Online-compatible unsupervised nonresonant anomaly detection

Vinicius Mikuni, Benjamin Nachman, and David Shih

Phys. Rev. D **105**, 055006 — Published 8 March 2022

DOI: 10.1103/PhysRevD.105.055006

Online-compatible Unsupervised Non-resonant Anomaly Detection

Vinicius Mikuni,[1,*] Benjamin Nachman,[2,3,†] and David Shih[4,‡]

[1] *National Energy Research Scientific Computing Center, Berkeley Lab, Berkeley, CA 94720, USA*
[2] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[3] *Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*
[4] *NHETC, Department of Physics & Astronomy, Rutgers University, Piscataway, NJ 08854, USA*

There is a growing need for anomaly detection methods that can broaden the search for new particles in a model-agnostic manner. Most proposals for new methods focus exclusively on signal sensitivity. However, it is not enough to select anomalous events - there must also be a strategy to provide context to the selected events. We propose the first complete strategy for unsupervised detection of non-resonant anomalies that includes both signal sensitivity and a data-driven method for background estimation. Our technique is built out of two simultaneously-trained autoencoders that are forced to be decorrelated from each other. This method can be deployed offline for non-resonant anomaly detection and is also the first complete online-compatible anomaly detection strategy. We show that our method achieves excellent performance on a variety of signals prepared for the ADC2021 data challenge.

I. INTRODUCTION

Despite the compelling indirect evidence for new fundamental particles from astrophysical and other observations, no direct discoveries have been confirmed since the identification of the Higgs boson [1, 2]. This means that the new physics is either rare, inaccessible, or we are looking in the wrong place for it. This last possibility has motivated a new anomaly detection research program at particle colliders by which search strategies are constructed with less model dependence than previous approaches. Many of these new methods employ modern machine learning to achieve broad sensitivity to unforeseen scenarios [3–60] (list from Ref. [61]).

A complete anomaly detection algorithm is required to have two attributes: it should be sensitive to anomalous events and it should be possible to estimate the rate of Standard Model (SM) events that are labeled as anomalous (false positive rate) [6]. Complete anomaly detection methods have so far primarily focused on *resonant* anomalies, where data sidebands can be used as reference samples to both construct signal-sensitive classifiers and to estimate the SM background [3–12, 18, 19].

Much less well-explored so far has been complete anomaly detection methods for *non-resonant* anomalies. One widely studied approach based on unsupervised learning that does not require the new physics to be resonant is the autoencoder [13–54]. The idea is to build models for compressing and uncompressing events, trained directly on the (mostly background) data. Events that have a low probability density tend to be poorly reconstructed when compressing and uncompressing compared with events that have a relatively higher probability density. If anomalous events are located in regions of low data probability density, then the reconstruction quality can be used as an anomaly score.

However, autoencoders by themselves are not a complete anomaly detection algorithm - they provide a method for achieving signal sensitivity, but they do not have a natural background estimation component. In the non-resonant case, one could compare the spectrum of anomalous events with background-only simulations, but this requires an excellent model of the background. Given that we expect the unexpected to occur in regions that are poorly modeled, this is unlikely to be a viable strategy in general.

In this paper, we introduce a new method for detecting non-resonant anomalies, based on autoencoders, that is complete in the sense that it includes both signal sensitivity and simulation-free background estimation. Instead of constructing one autoencoder, we advocate for training two or more autoencoders. The set of autoencoders are trained to be as independent of each other as possible. While many methods for decorrelating neural networks exist [62–78] and could be used here, we chose to employ the DisCo decorrelation method first developed in Ref. [69] and explored for simultaneous background estimation in Ref. [76]. Events are labeled as anomalous if the reconstruction quality is poor for all autoencoders. Events labeled as anomalous by one, but not all, of the autoencoders provide the context needed to estimate the Standard Model background in a model independent way, via the ABCD method.

An additional benefit of our method is that it can be run equally well online or offline; indeed this forms a second major motivation for our work. Typically, a key assumption is that anomalous events will be saved by the detectors for offline analysis. Due to the immense data rate at the Large Hadron Collider (LHC), it is not possible to save every collision event for offline processing. Instead, a system of triggers are used to save interesting events [79, 80]. The definition of interesting is model dependent and therefore the new physics may be thrown away in real time. It is therefore of utmost importance

* vmikuni@lbl.gov
† bpnachman@lbl.gov
‡ shih@physics.rutgers.edu

to design model independent strategies for saving anomalous events.

Autoencoders can be run online because they do not require comparing data to a reference sample [28–31]. However, no autoencoder-based trigger proposal so far has been complete in the sense introduced above. Many conventional triggers are complemented by *support* triggers which provide the context needed for data-driven background estimation offline. Our method provides the first complete anomaly detection strategy in a similar way to these conventional methods. By using two decorrelated autoencoders, we can trigger on potentially anomalous events and then additionally save (at a reduced rate) anti-tagged events in a way that background estimation is possible offline.

This paper is organized as follows. First, we introduce the technique of decorrelated autoencoders in Sec. II. Numerical results with the ADC2021 dataset are presented in Sec. III. By definition, this demonstration highlights an offline application of our approach. Section IV provides a discussion about the online-compatibility of our technique for experimental integration online. The paper ends with conclusions and outlook in Sec. V.

II. DECORRELATED AUTOENCODERS

A vanilla autoencoder is a composition of two functions, an encoder g and a decoder f. These two functions are parameterized as neural networks and are optimized to minimize the reconstruction loss:

$$L[f,g] = \sum_i (f(g(x_i)) - x_i)^2 \,, \tag{1}$$

where $x \in \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $f : \mathbb{R}^m \to \mathbb{R}^n$. In order to encourage compression, the latent space dimension is chosen such that $m < n$. A popular variation on this setup is the variational autoencoder [81, 82], whereby the encoding and decoding are probabilistic and the latent space has well-defined statistical properties. The methods proposed here are compatible with variational autoencoders, and while preliminary studies indicate that the results are similar, we leave a full exploration to future work.

