PROCEEDINGS OF SCIENCE

Event-Level Anomaly Detection for Multijet BSM Searches with Probabilistic Autoencoders

I-M. Dinu[a,b,c,d,*]

[a] *Laboratoire de Phisique De Clermont, Av. Blaise Pascal, 63178 Aubiere, France*

[b] *Horia Hulubei National Institute of Physics and Nuclear Engineering, Str. Reactorului 30, Măgurele, Romania*

[c] *Ecole Doctorale des Sciences Fondamentales, Universite Clermont Auvergne, 34 Av. Carnot, 63000 Clermont-Ferrand, France*

[d] *Faculty of Physics, University of Bucharest Str. Atomistilor 405, Măgurele, Romania*

E-mail: ioan-mihail.dinu@cern.ch

Although most of Beyond Standard Model (BSM) searches are targeting specific theory models, there has always been a keen interest in the development of model-independent methods amongst the High Energy Physics (HEP) community. Machine Learning (ML) based anomaly detection stands among the latest up-and-coming avenues for creating model-agnostic BSM searches. The focus of this research is the design of anomalous event taggers based on autoencoder models. Alongside the signal discrimination power, a high priority is placed on both signal-model and background-model independence. To this end, the autoencoder is used in conjunction with a Normalizing Flow model tasked with latent space density estimation. Both event reconstruction error and latent representation likelihood are combined to mitigate the bias of the resulting event anomaly score. Overall this method is showing promising anomaly detection performance without losing much in terms of generalization power. On the multijet LHC Olympics data, it is consistently able to identify BSM signals, even in the challenging scenarios posed by the Black Box datasets, where the signal content is unknown.

The Ninth Annual Conference on Large Hadron Collider Physics - LHCP2021
7-12 June 2021
Online

*Speaker

PoS(LHCP2021)340

1. Introduction and Motivation

The increasing amount of data from collider experiments can open the door to new approaches to physics analysis. When paired with state of the art Machine Learning techniques, it becomes possible to view the BSM search problem as an anomaly detection exercise. Patterns that are shared throughout the sizable amount of well-understood data may be utilized to discriminate for potentially interesting events originating from yet-unknown processes.

This work illustrates a summary and the initial results of ongoing research attempting to apply model-independent ML-based anomaly detection to jet physics. For this application, the LHC Olympics Challenge [1] datasets have been used as a benchmark for this method's performance. The datasets offer millions of multi-jet simulated events represented by the four-vectors of each associated constituent particle. Here, plenty QCD dijet events are given as background examples. On the other side, there are several "Black Box" datasets containing trace amounts of signal originating from an unknown BSM process. To make matters more difficult, the background events in those black boxes are modelled slightly differently than the ones available in the reference dataset. This slight disparity tries to account for potential variations in detector calibration and event reconstruction that are often encountered in real data.

Data preparation for this study consisted in applying jet-clustering algorithms followed by custom feature engineering. Thus, obtaining information about each jet's kinematics and substructure, alongside event-level features such as the jet multiplicity and the combined mass of the jets.

2. The Probabilistic Autoencoder

During the last decade, the autoencoder has been one of the prevailing neural network architectures for anomaly detection tasks. This framework attempts to learn a lower-dimensional representation of the training dataset and use it to reconstruct the original inputs. Any new data point can be fed to this model and the error in its reconstruction could be used as a metric for how *different* the input is from the training dataset.

On the other hand, one may also frame anomaly detection in terms of likelihood. Applying clustering methods to the training data yields the possibility of estimating its density (either implicitly or explicitly). This has been historically a computationally intensive task, especially for high dimensional data. However, recent developments in ML-based density estimation, such as the Normalizing Flow (NF) [2], have made it much more approachable and effective. Employing a chain of triangular maps (denoted by $\mathbf{b}_\gamma$ throughout this paper), the NF model learns a bijective transformation between the latent space and a multidimensional normal distribution. For any data point $\vec{z}$ there will be an equivalent point $\vec{u}$ belonging to a multivariate Gaussian distribution. With this bijection: $\mathbf{b}_\gamma(\vec{z}) = \vec{u}$, finding the data density in the original space only requires the determinant of its Jacobian: $p(\vec{z}) = p(\vec{u}) \det |\mathcal{J}_\gamma|^{-1}$.

