
Variational Autoencoders for New Physics Mining at the Large Hadron Collider

Olmo Cerri¹, Thong Q. Nguyen¹, Maurizio Pierini², Maria Spiropulu¹, and Jean-Roch Vlimant¹

¹California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125

²CERN, CH-1211 Geneva, Switzerland

Abstract

Using variational autoencoders trained on known physics processes, we develop a one-side p-value test to isolate previously unseen processes as outlier events. Since the autoencoder training does not depend on any specific new physics signature, the proposed procedure has a weak dependence on underlying assumptions about the nature of new physics. An event selection based on this algorithm would be complementary to classic LHC searches, typically based on model-dependent hypothesis testing. Such an algorithm would deliver a list of anomalous events, that the experimental collaborations could further scrutinize and even release as a catalog, similarly to what is typically done in other scientific domains. Repeated patterns in this dataset could motivate new scenarios for beyond-the-standard-model physics and inspire new searches, to be performed on future data with traditional supervised approaches. Running in the trigger system of the LHC experiments, such an application could identify anomalous events that would be otherwise lost, extending the scientific reach of the LHC.

1 Introduction

One of the main motivations behind the construction of the CERN Large Hadron Collider (LHC) is the exploration of the high-energy frontier in search for *new physics* phenomena. This new physics could answer some of the standing fundamental questions in particle physics, e.g., the nature of dark matter or the origin of electroweak symmetry breaking. In LHC experiments, searches for physics beyond the Standard Model (BSM) are typically carried on as fully-supervised data analyses: assuming a new physics scenario of some kind, a search is structured as a hypothesis test based on profiled likelihood ratios [1]. These searches are said to be *model dependent*, since they depend on considering a specific new physics model.

Assuming that one is testing the *right* model, this approach is very effective in discovering a signal, as demonstrated by the LHC searches for the Standard Model (SM) Higgs boson [2, 3]. On the other hand, given the (so far) negative outcome of many BSM searches at the LHC and at other particle-physics experiments, it is possible that a future BSM model, if any, is not among those typically tested. The problem is more profound if analyzed in the context of the LHC big-data problem: at the LHC, 40 million proton-beam collisions are produced every second, but only 1000 collision events/sec can be stored by the ATLAS and CMS experiments, due to limited bandwidth, processing, and storage resources. It is possible to imagine BSM scenarios that would escape detection, simply because the corresponding new physics events would be rejected by a typical set of online selection algorithms.

Establishing alternative search methodologies with reduced model dependence is an important aspect of future LHC runs. Traditionally, this issue was addressed with so-called model-independent

searches, performed at the Tevatron [4, 5], at HERA [6], and at the LHC [7, 8], as discussed in Section 2.

In this paper, we propose to address this need by deploying an unsupervised algorithm in the online selection system of the LHC experiments. This algorithm would be trained on known SM processes and could be able to identify BSM events as anomalies. The selected events could be stored in a special stream, scrutinized by experts (e.g., to exclude detector malfunctioning that could explain the anomalies), and even released outside the experimental collaborations, in the form of an open-access catalog. The final goal of this application is to identify anomalous event topologies and inspire future supervised searches on data collected afterwards.

As a proof of principle, we consider the case of a typical single-lepton data stream, selected by the hardware-based L1 trigger system. On this stream of data, a variational autoencoder (VAE) is trained to compress the input event representation into a low-dimension latent space and then decompressed to return the shape parameters describing the probability density function (pdf) of each input quantity, given a point in the compressed space. The event distribution in a proper test statistic, namely part of the VAE loss function, is used to perform a one-side p-value test, to associate to each incoming event the probability of originating from known SM processes. A p-value threshold is applied to decide which event should be included into a low-rate anomalous-event data stream. In this work, we set the threshold such that ~ 30 events could be collected every day under current LHC operation conditions. In particular, we took as a reference 8 months of data taking per year, with an integrated luminosity of $\sim 40 \text{ fb}^{-1}$, as in 2016. Assuming an LHC duty cycle of $2/3$, this corresponds to an average instantaneous luminosity of $\sim 2.8 \times 10^{33} \text{ cm}^{-2} \text{ s}^{-1}$.

We then measure the BSM production cross section that would correspond to a signal excess of 100 event/month, as well as the one that would give a signal yield $\sim 1/3$ of the daily SM yield. For this, we consider a set of low-mass BSM resonances, decaying to one or more leptons and light enough to be challenging for the currently employed LHC trigger algorithms.

This paper is structured as follows: we discuss related works in Section 2. Section 3 gives a brief description of the dataset used. Section 4 describes the VAE model used in the study, as well as a set of fully-supervised classifiers used for performance comparison. Results are discussed in Section 5. In Section 6 we discuss how such an application could be used in a typical LHC experimental environment. Conclusions are given in Section 7.

2 Related Work

Model-independent searches for new physics have been performed at the Tevatron [4, 5], at HERA [6], and the LHC [7, 8]. These searches are based on the comparison of a large set of binned distributions to the prediction from Monte Carlo simulation, in search for bins exhibiting a deviation larger than some predefined threshold. While the effectiveness of this strategy in establishing a discovery has been matter of discussion, a recent study by the ATLAS collaboration [8] has rephrased this model-independent search strategy into a tool to identify interesting excesses, on which traditional analysis techniques could be performed on independent datasets (e.g., the data collected after running the model-independent analysis). This change of scope has the advantage of reducing the trial factor (i.e., the so-called *look-elsewhere* effect [9, 10]), which washes out the significance of an observed excess.

Our strategy is similar to what is proposed in Ref. [8], with two substantial differences: (i) we aim to monitor also those events that could be discarded by the online selection, by running the algorithm in the trigger system; (ii) we do so exploiting deep-learning-based anomaly detection techniques.

Recent works [11, 12, 13, 14] have investigated the use of machine-learning techniques to setup new strategies for BSM searches with minimal or no assumption on the specific new-physics scenario under investigation. In this work, we use variational autoencoders based on high-level features as a baseline. Previously, autoencoders have been used in collider physics for detector monitoring [15, 16] and event generation [17]. Autoencoders have also been explored to define a jet tagger that would identify new physics events with anomalous jets [18, 19], with a strategy similar to what we apply to the full event in this work.

3 Data samples

The dataset used for this study is a refined version of the high-level-feature (HLF) dataset used in Ref. [20]. Proton-proton collisions are generated using the PYTHIA8 event-generation library [21], fixing the center-of-mass energy to the LHC Run-II value (13 TeV) and the average number of overlapping collisions per beam crossing (pileup) to ~ 20 . These beam conditions loosely correspond to the LHC operating conditions in 2016.

Events generated by PYTHIA8 are processed with the DELPHES library [22], to emulate detector efficiency and resolution effects. We take as benchmark detector description the upgraded design of the CMS detector, foreseen for the High-Luminosity LHC phase [23]. In particular, we use the CMS HL-LHC detector card distributed with DELPHES. We run the DELPHES *particle-flow* (PF) algorithm, which combines information from different detector components to derive a list of reconstructed particles, the so-called PF candidates. For each particle, the algorithm returns the measured energy and flight direction. Each particle is associated to one of three classes: charged particles, photons, and neutral hadrons. In addition, lists of reconstructed electrons and muons are given.