Instead of training a single autoencoder as in Eq. 1, we propose to train two (or more) *statistically independent* autoencoders at the same time, in order to enable data-driven background estimation. Following [69, 76], we achieve the decorrelation of the autoencoders by including in the training a regularizer term based on the distance correlation (DisCo) measure of statistical dependence. Focusing on the case of two autoencoders (f_1, g_1) and (f_2, g_2) for simplicity, we consider the following loss function:

$$\begin{aligned} L[f_1, f_2, g_1, g_2] = &\sum_i R_1(x_i)^2 + \sum_i R_2(x_i)^2 \\ &+ \lambda \,\mathrm{DisCo}^2[R_1(X), R_2(X)] \,, \end{aligned} \tag{2}$$

where $R_i(x) = (f_i(g_i(x)) - x)^2$, $\lambda > 0$ is a hyperparameter, and DisCo is the distance correlation [83–86]. DisCo is between 0 and 1 and is zero if and only if its arguments are independent. The capital X is used in the last term of Eq. 2 to indicate that the distance correlation is computed at the level of a batch of examples x, which are realizations of the random variable X. Given autoencoders trained via Eq. 2, we can define counts $N_{\lessgtr,\lessgtr}(\vec{c}) = \sum_i \mathbb{I}[R_1(x_i) \lessgtr c_1]\,\mathbb{I}[R_2(x_i) \lessgtr c_2]$, where $\vec{c} = (c_1, c_2)$ are given thresholds and $\mathbb{I}[\cdot]$ is the indicator function that is zero when its argument is false and one otherwise. The signal sensitive region is $N_{>,>}(\vec{c})$ and the other three regions can be used to estimate the background:

$$N^{\text{predicted}}_{>,>}(\vec{c}) = \frac{N_{>,<}(\vec{c}) N_{<,>}(\vec{c})}{N_{<,<}(\vec{c})} \,. \tag{3}$$

Equation 3 is known as the ABCD method and the $N_{>,>}(\vec{c})$ is exactly the background in the signal-sensitive region if there are enough events and if the two dimensions are effective at rejecting the background.

III. EMPIRICAL RESULTS

The performance of the double autoencoder and decorrelation strategy is tested on the ADC2021 dataset, which was created for unsupervised anomaly detection [31, 87]. In the dataset, proton-proton collisions at the LHC are simulated at center-of-mass energy of 13 TeV. Collision events are required to contain at least one electron (e) or muon (μ) with transverse momenta $p_T > 23$ GeV. A set of various Standard Model processes are generated with Pythia 8.240 generator [88, 89] with detector response modeled by Delphes 3.3.2 [90–92] using the Phase-II CMS detector card. During the training, 2 million events are used while results are reported using an independent validation set containing 800k SM events.

Four benchmark scenarios containing new physics processes are used to evaluate the performance of the algorithm: a leptoquark (LQ) with 80 GeV mass decaying to a b-quark and a τ lepton, a neutral scalar boson (A) of 50 GeV mass decaying to a pair of off-shell Z bosons, which in turn are forced to decay to leptons ($A \to 4l$), a scalar boson h^0 of 60 GeV mass decaying to a pair of τ leptons ($h^0 \to \tau\tau$), and a charged scalar boson $h^\pm$ with 60 GeV mass, decaying to a τ lepton and a neutrino ($h^\pm \to \tau\nu$). In the performance evaluation, each new physics scenario is considered independently, with total amount of events fixed to 0.1% of the total sample size.

The autoencoders are trained on a sample of pure background events. In practice, this corresponds to the case of training on simulation and testing on data. Differences between data and simulation (which are not modeled or taken into account in the ADC2021 dataset used here) may degrade the autoencoder performance if background data events are not reconstructed as well by the autoencoder as background simulation events. Fortunately, it is well-known from previous studies (see e.g. [13, 14, 29]) that autoencoder training is highly insensitive to low amounts of signal contamination. This means that autoencoders can be trained directly on data with a small amount of signal contamination without a significant change in the learned neural networks. We have explicitly verified this in the case of the decorrelated autoencoders using the $A \to 4l$ and $h^0 \to \tau\tau$ signals, with additional results shown in Appendix A.

Each autoencoder architecture is built using deep neural networks containing five fully connected layers. The encoders have 256, 128, 64, 32, and 5 hidden nodes, while the decoder is simply the mirrored version of the encoder. The inputs given to the training are the four-momenta of jets [93, 94] and leptons in (p_T, η, ϕ, m) coordinates, each normalized to 1 during data pre-processing. Only the first (sorted by p_T) four muons, four electrons, and 10 jets in the event are kept with zero padding to each particle applied if fewer objects are present. The implementation is carried out with TensorFlow [95] optimized with the Adam [96]. Even though all new physics scenarios considered here contain a mass resonance, no invariant mass information is directly used in the training process. The λ parameter from Eq. 2 is fixed to 100 and training batch size fixed to 10k to improve the decorrelation performance. The double autoencoder structure is then trained for a total of 1000 epochs, or stopped if the overall training loss does not improve in an independent testing set for 10 consecutive epochs. The complete model uses 230k trainable weights with a total of 460k floating point operations. The neural network architecture and training procedure were not extensively optimized, due in part to the unsupervised nature of this task.

The performance of each autoencoder for anomaly detection is assessed by using the reconstruction loss as the main discriminator. The significance improvement characteristic (SIC) curves are built for each new physics scenario shown in Fig. 1. The comparison with a single autoencoder trained without the decorrelation loss is also shown. We also show in Fig. 1 the combined performance of both autoencoders. The combined result uses a "diagonal" cut to select the same SM background efficiency for each autoencoder. We see that the signal sensitivity of the combined autoencoders is greater than (in fact roughly double) that of each autoencoder individually. This observation indicates that the decorrelation was successful and each autoencoder learned something independent about the BSM anomalies in question. For an additional comparison for the SIC curve at different selection thresholds see Appendix B.

