposed in [27, 32–37].

In this paper we perform a systematic comparison of a variety of anomaly detection techniques for LHC searches, including novel ideas such as training within the latent space of a variational autoencoder, and combining algorithms in various ways. Throughout, our aim is to define an anomaly-score variable that can be used on an *event-by-event* basis to classify a given event as a potential signal of new physics (yielding score ≈ 1) or stemming from the SM (yielding a score ≈ 0). The underlying assumption is that some signals of new physics are different in the kinematics (defined by 4-vectors) and object types from SM events. We first determine the extent to which a variety of unsupervised anomaly-detection techniques can detect anomalies when used on final state multiplicities and 4-vector information at the LHC. The considered techniques are an isolation forest (IF), a Gaussian mixture model (GMM), a static autoencoder (AE), and a variational autoencoder (VAE) architecture where we define the reconstruction loss as a weighted combination of regression and classification terms. We use a selection of supersymmetric benchmark models to assess the performance on the hypothetical signals, without optimising the hyperparameters of our unsupervised techniques. We then assess various "combination" techniques that explore ways to combine the results of each anomaly detection method, such as taking a logical OR or AND of the different anomaly scores, or taking the product/average of the scores. We investigate how the performance changes when the non-VAE techniques are run on the latent space variables of the VAE, supplemented by the VAE reconstruction loss. We perform the same combination techniques as in the earlier case, and compare the performance of combinations of algorithms trained on 4-vectors to algorithms trained on latent space representations. Taking one particular combination of the loss parameters, we find strong performance on both gluino pair-production and stop pair-production scenarios.

We emphasise that, although we test our techniques on supersymmetric scenarios, we do not optimise the performance on details of the kinematics of the final state particles

JHEP09(2021)024

in those models. It is thus to be expected that our techniques are easily generalisable to other new physics scenarios, and immediately applicable to new physics processes that populate similar final states to our particular benchmark choices. The scope of our paper is to provide a proof of principle that our techniques are viable for some realistic new physics scenarios. In this regard, we also note that some of the supersymmetric scenarios are already comfortably excluded by current dedicated ATLAS and CMS searches. It is to be expected that an dedicated approach would always outperform an unsupervised approach. However, one important question in our paper is "do unsupervised techniques have the power to pick up obviously discoverable new-physics scenarios?" If this were not true for past examples, it would not inspire confidence for future searches that use an unsupervised approach. It is also worth noting that this is not a detailed analysis within the tails of distributions at low event counts. Instead we emphasise that this is a systematic comparison of novel unsupervised anomaly detection techniques using realistic new physics scenarios.

This paper is structured as follows. In section 2, we provide the details of our SM data and the supersymmetric benchmark models. In section 3 we define and describe the VAE, isolation forest, Gaussian mixture model and static autoencoder algorithms. Results are presented in section 4 and a short summary is presented in section 4.3, before we present our conclusions in section 5.

2 Selection of processes and generation of events

Although our proposed techniques are designed to provide a model-independent approach to LHC searches, it is useful to test their performance on particular models of interest. To this end, we use a variety of supersymmetric benchmark models to test our techniques, using the supersymmetric signal and SM background processes from the dataset published and described in ref. [38].

The events are generated at leading order for a 13 TeV LHC centre of mass energy using `Madgraph v2.6.3` [39] with the NNPDF PDF LO set [40] in the five-flavour scheme. `Madgraph` was interfaced with `Pythia 8.2` [41] for the parton shower. Events are produced in the hard process with up to 2 jets, indicated by $(+2j)$ in table 2. To avoid double-counting by the parton shower, we employ MLM matching [42]. Detector effects are included using `Delphes 3` [43] with a modified version of the ATLAS detector card. Tau leptons are produced in decays of particles (i.e. $W^{\pm}$ decays), which then further decay into lighter leptons but not tagged in the final objects. The generation chain that was used, including the Madgraph commands that generated the signals, can be downloaded from [44]. A summary of the supersymmetric benchmark models and SM backgrounds used can be seen in table 1 and 2 respectively. The total number of events that were generated and the required number of events at 36 fb^{-1} are indicated in table 1 and 2 as well. As can be seen there, we have a weight that is larger than 1 for some of the SM processes that have a large cross section. We expect that the tails of the distributions are miss-modeled, however this is largely inconsequential as our algorithms train on the large bulk of the SM events and will not be affected by a few outliers. We apply a further minimal pre-selection of:

JHEP09(2021)024

Process	Process ID	σ (pb)	N_{tot} ($N_{10\text{fb}^{-1}}$)
$pp \to \tilde{g}\tilde{g}$ (1 TeV)	Gluino 01	0.20	50,000 (7,246)
$pp \to \tilde{g}\tilde{g}$ (1.2 TeV)	Gluino 02	0.05	50,000 (1,829)
$pp \to \tilde{g}\tilde{g}$ (1.4 TeV)	Gluino 03	0.014	50,000 (518)
$pp \to \tilde{g}\tilde{g}$ (1.6 TeV)	Gluino 04	0.004	50,000 (158)
$pp \to \tilde{g}\tilde{g}$ (1.8 TeV)	Gluino 05	0.001	50,000 (51)
$pp \to \tilde{g}\tilde{g}$ (2 TeV)	Gluino 06	4.8×10^{-4}	50,000 (18)
$pp \to \tilde{g}\tilde{g}$ (2.2 TeV)	Gluino 07	1.7×10^{-4}	50,000 (7)
$pp \to \tilde{t}_1\tilde{t}_1$ (220 GeV), $m_{\tilde{\chi}_1^0} = 20$ GeV	Stop 01	26.7	1,000,000 (9,629,78)
$pp \to \tilde{t}_1\tilde{t}_1$ (300 GeV), $m_{\tilde{\chi}_1^0} = 100$ GeV	Stop 02	5.7	1,000,000 (205,117)
$pp \to \tilde{t}_1\tilde{t}_1$ (400 GeV), $m_{\tilde{\chi}_1^0} = 100$ GeV	Stop 03	1.25	500,000 (44,938)
$pp \to \tilde{t}_1\tilde{t}_1$ (800 GeV), $m_{\tilde{\chi}_1^0} = 100$ GeV	Stop 04	0.02	500,000 (723)

Table 1. Summary of the supersymmetric benchmark models that are used to test our methods. The details include the production cross-section at $\sqrt{s} = 13$ TeV, the number of events that were generated, and the number of events expected in 36 fb^{-1} of LHC data [38].

- The missing transverse energy $E_T^{\text{miss}} \geq 150$ GeV,
- ≥ 4 jets with their transverse momenta $p_{T,j} > 20$ GeV,
- The scalar sum of the jet tranverse momenta $H_T \geq 600$ GeV.

We stress that in the generation of these events, we have not included the most-accurate prediction for each and every observable, or reweighted the cross section by higher-order/resummed corrections. In our work, we are interested in the comparison of methods. To that end, we produced a data set that approximates the SM up to a decent accuracy. Of course, including higher-order corrections (both strong and electro-weak corrections) or resummed results would change the distributions and the total cross sections, and hence the absolute significance of each of the techniques, but we expect that it does not change the relative performance of each method.

Secondly, in the present study, we have a 'perfect' simulator, as we compare signal+background events with background events generated by the same event generator. Therefore, the algorithms pick up anomalies that originate from the presence of new physics explicitly injected to our 'signal+background' data set. However, if any unsupervised anomaly-detection method trained on Monte-Carlo (MC) data would be used to tag anomalous events in *real* collider data, these events are not guaranteed to point at signs of new physics. When anomalous events would be found, we first would need to verify that these events do not originate from miss-modelling in the MC. Therefore, although our techniques are signal-independent, they definitely do depend on the (imperfect) modeling of the background.

JHEP09(2021)024

Process	Process ID	σ (pb)	N_{tot} ($N_{36\text{fb}^{-1}}$)
$pp \to jj$	njets	$19718_{H_T>600\,\text{GeV}}$	415,331,302 (709,844,904)
$pp \to \ell\nu_\ell(+2j)$	w_jets	$10537_{H_T>100\,\text{GeV}}$	135,692,164 (379,318,453)
$pp \to \gamma(+2j)$	gam_jets	$7927_{H_T>100\,\text{GeV}}$	123,709,226 (285,367,766)
$pp \to \ell^+\ell^-(+2j)$	z_jets	$3753_{H_T>100\,\text{GeV}}$	60,076,409 (135,106,531)
$pp \to t\bar{t}(+2j)$	ttbar	541	13,590,811 (19,483,873)
$pp \to W^+W^-(+2j)$	ww	244	17,740,278 (8,788,874)
$pp \to t$+jets$(+2j)$	single_top	130	7,223,883 (4,669,711)
$pp \to \bar{t}$+jets$(+2j)$	single_topbar	112	7,179,922 (4,019,025)
$pp \to W^\pm t(+2j)$	wtop	57.8	5,252,172 (2,079,148)
$pp \to W^\pm\bar{t}(+2j)$	wtopbar	57.8	4,723,206 (2,079,148)
$pp \to \gamma\gamma(+2j)$	2gam	47.1	17,464,818 (1,694,361)
$pp \to W^\pm\gamma(+2j)$	Wgam	45.1	18,633,683 (450,672)
$pp \to ZW^\pm(+2j)$	zw	31.6	13,847,321 (315,781)
$pp \to Z\gamma(+2j)$	Zgam	29.9	15,909,980 (299,439)
$pp \to ZZ(+2j)$	zz	9.91	7,118,820 (99092)
$pp \to h(+2j)$	single_higgs	1.94	2,596,158 (19,383)
$pp \to t\bar{t}\gamma(+1j)$	ttbarGam	1.55	95,217 (15,471)
$pp \to t\bar{t}Z$	ttbarZ	0.59	300,000 (5874)
$pp \to t\bar{t}h(+1j)$	ttbarHiggs	0.46	200,476 (4568)
$pp \to \gamma t(+2j)$	atop	0.39	2,776,166 (3947)
$pp \to t\bar{t}W^\pm$	ttbarW	0.35	279,365 (3495)
$pp \to \gamma\bar{t}(+2j)$	atopbar	0.27	477,0857 (2707)
$pp \to Zt(+2j)$	ztop	0.26	3,213,475 (2554)
$pp \to Z\bar{t}(+2j)$	ztopbar	0.15	2,741,276 (1524)
$pp \to t\bar{t}t\bar{t}$	4top	0.0097	399,999 (96)
$pp \to t\bar{t}W^+W^-$	ttbarWW	0.0085	150,000 (85)

Table 2. Summary of the background processes included in the analysis. The details include the production cross-section at $\sqrt{s} = 13\,\text{TeV}$, the number of events that were generated, and the number of events expected in $36\,\text{fb}^{-1}$ of LHC data [38]. To account for interference effects, the `njets` sample also contains electro-weak exchanges (i.e. with a Z, $W^\pm$ or photon). Then the remaining leptonic productions of Z/γ s-channel processes are included in $pp \to l^+l^-$.

JHEP09(2021)024

We stress here again that our algorithms are developed and trained on the SM background processes only, which will in general give worse performance than a supervised approach that assumes knowledge of the signal. However, it is important to stress that the question we wish to answer is "to what extent would our techniques be able to discover or exclude interesting models if one does not know about them in advance?". With this goal in mind we have created a selection of supersymmetric benchmark model points, some of which are already excluded by the dedicated ATLAS and CMS searches.

Our first set of BSM models involve supersymmetric gluino pair production, with each gluino subsequently decaying to a boosted top-quark pair and the lightest neutralino, which is stable by assuming R-parity conservation. The gluinos are assumed to have a mass of 1–2.2 TeV (in steps of 200 GeV), while the neutralinos have a mass of 1 GeV. The branching ratio of the decay process $\tilde{g} \to t\bar{t}\tilde{\chi}_1^0$ is taken to be 100%.

In the second scenario two stop quarks ($\tilde{t}_1$) are produced, with each stop decaying into an on-shell top quark and a lightest neutralino ($\tilde{t}_1 \to t\tilde{\chi}_1^0$). We have chosen to take four different benchmark scenarios. In the first model, the lightest neutralino has a mass of 20 GeV and the lightest stop has a mass of 220 GeV. In the other models, the mass of the lightest neutralino is 100 GeV and the stops have masses of 300, 400 and 800 GeV.