Events are filtered at generation requiring an electron, muon, or tau lepton with $p_T > 22$ GeV. Once detector effects are taken into account through the DELPHES simulation, events are further selected requiring the presence of one reconstructed electron or muon with transverse momentum $p_T > 23$ GeV and a loose isolation requirement $\text{ISO} < 0.45$, where the isolation is computed as:

$$\text{ISO} = \frac{\sum_{p \neq q} p_T^p}{p_T^q}, \quad (1)$$

and the sum extends over all the photons, charged and neutral hadrons within a cone of size $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$ from the lepton.¹

The 21 considered HLF quantities are:

- The absolute value of the isolated-lepton transverse momentum p_T^ℓ .
- The three isolation quantities (CHPFISO, NEUPFISO, GAMMAPFISO) for the isolated lepton, computed with respect to charged particles, neutral hadrons and photons, respectively.
- The lepton charge.
- A Boolean flag (ISELE) set to 1 when the trigger lepton is an electron, 0 otherwise.
- S_T , i.e. the scalar sum of the p_T of all the jets, leptons, and photons in the event with $p_T > 30$ GeV and $|\eta| < 2.6$. Jets are clustered from the reconstructed PF candidates, using the FASTJET [24] implementation of the anti- k_T jet algorithm [25], with jet-size parameter $R=0.4$.
- The number of jets entering the S_T sum (N_J).
- The invariant mass of the set of jets entering the S_T sum (M_J).
- The number of these jets being identified as originating from a b quark (N_b).
- The missing transverse momentum, decomposed into its parallel ($p_{T,\parallel}^{\text{miss}}$) and orthogonal ($p_{T,\perp}^{\text{miss}}$) components with respect to the isolated lepton direction. The missing transverse momentum is defined as the negative sum of the PF-candidate p_T vectors:

$$\vec{p}_T^{\text{miss}} = - \sum_q \vec{p}_T^q. \quad (2)$$

- The transverse mass, M_T , of the isolated lepton ℓ and the E_T^{miss} system, defined as:

$$M_T = \sqrt{2p_T^\ell E_T^{\text{miss}} (1 - \cos \Delta\phi)}, \quad (3)$$

with $\Delta\phi$ the azimuth separation between the \vec{p}_T^ℓ and \vec{p}_T^{miss} vectors, and E_T^{miss} the absolute value of \vec{p}_T^{miss} .

¹As common for collider physics, we use a Cartesian coordinate system with the z axis oriented along the beam axis, the x axis on the horizontal plane, and the y axis oriented upward. The x and y axes define the transverse plane, while the z axis identifies the longitudinal direction. The azimuth angle ϕ is computed from the x axis. The polar angle θ is used to compute the pseudorapidity $\eta = -\log(\tan(\theta/2))$. We fix units such that $c = \hbar = 1$.

Table 1: Acceptance and trigger efficiency of SM processes and corresponding values for BSM benchmark models. For SM processes, we report the total cross section before the trigger, the expected number of events per month and the fraction in the SM cocktail. For BSM models, we compute the production cross section corresponding to an average of 100 events per month passing the acceptance and trigger requirements. The monthly event yield is computed assuming the conditions discussed in Section 1.

Standard Model processes					
Process	Acceptance	Trigger efficiency	Cross section [nb]	Events fraction	Event /month
W	55.6%	68%	58	59.2%	110M
QCD	0.08%	9.6%	$1.6 \cdot 10^5$	33.8%	63M
Z	16%	77%	20	6.7%	12M
$t\bar{t}$	37%	49%	0.7	0.3%	0.6M
BSM benchmark processes					
Process	Acceptance	Trigger efficiency	Total efficiency	Cross-section	100 events/month
$A \rightarrow 4\ell$	5%	98%	5%		436 fb
$LQ \rightarrow b\tau$	19%	62%	12%		166 fb
$h^0 \rightarrow \tau\tau$	9%	70%	6%		335 fb
$h^\pm \rightarrow \tau\nu$	18%	69%	12%		163 fb

- The number of selected muons (N_μ).
- The invariant mass of this set of muons (M_μ).
- The absolute value of the total transverse momentum of these muons ($p_{T,TOT}^\mu$).
- The number of selected electrons (N_e).
- The invariant mass of this set of electrons (M_e).
- The absolute value of the total transverse momentum of these electrons ($p_{T,TOT}^e$).
- The number of reconstructed charged hadrons.
- The number of reconstructed neutral hadrons.

This list of HLF quantities is not defined having in mind a specific BSM scenario. Instead, it is conceived to include relevant information to discriminate the various SM processes populating the single-lepton data stream. On the other hand, it is generic enough to allow (at least in principle) the identification of a large set of new physics scenarios.

Many SM processes would contribute to the considered single-lepton dataset. For simplicity, we restrict the list of relevant SM processes to the four with highest production cross section, namely:

- Inclusive W production, with $W \rightarrow \ell\nu$ ($\ell = e, \mu, \tau$).
- Inclusive Z production, with $Z \rightarrow \ell\ell$ ($\ell = e, \mu, \tau$).
- $t\bar{t}$ production.
- QCD multijet production.²

These samples are mixed to provide a SM cocktail dataset, which is then used to train autoencoder models and to tune the threshold requirement that defines what we consider an anomaly. The cocktail is built scaling down the high-statistics samples ($t\bar{t}$, W , and Z) to the lowest-statistics one (QCD, whose generation is the most computing-expensive), according to their production cross-section values (estimated at leading order with PYTHIA) and selection efficiency (shown in Tab. 1).

In addition, we consider the following BSM models to benchmark anomaly-detection capabilities:

- A leptoquark LQ with mass 80 GeV, decaying to a b quark and a τ lepton.

²To speed up the generation process for QCD events, we require $\sqrt{\hat{s}} > 10$ GeV, the fraction of QCD events with $\sqrt{\hat{s}} < 10$ GeV and producing a lepton within acceptance being negligible but computationally expensive.

