Title
Anomaly Detection under Coordinate Transformations

Authors
Kasieczka, Gregor
Mastandrea, Radha
Mikuni, Vinicius
et al.

Publication Date
2022-09-13

Anomaly Detection under Coordinate Transformations

Gregor Kasieczka,[1,*] Radha Mastandrea,[2,3,†] Vinicius Mikuni,[4,‡]
Benjamin Nachman,[3,5,§] Mariel Pettee,[3,¶] and David Shih[6,**]

[1] *Institut für Experimentalphysik, Universität Hamburg, 22761 Hamburg, Germany*
[2] *Department of Physics, University of California, Berkeley, CA 94720, USA*
[3] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[4] *National Energy Research Scientific Computing Center, Berkeley Lab, Berkeley, CA 94720, USA*
[5] *Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*
[6] *New High Energy Theory Center, Rutgers University, Piscataway, NJ 08854, USA*

There is a growing need for machine learning-based anomaly detection strategies to broaden the search for Beyond-the-Standard-Model (BSM) physics at the Large Hadron Collider (LHC) and elsewhere. The first step of any anomaly detection approach is to specify observables and then use them to decide on a set of anomalous events. One common choice is to select events that have low probability density. It is a well-known fact that probability densities are not invariant under coordinate transformations, so the sensitivity can depend on the initial choice of coordinates. The broader machine learning community has recently connected coordinate sensitivity with anomaly detection and our goal is to bring awareness of this issue to the growing high energy physics literature on anomaly detection. In addition to analytical explanations, we provide numerical examples from simple random variables and from the LHC Olympics Dataset that show how using probability density as an anomaly score can lead to events being classified as anomalous or not depending on the coordinate frame.

I. INTRODUCTION

Given the vast parameter space of Beyond-the-Standard-Model (BSM) physics as well as the lack of recent discoveries, there has been a growing interest in new search techniques that reduce model dependence. A number of novel approaches based on machine learning (ML) have been proposed that can automatically identify anomalous regions of phase space [1–79]. These techniques can be categorized by their BSM hypotheses, which directly relates to their ML strategy. While most proposals do not make these hypotheses explicit, the existing approaches typically posit one of three possibilities: (i) the BSM is rare: $p_B(x)$ is small for background probability density p_B and for BSM data point x; (ii) the BSM is overdense: $p_S(x)/p_B(x)$ is large for signal probability density p_S; (iii) the BSM is more similar to known BSM models than to the background.

These three possibilities approximately map onto unsupervised, weakly supervised, and semisupervised machine learning methods, respectively. All three of these generic hypotheses are much weaker than the usual, fully supervised case where the hypothesis is very specific and involves assuming particular couplings, decay chains, masses, etc. The unsupervised methods learn implicitly or explicitly the probability density p_B, and then anomalous events are defined by those with a low density, usually through the use of an anomaly score that is a proxy for p_B (such as the loss function of an autoencoder - see Sec. II A). Weakly supervised methods learn likelihood ratios between a target dataset and a reference (mostly anomaly-free) dataset. Weakly supervised learning refers to learning with noisy labels – in this context, the target dataset has a noisy label of 'signal' while the reference dataset has a noisy label of 'background'. Semisupervised methods use a number of simulated signal models, often combined with (mostly anomaly-free) data. This categorization is not unique and the names used here are based on their meaning in the high energy physics (HEP) ML literature (see Ref. [80] for a recent review). Most papers on anomaly detection for HEP fall in the unsupervised category, although the only existing ML-based anomaly detection physics results use weakly supervised learning [21, 45]. For this paper, we consider only the unsupervised and weakly supervised cases, as they are the most commonly studied for HEP analyses.