Although the production cross-section for the lowest-mass stop quark pair production is the highest out of all assumed signal hypothesis, it is actually the most difficult to discover in traditional search methods. The mass difference of $\tilde{t}_1$ and $\tilde{\chi}_1^0$ is close to the mass of the top quark, which makes the production of top quark pairs an important irreducible background. The techniques described in the next section are designed to find anomalies, but this model does not result in an obviously anomalous signal. Therefore, we expect that the techniques will show least sensitivity to the 200 GeV stop scenario, although this might be compensated by the fact that its cross section is the highest. On the other hand, the gluino signals are more anomalous as they result in four top quarks and a sizable missing transverse energy. This is a rare final state for SM production, and since the 1 TeV gluino carries the highest production cross-section, we expect that this scenario will be the easiest.

All data is first zero-padded so every event has the same dimensionality. Next, the continuous data and the categorical data are split and the number of objects in the events are counted. From this, the following event structure is defined:

$$\mathbf{x} = \left(N, \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{\max} \end{bmatrix}, \begin{bmatrix} (p_T, \eta, \phi)_0 \\ (p_T, \eta, \phi)_1 \\ \vdots \\ (p_T, \eta, \phi)_{\max} \end{bmatrix} \right), \tag{2.1}$$

where max indicates the maximum number of objects the events. In this vector, N is the number of objects in the event, c_i is the object type as a one-hot encoded vector, p_T is the transverse momentum, η the pseudorapidity and ϕ the azimuthal angle of an object. This layout is used to train the unsupervised machine learning algorithms, detailed in section 3, on the 4-vector representations of the data. When we later use the non-VAE techniques on the latent space variables of the VAE, it is still true to say that the starting point for the analysis is this 4-vector representation.

JHEP09(2021)024

3 Search methods

Searches for new physics at the LHC can be divided into two main categories:

1. *Searches for visibly decaying new particles*, in which all decay products of a new particle are expected to be observed. In this case, one can often find the new physics by observing the invariant mass of the anticipated decay products of the new physics signal, although this becomes problematic in the case that the decay products themselves are unstable (e.g. new resonances decaying to top quarks or gauge bosons), in cases where the width of the resonance is expected to be large, or in cases where strong interference effects distort the shape of the invariant mass peak.

2. *Searches for semi-invisibly decaying new particles*, in which one cannot rely on the invariant mass to highlight the new physics, and must instead construct various functions of the final state four-vectors in events to try and discriminate the signal from the SM backgrounds. These searches are typically conducted within a given final state, characterised by the multiplicity of jets, b-jets, and leptons. This is useful both for scientific reasons (the SM backgrounds have an entirely different composition, and thus require a dedicated measurement, in each final state), and for political reasons (organisation of physics working groups by final state is an efficient way to parallelise search efforts). We therefore continue to assume in this paper that building an analysis within a given final state is a good goal, deferring the development of techniques that creatively use information across final states to further work.

In this paper, we focus on the second of these problems, using our unsupervised machine learning techniques to define an *anomaly score*. This is an event-by-event scalar that rates how anomalous an event is in the space of variables that the anomaly score algorithm was trained on. Our underlying assumption is that the new physics must be noticeably different from the SM background in the space constructed by the reconstructed 4-vectors and multiplicities of the final state objects. Since our approach works on 4-vector information (which, along with particle multiplicities, are amongst the most basic set of reconstructed properties in LHC events), we believe that it is more model-independent than high-level variables that target specific kinematic configurations. We here provide a brief overview of the techniques that are used in our study, as well as a short summary of traditional search methods.

3.1 Traditional methods

A traditional LHC search involves constructing useful physical variables that yield some separation between signal and background. These variables are then used to perform cuts that maximise the signal while minimising the background. A set of these constraints is known as a signal region. Different signals will appear in different signal regions, meaning that each signal region must be constructed specifically for a given search, though we can expect signals that are close in parameter space to be covered by similar signal regions. Some variables that are commonly used for semi-invisible particle searches are:

JHEP09(2021)024

- E_T^{miss}: missing energy is useful in identifying models with extra invisible particles, or anomalous production of SM neutrinos. Heavy supersymmetric particles decaying to the lightest neutralinos will yield a significantly broader E_T^{miss} distribution than the SM background.

- H_T: the scalar sum of the p_T of the jets. Its distribution is generally broader for events that produce heavy BSM particles compared to the SM events.

- m_{eff}: the scalar sum of the p_T of objects of interest plus the E_T^{miss} is a similarly useful variable, yielding a broader distribution for cases where the signal events produce much heavier particles than SM events.

- $m_T^{b,\text{min}}$: the transverse mass calculated from the E_T^{miss} and the b-tagged jet closest in ϕ to the p_T^{miss} direction is commonly used to reject events in which a W boson decays via a lepton and neutrino (see e.g. [45–47]). This is helpful to reject $t\bar{t}$ background events in searches for new particles that decay to top quarks (such as the supersymmetric top quarks we consider in our benchmark models).

3.2 Isolation Forests

As first outlined in ref. [48], the Isolation Forest (IF) is an unsupervised learning algorithm that assigns each point in a dataset a value based on the ease with which it is isolated from the other points in the dataset. It is attractive due to its simple concept, linear time complexity and low memory requirement.

Given a set of data $X = \{\vec{x}_1, \vec{x}_2, \ \ldots, \vec{x}_n\}$ from a multivariate distribution, where each $\vec{x}_i$ is a vector with d dimensions, one first chooses a feature $k \in \{1, \ldots, d\}$, and a "split value" p which lies between the maximum value and minimum value of the feature k. These are both chosen randomly using a flat prior. Then all $\vec{x}_i$ of the dataset with $x_{ik} < p$ are placed in a set of points called X_l while if $x_{ik} \geq p$, it is placed in a set called X_r. This process is repeated recursively, selecting a new k each time, until *one* of the following stopping conditions is met:

- every data point $\vec{x}_i$ is isolated in its own set,

- all $\vec{x}_i$ in a given set are equal,

- a limit imposed on the number of splits is reached.

The sequence of splits generated are called *trees*, and the number of splits in them is called the *path length* of the tree. Each split is a *node* of the tree. Nodes that do not begin or end trees are *internal*, and those which do are *external*. By randomly selecting batches of size m from the dataset and constructing a tree for the batch, we construct what is called a *forest*. The combination of many trees in this way improves stability and performance.

An anomaly is by definition an outlier, thus an anomaly should on average require a fewer number of splits to become isolated. The measure of anomalousness can then be defined via the average path length of the trees in the forest. This average path length is

JHEP09(2021)024

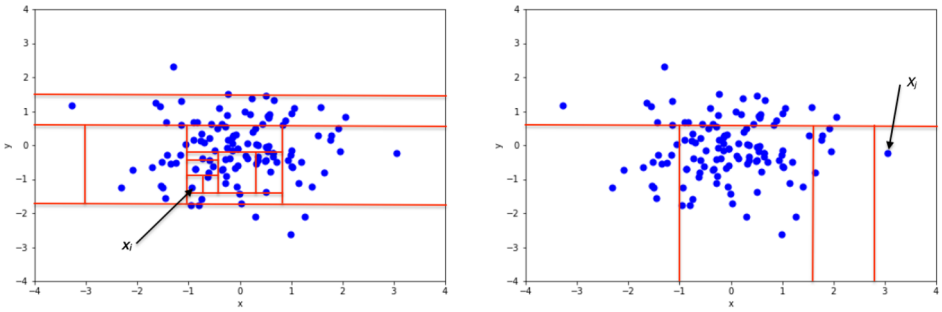

Figure 1. An example of two trees formed in the IF algorithm for an arbitrary 2D Gaussian distribution. Left: the isolation of a non-anomalous data point, which has path length 13. Right: the isolation of an anomalous data point, which has path length 4 [48].

normalized using the average path length of an unsuccessful search in a binary search tree (BST), as detailed in section 10.3.3 in ref. [49].:

$$c(m) = 2H(m-1) - \left(\frac{2(m-1)}{n}\right), \tag{3.1}$$

where n is the full dataset size, m is the size of a randomly sampled batch, and $H(x)$ indicates the harmonic number. The anomaly score of a point $\vec{x}_i$ is then defined as

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}, \tag{3.2}$$

where $h(x)$ is the path length and $E(h(x))$ is the mean path length of all trees constructed for x. This definition is convenient as it normalises the anomaly score between 0 and 1. It can be seen from eq. (3.2) that $s \approx 1$ implies a high level of anomalousness (since $E(h(x))$ would be small), $s \approx 0$ indicates no anomaly at all (since $E(h(x))$ would be large). If the whole sample generates $s \approx 0.5$, we find that the entire sample is likely devoid of anomaly. For our purposes, the anomaly score has been renormalised to fall between -1 and 1, with -1 indicating no anomaly at all, and 1 indicating a high level of anomalousness. An example of two trees, one for a non-anomalous point and one for an anomalous point, in two dimensions can be found in figure 1. Note that the non-anomalous point required thirteen nodes (or splits) to isolate, while the anomalous point required only four, showing that their path lengths are vastly different.

3.3 Gaussian mixture models

Datasets often have subsets that share a common characteristic. Mixture models are statistical models that approximate the statistical distributions of the characteristics of such datasets. This methodology allows one to approximate a set of statistical distributions that a set of data was most likely sampled from. Specifically, Gaussian mixture models (GMMs) are an implementation of this methodology where the individual statistical distributions being fitted are Gaussian distributions [50]. The statistical distribution of the entire dataset would then build up out of these Gaussian distributions.

JHEP09(2021)024

Let us define a set of data points as $\boldsymbol{X} = \{\vec{x}_1, \ \ldots, \vec{x}_n, \ \ldots, \vec{x}_N\}$, where each $\vec{x}_n$ is a vector with d features. Let $\vec{\mu}_k$, $\vec{\Sigma}_k$ with $k = 1, \ldots, K$ be the mean vectors and covariance matrices of a chosen number (K) of d-dimensional Gaussian distributions, initialized arbitrarily. For each data point we introduce a vector of latent variables, $\vec{z}_n$ representing that it belongs to a particular Gaussian: if the n^{th} data point belongs to the k^{th} Gaussian we set $z_{nk} = 1$, otherwise it is zero.

We can write the probability of observing a given data point $\vec{x}_n$ from its Gaussian as

$$p(\vec{x}_n|\vec{z}_n) = \prod_{k=1}^{K} \mathcal{N}(\vec{x}_n|\vec{\mu}_k, \vec{\Sigma}_k)^{z_{nk}}, \tag{3.3}$$

where $\mathcal{N}$ denotes a Gaussian distribution. Note that this product occurs over all Gaussians but the way we have constructed the latent vector $\vec{z}_n$ suppresses all but the Gaussian $\vec{x}_n$ belongs to. Now by Bayes rule and marginalization over all $\vec{z}$ we get

$$p(\vec{x}_n) = \sum_{k=1}^{K} p(\vec{x}_n|\vec{z})p(\vec{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\vec{x}_n|\vec{\mu}_k, \vec{\Sigma}_k), \tag{3.4}$$

where we have defined a mixing parameter $\pi_k \equiv p(z_k = 1)$. These represent the probability that an arbitrary point belongs to the k-th mixture component (the k-th Gaussian), and hence the sum of π_k over all k is 1.

We aim to maximize the probability that the observed data was sampled from the set of K Gaussians ($p(\boldsymbol{X})$) by updating the parameters of those Gaussians. The log-likelihood of this probability is given by

$$\log\left(p(\boldsymbol{X})\right) = \sum_{n=1}^{N} \log\left(p(\vec{x}_n)\right) = \sum_{n=1}^{N} \log\left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\vec{x}_n|\vec{\mu}_k, \vec{\Sigma}_k)\right]. \tag{3.5}$$

The optimization of this function can be performed using the Expectation-Maximization (EM) algorithm. There are two steps to the EM algorithm: the expectation step (E-step) and maximization step (M-step).