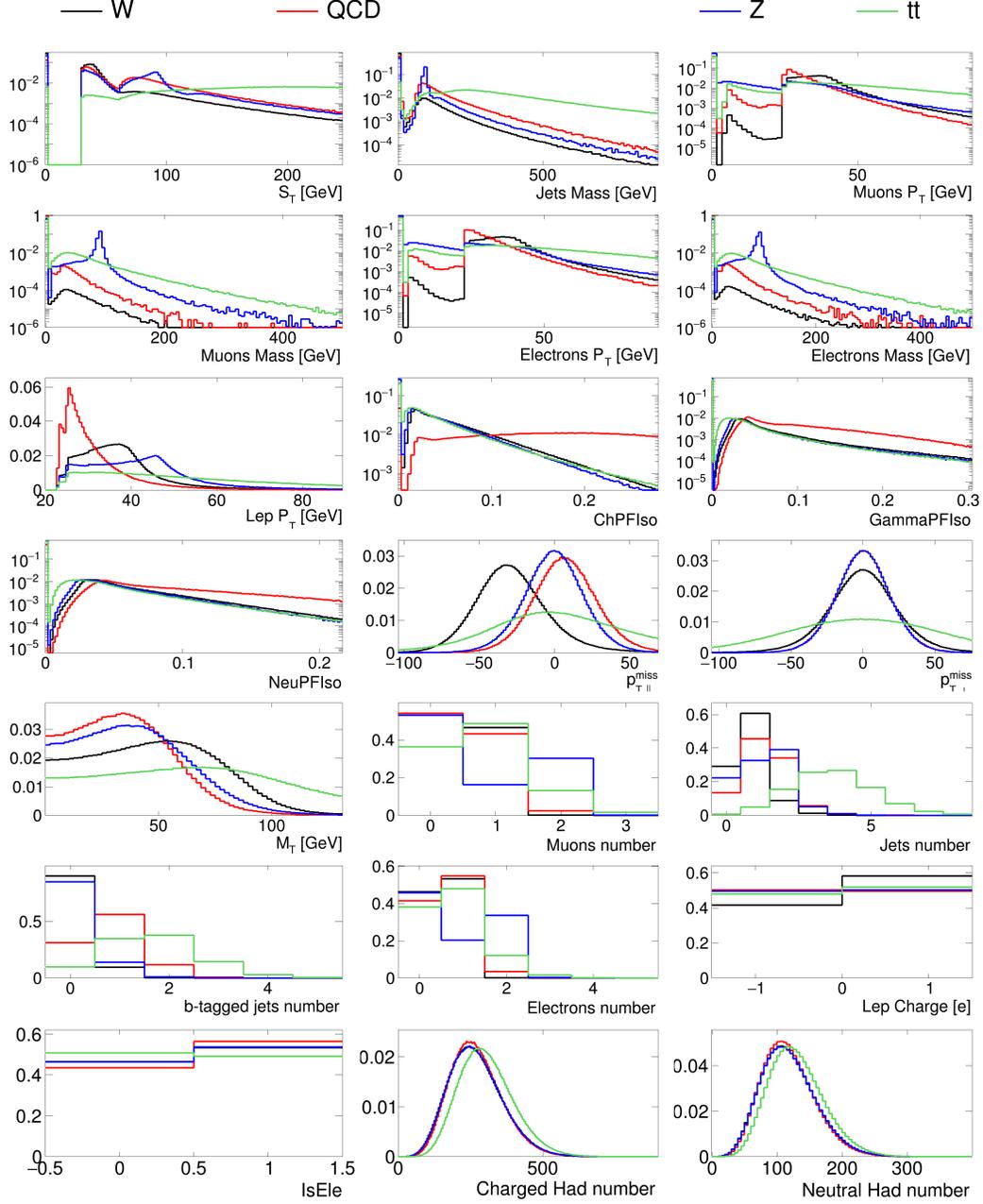


Figure 1: Distribution of the HLF quantities for the four considered SM processes.

- A neutral scalar boson with mass 50 GeV, decaying to two off-shell Z bosons, each forced to decay to two leptons: $A \rightarrow 4\ell$.
- A scalar boson with mass 60 GeV, decaying to two tau leptons: $h^0 \rightarrow \tau\tau$.
- A charged scalar boson with mass 60 GeV, decaying to a tau lepton and a neutrino: $h^\pm \rightarrow \tau\nu$.

For each BSM scenario, we consider any direct production mechanism implemented in PYTHIA8, including associate jet production. We list in Tab. 1 the leading-order production cross section and selection efficiency for each model.

Figures 1 and 2 show the distribution of HLF quantities for the SM processes and the BSM benchmark models, respectively.

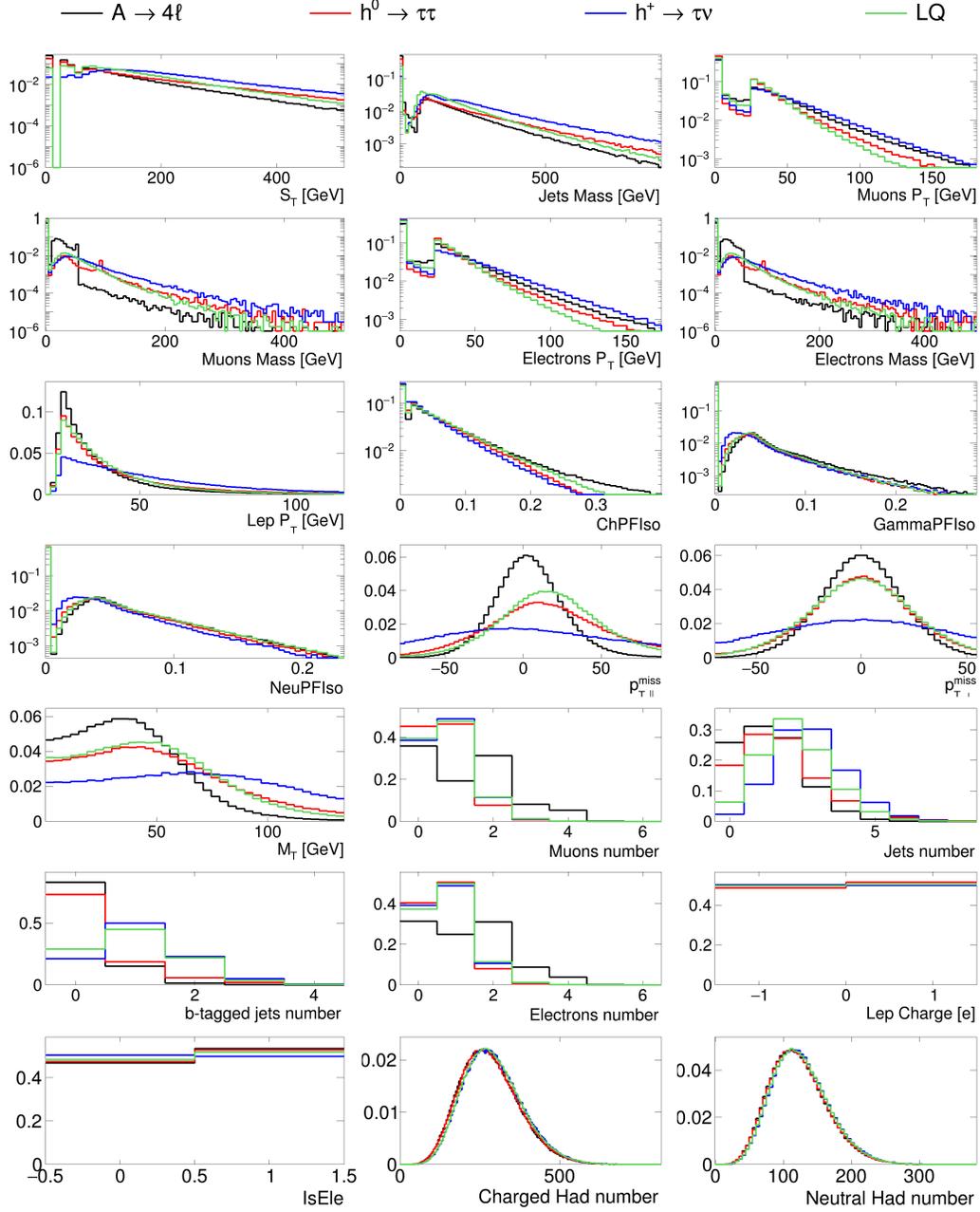


Figure 2: Distribution of the HLF quantities for the four considered BSM benchmark models.

4 Model description

We train Autoencoders (AEs) on the SM cocktail sample described in Section 3, taking as input the 21 HLF quantities listed there. The use of HLF quantities to represent events limits the model independence of the anomaly detection procedure. While the list of features is chosen to represent the main physics aspects of the considered SM processes and in no way tailored to specific BSM models, it is true that such a list might be more suitable for certain models than for others. In this respect, one cannot guarantee that the anomaly-detection performance observed on a given BSM model would generalize to any BSM scenario. We will address in a future work a possible solution to reduce the model carried by the input event representation.

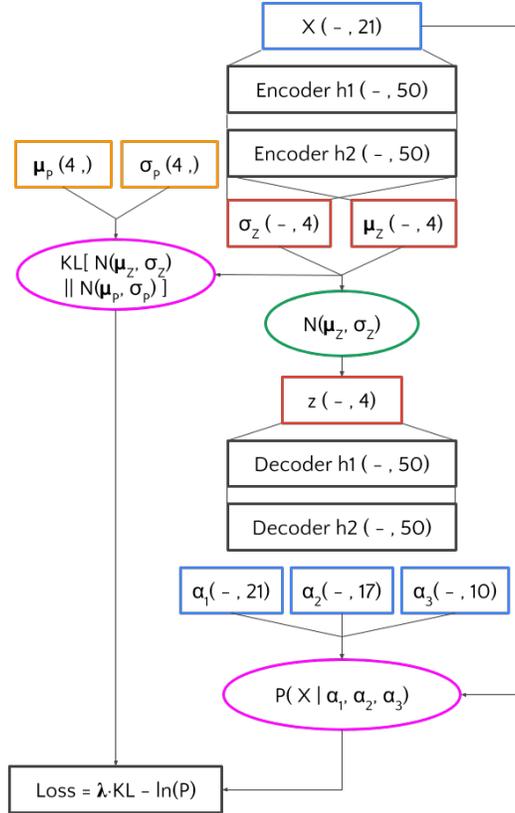


Figure 3: Schematics of the VAE used to perform anomaly detection, where X represent the input variables and z the latent space variables. The shape of each layer is reported in brackets.