Due to the ill-posed nature of attempting to identify *any* anomaly, there is no one method that will be more sensitive than all others[1]. Each approach has merits and disadvantages. For example, previous works have explored the tradeoffs between unsupervised and weakly supervised learning [24, 37, 42]. In particular, Ref. [42] pointed out that in the context of resonance searches, weakly supervised methods may outperform unsupervised methods for relatively higher signal fractions because they can explicitly use the presence of the anomalies to guide their performance. In contrast, unsupervised

* gregor.kasieczka@uni-hamburg.de
† rmastand@berkeley.edu
‡ vmikuni@lbl.gov
§ bpnachman@lbl.gov
¶ mpettee@lbl.gov
** dshih@physics.rutgers.edu

[1] In the limit of infinite statistics, flawless background estimation, and arbitrarily powerful ML model architecture/training, weakly supervised methods can be universally optimal. Of course, this is never true in practice; see Ref. [16], App. A.

approaches are nearly independent of the presence of signal and so can maintain performance even at low signal fraction. However, if the signal is in the bulk of the background distribution, then unsupervised methods may be unable to find it no matter how much signal is present.

Another core feature of anomaly detection approaches is their response to coordinate transformations. While likelihood ratios are independent of invertible coordinate transformations, the notion of an event being 'rare' is inherently coordinate-dependent. Since unsupervised methods cannot be guided by the presence of anomalies as in weakly supervised approaches, the selection of observables used for anomaly detection may be more important for unsupervised methods compared with weakly supervised approaches. The fact that probability densities are not invariant under coordinate transformations is well known and the connection to anomaly detection has recently been explored in the broader machine learning community [81]. Our goal is to bring awareness of this issue to HEP, where there are a growing number of proposals that make use of coordinate-dependent methods. While coordinate sensitivity is relevant for both achieving signal sensitivity and estimating the Standard Model background, we focus entirely on the former as it is usually the focus of recent anomaly detection proposals.

This paper is organized as follows. Section II provides a taxonomy of ML-based anomaly detection methods. The statistical properties of coordinate transformations of observables are described in Sec. III. Illustrative numerical examples are given in Sec. IV, first with a simple, analytic example and then a more realistic example based on a dijet search at the Large Hadron Collider (LHC). The paper ends with conclusions in Sec. V.

II. LANDSCAPE OF ANOMALY DETECTION METHODS

In this section, we provide a brief summary of unsupervised and weakly supervised anomaly detection methods. We also provide references to recent applications of these methods in the HEP field.

A. Unsupervised

One of the most popular approaches studied in the phenomenology literature is the autoencoder (AE). The first AE approaches [5, 6, 10] worked by simultaneously training two neural networks: an encoder network $f : \mathbb{R}^N \to \mathbb{R}^M$ and then a decoder network $g : \mathbb{R}^M \to \mathbb{R}^N$. The typical loss function is the mean squared error[2]: $\langle (g(f(X)) - X)^2 \rangle$. For arbitrarily flexible networks and training procedures, $f \circ g$ could approach the identity. To ensure this does not happen, the network capacities and training procedure are restricted and $M \ll N$. Anomalies are then characterised by high reconstruction loss $(g(f(x)) - x)^2$ compared to the background.

As with any compression algorithm, the autoencoder will maximize its efficiency if it dedicates its limited capacity based on the probability density of a given event. For this reason, the AE implicitly[3] estimates $p_B(x)$. Anomaly scores based on autoencoders can also be created to take advantage of the compressed latent space created by the algorithm. Those are often based on Variational Autoencoders (VAEs) [82, 83] or similar methods, trained to generate a latent space with useful statistical properties [8, 25, 31, 32, 41, 43, 58, 78, 84].

Beyond VAEs, other deep generative models proposed for unsupervised anomaly detection include Generative Adversarial Networks (GANs) [85, 86] and Normalizing Flows [87, 88]. In all of these cases, the generative model is implicitly (GANs and VAEs) or explicitly (Normalizing Flows (NFs)) learning $p_B(x)$, so anomaly scores are directly linked to the probability density. In the case of GANs, anomalies can be identified by combining the generative model with an autoencoder [20] and assigning an anomaly score to the reconstruction loss between the inputs and the generated outputs. With a direct estimate of the density, the output of a NF can be used directly as an anomaly score [48, 76, 79][4]. A detailed comparison of various generative models on benchmark BSM signals was studied in Ref. [48].