The E-step is performed by calculating the probability that each point was sampled from a particular Gaussian. This can be expressed in terms of the latent variables as $p(\vec{z}_k = 1|\vec{x}_n)$, which is often referred to as the *responsibility* of the distribution k for a given data point $\vec{x}_n$. Using Bayes law we can write [51]

$$p(\vec{z}_k = 1|\vec{x}_n) = \frac{p(\vec{x}_n|\vec{z}_k = 1)p(\vec{z}_k = 1)}{\Sigma_{j=1}^{K} p(\vec{x}_n|\vec{z}_j = 1)p(\vec{z}_j = 1)} = \frac{\pi_k \mathcal{N}(\vec{x}_n|\vec{\mu}_k, \vec{\Sigma}_k)}{\Sigma_{j=1}^{K} \pi_j \mathcal{N}(\vec{x}_n|\vec{\mu}_j, \vec{\Sigma}_j)} \equiv \gamma(z_{nk}). \tag{3.6}$$

Once we have calculated $\gamma(z_{nk})$ for all n and k we can undertake the M-step to estimate the updated parameters of each Gaussian. First one calculates the number of points N_k for which Gaussian k is responsible

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{3.7}$$

JHEP09(2021)024

With this value, we update the mean of Gaussian k by calculating the mean of the data points that belong to it, weighted by the responsibilities $\gamma(z_{nk})$

$$\vec{\mu}'_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \vec{x}_n. \tag{3.8}$$

Similarly, the updated covariances for Gaussian k are given by the covariance of the points that belong to Gaussian k with the updated mean $\vec{\mu}'_k$, weighted by the responsibilities

$$\vec{\Sigma}'_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\vec{x}_n - \vec{\mu}'_k)(\vec{x}_n - \vec{\mu}'_k)^T. \tag{3.9}$$

Finally, the mixing parameter π_k of Gaussian k is updated by calculating the percentage of the total dataset that belong to it

$$\pi'_k = \frac{N_k}{N}. \tag{3.10}$$

The new log-likelihood may be computed directly using eq. (3.5) with the new parameters for each Gaussian. The process is repeated iteratively, until we see convergence of the log-likelihood (parameterized by a tolerance), or when the maximum number of epochs is reached. The anomaly score of a given data point (event) is then given by the log-probability $\log p(\vec{x}_n)$, with $p(\vec{x}_n)$ defined in eq. (3.4).

3.4 Autoencoders

Autoencoders [52] (AEs) are a special class of neural networks where the input and output of the network are equal. This means that AEs can be trained without labels in unsupervised applications. The loss function typically is chosen to be the reconstruction loss, which is the difference between the output and input, quantified by, for example, the mean squared error on every dimension of the data. Generally, the number of hidden neurons in the neural network first decreases and then increases again, so the data needs to be squeezed in a lower dimensional representation. The lowest dimensional representation, usually in the middle of the network, is called the latent space. If the latent space dimensionality is too high, the neural network can simply learn the identity function to make the output equal to the input. When it is too low, assuming a fixed computational capacity, too much information needs to be removed in order to have a good reconstruction ability. The part of the network that transforms the input to latent space representation is called the encoder, while the part of the network that transforms the latent space representation to output is called the decoder.

If the latent space dimensionality is chosen sensibly, the input data is transformed into a lower dimensional representation, which contains only the relevant information that is required for reconstruction of the original input. If an AE is trained on a dataset without any anomalies and applied on a dataset with both normal and anomalous data, the AE will have a low reconstruction loss for the normal events and a high reconstruction loss for the anomalous events. This is because the anomalous events are different from the normal events, and thus are placed in unexpected locations in the latent space by the

JHEP09(2021)024

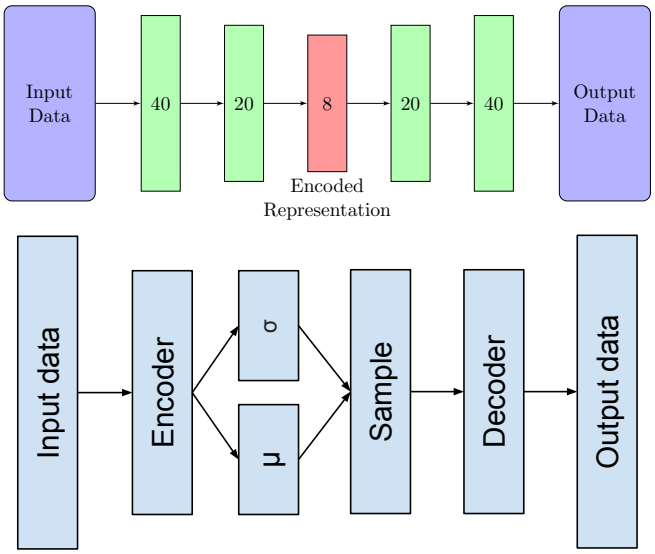

Figure 2. Schematic of our AE (top) and a VAE (bottom).

encoder, which the decoder cannot reconstruct well. These anomalous events are then reconstructed badly. An AE can thus be used as a anomaly detector where the anomaly score of a given event is defined as the reconstruction loss of that event [53].

In this work, we define an AE with 5 hidden layers that have 40, 20, 8, 20, and 40 nodes, as shown in figure 2 (top). This shape is modelled after [26]. Each layer uses a sigmoid activation function. The loss function used is a Sliced Wasserstein Distance Metric [54]. The Wasserstein Distance (sometimes referred to as "Earth Movers Distance") between two distributions u and v can be thought of as the minimum amount of energy required to transform u into v, where the energy is defined by a cost function given by the distribution weight multiplied by the distance to move the distribution. It is a useful tool as it metrizes the energy flow between two events. The Sliced Wasserstein Distance is the Wasserstein Distance between a projection of the data onto a 1-D distribution. It has similar properties to the Wasserstein Distance metric, and is more computationally efficient.

In addition to using the reconstruction loss as an outlier detection variable, one can also explore the latent space of an AE. If the latent space has ordering (similar events are clustered closely together in latent space) and the AE is trained to correctly reconstruct the standard model background, the latent space variables offer another representation of the standard model events. While the input space can have discontinuous and categorical data, the latent space only contains continuous data. This makes working with the latent space representation much more easy than working with 4-vector information, and one can define other outlier detection techniques on top of the latent space representation of the data.

3.5 Variational autoencoders

An AE does not have ordering in the latent space, because there is no term in the loss function that constrains the latent space. The variational autoencoder [55] (VAE), however, has this property, obtained by modifying the middle part of the neural network. The encoder

JHEP09(2021)024

outputs two numbers per latent space dimension, which represent the mean and standard deviation of a Gaussian distribution (see figure 2). The decoder takes a random sample of this distribution and decodes the sample into the original input. The loss function is augmented such that the KL-divergence [56] of these Gaussians and a standard normal distribution should be as low as possible. The loss function of a VAE then consists a function that encodes the ability to reconstruct the original data point, and a KL-divergence term. The former optimises for optimal reconstruction, while the KL-divergence term forces ordering in the latent space: all input should be encoded as close to $\vec{0}$ in the latent space as possible.

During training, a balance between the two contributions to the loss function should be found: if the KL-divergence term is zero, all input is encoded to $\mathcal{N}(0,1)$, which means there is no ability to reconstruct different points any more. The relative importance between the terms can be tuned. This was first done in ref. [57], where it was shown that if the KL-divergence term is more important than the reconstruction loss term, one will achieve disentanglement: every latent space dimension describes a different feature in the dataset. Then, ref. [58] showed that if the reconstruction loss term is more important than the KL-divergence term, ordering in the latent space still occurs while you can achieve a very good reconstruction loss. The relative importance of these two contributions is parameterised by β for the reconstruction loss and $(1-\beta)$ for the KL-divergence term, and we set $\beta = 3\cdot 10^{-3}$ in our work.

Using the event structure defined in eq. (2.1), the reconstruction loss is chosen to consist of three components: a-mean-squared error on the number of objects x_n, a-mean-squared error on the dimensionless regression variables ($\vec{x}_{r,i} = p_T/\text{MeV}, \eta$ or ϕ, see eq. (2.1)), and a categorical cross-entropy (see e.g. [59]) on the categorical variables $x_{c,i}$ that represent the types of the different objects in an event (see eq. (2.1)). These quantities are normalised between 0 and 1. The total loss function of the VAE is then defined as

$$\begin{aligned}\mathcal{L} = {} & 100\beta\,(x_n - \hat{x}_n)^2 \\ & + \frac{\beta}{d_r}\sum_i^{d_r}(x_{r,i} - \hat{x}_{r,i})^2 \\ & - \frac{10\beta}{d_c}\sum_i^{d_c}(x_{c,i}\log(\hat{x}_{c,i}) + (1-x_{c,i})\log(1-\hat{x}_{c,i})) \\ & + (1-\beta)\sum_i^{d_z}\mathrm{KL}\left(\mathcal{N}(\hat{\mu}_\mathrm{i},\hat{\sigma}_\mathrm{i}),\mathcal{N}(0,1)\right).\end{aligned} \tag{3.11}$$

Here, $\hat{x}_n$ the predicted number of objects, $\hat{x}_{r,i}$ the i-th predicted regression label, $\hat{x}_{c,i}$ the i-th predicted categorical label, d_r the number of regression variables, and d_c the dimensionality of the categorical data. All these variables are dimensionless, although they can represent a physical quantity that carries a dimension (e.g. the transverse momentum). The loss function just compares the dimensionless value that this transverse momentum has in a real event, with that obtained from the VAE. It assigns a big loss in cases these numbers are very different, in which case the VAE is not reconstructing the Monte-Carlo data accurately enough. The relative importance of these contributions to the loss function

JHEP09(2021)024

Algorithm	Anomaly-score definition
Isolation forest (IF, section 3.2)	Mean path length (eq. (3.2))
Gaussian mixture model (GMM, section 3.3)	Log probability (log of eq. (3.4))
Static autoencoder (AE, section 3.4)	Sliced Wasserstein Distance [54]
Variational autoencoder (VAE, section 3.5)	Reconstruction loss (first three lines of eq. (3.11))

Table 3. Summary of the considered ML algorithms and the definition of their anomaly scores.

is indicated by β. The first three components together form the total reconstruction loss, and the last component is the KL-divergence loss term. Because the three components for the reconstruction loss are not equally important, they are weighted with a numerical factor for the first (weight is 100) and the third (weight is 10) line. We chose these values for no other reason that these gave a sensible result, but that does not mean that other combinations of values for these weights (or the value of β) would not work.

Regarding the VAE architecture, we have used for the encoder and decoder 3 fully-connected hidden layers, each containing 512, 256 and 128 nodes respectively for the former, and 128, 256 and 512 nodes respectively for the latter. The activation function used between the hidden nodes is the exponential linear unit (ELU) [60]. The dimension of the latent space z is chosen to be 13. We have tested different combinations of these hyperparameters (β between 0 and 0.5, z between 5 and 30 and layer composition between 3 and 6 layers with varying amount of nodes between 32 and 1024) and found this combination to yield the best results on the loss. However, we did not do an extensive search for the optimal hyperparameter combination due to computational constraints [61]. It should be emphasised that the algorithm is being trained to reconstruct the SM and the best hyperparameter combination is selected based on the reconstruction loss, not on the performance on the test set. While there might be a combination of hyperparameters that could yield to even better results on the reconstruction loss, it might also lead to overfitting which would decrease the performance on the test set.

3.6 Concluding remarks and combinations of anomaly detection methods

Table 3 summarizes the anomaly scores for each algorithm that we have discussed. Since the VAE transforms the events into a lower-dimensional continuous space, it is believed that the other outlier detection techniques can find outliers more easily in the latent space of the VAE. Therefore, besides for exploring the above-discussed algorithms individually, we will also apply the IF, GMM, and the AE on the latent space of the VAE.

Moreover, if the anomaly scores yielded from each algorithm are minimally correlated with each other, there is information to be gained from combining them. In this paper we explore AND, OR, product, and averaging combinations. For a given event, let the anomaly score normalised to uniform background efficiency be x_i where i denotes the anomaly score algorithm. The AND anomaly score combination is given by $x^{\mathrm{AND}} = \min(x_i)$ for a given event, whereas OR anomaly score combination is given by $x^{\mathrm{OR}} = \max(x_i)$. The

JHEP09(2021)024

product combination is given by $x^{\text{product}} = \prod_i x_i$, and the average combination is given by $x^{\text{average}} = \frac{1}{N}\sum_i x_i$, where N is the number of algorithms being used, in this case, 4 (IF, GMM, AE, and the VAE).

The technique of combining algorithms is not guaranteed to always outperform a single algorithm. To this end, consider the following example. Imagine we train an algorithm (algorithm 1) that incorrectly classifies every background event as signal and vice-versa (i.e. 0 for every signal point and 1 for every background point). Consider a second algorithm (algorithm 2) that correctly classifies every background (signal) event as background (signal). The OR of these two algorithms will take the maximum value for each event — meaning every event will be classified as signal. This of course performs worse than algorithm 2. Now lets consider an AND combination, taking the minimum value for every event will classify every event as background, which also performs worse than algorithm 2. This shows that indeed the combination of algorithms is not guaranteed to outperform a single algorithm.