In this section, we present both the best-performing autoencoder model and a set of supervised classifiers, trained to distinguish each of the four BSM benchmark models from SM events. We use the classification performance of these supervised algorithms as an estimate of the best performance that the VAE could get to.

4.1 Autoencoders

Autoencoders are algorithms that compress a given set of inputs variables in a latent space (encoding) and then, starting from the latent space, reconstruct the HLF input values (decoding). Autoencoders are used in the context of anomaly detection, associating a p-value to a given event through a quantification of the encoding-decoding distance.

In this work we focus on VAEs [26]. Unlike traditional AEs, VAEs return the event pdf in the latent and original space, instead of decoded values of the input quantities and the encoded point in the latent space. The functional form of the pdfs is specified through the loss function *a priori* and the pdfs' shape parameters are the output of a trainable function of the inputs. Such a function is the VAE itself and is determined during training.

We consider the VAE architecture shown in Fig. 3, characterized by a four-dimensional latent space. Each latent dimension is associated to a Gaussian pdf and its two degrees of freedom (mean μ and variance σ^2). The input layer consists of 21 nodes, corresponding to the 21 HLF quantities described in Section 3. This layer is connected to the hidden space through two hidden dense layers, each consisting of 50 neurons with ReLU activation functions. Two four-neuron layers are connected to the second hidden layer. Linear activation functions are used for the first of these four-neuron layers. Its nodes are interpreted as the mean values μ_z of the latent-space Gaussian pdfs. The nodes of the

second layer are activated by the functions:

$$\text{p-ISRLu}(x) = 1 + 5 \cdot 10^{-3} + \Theta(x)x + \Theta(-x)\frac{x}{\sqrt{1+x^2}}. \quad (4)$$

This activation, inspired by [27], has been chosen to increase training stability since it's strictly positive defined, non linear but does not involve exponential which might create instabilities in early epochs. These four nodes are interpreted as the σ_z parameters of the latent-space four-dimensional Gaussian. After several trials, the dimension of the latent space has been set to 4 in order to keep a good training stability without impacting the VAE performances. The decoding step originates from a point in the latent space, sampled according to the predicted pdf (green oval in Fig. 3). The coordinates of this point in the latent space are fed into a sequence of two hidden dense layers, each consisting of 50 neurons with ReLU activation functions. The last of these layers is connected to three dense layers of 21, 17, and 10 neurons, activated by linear, p-ISRLu and clipped-tanh functions, respectively. The clipped-tanh function is written as:

$$C_{\tanh}(x) = \frac{1}{2}(1 + 0.999 \cdot \tanh x). \quad (5)$$

The 48 output nodes represent the parameters of the pdfs describing the input HLF quantities, which enter the loss function to be minimized.

The VAE loss function Loss_{Tot} is a weighted sum of two pieces: the probability of the inputs given the predicted output pdf parameters ($\text{Loss}_{\text{reco}}$) and the Kullback-Leibler divergence (D_{KL}) between the latent space pdf and the prior:

$$\text{Loss}_{\text{Tot}} = \text{Loss}_{\text{reco}} + \lambda D_{\text{KL}}, \quad (6)$$

where λ is a free parameter set to 0.3. The prior chosen for the latent space is a 4-dim Gaussian with a diagonal covariance matrix. The means (μ_P) and the diagonal terms of the covariance matrix (σ_P) are free parameters of the algorithm and are optimized during the back-propagation. The Kullback-Leibler divergence between two Gaussian distributions has an analytic form. Hence, for each batch, D_{KL} can be expressed as:

$$\begin{aligned} D_{\text{KL}} &= \frac{1}{k} \sum_i D_{\text{KL}}(N(\mu_z^i, \sigma_z^i) \parallel N(\mu_P, \sigma_P)) \\ &= \frac{1}{2k} \sum_{i,j} \left(\sigma_P^j \sigma_z^{i,j} \right)^2 + \left(\frac{\mu_P^j - \mu_z^{i,j}}{\sigma_P^j} \right)^2 + \ln \frac{\sigma_P^j}{\sigma_z^{i,j}} - 1, \end{aligned} \quad (7)$$

where k is the batch size, i runs over the samples and j over the latent space dimensions. Similarly, $\text{Loss}_{\text{reco}}$ is the average likelihood of the inputs given the predicted α values:

$$\begin{aligned} \text{Loss}_{\text{reco}} &= -\frac{1}{k} \sum_i \ln(P(x \mid \alpha_1, \alpha_2, \alpha_3)) \\ &= -\frac{1}{k} \sum_{i,j} \ln(f_j(x_{i,j} \mid \alpha_1^{i,j}, \alpha_2^{i,j}, \alpha_3^{i,j})), \end{aligned} \quad (8)$$

where j runs over the input space dimensions, f_j is the functional form chose to describe the pdf of the j -th input space variable and $\alpha_m^{i,j}$ are the parameter of the function. Different functional forms have been chosen for f_j , to properly describe different classes of HLF distributions:

- **Clipped Log-normal + δ function:** used to describe S_T , M_J , p_T^μ , M_μ , p_T^e , M_e , p_T^ℓ , ChPFIso, NeuPFIso and GammaPFIso:

$$P(x \mid \alpha_1, \alpha_2, \alpha_3) = \begin{cases} \alpha_3 \delta(x) + \frac{1-\alpha_3}{x\alpha_2\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \alpha_1)^2}{2\alpha_2^2}\right) & \text{for } x \geq 10^{-4} \\ 0 & \text{for } x < 10^{-4} \end{cases}. \quad (9)$$

- **Gaussian:** used for $p_{T,\parallel}^{\text{miss}}$ and $p_{T,\perp}^{\text{miss}}$:

$$P(x \mid \alpha_1, \alpha_2) = \frac{1}{\alpha_2\sqrt{2\pi}} \exp\left(-\frac{(x - \alpha_1)^2}{2\alpha_2^2}\right). \quad (10)$$

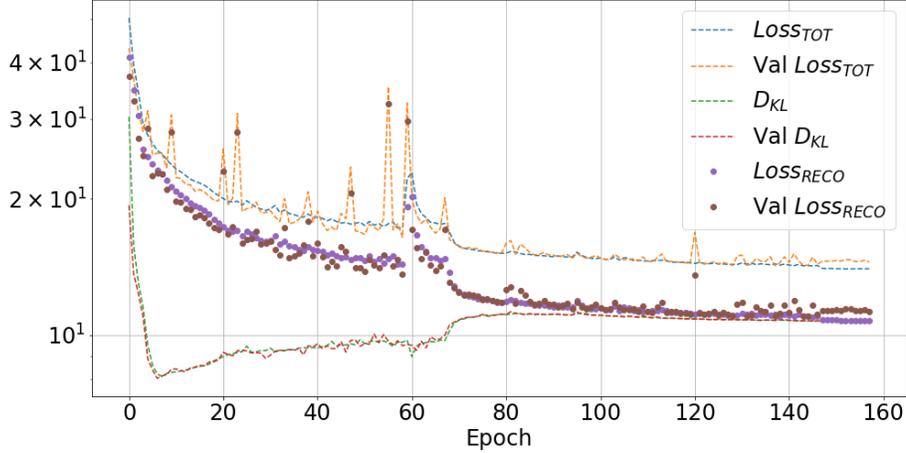


Figure 4: Training history for VAE. Total loss, reconstruction NLL ($Loss_{reoc}$) and KL divergence (D_{KL}) are shown separately for training and validation set though all the training epochs.