B. Weakly Supervised

In contrast to unsupervised methods, weakly supervised approaches require two datasets: a reference and a target. Some approaches emphasize the estimation of the reference sample [2, 3, 15, 16, 55, 73] and some approaches take the reference sample as given [1, 4, 38, 65, 70, 74, 90]. Strategies for determining the reference sample span a spectrum ranging from signal-model agnostic and background-model dependent approaches using simulations to resonance searches where sideband information can be directly used to estimate a background-only reference. Hybrid methods have also been proposed, as in the case of creating noisy labels for weak supervision using unsupervised autoencoders [24].

Once the reference sample is acquired, most methods estimate the likelihood ratio directly by training a classifier to distinguish examples from the target and reference datasets. It is well-known that the output of a classifier

[2] Capital letters represent random variables and lower case letters represent realizations of the random variables.

[3] Vanilla AEs have a strong dependence on the ML architecture and training procedure, which means that they may not be as precise at estimating the density as other approaches.

[4] Normalizing Flows have also been proposed for use as weakly supervised anomaly detection methods - see Ref. [16, 55, 73, 89].

trained with a standard loss function like binary cross-entropy is monotonically related to the likelihood ratio (see e.g. Ref. [91, 92]). Directly estimating probability densities and taking ratios has also been explored [16].

III. STATISTICS OF COORDINATE TRANSFORMATIONS

In this section, we will review some elementary facts about probability densities and their applications to anomaly detection[5].

Suppose that we have initial coordinates $X \in \mathbb{R}^N$ and coordinate transformation $Y = f(X)$, where f is an invertible and differentiable function[6]. If a point in phase space x has probability density p_X, then the corresponding point $y = f(x)$ has probability density:

$$p_Y(y) = p_X(f^{-1}(y)) \left| \frac{d}{dy} f^{-1}(y) \right| , \tag{1}$$

where the last term is the Jacobian determinant of f^{-1} evaluated at y. If f is a linear transformation, then the Jacobian determinant is independent of x. This means that if we order events by density, then the ordering is unchanged. As an example, consider the linear function $y = ax + b$. By the above equation, we have $p_Y(y) = p_X(\frac{y-b}{a}) \left|\frac{1}{a}\right|$. Coordinate changes of these types produce a simple shift and rescaling of the probability distribution p_X, as shown in Fig. 1(a). Note that this includes standardization where the mean is subtracted and then the data are divided by the standard deviation.

In contrast, if f is non-linear, then the Jacobian determinant can depend on x. As an example, the non-linear function $y = e^{-x}$ yields the probability density $p_Y(y) = p_X(-\ln(y)) \left|\frac{1}{y}\right|$, so the Jacobian determinant is still a function of y and therefore also of x. Since the Jacobian determinant is non-constant, this choice of coordinate transformation can dramatically affect the density-ranked order of events, as shown in Fig. 1: low-density values of X are mapped to high-density values of Y.

One popular anomaly detection protocol would be to take events that are 'rare' in an absolute sense: $p_X(x) < c$ for some threshold c. If c is fixed, then the events selected would change under coordinate transformations due to the Jacobian factor in Eq. 1. An alternative protocol that is more robust (but still sensitive) to coordinate transformations would consider 'rare' in a relative sense so that the Jacobian factors cancel. In particular, instead of comparing densities to an absolute threshold, we could compare the density of one event to the density of other events. A protocol in this direction would be to take a fraction q of the 'rarest' events.