4 Results

In the following, we investigate two different ways of applying our techniques to particle physics data. In our first approach, we train the IF, GMM, AE and VAE directly on the 4-vector representation of the events that pass the pre-selection, and compare the relative performance of each technique. We then also assess the performance of the various combination techniques described in section 3.6. The full list of input variables is $(E_T, \phi)_{\text{miss}}$, $(E, p_T, \eta, \phi)_{\text{jets}}$, $(E, p_T, \eta, \phi)_{\text{bjets}}$, $(E, p_T, \eta, \phi)_{\text{leptons}}$, $(E, p_T, \eta, \phi)_{\text{photons}}$. Here, leptons can be positively or negatively charged electrons or muons. In our second approach, we train the VAE in the same way, but apply the IF, GMM and AE algorithms to the latent space variables defined by the VAE. The combination of techniques applied to these anomaly score methods end up providing the optimum results, as we will see in what follows.

4.1 Comparison of techniques with training on original 4-vectors

Let us first deal with the case of applying the techniques directly on the 4-vectors of the selected events. Having used each technique to define an anomaly score, we show in figure 3 the ROC curves for each algorithm that result from applying anomaly score cuts to the gluino signals. Given a particular selection on the anomaly score (see table 3), the true positive rate is calculated as the proportion of signal events to the right of the cut, whilst the false positive rate is calculated as the proportion of background events to the right of the cut. For the use of a physical variable (the effective mass for gluino events, and $m_T^{b,\min}$ for the stop events), the ROC curve is obtained by producing a series of selection cuts on this variable. We classify all events below (above) that certain cut as background (signal), from which we obtain the true/false positive rates as described above.

We also quote significance values, calculated for the choice of anomaly score cut that results in 100 background events being selected in $36\,\text{fb}^{-1}$ of LHC data. The inset on each

JHEP09(2021)024

panel of the figures shows the ROC curves near this region, where the black dashed line indicates the 100 background event cut. Significance values are calculated using

$$Z_B = \frac{S}{\sqrt{S + B + (\sigma_B B)^2}}, \tag{4.1}$$

where S is the number of signal events, B is the number of background events (100 in this case), and σ_B is the assumed systematic uncertainty. We use Z_{100} to compare the algorithms in what follows. We start our discussion by assuming zero systematic uncertainty ($\sigma_B = 0$ (figures 3, 4, 5, and 9). Plots comparing the significance values yielded from the 4-vector and latent space representations by performing a cut at 100 background events with and without a 15% relative systematic uncertainty are displayed and discussed in section 4.3.

In figure 3 we compare the performance of the IF, GMM, AE, VAE, the effective mass (defined below), and the combinations detailed in section 3.6 on various gluino signals, the details of which are contained in table 1. The effective mass is defined as $m_{\text{eff}} = E_T^{\text{miss}} + \sum_{\text{jets}} p_T$ and is a common discriminating variable in conventional gluino searches. Using Z_{100}, we see that the VAE provides the strongest separation between signal and background for all gluino signal models. The OR combination also gives a good separation.

However, the OR combination is not as effective as the VAE. The relatively poor performance of the IF algorithm can be explained by the fact that dividing the space by placing cuts in the space of the 4-vector components does not obviously isolate outlying events, since the anomalies are more likely to appear in non-trivial functions of the four-vector components. The AE and VAE perform better because their attempts to reproduce the structure of the background events involve defining non-linear functions of the input variables that do not then generalise well to the case of the signal events. Past the 100-background-event cut, the product and average combinations excel based off the area under the curve for each. Their relatively poor performance at the 100 event cut is due to the poor performance of the IF, GMM, and AE in these low background efficiency regions. For models with a higher gluino mass, our techniques lose discovery and exclusion potential despite the more anomalous kinematics of these models. This is caused by the reduced production cross-section in each case, resulting in a smaller value for S. It is possible that our anomaly score could be supplemented by traditional kinematic selections, and we return to this point in the next subsection.

Figure 4 displays the ROC curves that result from applying each algorithm applied to the various stop signals, the details of which are contained in table 1. The algorithms are compared to $m_T^{b,\text{min}} = \sqrt{2p_T^b E_T^{\text{miss}}[1 - \cos\Delta\phi(p_T^b, p_T^{\text{miss}})]}$, which is a common discriminating variable for stop signals. The first of these signals is particularly difficult to differentiate from the background, since it is kinematically very similar to the dominant background processes. Using Z_{100}, the OR combination (marginally) yields the strongest separation between signal and background. The VAE also provides a fair separation for all signals, while the GMM is comparable, although definitely a poorer anomaly detector for the stop case. The IF and the AE give much worse performance, which is easy to rationalise in the case of the IF using a similar argument to that provided for the gluino results. The AE

JHEP09(2021)024

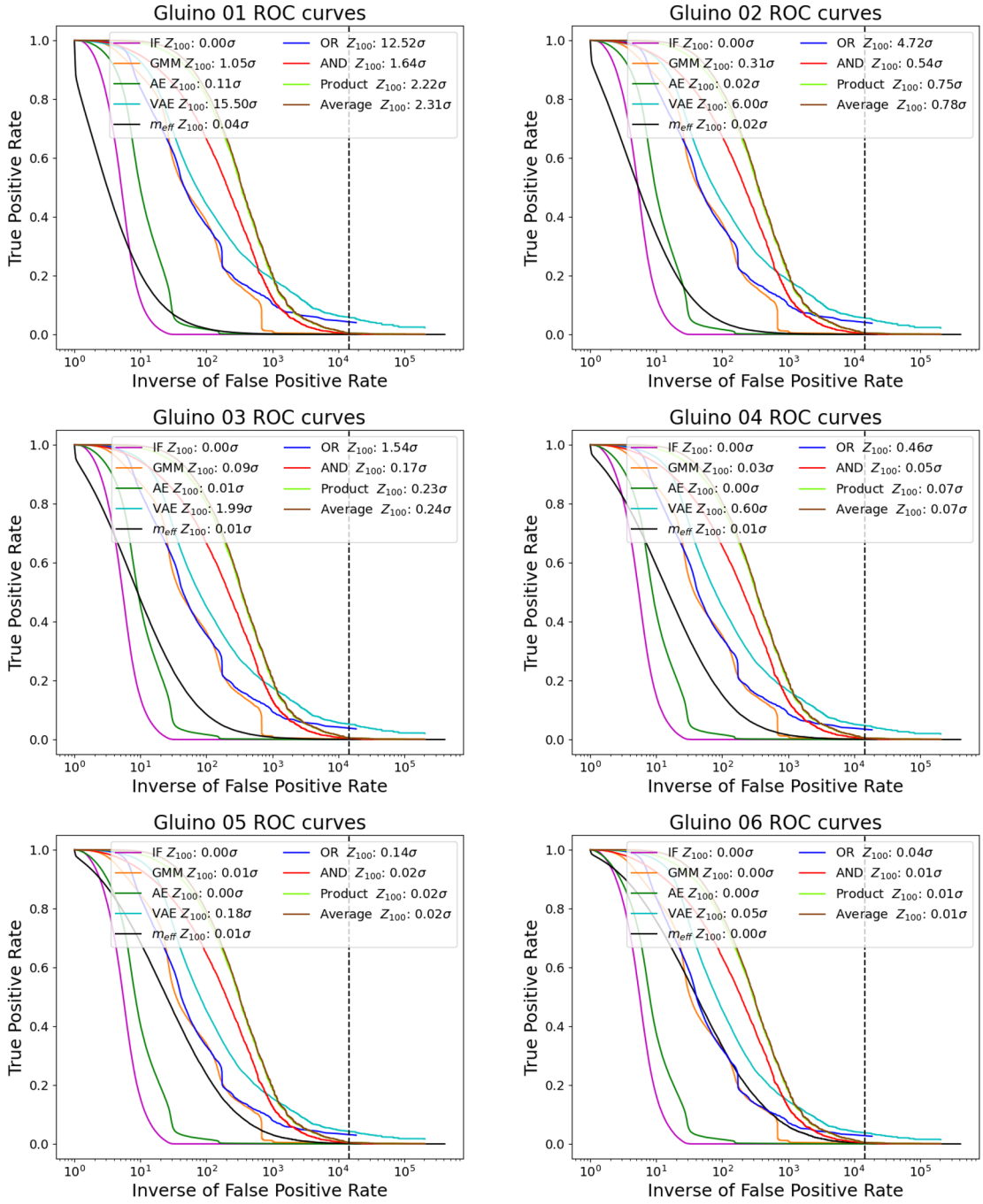

Figure 3. ROC curves for the gluino signals (table 1) for the algorithms applied on 4-vector representations, with on the horizontal (vertical) axis the inverted false-positive (true-positive) rate. The ROC curves of IF, GMM, AE and VAE (table 3) are shown in pink, orange, dark green and cyan respectively. The effective mass m_{eff} is shown in black, and combinations of the models are shown in blue (OR), red (AND), light-green (Product) and brown (Average). The black dashed line indicates the inverse false-positive rate at which $B = 100$.

JHEP09(2021)024

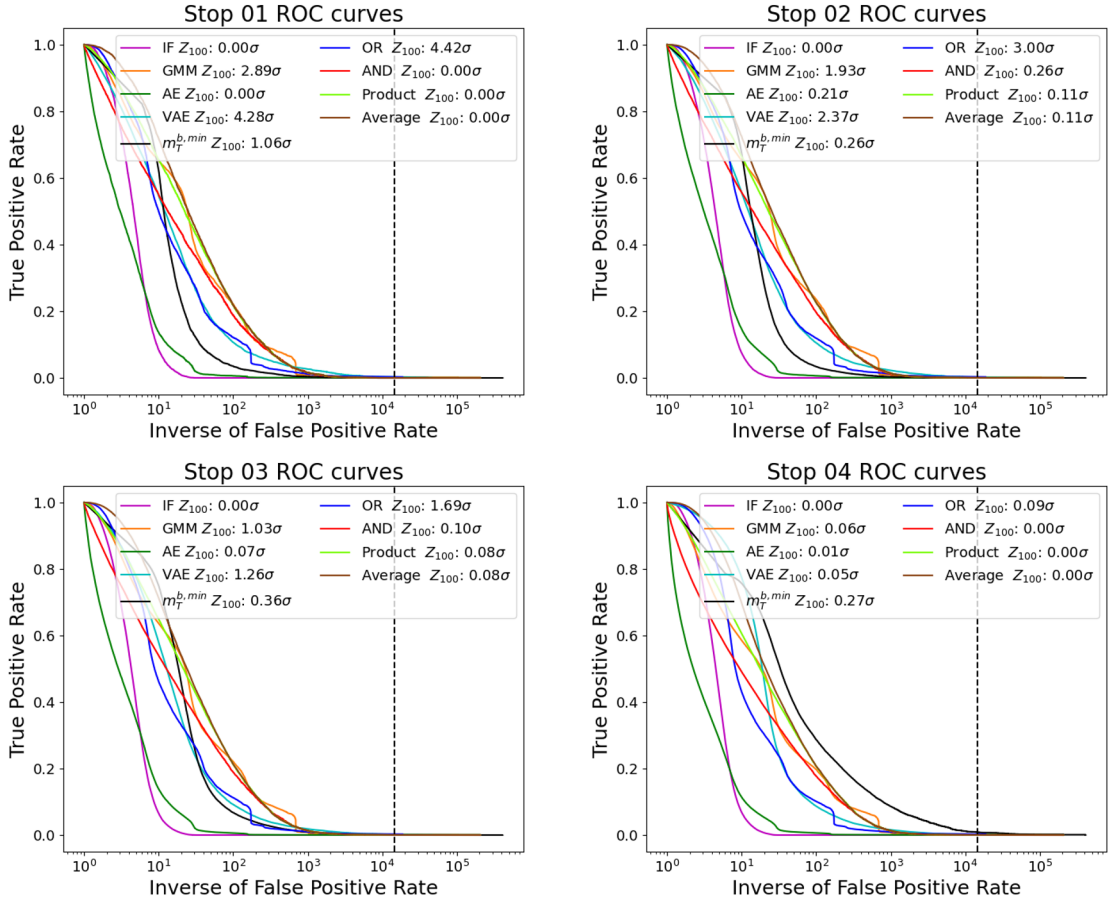

Figure 4. ROC curves for the stop signals (table 1) for the algorithms applied on 4-vector representations. For further information see figure 3. The physical variable that is used here is $m_{\mathrm{T}}^{\mathrm{b,min}}$.

is less flexible than the VAE, which explains its poor performance. It remains to be seen whether a more complex AE would improve the performance (which we will not pursue here by virtue of having defined the VAE). Surprisingly, the values for Z_{100} show the most promising results for the lower stop mass case despite being the most standard model-like. The reason that this happens is because of its cross section, which is significantly higher than those of the other tested stop scenarios. For higher stop masses it is clear that our techniques would not deliver discovery or exclusion potential despite the more anomalous kinematics, which is again driven by the falling production cross-section as the stop mass increases. Ultimately, this results from the fact that the anomaly score alone is not an effective discriminant between the signal and background for stop models. For the stop 04 signal, the variable $m_T^{b,\mathrm{min}}$ outperforms the algorithms, and additionally the sensitivity is very low. This is an indication that the unsupervised anomaly detection methods can reproduce the signal events roughly as well as the background events (either very well or very poorly). If an unsupervised anomaly detection is to select events that stem from new physics, the training set needs to be reconstructed to a sufficient level, while events outside the training set need to be reconstructed poorly.