- **Truncated Gaussian:** a Gaussian function truncated for negative values and normalized to unit area for $X > 0$. Used to model M_T :

$$P(x | \alpha_1, \alpha_2) = \Theta(x) \cdot \frac{1 + 0.5 \cdot \left(1 + \operatorname{erf} \frac{-\alpha_1}{\alpha_2 \sqrt{2}}\right)}{\alpha_2 \sqrt{2\pi}} \exp\left(-\frac{(x - \alpha_1)^2}{2\alpha_2^2}\right). \quad (11)$$

- **Discrete truncated Gaussian:** like the truncated Gaussian, but normalized to be evaluated on integers (i.e. $\sum_{n=0}^{\infty} P(n) = 1$). This function is used to describe N_μ , N_e , N_b and N_J . It is written as:

$$P(n | \alpha_1, \alpha_2) = \Theta(x) \left[\operatorname{erf}\left(\frac{n + 0.5 - \alpha_1}{\alpha_2 \sqrt{2}}\right) - \operatorname{erf}\left(\frac{n - 0.5 - \alpha_1}{\alpha_2 \sqrt{2}}\right) \right] \mathcal{N}, \quad (12)$$

where the normalization factor \mathcal{N} is set to:

$$\mathcal{N} = 1 + \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{-0.5 - \alpha_1}{\alpha_2 \sqrt{2}}\right) \right). \quad (13)$$

- **Binomial:** used for IsE1e and lepton charge:

$$P(n | p) = \delta_{n,m} p + \delta_{n,l} (1 - p) \quad (14)$$

where m and l are the two possible values of the variable (0 or 1 for IsE1e and -1 or 1 for lepton charge) and $p = C_{\tanh}(\alpha_1)$

- **Poisson:** used for charged-particle and neutral-hadron multiplicities:

$$P(n | \mu) = \frac{\mu^n e^{-\mu}}{\Gamma(n + 1)} \quad (15)$$

where $\mu = \text{p-ISRLu}(\alpha_1)$.

The model is implemented in KERAS+TENSORFLOW [28, 29], trained with the Adam optimizer [30] on a SM dataset of 3.45M samples, equivalent to an integrated luminosity of $\sim 100 \text{ pb}^{-1}$. The SM validation dataset is made of 3.45M of statistically independent samples. Such a sample would be collected in about ten hours of continuous run, under the assumptions made in this study (see Section 1). In training, we fix the batch size to 1000. We use early stopping with patience set to 20 and $\delta_{\min} = 0.005$, and we progressively reduce the learning rate on plateau, with patience set to 8 and $\delta_{\min} = 0.01$.

Table 2: Classification performance of the four BDT classifiers, trained on the considered BSM benchmark models: area under ROC curve (AUC), and true positive rate (TPR) corresponding to a SM false positive rate $\epsilon_{SM} = 5.4 \cdot 10^{-6}$, equivalent to the acceptance rate chosen for the VAE.

Process	AUC	TPR [%]
$A \rightarrow 4\ell$	0.98	5.4
$LQ \rightarrow b\tau$	0.94	0.2
$h^0 \rightarrow \tau\tau$	0.90	0.1
$h^\pm \rightarrow \tau\nu$	0.97	0.3

While optimizing anomaly-detection performance, alternative architectures were tested. For instance, we increased or decreased the dimensionality of the latent space, we changed the value of λ in Eq.(6), we changed the number of neurons in the hidden layers, tried the RMSprop optimizer, and used plain Gaussian priors for the 21 input features. In addition, we tested the use of a vampprior [31]. While some of these alternative models improved the encoding-decoding capability of the VAE, no sizable improvement in anomaly-detection performance was observed. For simplicity, we limited our study to the architecture in Fig. 3 and dropped these alternative models.

The model’s training history is shown in Fig. 4. Figure 5 shows the comparison of the input and output distributions for the 21 HLF quantities in the validation dataset. While discrepancies are observed in some tail, good agreement is observed on the bulk of the distributions.

4.2 Supervised classifiers

For each of the four BSM benchmark models, we train a fully-supervised classifier, based on a Boosted Decision Tree (BDT). Each BDT receives as input the same 21 features used by the VAE and is trained on a labelled dataset consisting of the SM cocktail (the background) and one of the four BSM benchmark models (the signal). The implementation is done through the Gradient Boosted Regressor of scikit-learn library [32] with up to 150 estimators, minimum samples per leaf and maximum depth equal to 3 a learning rate of 0.1 and a tolerance of 10^{-4} on the validation loss function (choose to be the default deviance). Each BDT, tailored to a specific BSM model, is trained on 3.45M SM events and about 0.5M BSM events, consistently up-weighted in order to have the same impact on the loss function (i.e. the weights are 1 for SM events and ~ 7 for BSM ones, depending on the actual size of the BSM sample used). In addition, we experimented with fully-connected deep neural networks (DNNs) with two hidden layers. Despite trying different architectures, we didn’t find a configuration in which DNNs outperformed BDTs. We then decided to use the BDTs as a reference of fully-supervised discrimination capabilities.

Figure 6 shows the ROC curves obtained for the four BDTs. We summarize in Tab. 2 the classification performance of the four supervised BDTs, which set a qualitative upper limit for VAE’s results. Overall, the four models can be discriminated with good accuracy, with some loss of performance for those models sharing similarities with specific SM processes (e.g., $h^0 \rightarrow \tau\tau$ exhibiting single- and double-lepton topology with missing transverse energy, typical of $t\bar{t}$ events). In the table, we also quote the true-positive rate (TPR) corresponding to a SM false positive rate $\epsilon_{SM} = 5.4 \cdot 10^{-6}$. This value of the efficiency is the one needed for an average of 1000 SM events per month.

5 Results with VAE

An event is classified as anomalous whenever the associated loss, computed from the VAE output, is above a given threshold. Since no BSM signal was observed so far, it is reasonable to expect that a new-physics signal, if any, would be characterized by a low production cross section and/or features very similar to those of a SM process. In view of this, we decided to use a tight threshold value, in order to reduce as much as possible any SM contribution.

Figure 7 shows the distribution of $\text{Loss}_{\text{reco}}$ and D_{KL} loss components for the validation dataset. In both plots, the vertical line represents a lower threshold such that a $5.4 \cdot 10^{-6}$ of the SM events would be retained. This threshold value would result in ~ 1500 SM events to be selected every month, i.e., a daily rate of ~ 50 events, as illustrated in Table 3. The acceptance rate is calculated assuming