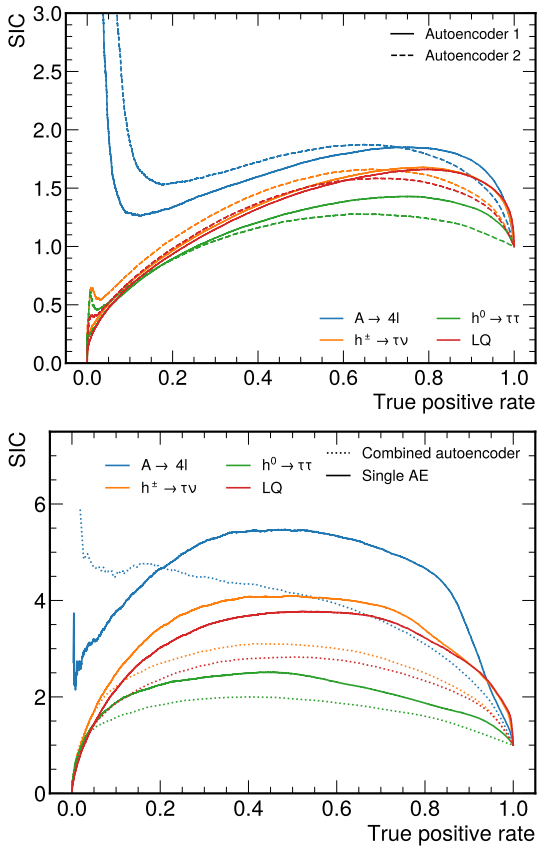

FIG. 1. Top: significance improvement characteristic (SIC) curves of the individual, decorrelated autoencoders, for each new physics benchmark scenario. Bottom: SIC curves for the "diagonal" combination of both decorrelated autoencoders, compared with that of a single autoencoder (obviously trained without any decorrelation). In both panels, the SIC curves are cut off at lower true positive rates due to low background yields.

The independence between reconstruction losses is more fully validated by estimating the difference between the background predicted using the ABCD method (Eq. 3) and the real number of background events in the region of interest. The ratio between the two quantities is shown in Fig. 2. Multiple choices of $\vec{c}$ yielding the same SM efficiency are tested (represented as multiple entries in Fig. 2) for samples containing only SM processes (blue) and mixtures of SM and a new physics process. At lower SM efficiencies, departures from unity (overestimated background [76]) are observed in all mixed samples while the highest variation for a sample containing only SM events is 2.5%, compatible the statistical uncertainty of the sample.

These differences can also be quantified in terms of the signal significance for each benchmark process by comparing the observed and predicted number of background events from the ABCD method. Given N observations in the region of interest with predicted number of background events B, the significance is defined as $\frac{N-B}{\sqrt{N}}$ if

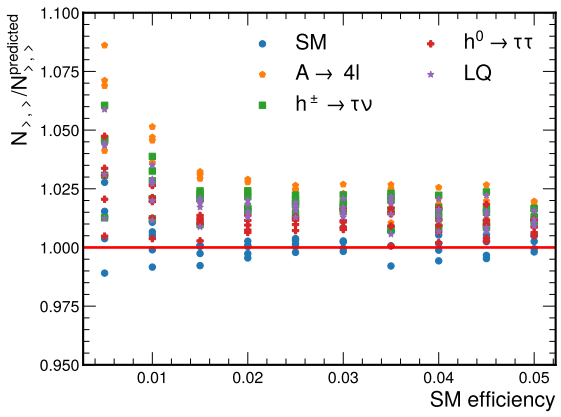

FIG. 2. Closure test of the ABCD estimation method for different SM efficiencies and benchmark scenarios. Different selection combinations yielding the same background efficiency are shown as independent entries.

$N - B > 0$ and 0 otherwise. With the initial signal fraction fixed, the total sample significance is around 0.8 prior to the application of the method. As pointed out in Ref. [76], unaccounted contamination from the signal of interest in the ABCD sidebands may result in different significance values when compared to the correct estimation of the background. While this issue can be accounted when performing model specific exclusion limits, we also show in Fig. 3 (top) the significance obtained using the ABCD method with and without correcting the number of background events. To avoid fine tuning, the threshold applied to each autoencoder reconstruction loss is the one where both autoencoders have the same SM rejection efficiency.

In all new physics benchmark scenarios, the uncorrected significance for SM efficiencies above 1.5% is lower than the corrected for SM efficiencies. Nevertheless, all new physics scenarios show significance between 1 to 4 while the SM only sample has a maximum deviation below 1. We have also probed the stability of the method by performing five independent trainings with different random weight initialization. The standard variation of the average significance was below 6% for all benchmark scenarios tested.

The additional distance correlation loss leads to increased reconstruction loss in the background training sample, resulting in decreased performance compared to a single autoencoder training. This difference is illustrated in Fig. 3 (bottom) where the significance is compared with the values obtained from training a single autoencoder with same network architecture. Since the ABCD method is only applicable in the double autoencoder case, no background estimation method is used in the comparison. In all cases, the difference in significance of the double and single autoencoders is less than 30%. While the single autoencoder consistently outperforms the double autoencoder, the lack of dedicated background estimation might lead to unattainable performances when applied to real particle collisions.