JHEP09(2021)024

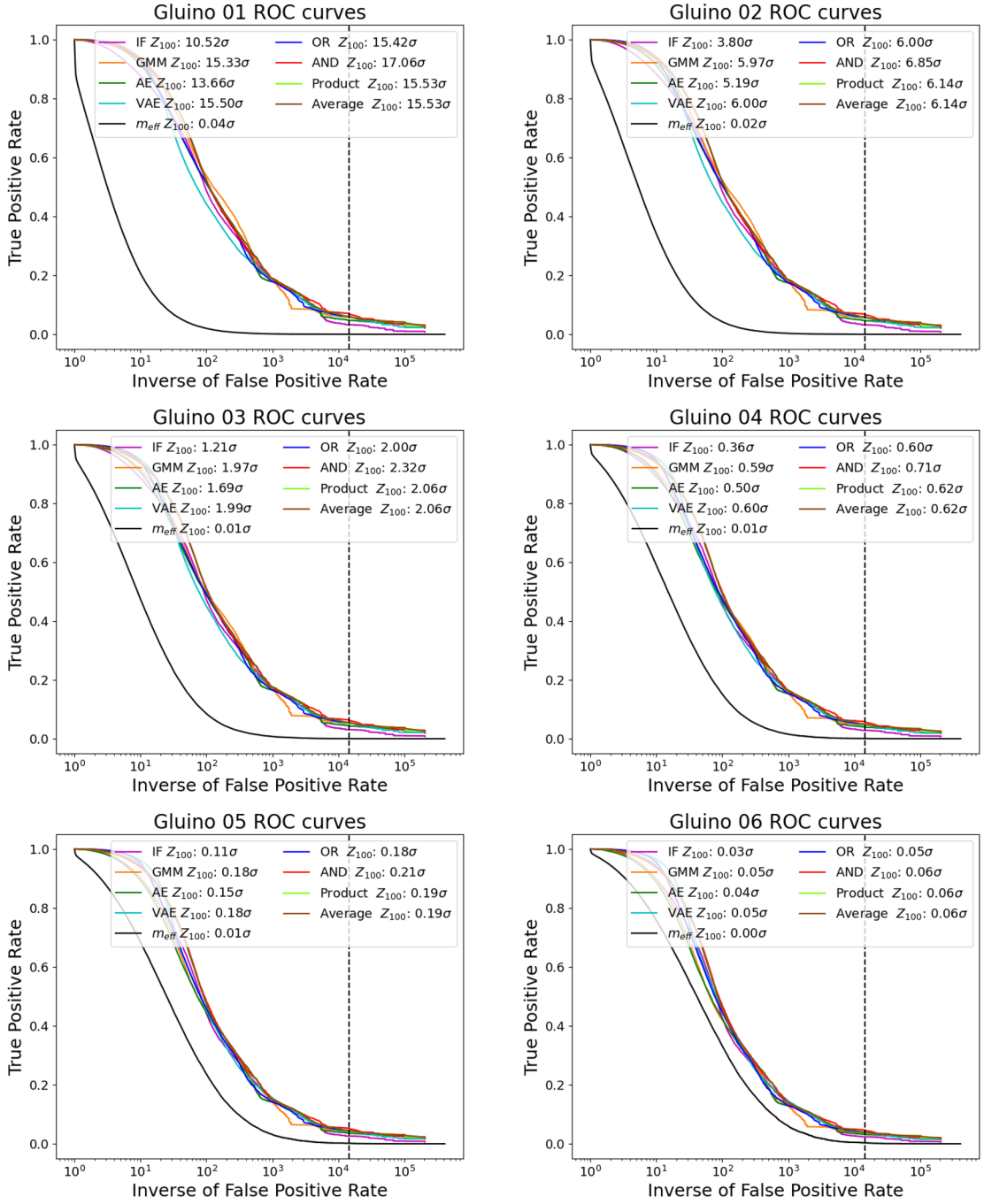

Figure 5. ROC curves for the gluino signals (table 1) for the algorithms applied on latent space representations. For further information see figure 3.

4.2 Comparison of techniques with training on latent space variables

Let us now consider the approach of training the IF, GMM and AE on the latent space representation of the SM events obtained from the VAE. The input variables for our VAE in this approach are the same as those in the previous section, and we continue to use the reconstruction loss to define the VAE anomaly score. However, the IF, GMM, and AE are

JHEP09(2021)024

Signal	Gluino 01			Gluino 02			Gluino 03			Gluino 04			Gluino 05			Gluino 06		
Variable	Z	S	σ	Z	S	σ	Z	S	σ	Z	S	σ	Z	S	σ	Z	S	σ
VAE	15.5	316	0.19	6.00	81.0	0.19	1.99	22.1	0.19	0.60	6.19	0.19	0.18	1.82	0.19	0.05	0.54	0.19
IF	10.5	174	0.27	3.80	46.0	0.27	1.21	12.9	0.27	0.36	3.64	0.27	0.11	1.09	0.27	0.03	0.33	0.27
GMM	15.3	311	0.27	5.97	80.3	0.27	1.97	21.9	0.27	0.59	6.08	0.27	0.18	1.79	0.27	0.05	0.53	0.27
AE	13.7	259	0.27	5.19	67.2	0.27	1.69	18.4	0.27	0.50	5.13	0.27	0.15	1.52	0.27	0.04	0.45	0.27
OR	15.4	314	0.27	6.00	80.8	0.27	2.00	22.2	0.27	0.60	6.19	0.27	0.18	1.83	0.27	0.05	0.55	0.27
AND	17.1	370	0.27	6.85	96.3	0.27	2.32	26.2	0.27	0.71	7.34	0.27	0.21	2.16	0.27	0.06	0.65	0.27
Prod	15.5	313	0.28	6.14	81.2	0.29	2.06	22.2	0.28	0.62	6.21	0.28	0.19	1.83	0.28	0.06	0.54	0.28
Avg	15.5	313	0.28	6.14	81.2	0.29	2.06	22.2	0.28	0.62	6.21	0.28	0.19	1.83	0.28	0.06	0.54	0.28

Table 4. Table for algorithms applied on latent space representations. It shows Z-scores for the gluino signals taken at 100 background events with no systematic uncertainty applied, the number of signal events (S) at said background cut, and the uncertainty on the background cut (σ).

trained on the 13 latent space variables defined by the VAE, which are non-linear functions of the original 4-vector variables, supplemented by the reconstruction loss of the VAE.

In figure 5, we show the ROC curves for the gluino signals, which demonstrate that the performance of our non-VAE techniques has now improved dramatically in each case. The effective mass has been left in these plots for further comparison. However, we see that their performance still does not exceed that of the VAE. The resulting Z_B-values for $B = 100$, the number of signal events that remain after taking the cut and the uncertainty on the cut σ are indicated in table 4. As mentioned before, the reduced production cross-section for higher-mass gluino models results in a smaller value of S. This means that for Gluino 05 and Gluino 06, σ exceeds the size of Z.

When compared to the effective mass $m_{\text{eff}} = E_T^{\text{miss}} + \sum_{\text{jets}} p_T$, all anomaly score definitions outperform it by a considerable margin. Once more, we provide significance values that are calculated for an anomaly score cut that leaves 100 background events in the selected sample, assuming $36\,\text{fb}^{-1}$ of LHC data. We can see that the AND combination anomaly score outperforms all other anomaly scores, though the OR, Product and Average combinations also perform at least as well as the VAE at the 100 event cut. This can be explained by the observation that the anomaly scores yielded from the VAE and the other algorithms are minimally correlated. This may be observed in figure 6, where a strong correlation would show up as a diagonal yellow line that goes from 0 to 1. Since this is not seen, this implies that there is further information to be gained by performing these combinations.

In figure 7, we show histograms of the anomaly score itself, for both the SM background and Gluino 01. The signal is plotted separately from the background in order to better show the difference in the shape of the anomaly score distribution for each algorithm. The final bin is an overflow bin containing all events to the right of it, chosen as the first bin containing less than 10 events. These plots show a clear separation between background and signal, with the signal being more clustered in the high anomaly score region as expected.

JHEP09(2021)024

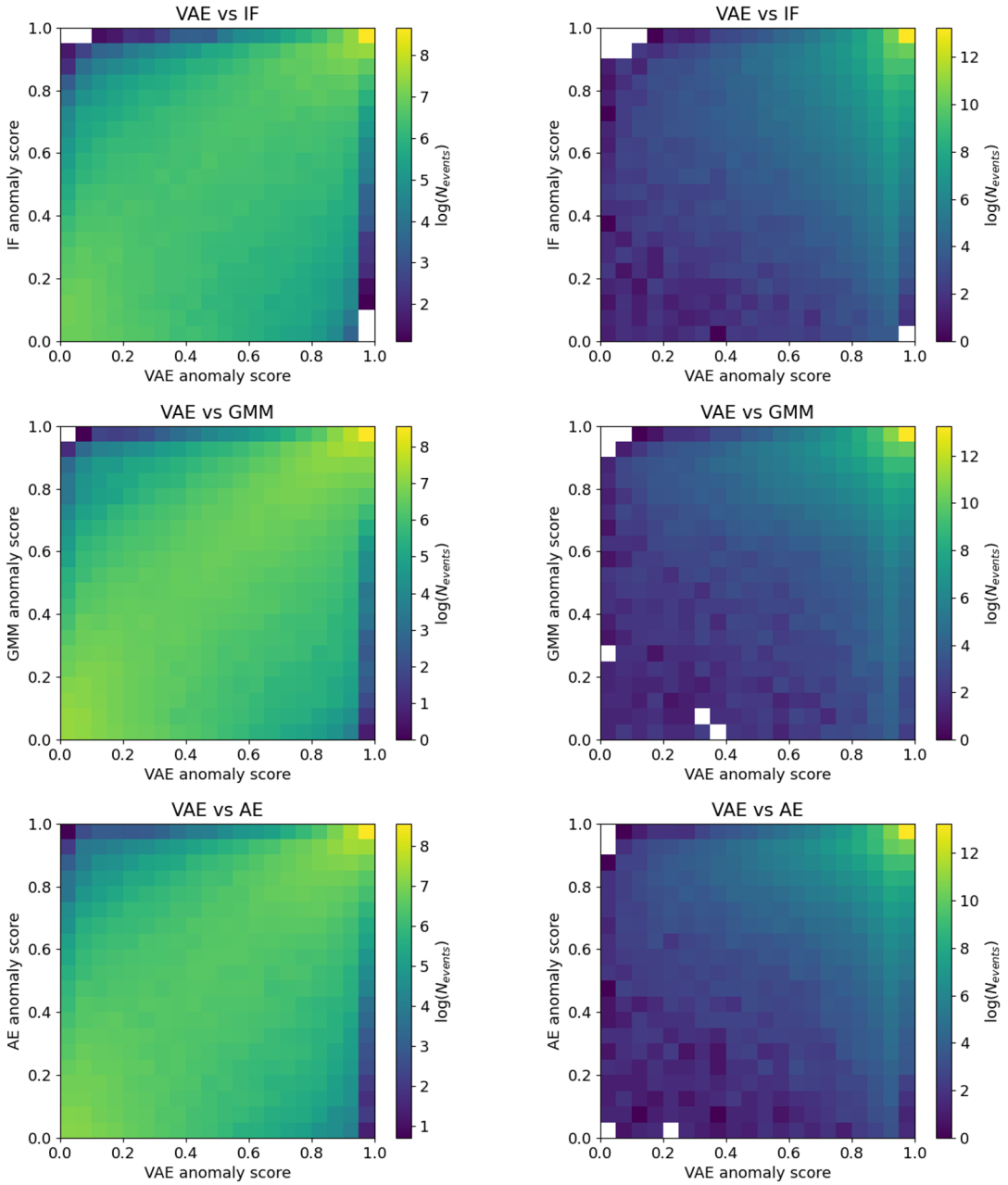

Figure 6. 2D correlation plots of various anomaly scores with normalised background efficiency compared to the VAE loss for the background (left) and gluino 01 signal (right). The colour coding represents $\log N_{events}$.

