EPJ Web of Conferences **245**, 06039 (2020) https://doi.org/10.1051/epjconf/202024506039
CHEP 2019

New Physics Agnostic Selections For New Physics Searches

Kinga Anna Woźniak[1,2,*], *Olmo* Cerri[3], *Javier M.* Duarte[4], *Torsten* Möller[2], *Jennifer* Ngadiuba[1], *Thong Q.* Nguyen[3], *Maurizio* Pierini[1], *Maria* Spiropulu[3], and *Jean-Roch* Vlimant[3]

[1]CERN, CH-1211 Geneva, Switzerland
[2]University of Vienna, Austria
[3]California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125, United States
[4]Fermi National Accelerator Laboratory, Batavia, IL 60510, United States

Abstract. We discuss a model-independent strategy for boosting new physics searches with the help of an unsupervised anomaly detection algorithm. Prior to a search, each input event is preprocessed by the algorithm - a variational autoencoder (VAE). Based on the loss assigned to each event, input data can be split into a background control sample and a signal enriched sample. Following this strategy, one can enhance the sensitivity to new physics with no assumption on the underlying new physics signature. Our results show that a typical BSM search on the signal enriched group is more sensitive than an equivalent search on the original dataset.

1 Introduction

The search for physics beyond the Standard Model of particle physics (BSM) is one of the most important aspects of the Large Hadron Collider (LHC) physics program. Several model-independent strategies to search for new physics have been studied [1–9]. Inspired by these studies, we propose a strategy to boost BSM searches by applying a machine learning (ML) method to preprocess data in an unsupervised approach. The input to a BSM analysis usually has a large share of background events and a small share of potential signal events.
In our procedure, events are filtered before the analysis, such that a signal event is more likely to pass the selection than a background event. This way we enhance the signal-over-background ratio where the filter threshold defines how "anomalous" an event looks like. In our study, the definition of anomaly is learned training an unsupervised algorithm on a data sideband. This allows to limit the amount of assumptions on the underlying new physics.

2 Anomaly Selection with Variational Autoencoding

Basic idea of Autoencoding (AE)

The mechanism of an autoencoder can be described in three steps: First, one maps an input X of dimension $\mathbb{R}^N$ to a latent representation Z of dimension $\mathbb{R}^M$ where $M << N$ (bottleneck).

*e-mail: kinga.anna.wozniak@cern.ch

Figure 1. Basic autoencoding (left). Structure of a variational autoencoder (right).

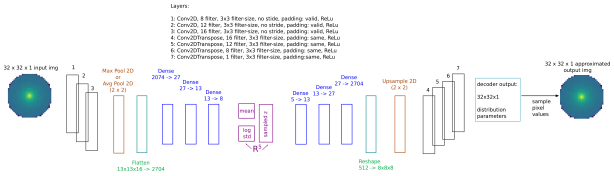

Figure 2. VAE architecure and implementation details.

Second, one reconstructs the input from that latent representation (Figure 1 left). Third, one computes the distance (in some metric) between input and reconstruction. If the distance of a sample i is small, the reconstruction worked well, i.e. sample i was "easy" to reconstruct. It is then considered a typical example of the kind of events processed by the AE at training time. Instead, if the distance is large, the sample was hard to reconstruct. This is interpreted as an indication that the example could be anomalous. No assumptions are made on *why* a distance is large, i.e. on *why* the sample is anomalous. On one hand, this guarantees a model-independent selection and could allow to retain sensitivity to untested scenarios. On the other hand, one retains little control on the physics behind the event selection. This procedure should be used to test unthought scenarios and weakly defined BSM frameworks.

Variational Autoencoder (VAE)

In our work we consider a *variational* autoencoder [10, 11] (Figure 1 right). Its encoder and decoder parts are *stochastic*, meaning that their outputs are probability distributions rather than pointwise values. The encoder maps an input X to a distribution in the latent space $q_\Phi(z|x)$. A pointwise latent variable value is then sampled from this distribution $z \sim q_\Phi$. The obtained z is then fed to the decoder, which maps it back to a distribution in the input space $p_\Theta(x|z)$. A pointwise prediction for the reconstruction can be obtained by sampling from that distribution $x' \sim p_\Theta$. A prior probability distribution is imposed on the latent space, giving the opportunity to add any deviation from the prior to the loss function. The architecture of the VAE is illustrated in Figure 2.