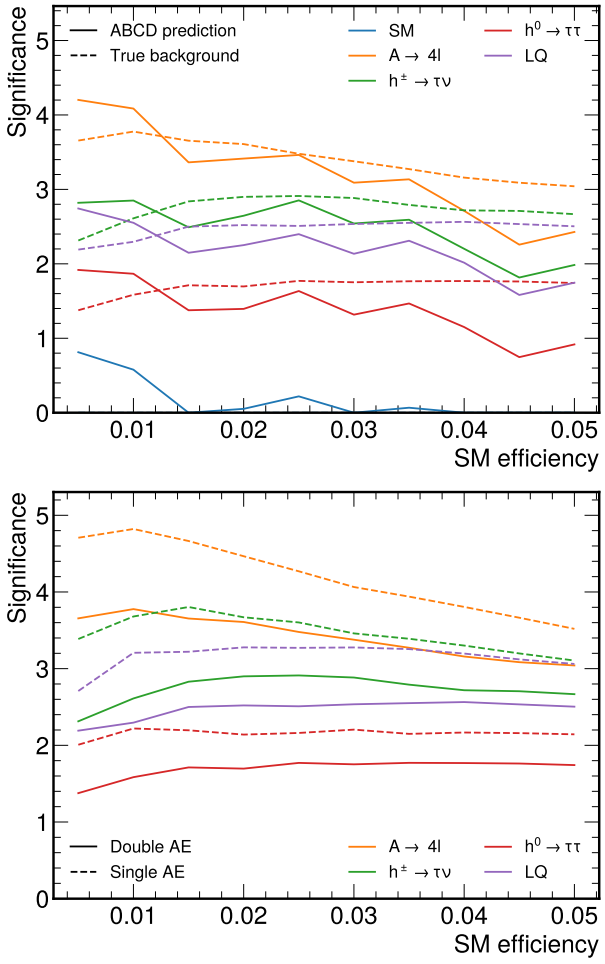

FIG. 3. Signal significance for each benchmark scenario (top) when the ABCD method is used to predict the background level (solid lines) compared to the real significance value (dashed lines). In the bottom panel, the comparison of the significance between a single autoencoder (dashed lines) and the double autoencoder (solid lines) is shown. In this case, no background estimation method is used.

IV. ANOMALY DETECTION ONLINE

The discussion so far has demonstrated that the decorrelated autoencoder protocol is an effective tool for simulation-free, non-resonant anomaly detection. This section briefly describes how this technique is also online compatible. We envision that in an actual trigger system, we would save all events in the signal sensitive region defined by the two autoencoders and then save a random fraction ('prescale') of events in the three other regions for offline background estimation (similar to existing 'support triggers' for certain background processes). The prescale would be set so that the statistical uncer-

tainty on the background prediction is smaller than the statistical uncertainty from events in the signal region. If the SM efficiency in the signal region is ϵ, the trigger rate would scale approximately as 4ϵ, including events saved from the background-dominated regions. The autoencoders themselves could be trained directly on data. These data could be from a previous run or from earlier in a given run. We note that this is the first complete online compatible anomaly detection protocol to be proposed - previous proposals have used single autoencoders and do not come with a method for estimating the background.

Moreover, each of our autoencoders is built using only a set of fully connected layers to allow for a memory and time efficient implementation. There have been many recent demonstrations of ultra low-latency implementations of these and related architectures on Field Programmable Gate Arrays (FPGAs) [30, 97–102]. For studies with more computational resources available, the baseline performance of each autoencoder may be enhanced using more complex reconstruction strategies, as studied in [34–36, 103, 104].

The ADC2021 community challenge dataset was used in part because it was created for the purpose of developing online methods [31, 87] as summarized by the challenge title: *Unsupervised New Physics detection at 40 MHz*. However, there are some features of this dataset which limit direct connection to online algorithms. For example, ATLAS and CMS have single lepton triggers that would likely save all of the challenge events for offline processing. Figure 3 indicates that our decorrelated autoencoder trigger reduces the bandwidth by nearly two orders of magnitude. This is not necessarily relevant for the lepton-triggered data, but it is a common reduction for dedicated triggers. Another issue is, as we have already noted above, that the ADC2021 dataset does not distinguish between "data" and "simulation"; this could be an issue for machine learning methods, which generally require a representative sample for the training data. Expanding the online challenge to other datasets would be interesting for the future.

V. CONCLUSIONS AND OUTLOOK

We have proposed a first complete online-compatible unsupervised non-resonant anomaly detection method that achieves signal sensitivity and can be used to estimate the SM background[1]. Autoencoders, a popular choice of anomaly detection algorithm, are used to identify anomalous events through a reconstruction loss. We advocate for the combination of two or more autoencoders that are trained simultaneously while a distance correlation (DisCo) regularizer term is added to make their reconstruction losses statistically independent. In this strategy, the background from SM events can be estimated with the ABCD method. In the absence of new physics, the method shows a good agreement between the predicted and observed amount of background events. In the presence of new physics, the signal significance varies between 1 and 4 for multiple new physics scenarios with initial contribution amounting to only 0.1% of all events. Given that our method is architecture agnostic, it can be readily generalized for other anomaly detection methods whose output is an anomalous score capable of discerning new physics scenarios from background events.

ACKNOWLEDGMENTS

We thank Barry Dillon, Gregor Kasieczka, and Matt Schwartz for feedback on the manuscript. VM and BN are supported by the U.S. Department of Energy (DOE), Office of Science under contract DE-AC02-05CH11231. The work of DS was supported by DOE grant DOE-SC0010008.