Figure 8 displays 2D correlation plots comparing the AND anomaly score to various physical variables for Gluino 01. If a significant correlation would exist between such a physical variable and the anomaly score, one would expect to see a diagonal line. However, the structure that we see stretches out horizontally, and gets marginally broader at higher values of the anomaly. It follows that there is minimal linear correlation between the anomaly score and any of the physical variables that are shown in figure 8. The Pearson

JHEP09(2021)024

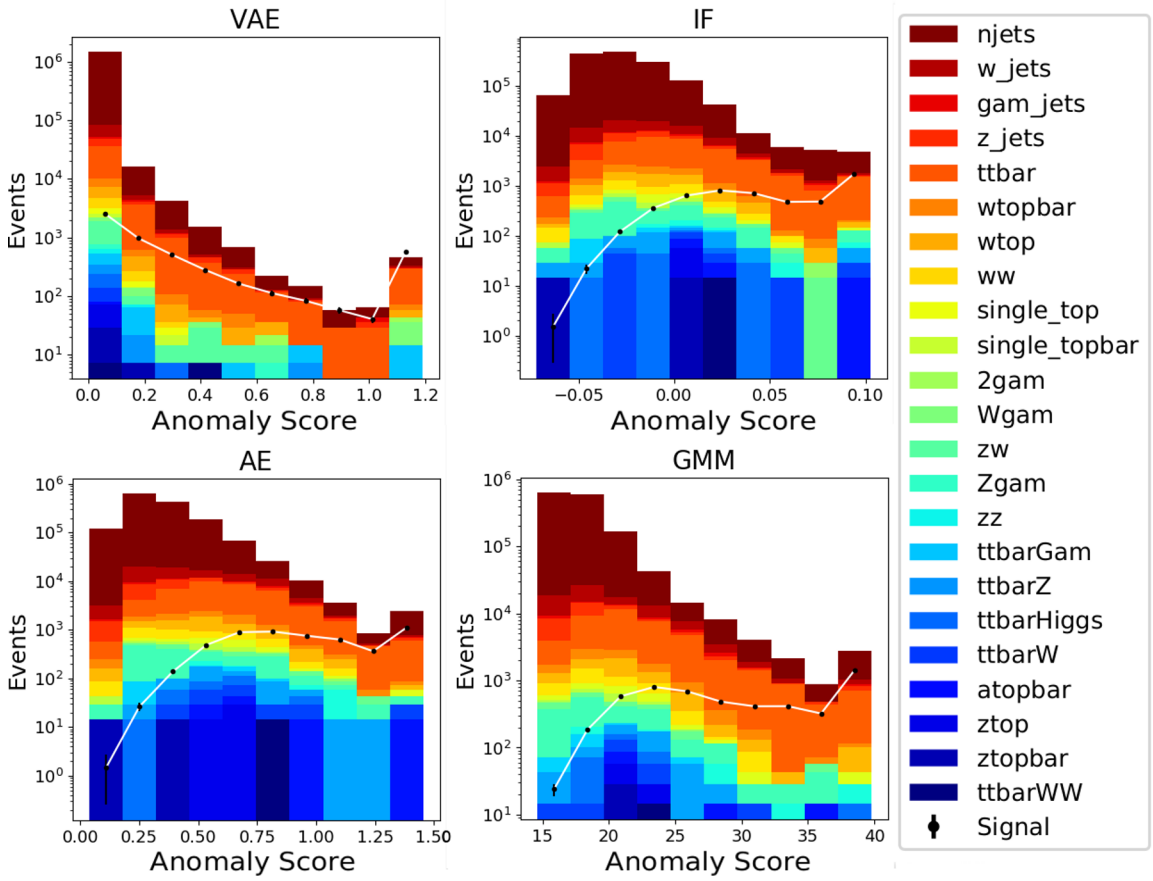

Figure 7. Anomaly score histograms derived from various algorithms for Gluino 01. The horizontal axis shows the anomaly score, and the histogram counts the number of events normalized to 36 fb^{-1} in each bin. The various colours indicate different backgrounds, while the black data points show the signal.

correlation coefficients are displayed in table 5. These values indicate that there is minimal linear correlation found between the variables. So long as the linear correlation is not 100%, one could use the anomaly score as the first selection of an LHC analysis (for example adding it to the high-level trigger menu, although this would have to be explored further), and then use conventional variables to enhance sensitivity to particular signals in the usual way. This hybrid approach reintroduces model-dependence through the later kinematic selections, but starts with very few signal assumptions.

Figure 9 displays ROC curves for the same algorithms being used on latent-space variables for the various stop signals. The resulting Z_B-values for $B = 100$, the number of signals that remain after taking the cut and the uncertainty on the cut σ are indicated in table 6. The significance numbers show promise, but it remains hard to isolate the stop signal using only a selection on the anomaly score. The IF, GMM, AE, and combination methods are much more effective when applied to the latent space. Again, the AND combination outperforms all other anomaly score definitions, except for the stop 04 signal, for which the traditional variable $m_T^{b,\text{min}}$ is the most effective. This suggests that, again,

JHEP09(2021)024

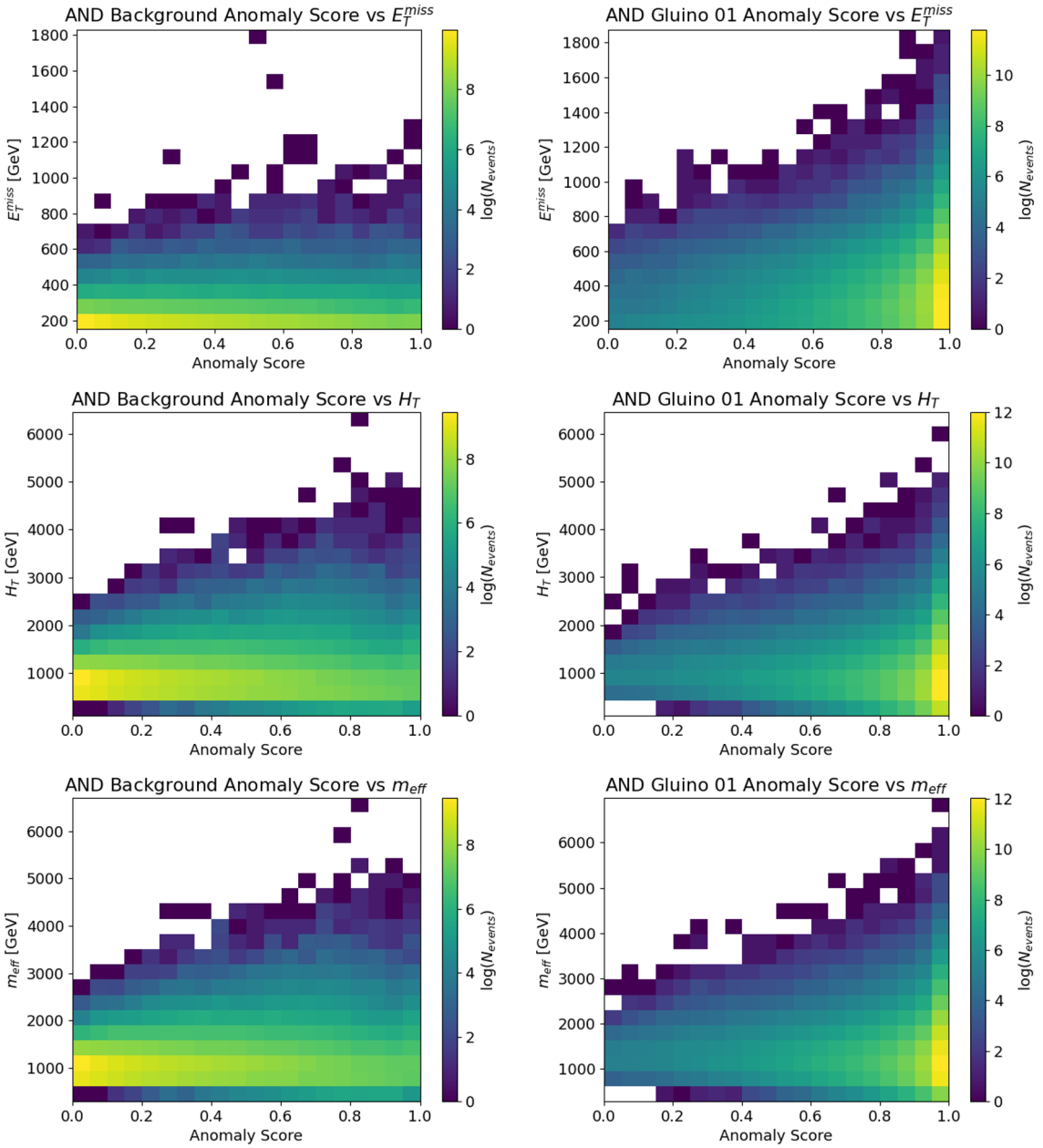

Figure 8. 2D histograms associated with Gluino 01 for background (left) and signal (right). Various physical variables are plotted on the y-axis, with the anomaly score generated from the AND combination applied in the latent space on the x-axis. The z-axis is $\log N_{\text{EVENTS}}$.

JHEP09(2021)024

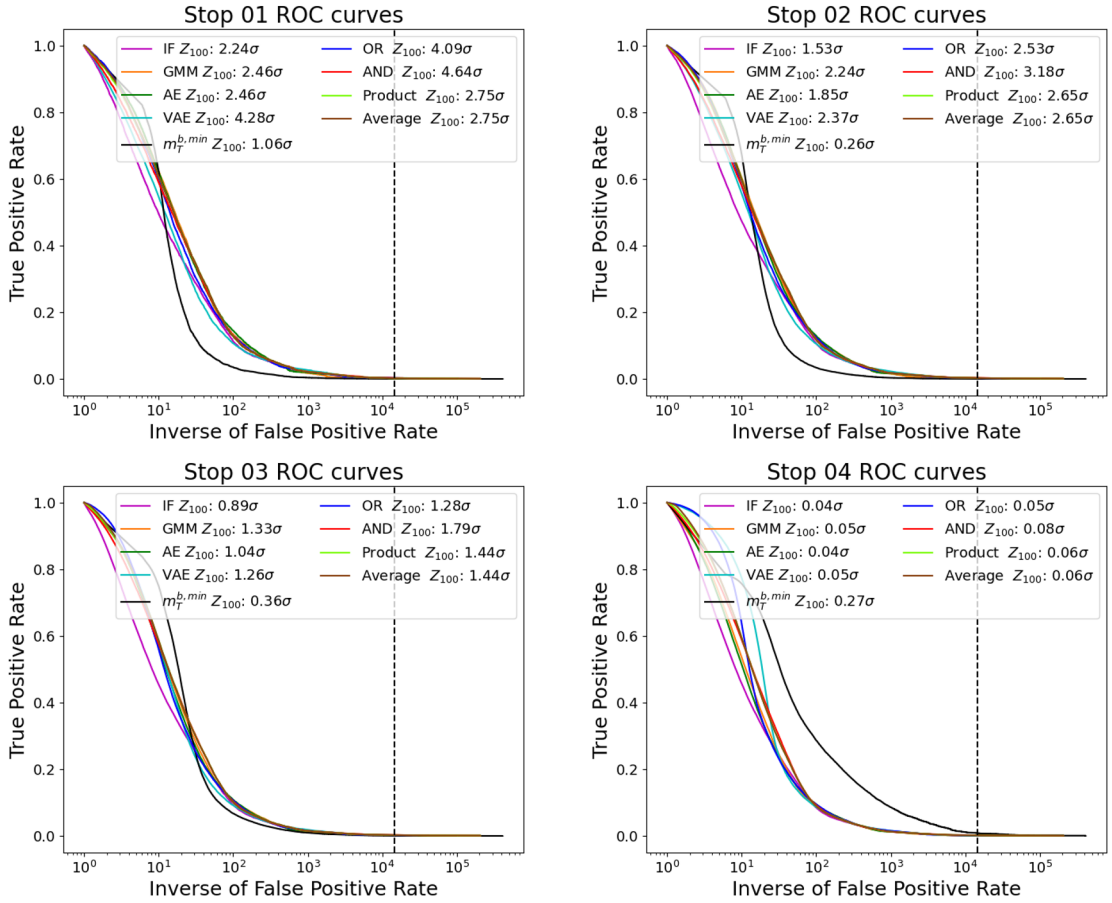

Figure 9. ROC curves for the stop signals (table 1) for the algorithms applied on latent space representations. Labeling is the same as in figure 3.