2.1 Optimization

The loss function of our VAE model consists of two terms (Eq. 1): A reconstruction loss and a loss on the latent space. The latent space penalty is the Kulback Leibler (KL) divergence [10] between the imposed prior and the encoder output. The term is scaled by a factor β [12] (in our model $\beta = 5 \cdot 10^{-4}$):

$$\text{Loss} = L_{\text{RECO}} + \beta \cdot D_{KL} \tag{1}$$

The reconstruction loss is defined on a Gaussian distribution of the values x reconstructed from the latent space z: $p_\theta(x|z) = \mathcal{N}(\mu(z), I)$, giving the minimization problem in Eq. 2.

$$L_{\text{RECO}} = -log\ p(X|\mu(z)) = -\sum_i log(e^{-\frac{(x_i-\mu_i(z))^2}{2}}) = \sum (x_i - \mu_i(z))^2 \tag{2}$$

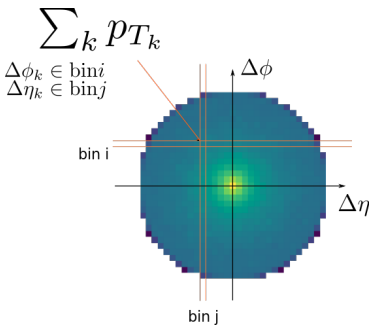

Figure 3. Illustration of a jet image used as the input data format.

The KL loss is defined on a Gaussian prior for variables in the latent space and computes the relative entropy between the prior and the approximated distribution (Eq. 3)

$$D_{KL}(q_\phi(z|x)||p_\lambda(z)) = \int p(z) \log \frac{q(z|x)}{p(z)} dz \tag{3}$$

2.2 Training

The training dataset consists of a sample of multijet events, generated with `PYTHIA8` [13] and processed with `DELPHES` [14] to emulate detector resolution effects and reconstruction efficiencies. The CMS Phase II [15] detector performances are assumed. The event reconstruction is performed running the DELPHES Particle Flow implementation. A Poisson profile is assumed for the pileup distribution, with the average number of collisions set to 40. Jets are clustered with the FASTJET [16] implementation of the anti-k_T jet algorithmm [17], with jet-size parameter R=0.8.
We train the VAE on a set of SM multijet events. The goal is to teach it what a SM event looks like, such that it returns a large loss when processing BSM events.
In our study we focus on dijet events. However, the VAE is trained on each jet independently and the procedure could be generalized to other event topologies with jets, with no need to retrain the model. We consider events with at least two jets having $pT > 30$ GeV and $|\eta| < 2.4$. Of those, we consider the first (jet 1) and second (jet 2) highest-pT jet in the event. The training set is built mixing jet 1 and jet 2 candidates from events belonging to the $|\Delta\eta|$ sideband, defined requiring $|\Delta\eta| > 1.4$ between the two jets. The information that they both belong to the same event is discarded for training (and inference) and will be needed only when their losses are recombined as described in Section 2.3.

Input format

From each jet sample in the input set, we take the 100 constituents with highest p_T. We construct a jet image, binned in $\Delta\eta$ and $\Delta\phi$ where each bin contains the corresponding summed particle p_T contributions as illustrated in Figure 3. Each image is made of 32x32 bins in a 0.8x0.8 wide window.

2.3 From Inference to Selection

Once trained, the VAE is applied to jet 1 and jet 2 of signal region's events ($|\Delta\eta| < 1.4$). We then collect the loss for the first jet L_1 and the loss for the second jet L_2. Both losses can be

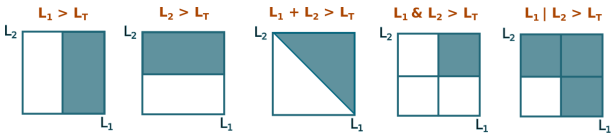

Figure 4. Loss combination strategies: Require only the loss of jet 1 or only the loss of jet 2 to be above the loss threshold L_T (first and second figure), require the *sum* of both jet losses to exceed L_T (third figure), require *both* losses to be above L_T (fourth figure) or require that *either* the loss of jet 1 *or* the loss of jet 2 be above L_T.