[1] G. Aad *et al.* (ATLAS), 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**, 1 (2012), arXiv:1207.7214 [hep-ex].
[2] S. Chatrchyan *et al.* (CMS), Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC, Phys. Lett. B **716**, 30 (2012), arXiv:1207.7235 [hep-ex].
[3] J. H. Collins, K. Howe, and B. Nachman, Anomaly Detection for Resonant New Physics with Machine Learning, Phys. Rev. Lett. **121**, 241803 (2018), arXiv:1805.02664 [hep-ph].
[4] J. H. Collins, K. Howe, and B. Nachman, Extending the search for new resonances with machine learning, Phys. Rev. **D99**, 014038 (2019), arXiv:1902.02634 [hep-ph].
[5] A. Andreassen, B. Nachman, and D. Shih, Simulation Assisted Likelihood-free Anomaly Detection, Phys. Rev. D **101**, 095004 (2020), arXiv:2001.05001 [hep-ph].
[6] B. Nachmanand D. Shih, Anomaly Detection with Density Estimation, Phys. Rev. D **101**, 075042 (2020), arXiv:2001.04990 [hep-ph].
[7] ATLAS Collaboration, Dijet resonance search with weak supervision using 13 TeV pp collisions in the ATLAS detector 10.1103/PhysRevLett.125.131801 (2020), arXiv:2005.02983 [hep-ex].
[8] K. Benkendorfer, L. L. Pottier, and B. Nachman, Simulation-Assisted Decorrelation for Resonant Anomaly Detection, (2020), arXiv:2009.02205 [hep-ph].
[9] G. Stein, U. Seljak, and B. Dai, Unsupervised in-distribution anomaly detection of new physics

[1] The scripts used to produce the results shown in this work are available at: https://github.com/ViniciusMikuni/DoubleAE

through conditional density estimation, (2020), arXiv:2012.11638 [cs.LG].
[10] G. Kasieczka, B. Nachman, and D. Shih, New Methods and Datasets for Group Anomaly Detection From Fundamental Physics (2021) arXiv:2107.02821 [stat.ML].
[11] D. Shih, M. R. Buckley, L. Necib, and J. Tamanas, Via Machinae: Searching for Stellar Streams using Unsupervised Machine Learning, (2021), arXiv:2104.12789 [astro-ph.GA].
[12] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, Classifying Anomalies THrough Outer Density Estimation (CATHODE), (2021), arXiv:2109.00546 [hep-ph].
[13] M. Farina, Y. Nakai, and D. Shih, Searching for New Physics with Deep Autoencoders 10.1103/PhysRevD.101.075021 (2018), arXiv:1808.08992 [hep-ph].
[14] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or What?, SciPost Phys. **6**, 030 (2019), arXiv:1808.08979 [hep-ph].
[15] T. S. Royand A. H. Vijay, A robust anomaly finder based on autoencoder, (2019), arXiv:1903.02032 [hep-ph].
[16] A. Blance, M. Spannowsky, and P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches, JHEP **10**, 047, arXiv:1905.10384 [hep-ph].
[17] J. H. Collins, P. Martín-Ramiro, B. Nachman, and D. Shih, Comparing Weak- and Unsupervised Methods for Resonant Anomaly Detection, (2021), arXiv:2104.02092 [hep-ph].
[18] A. Kahn, J. Gonski, I. Ochoa, D. Williams, and G. Brooijmans, Anomalous Jet Identification via Sequence Modeling, (2021), arXiv:2105.09274 [hep-ph].
[19] O. Amramand C. M. Suarez, Tag N' Train: A Technique to Train Improved Classifiers on Unlabeled Data 10.1007/JHEP01(2021)153 (2020), arXiv:2002.12376 [hep-ph].
[20] G. Kasieczka *et al.*, The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics, (2021), arXiv:2101.08320 [hep-ph].
[21] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, Novelty Detection Meets Collider Physics 10.1103/PhysRevD.101.076015 (2018), arXiv:1807.10261 [hep-ph].
[22] A. De Simoneand T. Jacques, Guiding New Physics Searches with Unsupervised Learning, Eur. Phys. J. **C79**, 289 (2019), arXiv:1807.06038 [hep-ph].
[23] A. Mullin, H. Pacey, M. Parker, M. White, and S. Williams, Does SUSY have friends? A new approach for LHC event analysis 10.1007/JHEP02(2021)160 (2019), arXiv:1912.10625 [hep-ph].
[24] G. M. Alessandro Casa, Nonparametric semisupervised classification for signal detection in high energy physics, (2019), arXiv:1809.02977 [hep-ex].
[25] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, Uncovering latent jet substructure, Phys. Rev. **D100**, 056002 (2019), arXiv:1904.04200 [hep-ph].
[26] M. Romo Crispim, N. Castro, R. Pedro, and T. Vale, Transferability of Deep Learning Models in Searches for New Physics at Colliders, Phys. Rev. D **101**, 035042 (2020), arXiv:1912.04220 [hep-ph].
[27] M. C. Romao, N. Castro, J. Milhano, R. Pedro, and T. Vale, Use of a Generalized Energy Mover's Distance in the Search for Rare Phenomena at Colliders 10.1140/epjc/s10052-021-08891-6 (2020), arXiv:2004.09360 [hep-ph].
[28] O. Knapp, G. Dissertori, O. Cerri, T. Q. Nguyen, J.-R. Vlimant, and M. Pierini, Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark 10.1140/epjp/s13360-021-01109-4 (2020), arXiv:2005.01598 [hep-ex].
[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**, 036, arXiv:1811.10276 [hep-ex].
[30] E. Govorkova *et al.*, Autoencoders on FPGAs for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider, (2021), arXiv:2108.03986 [physics.ins-det].
[31] E. Govorkova, E. Puljak, T. Aarrestad, M. Pierini, K. A. Woźniak, and J. Ngadiuba, LHC physics dataset for unsupervised New Physics detection at 40 MHz, (2021), arXiv:2107.02157 [physics.data-an].
[32] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc, Learning the latent structure of collider events 10.1007/JHEP10(2020)206 (2020), arXiv:2005.12319 [hep-ph].
[33] M. C. Romao, N. Castro, and R. Pedro, Finding New Physics without learning about it: Anomaly Detection as a tool for Searches at Colliders 10.1140/epjc/s10052-020-08807-w (2020), arXiv:2006.05432 [hep-ph].
[34] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, and T. Golling, Variational Autoencoders for Anomalous Jet Tagging, (2020), arXiv:2007.01850 [hep-ph].
[35] Adrian Alan Pol and Victor Berger and Gianluca Cerminara and Cecile Germain and Maurizio Pierini, Anomaly Detection With Conditional Variational Autoencoders, (2020), arXiv:2010.05531 [cs.LG].
[36] O. Atkinson, A. Bhardwaj, C. Englert, V. S. Ngairangbam, and M. Spannowsky, Anomaly detection with Convolutional Graph Neural Networks, (2021), arXiv:2105.07988 [hep-ph].
[37] C. K. Khosaand V. Sanz, Anomaly Awareness, (2020), arXiv:2007.14462 [cs.LG].
[38] P. Thaprasop, K. Zhou, J. Steinheimer, and C. Herold, Unsupervised Outlier Detection in Heavy-Ion Collisions, (2020), arXiv:2007.15830 [hep-ex].
[39] S. Alexander, S. Gleyzer, H. Parul, P. Reddy, M. W. Toomey, E. Usai, and R. Von Klar, Decoding Dark Matter Substructure without Supervision, (2020), arXiv:2008.12731 [astro-ph.CO].
[40] J. A. Aguilar-Saavedra, F. R. Joaquim, and J. F. Seabra, Mass Unspecific Supervised Tagging (MUST) for boosted jets 10.1007/JHEP03(2021)012 (2020), arXiv:2008.12792 [hep-ph].
[41] M. van Beekveld, S. Caron, L. Hendriks, P. Jackson, A. Leinweber, S. Otten, R. Patrick, R. Ruiz de Austri, M. Santoni, and M. White, Combining outlier analysis algorithms to identify new physics at the LHC, (2020), arXiv:2010.07940 [hep-ph].
[42] S. E. Park, D. Rankin, S.-M. Udrescu, M. Yunus, and P. Harris, Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge, (2020), arXiv:2011.03550 [hep-ph].
[43] D. A. Faroughy, Uncovering hidden patterns in collider events with Bayesian probabilistic models, (2020), arXiv:2012.08579 [hep-ph].