For example, in one dimension, this corresponds to using a threshold c given by the q quantile of the density. Symbolically, the quantile in X for a one-dimensional random variable are given by:[7]

$$q = \int_c^\infty p_X(x) dx \, . \tag{2}$$

Since c is now defined by an integral over a density and not a bare density, one may hope that it is more robust to coordinate transformations. Ideally, if we compute the threshold c' after transforming into $y = f(x)$, we would have $f(c) = c'$. In reality:

$$\begin{aligned} q &= \int_{c'}^\infty p_Y(y) dy & (3) \\ &= \int_{c'}^\infty p_X(f^{-1}(y)) \left| \frac{d}{dy} f^{-1}(y) \right| dy \\ &= \int_{f^{-1}(c')}^{f^{-1}(\infty)} p_X(x) \left| \frac{df}{dx} \right|^{-1} \frac{df}{dx} \, dx. & (4) \end{aligned}$$

The features X over which the quantile is computed could be the original observables or one could first map to the anomaly score and consider the most anomalous events. If the Jacobian in Eq. 4 is non-negative, then the two penultimate terms cancel and $f^{-1}(c') = c$, so the same events are selected before and after the coordinate transformation. However, if the Jacobian takes on negative values, the order of events under f is reversed and then different events can be selected ($f^{-1}(c') \neq c$). For example, if $f(x) = -x$, then the lowest and highest quantiles are completely reversed. Another extreme example is when f is the Cumulative Distribution Function (CDF). In this case, $f(X)$ is uniformly distributed between 0 and 1 so no point is rarer than any other.

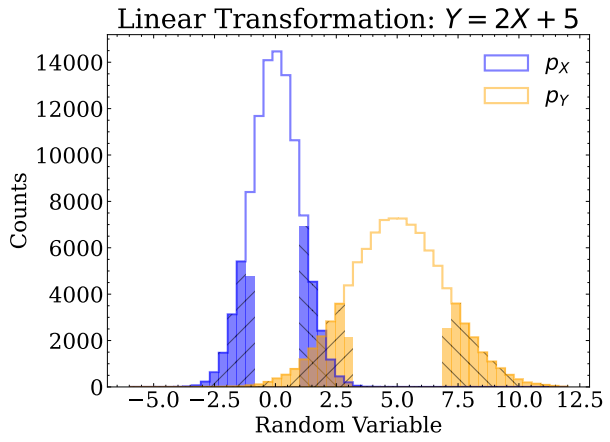
In contrast, likelihood ratio methods are invariant under coordinate transformations because the Jacobian determinant in Eq. 1 is the same for the target probability density and the reference probability density (and thus drops out in the ratio). This is strictly only true when f is bijective (as assumed above), but it may be approximately true even if this is not the case. Note that even though likelihood ratios are formally invariant under coordinate transformations, it may be that practical approaches benefit from a judicial choice of coordinates. For example, observable standardization is often essential in enabling effective ML training.

Equation 1 is a well-known fact found in textbooks of probability and statistics. Its connection with anomaly

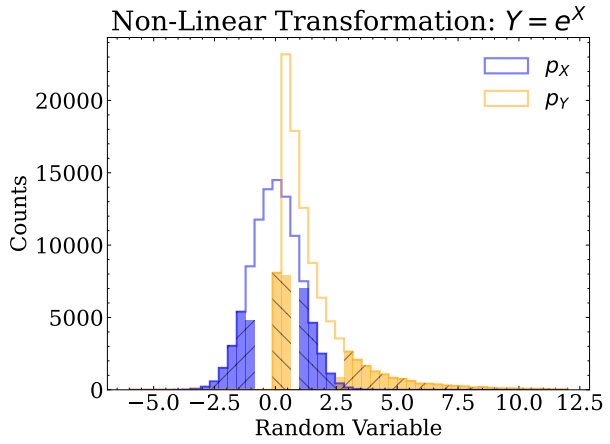
[5] Note that methods that do not exactly learn the density like vanilla autoencoders may have additional susceptibilities to variable transformations.

[6] Much of the discussion also still applies if this is not true everywhere, but the bookkeeping becomes significantly more complex, so we focus on this case.