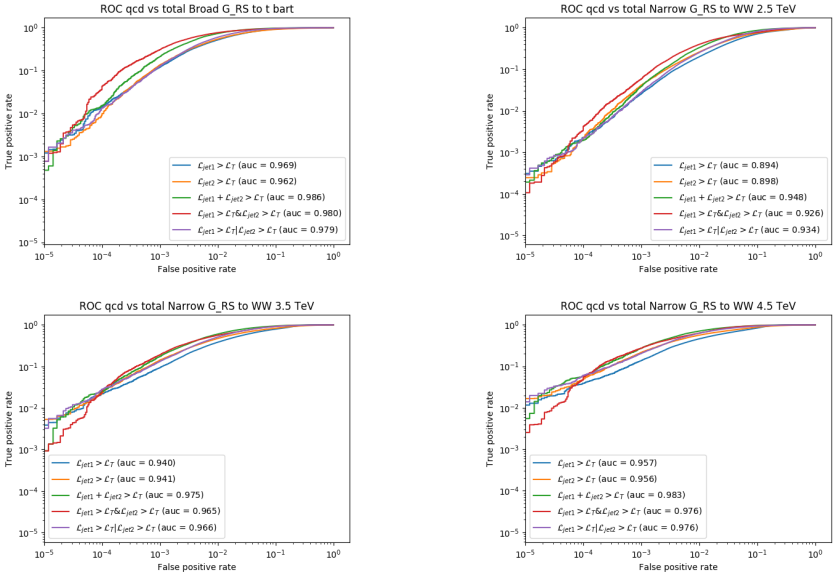

Figure 5. ROC curves for $G_{RS} \rightarrow t\bar{t}$ (top left), $G_{RS} \rightarrow WW2.5TeV$ (top right), $G_{RS} \rightarrow WW3.5TeV$ (bottom left) and $G_{RS} \rightarrow WW4.5TeV$ (bottom right) with different loss combination strategies (distinct lines in each plot)

combined to a total loss L in different ways. To decide if the event is normal or anomalous its total loss must be beyond a threshold L_T. The loss combination strategies are illustrated in Figure 4.

Given multiple strategies, we pick the one with the best ROC curve for a set of BSM benchmarks. In our study we evaluated several Randall-Sundrum Graviton [18] processes: $G_{RS} \rightarrow WW$ with $m_{jj} = \{1.5, 2.5, 3.5, 4.5\}$ TeV and $G_{RS} \rightarrow t\bar{t}$, broad and narrow resonance with $m_{jj} = 13$ TeV. ROC curves for some of those benchmarks with different loss combination strategies are given in Figure 5. The results show, that the loss strategy where both jet-losses are required to be above L_T (red line) yields the best outcome for signal efficiencies above 10^{-4}.

This procedure marks an event as normal or anomalous without the need for a prior definition of *anomaly*.

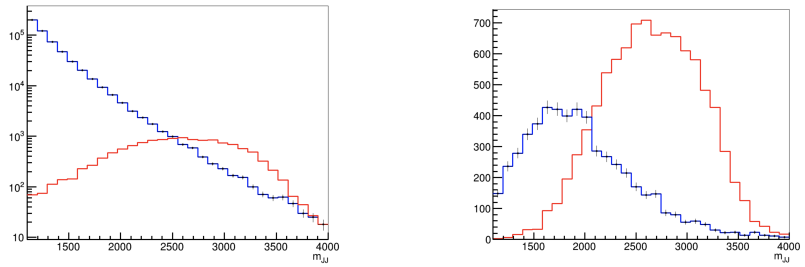

Figure 6. $G_{RS} \rightarrow t\bar{t}$ broad benchmark with $\sigma = 10$pb: Dijet mass spectrum before VAE cut (left) and dijet mass spectrum after VAE cut (right)

3 Application to supervised BSM Search

3.1 From Selection and Filtering to Analysis

Given an input dataset processed by the VAE, we split it into a background enriched and a signal enriched group: Events whose total loss is below the loss threshold L_T is considered to be the group enriched with standard physics. Events whose total loss is above L_T is considered to be the group enriched with anomalous events. The loss threshold L_T could be a fixed value or a function of some application-specific discriminating quantity, such as the dijet mass mjj for the case we consider here. Then we perform classic statistical analysis on the signal enriched group, taking advantage of the improved signal to background ratio.
In the results given in the following sections, the input dataset for training the VAE is a QCD simulation in a control region defined by $|\Delta\eta| > 1.4$. For analysis and inference, the signal region defined by $|\Delta\eta| < 1.4$ is used as described in Section 2. The variable of interest is picked to be the dijet mass m_{jj}.