Each of these methods has its advantages and shortcomings and there is no decisive argument to be made in using one over the other for those particular datasets. Those methods, however, don't have to be mutually exclusive. Recent research [3] suggests that better results could be obtained by combining such approaches. The Probabilistic Autoencoder (PAE) [4] attempts to do just that and achieves it by adding a density estimation model to the latent space of a standard autoencoder.

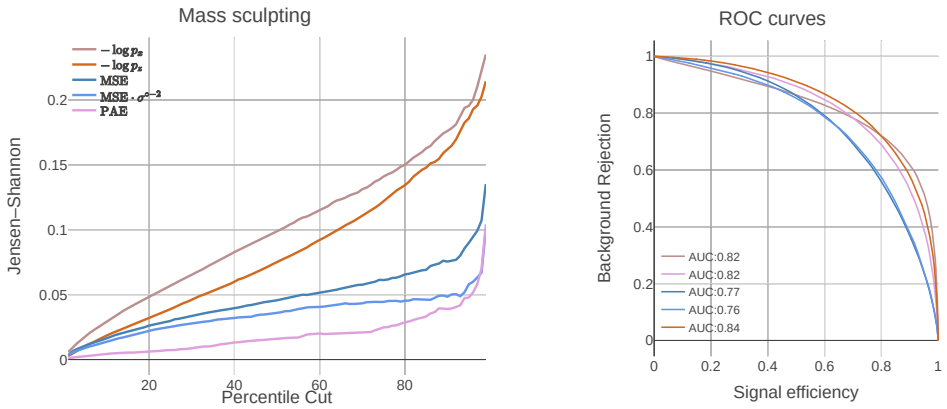
With such a framework, the likelihood of the inputs is computable, through approximation, while exploiting both reconstruction error information and latent representation likelihood. Equation 1 shows the full expression of the PAE log-likelihood, which the original paper [4] suggests as a good anomaly score.

$$\ln p(\vec{x}) \approx -\frac{1}{2}||\vec{x} - \vec{x}'||^2 \vec{\sigma}^{\circ -2} - \frac{1}{2} b_\gamma(\vec{z})^2 + \ln|\det \mathcal{J}_\gamma| \tag{1}$$

Equation 1's first term represents the reconstruction error relative to the average training reconstruction error $\vec{\sigma}$. The following terms encompass the likelihood of the latent representation, as estimated by the NF model.

The general strategy is to apply a high threshold cut on the anomaly score, keeping only the high-end of this distribution. Events passing this cut undergo further inspection. In the case of dijet events, the dijet mass (m_{jj}) distribution is compared to its pre-cut state in search for a peak, which would indicate the presence of a signal.

In order to avoid biasing the anomaly score, the training data is transformed, making all the features uniformly distributed while keeping the correlations between them. At the same time, since m_{jj} is used for bump hunting, all training events are weighted based on this feature. This normalization counterbalances the model's tendency towards favouring the mass range with the highest abundance of data points.



(a) The JS-divergence mass sculpting computed at every percentile cut on the anomaly score

(b) Signal efficiency vs background rejection plot. The legend label shows the area under each curve.

Figure 1: Performance metrics for several potential anomaly scores. The light brown curve is associated with the log-likelihood of a NF model trained directly on the input features, while, for the dark orange curve, the NF model was trained on the latent space. Blue shades relate to the Mean Squared Error (MSE) of the autoencoder, where dark blue is the MSE relative to the average training reconstruction error. Finally, in purple, there is the PAE log-likelihood anomaly score, shown previously in equation 1.