Dataset	E_T^{miss}	H_T	m_{eff}
Background	0.12	0.14	0.15
Gluino 01	0.032	-0.030	-0.017
Gluino 02	0.038	-0.057	-0.039
Gluino 03	0.041	-0.087	-0.063
Gluino 04	0.042	-0.11	-0.084
Gluino 05	0.043	-0.14	-0.11
Gluino 06	0.046	-0.16	-0.12
Stop 01	0.082	-0.0026	0.015
Stop 02	0.13	0.032	0.061
Stop 03	0.096	-0.029	0.0053
Stop 04	0.07	-0.10	-0.056

Table 5. Pearson correlation coefficients between the AND anomaly score and various physical variables for the background and signal datasets. A value of 0 implies no correlation, and a value of ±1 implies perfect positive/negative correlation. From these values we see minimal correlation between the AND anomaly score and these physical variables.

JHEP09(2021)024

Signal	Stop 01			Stop 02			Stop 03			Stop 04		
Variable	Z	S	σ	Z	S	σ	Z	S	σ	Z	S	σ
VAE	4.28	53.1	0.30	2.37	26.7	0.24	1.26	13.5	0.21	0.05	0.53	0.20
IF	2.24	25.1	0.43	1.53	16.6	0.32	0.89	9.35	0.29	0.04	0.37	0.28
GMM	2.46	27.9	0.41	2.24	25.1	0.30	1.33	14.3	0.28	0.05	0.53	0.28
AE	2.46	27.9	0.41	1.85	20.3	0.31	1.04	11.0	0.29	0.04	0.44	0.28
OR	4.09	50.3	0.36	2.53	28.9	0.30	1.28	13.8	0.29	0.05	0.53	0.28
AND	4.64	58.6	0.35	3.18	37.4	0.29	1.79	19.7	0.28	0.08	0.76	0.27
Prod	2.75	30.7	0.41	2.65	29.4	0.31	1.44	15.0	0.29	0.06	0.58	0.29
Avg	2.75	30.7	0.41	2.65	29.4	0.31	1.44	15.0	0.29	0.06	0.58	0.29

Table 6. Table for algorithms applied on latent space representations. It shows Z-scores for the stop signals taken at 100 background events with no systematic uncertainty applied, the number of signal events (S) at said background cut, and the uncertainty on the background cut (σ).

the anomaly detection methods are not able to distinguish data points that are similar or very different from the dataset.

Figure 10 displays the anomaly-score histograms for the background and Stop 01. The background and signal shapes appear much more similar than in the gluino case, though there is a tendency to push the signal to the right. Higher stop masses appear more anomalous but have a lower cross section and are thus more difficult to see at a 100 background event cut.

4.3 Summary and the inclusion of a systematic uncertainty

So far, we have assessed the performance of each algorithm by considering their ROC curves. Figure 11 displays side-by-side the significance values obtained using 4-vector components, and those obtained using the latent space variables. We can see that the performance of the IF, GMM, and AE increase when trained on the latent-space variables. Ultimately, training on latent space representations and performing an AND-combination of the normalised IF, GMM, AE and VAE anomaly scores yields the best performance for each considered signal of any technique detailed in this paper. Figure 12 displays the same significance values with a 15% assumed systematic applied. The significances for all except the Gluino 01 signal are not high enough for discovery potential, although Stop 01, Stop 02, Gluino 01, and Gluino 02 are above the exclusion limit. This lack of discovery potential is not unexpected as these techniques do not optimise on the signal, and many of the higher mass signals are difficult to find even by conventional analyses. However, these results are quite promising for use as a preliminary step in a conventional analysis, especially since every anomaly score used in this analysis is minimally correlated with commonly used physical variables, however it is possible there are algorithms that yield anomaly scores that are correlated with these physical variables.

JHEP09(2021)024

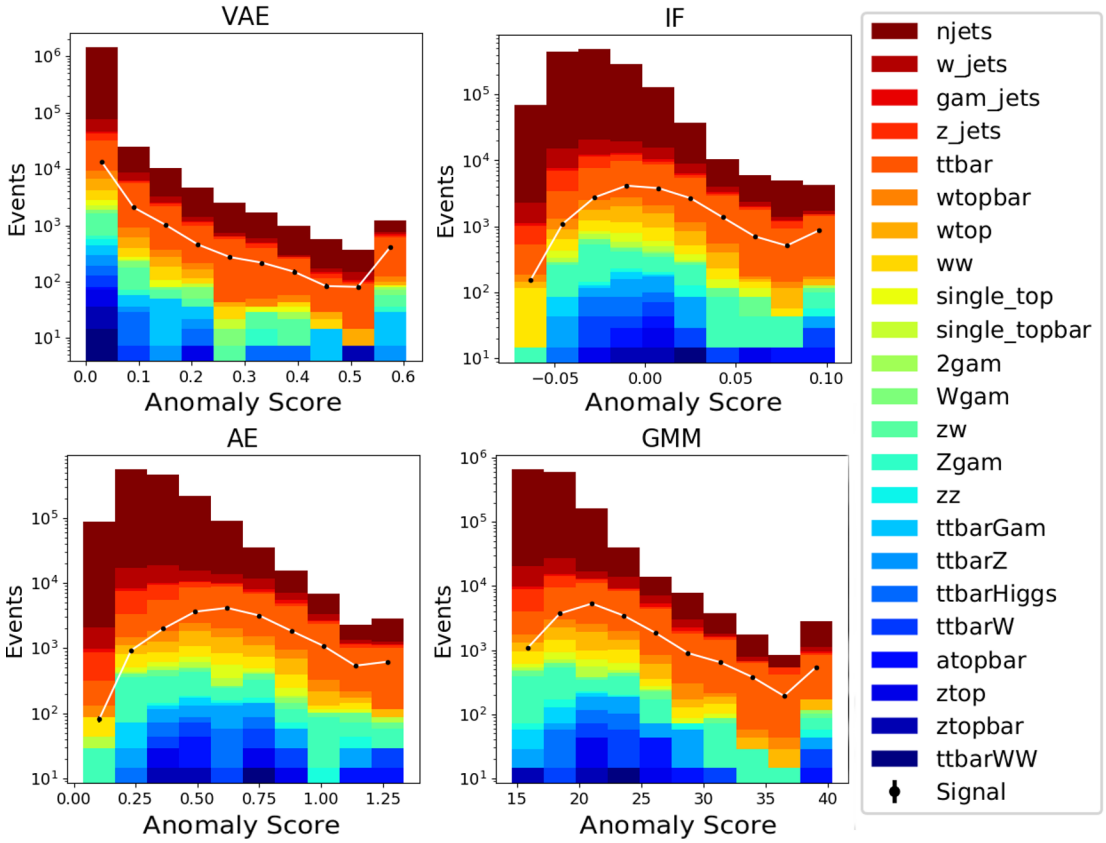

Figure 10. Anomaly score histograms derived from various algorithms for Stop 01.

5 Conclusions

In this paper, we have examined a variety of unsupervised ML methods to perform anomaly detection in LHC searches. Our aim is to specify an anomaly score on an event-by-event basis that indicates how likely it is that the given event originates from new physics. We only address the problem of detecting *anomalous* events, that is, events that are unlikely to have been produced by a SM process. The methods discussed in this paper are therefore not usable when the new physics shows up as an overproduction of a certain final state with similar kinematics to its respective SM background.

The studied ML methods are the isolation forest (section 3.2), the Gaussian mixture model (section 3.3), the static autoencoder (section 3.4), and the variational autoencoder (section 3.5). We have defined an anomaly score for each of these techniques (summarized in table 3), and assessed their performance in determining whether an event is anomalous by training them on the SM dataset published in ref. [38]. We have then tested this against a collection of supersymmetric benchmark scenarios, summarized in table 1. The performance of each model is represented by the significance measure Z_{100} (eq. (4.1)), which is the significance one obtains after cutting on the anomaly score that selects 100 background events. We note that our conclusions do not change within the latent space for a lower background event cut.

JHEP09(2021)024

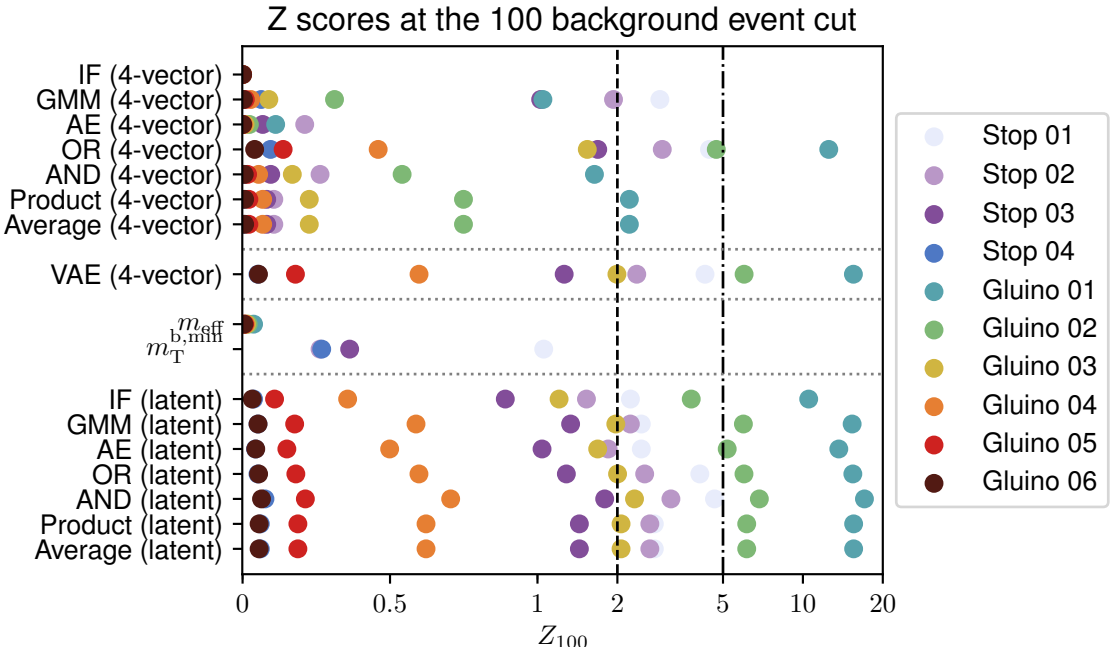

Figure 11. Z_{100} yielded from various algorithms applied to 4-vector components and latent space representations. See table 1 for the signal definitions, and table 3 for the definitions of the algorithms.

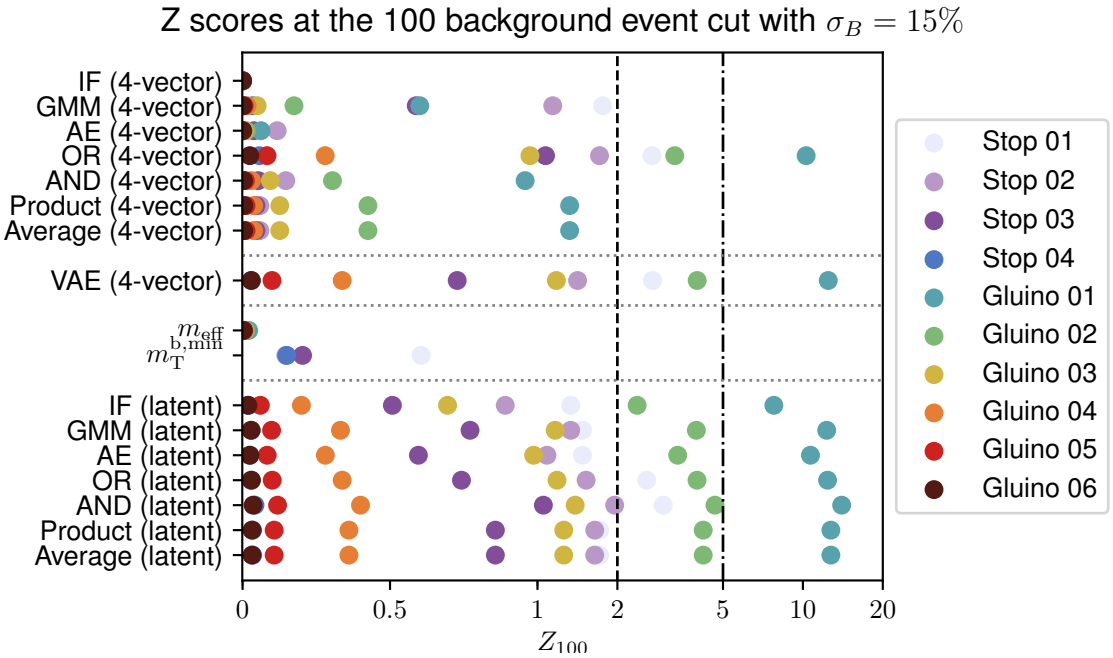

Figure 12. Z scores yielded from various algorithms applied to 4-vector components and latent space representations with a 15% relative systematic uncertainty applied. See table 1 for the signal definitions, and table 3 for the definitions of the algorithms.