3.2 VAE Boosted Supervised Searches on Tails

In this approach we use a fixed loss cut $L > L_T$ to preprocess events. We define L_T such that some defined fraction of events passes the filter (in the examples below it is 1%).

The advantage of this approach is that the signal to background ratio is indeed enhanced. However, the shape and the position of the background bulk region might render the search for a small signal statistically challenging in that mass range. Therefore, we perform a resonance search if the signal excess is not located in the bulk region of the background.

We tested the above described approach on a 'data mixture' of broad $G_{RS} \rightarrow t\bar{t}$ with $m_G = 3.5$TeV mass and QCD at different cross sections. At each cross section value, we performed two analyses: A traditional resonance search on the untreated input data mixture and a resonance search on the data mixture filtered by VAE loss cut. The results are shown in Figure 7: In this scenario, a 3σ excess could be turned into 5σ discovery. Note that the plots show a simple template fit assuming that we know the background shape. Our comparison thus is meaningful but we obtain an overestimated significance.

3.3 VAE Boosted Supervised Searches on bulks

In this approach we select events with a loss cut that is dependent on the variable of interest, in our case the dijet mass m_{jj}. For each dijet-mass value, a loss threshold $L_T(m_{jj})$ is identified,

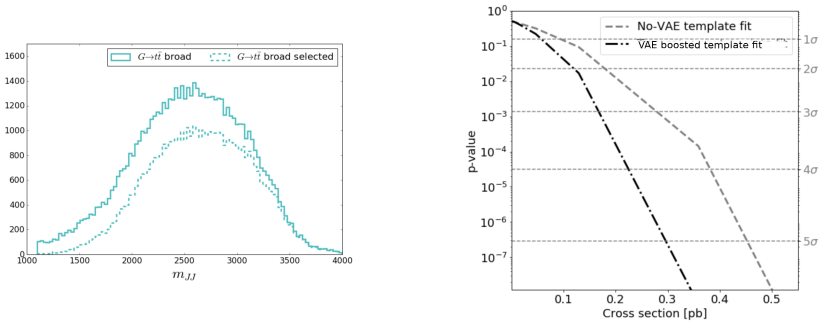

Figure 7. $G_{RS} \rightarrow t\bar{t}$ broad benchmark. Left plot: Dijet mass spectrum of total (solid line) vs selected (dashed line) events after loss cut. Right plot: Cross section and p-value significance for analysis on original dataset (grey line) and VAE loss cut filtered dataset (black line).

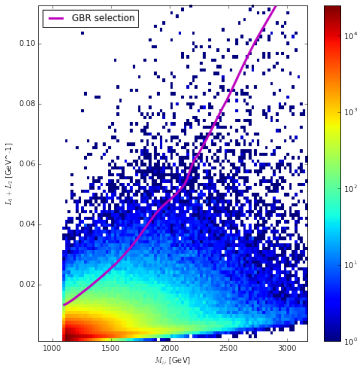

Figure 8. Quantile regression for loss threshold identification.

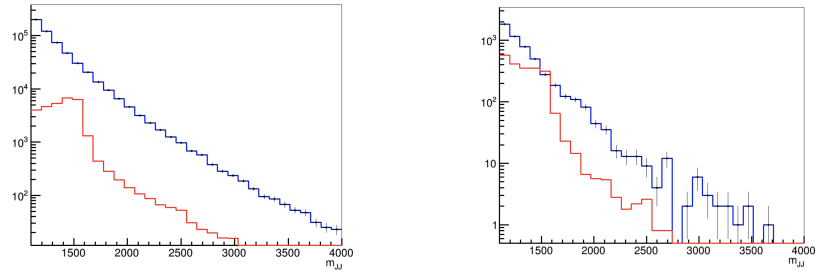

Figure 9. $G_{RS} \rightarrow WW$ broad benchmark with σ = 40pb: Dijet mass spectrum before VAE cut (left) and dijet mass spectrum after VAE cut (right).