[44] P. Chakravarti, M. Kuusela, J. Lei, and L. Wasserman, Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests, (2021), arXiv:2102.07679 [stat.AP].

[45] J. Batson, C. G. Haaf, Y. Kahn, and D. A. Roberts, Topological Obstructions to Autoencoding, (2021), arXiv:2102.08380 [hep-ph].

[46] A. Blanceand M. Spannowsky, Unsupervised Event Classification with Graphs on Classical and Photonic Quantum Computers, (2021), arXiv:2103.03897 [hep-ph].

[47] B. Bortolato, B. M. Dillon, J. F. Kamenik, and A. Smolkovič, Bump Hunting in Latent Space, (2021), arXiv:2103.06595 [hep-ph].

[48] B. M. Dillon, T. Plehn, C. Sauer, and P. Sorrenson, Better Latent Spaces for Better Autoencoders, (2021), arXiv:2104.08291 [hep-ph].

[49] T. Finke, M. Krämer, A. Morandini, A. Mück, and I. Oleksiyuk, Autoencoders for unsupervised anomaly detection in high energy physics, (2021), arXiv:2104.09051 [physics.data-an].

[50] T. Aarrestad *et al.*, The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider, (2021), arXiv:2105.14027 [hep-ph].

[51] S. Caron, L. Hendriks, and R. Verheyen, Rare and Different: Anomaly Scores from a combination of likelihood and out-of-distribution models to detect new physics at the LHC, (2021), arXiv:2106.10164 [hep-ph].

[52] S. Volkovich, F. D. V. Halevy, and S. Bressler, The Data-Directed Paradigm for BSM searches, (2021), arXiv:2107.11573 [hep-ex].

[53] B. Ostdiek, Deep Set Auto Encoders for Anomaly Detection in Particle Physics, (2021), arXiv:2109.01695 [hep-ph].

[54] K. Fraser, S. Homiller, R. K. Mishra, B. Ostdiek, and M. D. Schwartz, Challenges for Unsupervised Anomaly Detection in Particle Physics, (2021), arXiv:2110.06948 [cs.LG].

[55] R. T. D'Agnoloand A. Wulzer, Learning New Physics from a Machine, Phys. Rev. **D99**, 015014 (2019), arXiv:1806.02350 [hep-ph].

[56] R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, Learning Multivariate New Physics 10.1140/epjc/s10052-021-08853-y (2019), arXiv:1912.12155 [hep-ph].

[57] T. Dorigo, M. Fumanelli, C. Maccani, M. Mojsovska, G. C. Strong, and B. Scarpa, RanBox: Anomaly Detection in the Copula Space, (2021), arXiv:2106.05747 [physics.data-an].

[58] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, A generic anti-QCD jet tagger, JHEP **11**, 163, arXiv:1709.01087 [hep-ph].

[59] V. Mikuniand F. Canelli, Unsupervised clustering for collider physics, (2020), arXiv:2010.07106 [physics.data-an].

[60] J. A. Aguilar-Saavedra, Anomaly detection from mass unspecific jet tagging, (2021), arXiv:2111.02647 [hep-ph].

[61] M. Feickertand B. Nachman, A Living Review of Machine Learning for Particle Physics, (2021), arXiv:2102.02770 [hep-ph].

[62] G. Louppe, M. Kagan, and K. Cranmer, Learning to Pivot with Adversarial Networks, in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017) arXiv:1611.01046 [stat.ME].

[63] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure, JHEP **05**, 156, arXiv:1603.00027 [hep-ph].

[64] I. Moult, B. Nachman, and D. Neill, Convolved Substructure: Analytically Decorrelating Jet Substructure Observables, JHEP **05**, 002, arXiv:1710.06859 [hep-ph].

[65] J. Stevensand M. Williams, uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers, JINST **8**, P12013, arXiv:1305.7248 [nucl-ex].