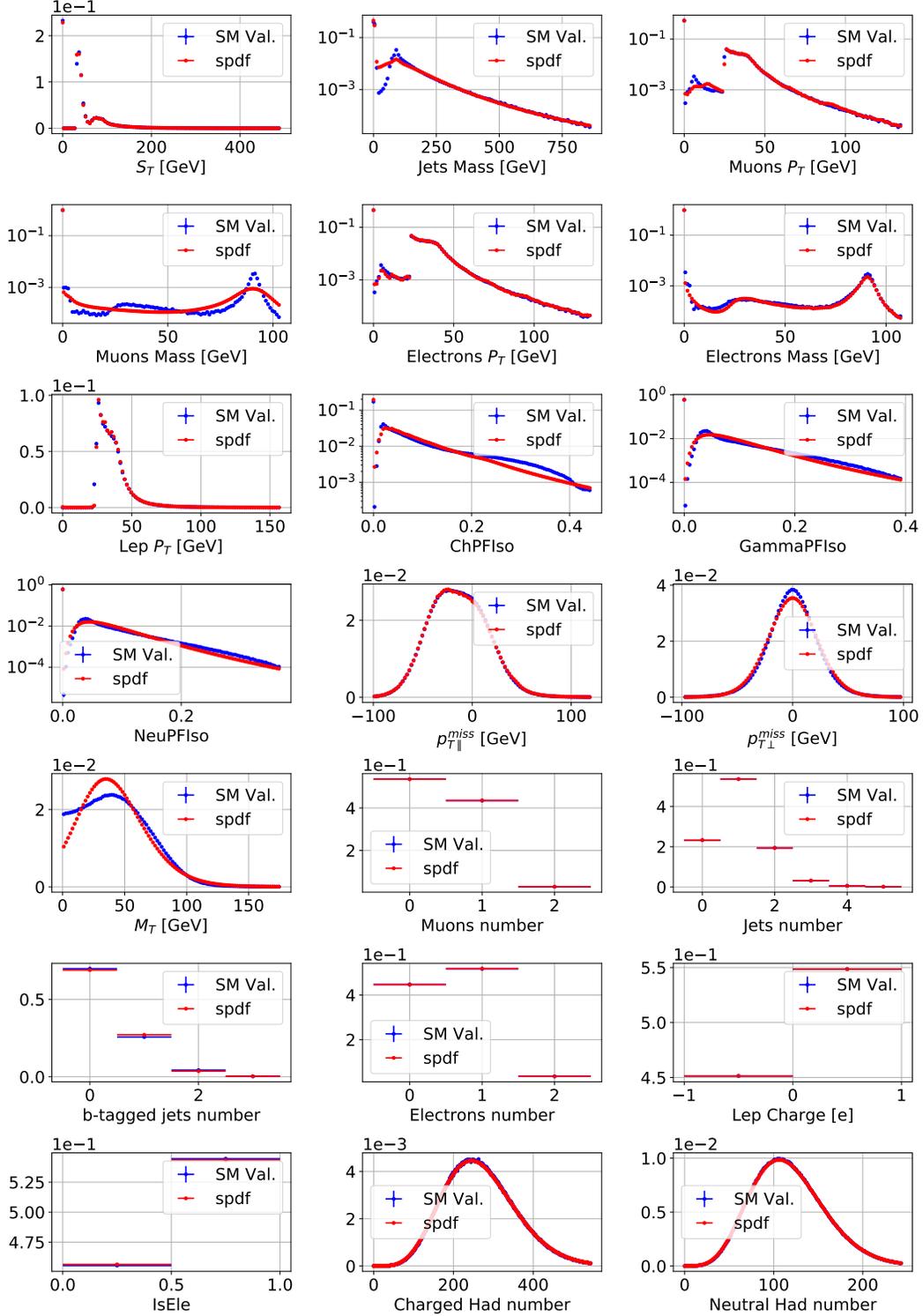


Figure 5: Comparison of input (blue) and output (red) probability distributions for the HLF quantities in the validation sample. Output distributions are obtained adding the predicted pdf for each event properly normalized.

the LHC running conditions listed in Section 1. Table 3 also reports the by-process VAE selection efficiency and the relative background composition of the selected sample.

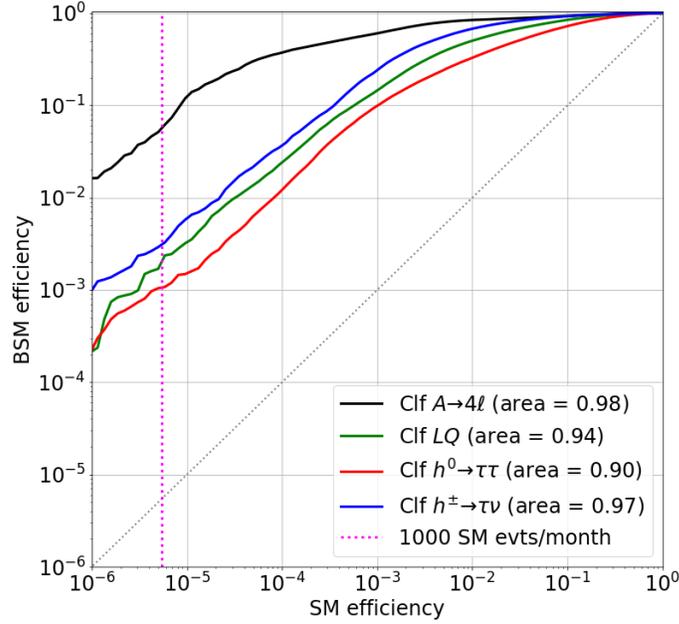


Figure 6: ROC curves for the fully-supervised BDT classifiers, optimized to separate each of the four BSM benchmark models from the SM cocktail dataset.

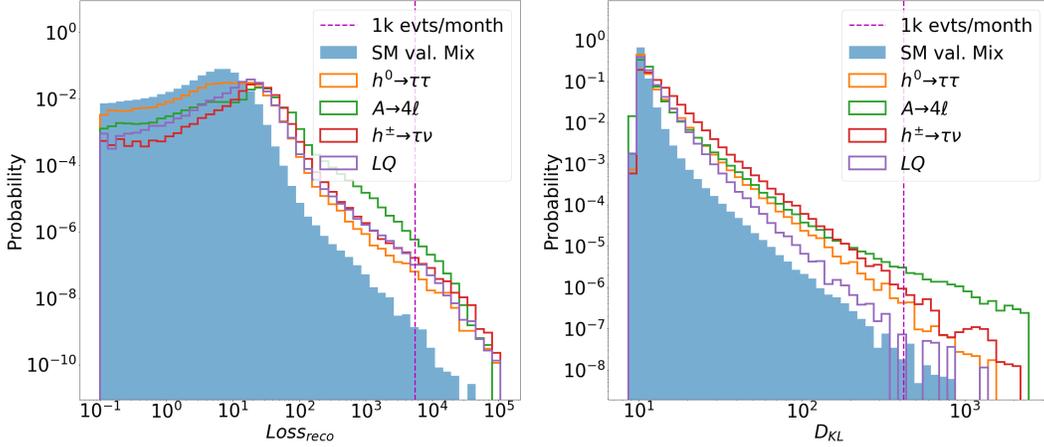


Figure 7: Distribution of the loss components: $Loss_{reco}$ (left) and D_{KL} (right) for the validation dataset. For comparison, the corresponding distribution for the SM processes and the four benchmark BSM models are shown. The vertical line represents a lower threshold such that $5.4 \cdot 10^{-6}$ of the SM events would be retained, equivalent to ~ 1500 expected SM events per month.

Figure 7 also shows $Loss_{reco}$ and D_{KL} distribution for the four benchmark BSM models. We observe that the discrimination power, loosely quantify by the integral of these distributions above threshold, is better for $Loss_{reco}$ than D_{KL} and that the impact of the D_{KL} term on $Loss_{Tot}$ discrimination is negligible. Anomalies are then defined as events laying on the right tail of the expected $Loss_{reco}$ distribution.

The left plot in Fig. 8 shows the ROC curves obtained from the $Loss_{reco}$ distribution of the four BSM benchmark models and the SM cocktail, compared to the corresponding BDT curves of Section 4.2.

Table 3: By-process acceptance rate for the anomaly detection algorithm described in the text, computed applying the lower threshold on $Loss_{reco}$ shown in Figure 7. The threshold is tuned such that a fraction of about $\epsilon_{SM} = 5.4 \cdot 10^{-6}$ of SM events would be accepted, corresponding to ~ 30 events/day and ~ 1000 events/month (assuming an average luminosity per month of 5 fb^{-1}). The sample composition refers to the subset of SM events accepted by the anomaly detection algorithm. All quoted uncertainties refer to 95% CL regions.