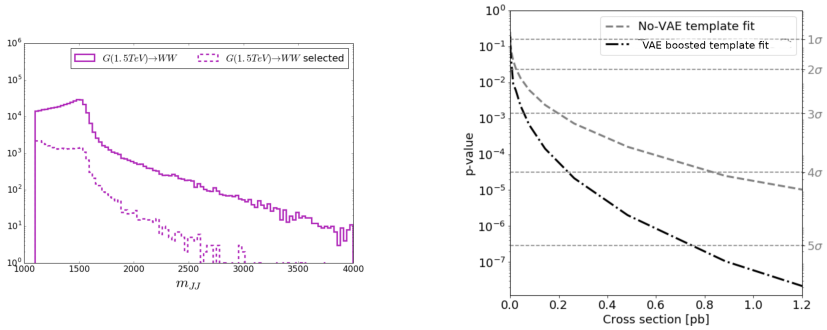

Figure 10. $G_{RS} \to WW$ benchmark. Left plot: Dijet mass spectrum of total (solid line) vs selected (dashed line) events after loss cut. Right plot: Cross section and p-value significance for analysis on original dataset (grey line) and VAE loss cut filtered dataset (black line).

that accepts a given constant fraction of events as anomalous if $L > L_T(m_{jj})$. We identify those thresholds by means of a quantile regression as shown in Figure 8.

The advantage of this approach is that it keeps the background unbiased. However, it reshapes the signal in an unfavorable way by penalising the tail. Therefore, we perform a resonance search if the excess is located in the bulk region of the distribution.

We tested the above described approach on a data mixture of $G_{RS} \to WW$ with m_G = 1.5TeV and QCD events at different cross sections. Again, we performed two analyses: A resonance search on the original and the VAE prefiltered dataset. The results are shown in Figure 10: In this scenario, a 3σ excess could be turned into 4σ. As in section 3.2, the resonance search fits are performed assuming that the background shape is known.

4 Conclusion

In this work we have discussed the use of unsupervised techniques in a typical BSM search at the LHC. We have shown that unsupervised machine learning can be utilized to select anomalous dijet BSM events in an input dataset mainly consisting of SM processes. We employed a variational autoencoder (VAE) that learns to identify previously unseen events and to distinguish them from processes that were used in the training procedure. We performed several statistical inference tests for observing processes induced by the decay of a hypothetical graviton in the presence of a dominating QCD background. Our studies show that the observation significance increases when the search is performed on a dataset that was previously filtered by cutting on the VAE loss. We achieve good sensitivity on the tail of dijet mass distributions which could benefit new physics studies focused on resonance searches. In the distribution bulks we observe a shape bias when applying our method. However, the sensitivity can be recovered with a background biased selection.

References

[1] T. Aaltonen et al. (CDF), Phys. Rev. **D79**, 011101 (2009), 0809.3781
[2] V.M. Abazov et al. (D0), Phys. Rev. **D85**, 092015 (2012), 1108.5362
[3] F.D. Aaron et al. (H1), Phys. Lett. **B674**, 257 (2009), 0901.0507
[4] Tech. Rep. CMS-PAS-EXO-14-016, CERN, Geneva (2017), https://cds.cern.ch/record/2256653

[5] M. Aaboud et al. (ATLAS), Eur. Phys. J. **C79**, 120 (2019), 1807.07447
[6] R. T. D'Agnolo and A. Wulzer, Phys. Rev. **D99**, 015014 (2019), 1806.02350
[7] A. De Simone and T. Jacques (2018), 1807.06038
[8] J.H. Collins, K. Howe, B. Nachman, Phys. Rev. Lett. **121**, 241803 (2018), 1805.02664
[9] J.H. Collins, K. Howe, B. Nachman, Phys. Rev. **D99**, 014038 (2019), 1902.02634
[10] D.P. Kingma, M. Welling, ArXiv e-prints (2013), 1312.6114
[11] J. An, S. Cho, Special Lecture on IE **2**, 1 (2015)
[12] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (OpenReview.net, 2017), https://openreview.net/forum?id=Sy2fzU9gl
[13] T. Sjöstrand et al., Comput. Phys. Commun. **191**, 159 (2015), 1410.3012
[14] J. de Favereau et al. (DELPHES 3), JHEP **02**, 057 (2014), 1307.6346
[15] D. Contardo, M. Klute, J. Mans, L. Silvestris, J. Butler, Tech. Rep. CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, CERN, Geneva (2015), https://cds.cern.ch/record/2020886
[16] M. Cacciari, G.P. Salam, G. Soyez, Eur. Phys. J. **C72**, 1896 (2012), 1111.6097
[17] M. Cacciari, G.P. Salam, G. Soyez, JHEP **04**, 063 (2008), 0802.1189
[18] L. Randall, R. Sundrum, Phys. Rev. Lett. **83**, 3370 (1999), hep-ph/9905221