[7] It may be useful to consider both the highest and lowest quantiles, although for ordering by anomaly score, presumably only the most anomalous events should be considered (not the least).



(a)



(b)

FIG. 1. Histograms for a random normal variable X with a (a) linear transform $Y = 2X + 5$, (b) non-linear transform $Y = e^{-X}$. The shaded regions mark where $|X| > 1$. Variables that originate from low-density values of X are hatched ("/" for $X < 1$ and "\" for $X > 1$). For the linear transformation, low-density values of X map to low-density values of Y. For the non-linear transformation, however, the low-density values originating from $X > 1$ are mapped to high-density values of Y.

detection was recently made by the machine learning community [81]. In the following section, we provide an illustrative Gaussian example and then make an explicit connection with HEP, both using the relative threshold protocol.

IV. NUMERICAL EXAMPLES

A. Analytic Case

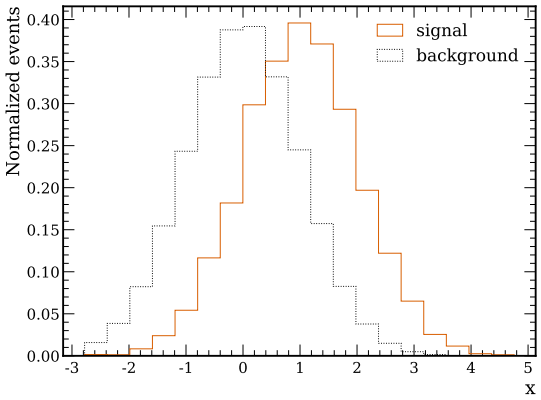
To clearly illustrate the ideas discussed in the previous section, we will construct a simple example to demonstrate a dramatic consequence of this sensitivity to coordinate transformations. Let $X_b \sim \mathcal{N}(0,1)$ represent a set of background observables, and let $X_s \sim \mathcal{N}(1,1)$ represent a set of signal observables. This scenario is illustrated in Fig. 2(a). A density estimation-based search for anomalies would consist of learning the density of the background p_{X_b}, then making a cut where the density is low. This would designate the two tails of X_b as rare, and a search for anomalies would then successfully pick up the signal events X_s overlapping with the right-tail phase space of the background.

Now, suppose that instead of the variables X_b and X_s, we used $Y_b = f(X_b)$ and $Y_s = f(X_s)$, where f is the CDF of a standard normal random variable. This scenario is illustrated in Fig. 2(b). In this case, Y_b (but not Y_s) would be distributed uniformly from 0 to 1. A density estimation-based anomaly detection search would then fail: while the signal is mapped to high values under the transformation $Y_s = f(X_s)$, there are no anomalous (i.e. low-density) regions of the background variable Y_b that would be identified and probed for signal.

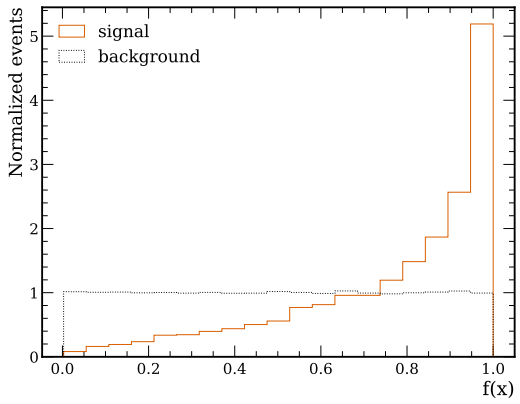
One could imagine even less optimal transformations that produce high background densities where there are high signal densities and low background densities where there are low (or zero) signal densities. One such scenario is illustrated in Fig. 2(c) for the transformation $Y_{b,s} = g(X_{b,s}) = \tanh(X_{b,s} + 2)$. Anomaly detection through density estimation would fail for such a transformation of variables due to the background distribution aligning closely with the signal distribution.