Standard Model processes			
Process	VAE selection	Sample composition	Event/month
W	$3.6 \pm 0.7 \cdot 10^{-6}$	32%	379 ± 74
QCD	$6.0 \pm 2.3 \cdot 10^{-6}$	29%	357 ± 143
Z	$21 \pm 3.5 \cdot 10^{-6}$	21%	256 ± 43
$t\bar{t}$	$400 \pm 9 \cdot 10^{-6}$	18%	212 ± 5
Tot			1204 ± 167

The right plot in Fig. 8 shows the p-value computed from the cocktail SM distribution, both for the SM events themselves (flat by construction) and for the four BSM processes. As the plot shows, BSM processes tend to concentrate at small p-values, which allows their identification as anomalies.

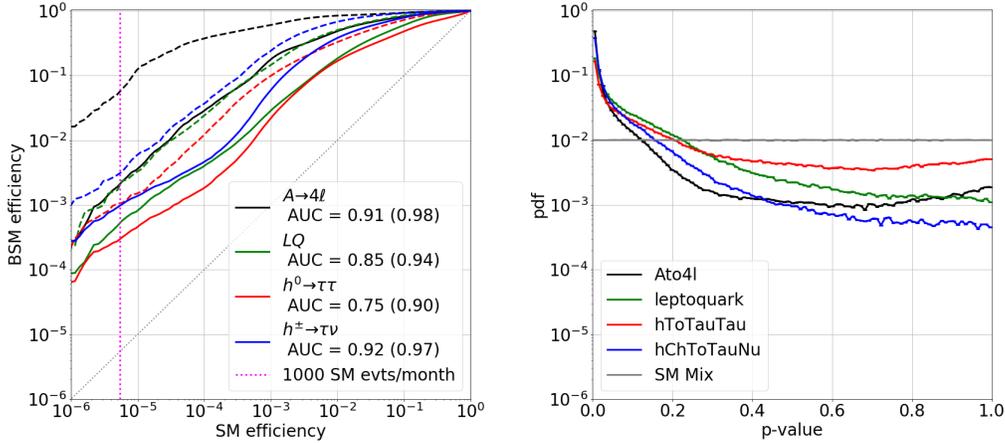


Figure 8: Left: ROC curves for the VAE trained only on SM mix (solid), compared to the corresponding curves for the four supervised BDT models (dashed) described in Section 4.2. Right: p-value distribution for the SM cocktail events and the four BSM benchmark processes.

Table 4 summarizes VAE's performance on the four BSM benchmark models. Together with the selection efficiency corresponding to $\epsilon_{SM} = 5.4 \cdot 10^{-6}$, the table reports the effective cross section (cross section after applying the trigger requirements) that would correspond to 100 selected events in a month (assuming an integrated luminosity of 5 fb^{-1}). Similarly, we quote the cross section that would result in a signal-to-background ratio of 1/3 on the sample of events selected by the VAE. The VAE can probe the four models down to relatively low cross section values, comparable to those that are typically probed in dedicated fully-supervised searches. As a comparison, Ref. [33] excludes a $LQ \rightarrow \tau b$ with a mass of 150 GeV and production cross section larger than $\sim 10 \text{ pb}$, using 4.8 fb^{-1} at a center-of-mass energy of 7 TeV, while most recent searches [34] only cover larger mass values.

6 How to deploy a VAE for BSM detection

The work presented in this paper suggests the possibility of deploying a VAE as a trigger algorithm associated to dedicated data streams. These trigger would isolate anomalous events, similarly to what was done by the CMS experiment at the beginning of the first LHC run. At that time, with early new physics signal being a possibility, the CMS experiment deployed online a set of algorithms

Table 4: Breakdown of BSM processes efficiency, and cross section values corresponding to 100 selected events in a month and to a signal-over-background ratio of 1/3. The monthly event yield is computed assuming an average luminosity per month of 5 fb^{-1} , computing by taking the LHC 2016 data delivery ($\sim 40 \text{ fb}^{-1}$ collected in 8 months). All quoted efficiencies are computed fixing the VAE loss threshold $\epsilon_{SM} = 5.4 \cdot 10^{-6}$. The quoted uncertainties correspond to a 95% CL region.

BSM benchmark processes			
Process	VAE selection efficiency	Cross-section 100 events/month [pb]	Cross-section S/B = 1/3 [pb]
$A \rightarrow 4\ell$	$2.8 \cdot 10^{-3}$	7.1	27
$LQ \rightarrow b\tau$	$6.7 \cdot 10^{-4}$	30	110
$h^0 \rightarrow \tau\tau$	$3.6 \cdot 10^{-4}$	55	210
$h^\pm \rightarrow \tau\nu$	$1.2 \cdot 10^{-3}$	17	65

(collectively called *hot line*) to select potentially interesting new-physics candidates. At that time, anomalies were characterized as events with high- p_T particles or high particle multiplicities, in line with the kind of early-discovery new physics scenarios considered at that time. The events populating the hot-line stream were immediately processed at the CERN computing center (as opposed to traditional physics streams, that are processed after 48 hours). The hot-line algorithms were tuned to collect O(10) events per day, which were then visually inspected by experts.

While the focus of the work presented in this paper is not an early discovery, the spirit of the application we propose would be similar: a set of VAEs deployed online would select a limited number of events every day. These events would be collected in a dedicated dataset and further analyzed. The analysis technique could go from visual inspection of the collisions to detailed studies of reconstructed objects, up to some kind of model-independent analysis of the collected dataset, e.g. a deep-learning implementation of a model-independent hypothesis testing [11] directly on the Loss distribution (provided a reliable sample of background-only data).

While a pure SM sample to train VAEs could only be obtained from Monte Carlo simulation, the presence of outlier contamination in the training sample has typically a tiny impact on performance. One could then imagine to train the VAE models on so-far collected data and use them on much larger dataset. In our study, we consider a training dataset of $\sim 100 \text{ pb}^{-1}$ and applied the VAE to a $\times 50$ larger dataset. One could even envision more frequent re-trainings (e.g., every factor $\times 10$ increase in integrated luminosity or in presence of substantial detector and/or accelerator condition changes). Such a training could happen offline on a dedicated dataset, e.g., deploying triggers randomly selecting events entering the last stage of the trigger system. The training could even happen online, assuming the availability of sufficient computing resources.

To demonstrate the feasibility of a train-on-data strategy, we enrich the dataset used in Section 4 with a signal contamination of $A \rightarrow 4\ell$ events. As a starting point, the amount of injected signal is tuned to a luminosity of 100 pb^{-1} and a cross section of 7.1 fb , corresponding to the value at which the VAE in Section 4 would select 100 $A \rightarrow 4\ell$ events in 5 fb^{-1} . This result into about 700 $A \rightarrow 4\ell$ events added to the training sample. The VAE is trained following the procedure outlined in Section 4 and its performance is compared to that obtained on a signal-free dataset of the same size. The comparison of the ROC curves for the two models is shown in Fig. 9. In the same figure, we show similar results, derived injecting a $\times 10$ and $\times 100$ signal contamination. A degradation of VAE’s performance is observed once the signal cross section is set to 710 pb (i.e., 100 times the sensitivity value found in Section 4). At that point, the contamination is so large that the signal becomes as abundant as $t\bar{t}$ events and would have easily detectable consequences. For comparison, at a production cross section of 27 pb a third of the events selected by the VAE in Section 4 would come from $A \rightarrow 4\ell$ production (see Table 4). And this would have negligible consequences on the training quality. This test shows that a robust anomaly-detecting VAE could be trained directly on data, even in presence of previously undetected (e.g., at Tevatron, 7 TeV and 8-TeV LHC) BSM signals.