JHEP09(2021)024

In our training, we have first employed the 4-vectors of the events as inputs for the ML algorithms. Our results are summarized in figure 11 assuming a non-existent systematic uncertainty. The IF, GMM and the AE on their own show a rather poor separation across all models. The VAE is unique due to its clustering of the information contained in the 4-vectors in its latent space, which groups non-linear combinations of the 4-vectors. These non-linear combinations may be viewed as new observables, therefore, the IF, GMM and AE algorithms may also be trained on the latent-space variables. By doing this, their performance dramatically increases, however, their performance does not exceed that of the VAE itself.

In addition to assessing the performance of individual ML methods, we also have explored combining these techniques in various ways. To this end, we have considered four different ways to combine their anomaly scores (section 3.6): AND, OR, product and averaging combinations. The performance of these combination depends on the signal and input representation (4-vector or latent-space variables). Using the 4-vector input representation, we find the best performance for the stop quark cases using the OR combination, while for the gluino events, the VAE gives the best result. When trained on the latent-space variables, this signal dependence drops out, and we find that the AND combination outperforms the other algorithms individually and the other combinations. The method that gives the best performance in the most signal-independent way is then:

- Train a VAE on 4-vectors of SM background events.
- Train a selection of ML techniques (which does not have to be limited to the techniques discussed in this work) on the latent-space representations of the 4-vectors of the SM events.
- Normalise their anomaly scores x_i and use $x^{\text{AND}} = \min(x_i)$ to determine the anomaly score for a given event.

We have compared our results to using the physical variable m_{eff} ($m_T^{b,\text{min}}$), which is often used in gluino (stop) searches to discriminate background from signal events. The techniques outlined in this paper outperform the use of this observable.

Acknowledgments

MvB acknowledge support from the Dutch NWO-I program 156, "Higgs as Probe and Portal", and the Christine Mohrmann Stipendium. R. RdA acknowledges partial funding/support from the Elusives ITN (Marie Skłodowska-Curie grant agreement No. 674896), and the Spanish MINECO grant "SOM Sabor y origen de la Materia" (FPA 2017-85985-P). MW and AL are supported by the ARC Discovery Project DP180102209.

JHEP09(2021)024

References

[1] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1 [arXiv:1207.7214] [INSPIRE].

[2] CMS collaboration, *Observation of a New Boson at a Mass of* 125 *GeV with the CMS Experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30 [arXiv:1207.7235] [INSPIRE].

[3] ATLAS collaboration, *Summary of the ATLAS experiment's sensitivity to supersymmetry after LHC Run 1 — interpreted in the phenomenological MSSM*, *JHEP* **10** (2015) 134 [arXiv:1508.06608] [INSPIRE].

[4] CMS collaboration, *Phenomenological MSSM interpretation of CMS searches in pp collisions at* $\sqrt{s} = 7$ *and* 8 *TeV*, *JHEP* **10** (2016) 129 [arXiv:1606.03577] [INSPIRE].

[5] GAMBIT collaboration, *Global fits of GUT-scale SUSY models with GAMBIT*, *Eur. Phys. J. C* **77** (2017) 824 [arXiv:1705.07935] [INSPIRE].

[6] GAMBIT collaboration, *A global fit of the MSSM with GAMBIT*, *Eur. Phys. J. C* **77** (2017) 879 [arXiv:1705.07917] [INSPIRE].

[7] GAMBIT collaboration, *Combined collider constraints on neutralinos and charginos*, *Eur. Phys. J. C* **79** (2019) 395 [arXiv:1809.02097] [INSPIRE].

[8] D0 collaboration, *Search for new physics in* $e\mu X$ *data at DØ using SLEUTH: A quasi-model-independent search strategy for new physics*, *Phys. Rev. D* **62** (2000) 092004 [hep-ex/0006011] [INSPIRE].

[9] D0 collaboration, *A Quasi model independent search for new physics at large transverse momentum*, *Phys. Rev. D* **64** (2001) 012004 [hep-ex/0011067] [INSPIRE].

[10] D0 collaboration, *A quasi-model-independent search for new high* p_T *physics at DO*, *Phys. Rev. Lett.* **86** (2001) 3712 [hep-ex/0011071] [INSPIRE].

[11] D0 collaboration, *Model independent search for new phenomena in* $p\bar{p}$ *collisions at* $\sqrt{s} = 1.96$ *TeV*, *Phys. Rev. D* **85** (2012) 092015 [arXiv:1108.5362] [INSPIRE].

[12] H1 collaboration, *A General search for new phenomena in ep scattering at HERA*, *Phys. Lett. B* **602** (2004) 14 [hep-ex/0408044] [INSPIRE].

[13] H1 collaboration, *A General Search for New Phenomena at HERA*, *Phys. Lett. B* **674** (2009) 257 [arXiv:0901.0507] [INSPIRE].

[14] CDF collaboration, *Model-Independent and Quasi-Model-Independent Search for New Physics at CDF*, *Phys. Rev. D* **78** (2008) 012002 [arXiv:0712.1311] [INSPIRE].

[15] CDF collaboration, *Global Search for New Physics with* $2.0\,fb^{-1}$ *at CDF*, *Phys. Rev. D* **79** (2009) 011101 [arXiv:0809.3781] [INSPIRE].

[16] G. Choudalakis, *On hypothesis testing, trials factor, hypertests and the BumpHunter*, in *PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, CERN, Geneva, Switzerland, 17–20 January 2011 [arXiv:1101.0390] [INSPIRE].

[17] 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** (2019) 120 [arXiv:1807.07447] [INSPIRE].

JHEP09(2021)024

[18] CMS collaboration, *MUSIC — An Automated Scan for Deviations between Data and Monte Carlo Simulation*, CMS-PAS-EXO-08-005 (2008).

[19] ATLAS collaboration, *A model independent general search for new phenomena with the ATLAS detector at* $\sqrt{s} = 13$ *TeV*, ATLAS-CONF-2017-001 (2017).

[20] P. Asadi, M.R. Buckley, A. DiFranzo, A. Monteux and D. Shih, *Digging Deeper for New Physics in the LHC Data*, *JHEP* **11** (2017) 194 [arXiv:1707.05783] [INSPIRE].

[21] CMS collaboration, *Model Unspecific Search for New Physics in pp Collisions at* $\sqrt{s} = 7$ *TeV*, CMS-PAS-EXO-10-021 (2011).

[22] R.T. D'Agnolo and A. Wulzer, *Learning New Physics from a Machine*, *Phys. Rev. D* **99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].

[23] R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer and M. Zanetti, *Learning multivariate new physics*, *Eur. Phys. J. C* **81** (2021) 89 [arXiv:1912.12155] [INSPIRE].

[24] M. Farina, Y. Nakai and D. Shih, *Searching for New Physics with Deep Autoencoders*, *Phys. Rev. D* **101** (2020) 075021 [arXiv:1808.08992] [INSPIRE].

[25] T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson, *QCD or What?*, *SciPost Phys.* **6** (2019) 030 [arXiv:1808.08979] [INSPIRE].

[26] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty Detection Meets Collider Physics*, *Phys. Rev. D* **101** (2020) 076015 [arXiv:1807.10261] [INSPIRE].

[27] 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** (2021) 27 [arXiv:2006.05432] [INSPIRE].

[28] S. Kim, Y. Choi and M. Lee, *Deep learning with support vector data description*, *Neurocomputing* **165** (2015) 111.

[29] O. Cerri, T.Q. Nguyen, M. Pierini, M. Spiropulu and J.-R. Vlimant, *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, *JHEP* **05** (2019) 036 [arXiv:1811.10276] [INSPIRE].

[30] O. Knapp, O. Cerri, G. Dissertori, T.Q. Nguyen, M. Pierini and J.-R. Vlimant, *Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark*, *Eur. Phys. J. Plus* **136** (2021) 236 [arXiv:2005.01598] [INSPIRE].

[31] M. Kuusela, T. Vatanen, E. Malmi, T. Raiko, T. Aaltonen and Y. Nagai, *Semi-Supervised Anomaly Detection — Towards Model-Independent Searches of New Physics*, *J. Phys. Conf. Ser.* **368** (2012) 012032 [arXiv:1112.3329] [INSPIRE].

[32] A. Andreassen, B. Nachman and D. Shih, *Simulation Assisted Likelihood-free Anomaly Detection*, *Phys. Rev. D* **101** (2020) 095004 [arXiv:2001.05001] [INSPIRE].

[33] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*, *Phys. Rev. D* **101** (2020) 075042 [arXiv:2001.04990] [INSPIRE].

[34] J.H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, *Phys. Rev. D* **99** (2019) 014038 [arXiv:1902.02634] [INSPIRE].

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

[36] J.H. Collins, K. Howe and B. Nachman, *Anomaly Detection for Resonant New Physics with Machine Learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

JHEP09(2021)024

[37] K. Benkendorfer, L.L. Pottier and B. Nachman, *Simulation-assisted decorrelation for resonant anomaly detection*, *Phys. Rev. D* **104** (2021) 035003 [arXiv:2009.02205] [INSPIRE].

[38] G. Brooijmans et al., *Les Houches 2019 Physics at TeV Colliders: New Physics Working Group Report*, in *11th Les Houches Workshop on Physics at TeV Colliders: PhysTeV Les Houches*, (2020) [arXiv:2002.12220] [INSPIRE].

[39] 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** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[40] NNPDF collaboration, *Parton distributions for the LHC Run II*, *JHEP* **04** (2015) 040 [arXiv:1410.8849] [INSPIRE].

[41] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[42] M.L. Mangano, M. Moretti, F. Piccinini and M. Treccani, *Matching matrix elements and shower evolution for top-quark production in hadronic collisions*, *JHEP* **01** (2007) 013 [hep-ph/0611129] [INSPIRE].

[43] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].

[44] https://github.com/melli1992/unsupervised_darkmachines.

[45] ATLAS collaboration, *Search for a scalar partner of the top quark in the all-hadronic $t\bar{t}$ plus missing transverse momentum final state at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **80** (2020) 737 [arXiv:2004.14060] [INSPIRE].

[46] ATLAS collaboration, *Search for a scalar partner of the top quark in the jets plus missing transverse momentum final state at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *JHEP* **12** (2017) 085 [arXiv:1709.04183] [INSPIRE].

[47] CMS collaboration, *Search for direct top squark pair production in events with one lepton, jets, and missing transverse momentum at 13 TeV with the CMS experiment*, *JHEP* **05** (2020) 032 [arXiv:1912.08887] [INSPIRE].

[48] F.T. Liu, K.M. Ting and Z. Zhou, *Isolation forest*, in *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422 (2008) [DOI].

[49] B.R. Preiss, *Data Structure and Algorithms: With Object-oriented Design Patterns in Java*, John Wiley & Sons (1999).

[50] G.J. McLachlan and D. Peel, *Finite mixture models*, John Wiley & Sons (2004).

[51] A.P. Dempster, N.M. Laird and D.B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, *J. Roy. Statist. Soc. B* **39** (1977) 1.

[52] P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, *Extracting and composing robust features with denoising autoencoders*, in *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, New York, NY, U.S.A., pp. 1096-1103, Association for Computing Machinery (2008) [DOI].

[53] M. Sakurada and T. Yairi, *Anomaly detection using autoencoders with nonlinear dimensionality reduction*, in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, MLSDA'14*, New York, NY, U.S.A., pp. 4–11, Association for Computing Machinery (2014) [DOI].

JHEP09(2021)024

[54] S. Kolouri, P.E. Pope, C.E. Martin and G.K. Rohde, *Sliced-wasserstein autoencoder: An embarrassingly simple generative model*, arXiv:1804.01947.

[55] D.P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, arXiv:1312.6114 [INSPIRE].

[56] S. Kullback, *Information Theory and Statistics*, Wiley, New York (1959).

[57] I. Higgins et al., *β-VAE: Learning basic visual concepts with a constrained variational framework*, in *ICLR 2017*, Toulon, France (2017).

[58] S. Otten et al., *Event Generation and Statistical Sampling for Physics with Deep Generative Models and a Density Information Buffer*, *Nature Commun.* **12** (2021) 2985 [arXiv:1901.00875] [INSPIRE].

[59] K.P. Murphy, *Machine learning: a probabilistic perspective*, MIT Press, Cambridge (2013).

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

[61] T. Yu and H. Zhu, *Hyper-parameter optimization: A review of algorithms and applications*, arXiv:2003.05689.

JHEP09(2021)024