We also illustrate the impact of a change of coordinates when popular anomaly detection algorithms are used to identify the anomalies. We train an Autoencoder, a Normalizing Flow, and a weakly-supervised model based on the Classification Without Labels (CWoLa) paradigm [2, 3, 93]. The dataset before the change of coordinates consists of two-dimensional distributions of background $X_b \sim \mathcal{N}(0,1)$ and signal $X_s \sim \mathcal{N}(1,1)$, with each dimension independent and identically distributed. The two-dimensional dataset is used to ensure the bottleneck layer of the Autoencoder is lower dimensional than the input. The two functions used are the same ones introduced previously: f, i.e. the CDF of a standard normal random variable, and $g(x) = \tanh(x + 2)$.

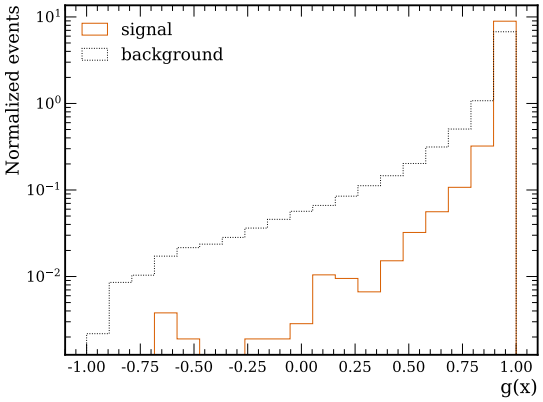
The Autoencoder compresses the two-dimensional data into a one-dimensional latent space using fully-connected layers of sizes (50, 20, 10) and ReLU activation functions before the bottleneck layer of size 1. The decoder is simply the mirrored version of the encoder architecture. Only background events are used during training, and the anomaly score is then defined by the reconstruction loss. The Normalizing Flow is built using a continuous Normalizing Flow [94] with a backbone neural network defined by two stacked fully-connected models with layer sizes (50, 20, 10) and tanh activation. The background-only density is estimated with anomaly score defined as minus the probability density of a single event. Finally, the weakly-supervised model based on CWoLa is trained using a classifier consisting of six fully-connected layers of sizes (50, 50, 20, 20,



(a)



(b)



(c)

FIG. 2. Histograms for the background and signal events in the Gaussian example for (a) the nominal coordinates, (b) after the coordinate transformation $Y = f(X)$ where f is the CDF of a standard normal, and (c) after the coordinate transformation $Y = g(X) = \tanh(X + 2)$.

10, 10) and the ReLU activation function. The model is trained to separate a reference sample of background only events from a mixed sample of signal plus background events, with signal events representing 10% of the overall dataset size. The anomaly score is taken as the ratio $h(x)/(1-h(x))$, where $h(x)$ is the classifier output after a sigmoid activation function. A summary of the anomaly detection methods and the anomaly scores is given in Tab. I with the different model architectures shown in Fig. 3. All methods are implemented using TensorFlow [95] and Adam [96] optimizer with learning rate of 0.001 for 500 epochs or until the validation loss, assessed using an independent dataset, does not improve for 10 consecutive epochs.

TABLE I. Choice of anomaly detection methods and anomaly scores used in this work.

Algorithm	Anomaly score
Autoencoder	$(g(f(x)) - x)^2$
Normalizing flow	$-p_b(x)$
Weakly-supervised	$h(x)/(1-h(x))$