7 Conclusions

We present a strategy to isolate potential BSM events produced by the LHC, using variational autoencoders trained on a reference SM sample. Such an algorithm could be used in the trigger

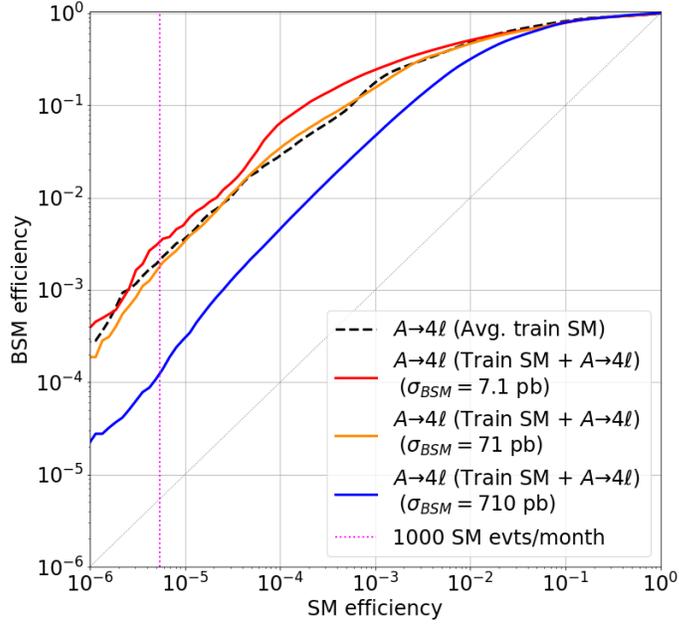


Figure 9: ROC curves for the VAE trained on SM contaminated with and without $A \rightarrow 4\mu$ contamination. Different level of contamination are reported corresponding to 0.02% ($\sigma = 7.15 \text{ pb}$ - equal to the estimated one to have 100 events per month), 0.19% ($\sigma = 71.5 \text{ pb}$) and 1.89% ($\sigma = 715 \text{ pb}$) of the training sample.

system of general-purpose LHC experiments to isolate recurrent anomalies, which might otherwise escape observation (e.g., being filtered out by a typical trigger selection). Taking as an example a single-lepton data stream, we show how such an algorithm could select datasets enriched with events originating from challenging BSM scenarios. We also discuss how the model training could happen directly on data, with no sizable performance loss.

The final outcome of the analysis would be a list of anomalous events, that the experimental collaborations could further scrutinize and even release as a catalog, similarly to what is typically done in other scientific domains. Repeated patterns in these events could motivate new scenarios for beyond-the-standard-model physics and inspire new searches, to be performed on future data with traditional supervised approaches.

We believe that such an application could help extending the physics reach of the current and next stages of the CERN LHC.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement n^o 772369) and the United States Department of Energy, Office of High Energy Physics Research under Caltech Contract No. de-sc0011925. This work was conducted at "iBanks", the AI GPU cluster at Caltech. We acknowledge NVIDIA, SuperMicro and the Kavli Foundation for their support of "iBanks".

References

- [1] ATLAS, CMS, LHC Higgs Combination Group Collaboration, *Procedure for the LHC Higgs boson search combination in summer 2011*, .

- [2] **ATLAS** Collaboration, G. Aad et al., *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett.* **B716** (2012) 1–29, [arXiv:1207.7214].
- [3] **CMS** Collaboration, S. Chatrchyan et al., *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys. Lett.* **B716** (2012) 30–61, [arXiv:1207.7235].
- [4] **CDF** Collaboration, T. Aaltonen et al., *Global Search for New Physics with 2.0 fb⁻¹ at CDF*, *Phys. Rev.* **D79** (2009) 011101, [arXiv:0809.3781].
- [5] **D0** Collaboration, V. M. Abazov et al., *Model independent search for new phenomena in pp collisions at $\sqrt{s} = 1.96$ TeV*, *Phys. Rev.* **D85** (2012) 092015, [arXiv:1108.5362].
- [6] **H1** Collaboration, F. D. Aaron et al., *A General Search for New Phenomena at HERA*, *Phys. Lett.* **B674** (2009) 257–268, [arXiv:0901.0507].
- [7] **CMS** Collaboration, *MUSiC, a Model Unspecific Search for New Physics, in pp Collisions at $\sqrt{s} = 8$ TeV*, Tech. Rep. CMS-PAS-EXO-14-016, CERN, Geneva, 2017.
- [8] **ATLAS** Collaboration, M. Aaboud et al., *A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment*, Submitted to: *Eur. Phys. J.* (2018) [arXiv:1807.07447].
- [9] L. Lyons, *Open statistical issues in particle physics*, *ArXiv e-prints* (Nov., 2008) [arXiv:0811.1663].
- [10] E. Gross and O. Vitells, *Trial factors for the look elsewhere effect in high energy physics*, *Eur. Phys. J.* **C70** (2010) 525–530, [arXiv:1005.1891].
- [11] R. T. D’Agnolo and A. Wulzer, *Learning New Physics from a Machine*, arXiv:1806.02350.
- [12] J. H. Collins, K. Howe, and B. Nachman, *CWoLa Hunting: Extending the Bump Hunt with Machine Learning*, arXiv:1805.02664.
- [13] A. De Simone and T. Jacques, *Guiding New Physics Searches with Unsupervised Learning*, arXiv:1807.06038.
- [14] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, *Novelty Detection Meets Collider Physics*, arXiv:1807.10261.
- [15] A. A. Pol, G. Cerminara, C. Germain, M. Pierini, and A. Seth, *Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider*, arXiv:1808.00911.
- [16] **CMS** Collaboration, *Anomaly detection using Deep Autoencoders for the assessment of the quality of the data acquired by the CMS experiment*, tech. rep., CERN, Geneva, Jul, 2018.
- [17] **ATLAS** Collaboration, *Deep generative models for fast shower simulation in ATLAS*, Tech. Rep. ATL-SOFT-PUB-2018-001, CERN, Geneva, Jul, 2018.
- [18] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, *QCD or What?*, arXiv:1808.08979.
- [19] M. Farina, Y. Nakai, and D. Shih, *Searching for New Physics with Deep Autoencoders*, arXiv:1808.08992.
- [20] T. Q. Nguyen et al., *Topology classification with deep learning to improve real-time event selection at the LHC*, arXiv:1807.00083.
- [21] T. Sjöstrand et al., *An Introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159–177, [arXiv:1410.3012].
- [22] **DELPHES 3** Collaboration, J. de Favereau et al., *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057, [arXiv:1307.6346].
- [23] **CMS** Collaboration, V. Khachatryan et al., *Technical Proposal for the Phase-II Upgrade of the CMS Detector*, Tech. Rep. CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, Geneva, Jun, 2015.
- [24] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, *Eur. Phys. J.* **C72** (2012) 1896, [arXiv:1111.6097].
- [25] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063, [arXiv:0802.1189].

- [26] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, *ArXiv e-prints* (Dec., 2013) [arXiv:1312.6114].
- [27] Wikipedia contributors, *Activation function — Wikipedia, the free encyclopedia*, 2018. [Online; accessed 25-November-2018].
- [28] F. Chollet et al., “Keras.” <https://github.com/fchollet/keras>, 2015.
- [29] M. Abadi et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. Software available from tensorflow.org.
- [30] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, *ArXiv e-prints* (Dec., 2014) [arXiv:1412.6980].
- [31] J. M. Tomczak and M. Welling, *VAE with a vampprior*, *CoRR* **abs/1705.07120** (2017) [arXiv:1705.07120].
- [32] F. Pedregosa et al., *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825–2830.
- [33] CMS Collaboration, S. Chatrchyan et al., *Search for pair production of third-generation leptoquarks and top squarks in pp collisions at $\sqrt{s} = 7$ TeV*, *Phys. Rev. Lett.* **110** (2013), no. 8 081801, [arXiv:1210.5629].
- [34] CMS Collaboration, A. M. Sirunyan et al., *Search for third-generation scalar leptoquarks and heavy right-handed neutrinos in final states with two tau leptons and two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **07** (2017) 121, [arXiv:1703.03995].