3. Results

Several PAE model configurations have been benchmarked for anomaly detection performance and the amount of m_{jj} bias. Mass sculpting was quantified using the Jensen-Shannon (JS) divergence between the mass spectra before and after the anomaly score cut. The lower the JS-divergence,

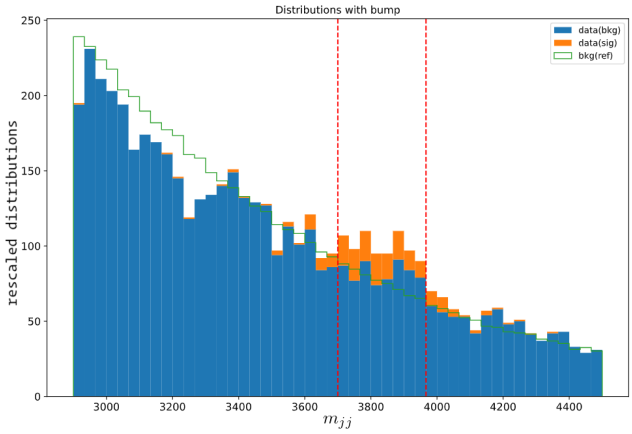


Figure 2: Results on the "Black Box 1" dataset from the LHC Olympics challenge. The m_{jj} distribution before the cut is shown with a green line, while the solid colour histogram shows what happens after the cut. Signal content is represented in orange and background in blue Analysis was performed in on the blinded dataset, signal labels only being accessed for validation.

the more confidence can be invested in any resonance found with this method. Given the flexibility of the PAE, there are actually several anomaly score candidates to be considered. A slew of those candidates was tested, under the same conditions, with results being revealed in figure 1.

As figure 1a shows, the PAE's approximation of the log-likelihood introduces the least amount of bias, thus allowing better sensitivity to lower amounts of signal. When considering the area under the receiver operating characteristic (ROC) curves, in figure 1b, it is apparent that none of those anomaly scores offers discrimination power that would be on-par with a supervised method. But, the PAE anomaly score manages to stay close to the better-performing alternatives while avoiding the significant bias issues that they raise.

The entire workflow has been applied to the "Black Box 1" dataset. With a cut on the 99th percentile of the PAE score and using the "bump hunter" method [5, 6], the outcome is presented in figure 2. The method correctly identifies the signal resonance at 3.8 TeV. Cutting on the anomaly score improves the initial $S/B \approx 0.08\%$ by a factor of **14**. The tightness of this cut introduces notable fluctuations of the m_{jj} distribution, but loosening the cut makes the signal peak increasingly more difficult to find.

4. Conclusions

In an attempt to combine both reconstruction-based and density-based anomaly detection, this study has evaluated the performance of the Probabilistic Autoencoder neural network ensemble in the context of event-level anomalous jet tagging. Among the many potential anomaly scores allowed by this model, the PAE's log-likelihood estimation showed the best compromise between higher performance and lower bias. When benchmarked on the difficult scenarios posed by the datasets from the LHC Olympics Challenge, this method was able to successfully identify a very faint unknown signal. The results obtained so far show real promise and, with a bit more polish, this method could become a powerful tool in a model-independent BSM search.

References

[1] G. Kasieczka, B. Nachman, D. Shih, O. Amram, A. Andreassen, K. Benkendorfer et al., *The lhc olympics 2020: A community challenge for anomaly detection in high energy physics*, 2101.08320.

[2] D.J. Rezende and S. Mohamed, *Variational inference with normalizing flows*, in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 1530–1538, JMLR.org, 2015 [1505.05770].

[3] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty detection meets collider physics*, *Physical Review D* **101** (2020) .

[4] V. Böhm and U. Seljak, *Probabilistic auto-encoder*, *CoRR* **abs/2006.05479** (2020) [2006.05479].

[5] G. Choudalakis, *On hypothesis testing, trials factor, hypertests and the BumpHunter*, in *PHYSTAT 2011*, 1, 2011 [1101.0390].

[6] L. VASLIN, "pybumphunter." https://github.com/scikit-hep/pyBumpHunter, 2021.

PoS(LHCP2021)340