[66] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Sgaard, Decorrelated Jet Substructure Tagging using Adversarial Neural Networks 10.1103/PhysRevD.96.074034 (2017), arXiv:1703.03507 [hep-ex].

[67] L. Bradshaw, R. K. Mishra, A. Mitridate, and B. Ostdiek, Mass Agnostic Jet Taggers 10.21468/SciPostPhys.8.1.011 (2019), arXiv:1908.08959 [hep-ph].

[68] Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS, ATL-PHYS-PUB-2018-014 (2018).

[69] G. Kasieczkaand D. Shih, DisCo Fever: Robust Networks Through Distance Correlation 10.1103/PhysRevLett.125.122001 (2020), arXiv:2001.05310 [hep-ph].

[70] L.-G. Xia, QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics, Nucl. Instrum. Meth. **A930**, 15 (2019), arXiv:1810.08387 [physics.data-an].

[71] C. Englert, P. Galler, P. Harris, and M. Spannowsky, Machine Learning Uncertainties with Adversarial Neural Networks, Eur. Phys. J. **C79**, 4 (2019), arXiv:1807.08763 [hep-ph].

[72] S. Wunsch, S. Jórger, R. Wolf, and G. Quast, Reducing the dependence of the neural network function to systematic uncertainties in the input space 10.1007/s41781-020-00037-9 (2019), arXiv:1907.11674 [physics.data-an].

[73] A. Rogozhnikov, A. Bukva, V. V. Gligorov, A. Ustyuzhanin, and M. Williams, New approaches for boosting to uniformity, JINST **10** (03), T03002, arXiv:1410.4140 [hep-ex].

[74] C. Collaboration, A deep neural network to search for new long-lived particles decaying to jets, Machine Learning: Science and Technology 10.1088/2632-2153/ab9023 (2020), 1912.12238.

[75] J. M. Clavijo, P. Glaysher, and J. M. Katzy, Adversarial domain adaptation to reduce sample bias of a high energy physics classifier, (2020), arXiv:2005.00568 [stat.ML].

[76] G. Kasieczka, B. Nachman, M. D. Schwartz, and D. Shih, ABCDisCo: Automating the ABCD Method with Machine Learning 10.1103/PhysRevD.103.035021 (2020), arXiv:2007.14400 [hep-ph].

[77] O. Kitouni, B. Nachman, C. Weisser, and M. Williams, Enhancing searches for resonances with machine learning and moment decomposition, (2020), arXiv:2010.09745 [hep-ph].

[78] A. Ghoshand B. Nachman, A Cautionary Tale of Decorrelating Theory Uncertainties, (2021), arXiv:2109.08159 [hep-ph].

[79] G. Aad *et al.* (ATLAS), Operation of the ATLAS trigger system in Run 2, JINST **15** (10), P10004, arXiv:2007.12539 [physics.ins-det].

[80] V. Khachatryan *et al.* (CMS), The CMS trigger system, JINST **12** (01), P01020, arXiv:1609.02366 [physics.ins-det].

[81] D. P. Kingmaand M. Welling, Auto-encoding variational bayes, (2014), arXiv:1312.6114 [stat.ML].

[82] D. P. Kingmaand M. Welling, An Introduction to Variational Autoencoders, Foundations and Trends in Machine Learning **12**, 307 (2019).

[83] G. J. Szkely, M. L. Rizzo, and N. K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Statist. **35**, 2769 (2007).

[84] G. J. Szkelyand M. L. Rizzo, Brownian distance covariance, Ann. Appl. Stat. **3**, 1236 (2009).

[85] G. J. Székelyand M. L. Rizzo, The distance correlation t-test of independence in high dimension, J. Multivar. Anal. **117**, 193 (2013).

[86] G. J. Szkelyand M. L. Rizzo, Partial distance correlation with methods for dissimilarities, Ann. Statist. **42**, 2382 (2014).

[87] E. Govorkova *et al.*, Unsupervised new physics detection at 40 mhz, `https://mpp-hep.github.io/ADC2021/` (2021).

[88] T. Sjostrand, S. Mrenna, and P. Z. Skands, PYTHIA 6.4 Physics and Manual, JHEP **05**, 026, arXiv:hep-ph/0603175.

[89] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015), arXiv:1410.3012 [hep-ph].

[90] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, JHEP **02**, 057, arXiv:1307.6346 [hep-ex].

[91] M. Selvaggi, DELPHES 3: A modular framework for fast-simulation of generic collider experiments, J. Phys. Conf. Ser. **523**, 012033 (2014).

[92] A. Mertens, New features in Delphes 3, J. Phys. Conf. Ser. **608**, 012045 (2015).

[93] M. Cacciari, G. P. Salam, and G. Soyez, FastJet User Manual, Eur. Phys. J. C **72**, 1896 (2012), arXiv:1111.6097 [hep-ph].

[94] M. Cacciariand G. P. Salam, Dispelling the N^3 myth for the k_t jet-finder, Phys. Lett. B **641**, 57 (2006), arXiv:hep-ph/0512210.

[95] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, Tensorflow: A system for large-scale machine learning., in *OSDI*, Vol. 16 (2016) pp. 265–283.

[96] D. Kingmaand J. Ba, Adam: A method for stochastic optimization, (2014), arXiv:1412.6980 [cs].

[97] J. Duarte *et al.*, Fast inference of deep neural networks in FPGAs for particle physics, JINST **13** (07), P07027, arXiv:1804.06913 [physics.ins-det].

[98] C. N. Coelho, A. Kuusela, S. Li, H. Zhuang, T. Aarrestad, V. Loncar, J. Ngadiuba, M. Pierini, A. A. Pol, and S. Summers, Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors 10.1038/s42256-021-00356-5 (2020), arXiv:2006.10159 [physics.ins-det].

[99] T. Aarrestad *et al.*, Fast convolutional neural networks on FPGAs with hls4ml, (2021), arXiv:2101.05108 [cs.LG].

[100] A. Heintz *et al.*, Accelerated Charged Particle Tracking with Graph Neural Networks on FPGAs, 34th Conference on Neural Information Processing Systems (2020), arXiv:2012.01563 [physics.ins-det].

[101] J. S. John *et al.*, Real-time Artificial Intelligence for Accelerator Control: A Study at the Fermilab Booster, (2020), arXiv:2011.07371 [physics.acc-ph].

[102] T. M. Hong, B. T. Carlson, B. R. Eubanks, S. T. Racz, S. T. Roche, J. Stelzer, and D. C. Stumpp, Nanosecond machine learning event classification with boosted decision trees in FPGA for high energy physics, (2021), arXiv:2104.03408 [hep-ex].

[103] J. H. Collins, An Exploration of Learnt Representations of W Jets (2021) arXiv:2109.10919 [hep-ph].

[104] B. Orzari, T. Tomei, M. Pierini, M. Touranakou, J. Duarte, R. Kansal, J.-R. Vlimant, and D. Gunopulos, Sparse Data Generation for Particle-Based Simulation of Hadronic Jets in the LHC, in *38th International Conference on Machine Learning Conference* (2021) arXiv:2109.15197 [physics.data-an].

Appendix A: Signal significance with signal contamination

We verify the changes in the significance obtained by training two additional models on datasets containing signal contamination of 0.1%, the same threshold used to derive the results reported in Sec. III. In the alternative setup we consider independently the contamination from the $A \to 4l$ and $h^0 \to \tau\tau$ signals, the ones with highest and lowest reported significance. In Fig. 4, results obtained in each of these scenarios are shown and compared with the results from the background-only training.

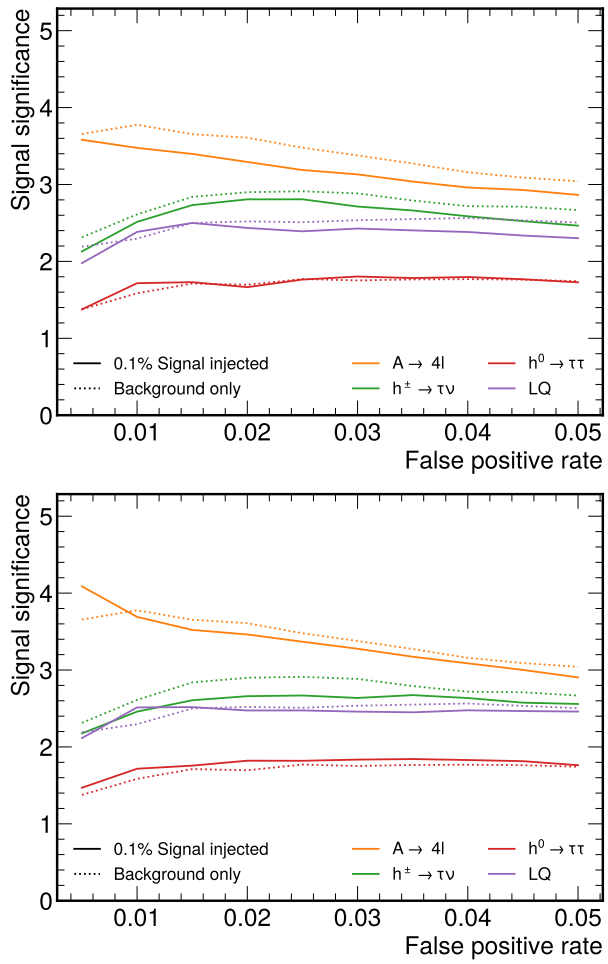

FIG. 4. Signal significance obtained when training the model in a dataset containing the signal contamination from $A \to 4l$ (top) and $h^0 \to \tau\tau$ (bottom). Results from the background-only training are shown for comparison.

Appendix B: Significance improvement characteristic for different selection thresholds

The studies presented in this work use the "diagonal" cut as the representative selection for the combined performance. Different choices, leading to different results, can be used when a particular new physics scenario is under study. To exemplify this difference, we show in Fig. 5 the SIC curve for different selections applied to the reconstruction loss of each autoencoder. While a symmetric selection results in maximum SIC values for all benchmarks, the exact threshold resulting in maximum SIC is different for each benchmark scenario.

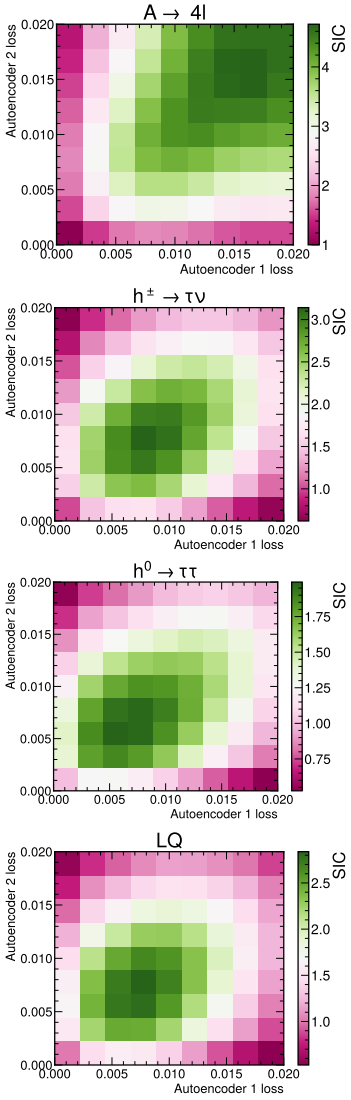

FIG. 5. Significance improvement characteristic for different new physics benchmark scenarios. The lower edges of each bin represents the selection threshold applied for each autoencoder loss function.