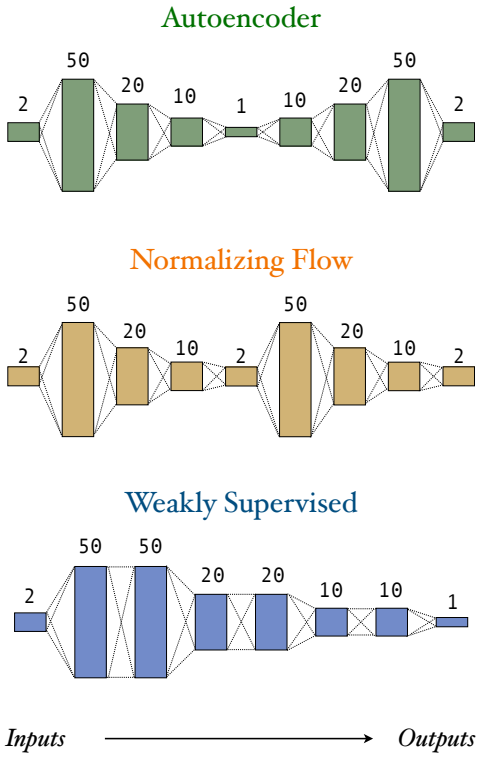


FIG. 3. Network architectures used to implement the anomaly detection methods.

We evaluate the performance for each algorithm using the Receiver Operating Characteristic (ROC) curve for signal and background events, as shown in Fig. 4.

Both the Autoencoder and Normalizing Flow show reasonable performance in the Gaussian example, but both

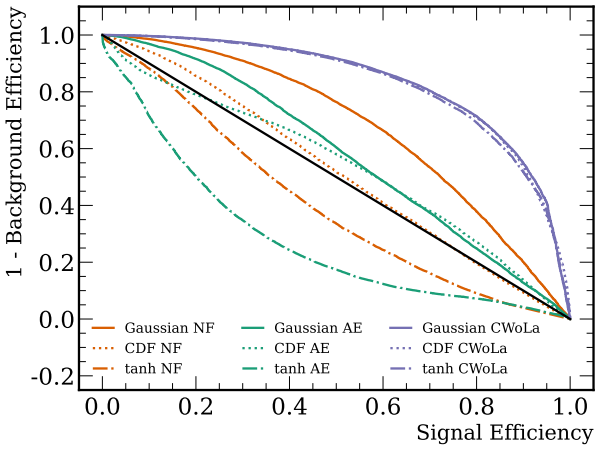


FIG. 4. Receiver operating characteristic (ROC) curve for different anomaly detection algorithms trained using the Gaussian dataset in the original coordinate system (Gaussian) and after the transformations $f(x) = \Phi(x)$ (CDF) and $f(x) = \tanh(x + 2)$ (tanh). The black line denotes the expected ROC curve for a random algorithm.

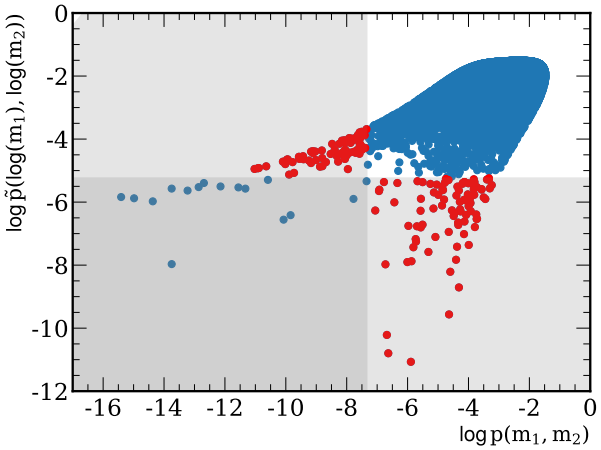


FIG. 5. A comparison of the two probability densities for the same events. The shaded regions and the red dots indicate the 1% most anomalous events.

fail to identify the anomaly after the CDF change of coordinates and even end up systematically removing signal events after the hyperbolic tangent transformation. The weakly-supervised algorithm, on the other hand, identifies the anomaly and shows the same performance for all choices of coordinate systems.

B. LHC Olympics Case

The examples in the previous section were contrived in order to demonstrate the most extreme cases. This section uses realistic HEP observables where the impact is not as dramatic, but the effects of coordinate transformations are still non-negligible.

The dataset used here was originally developed for the LHC Olympics [37] and is briefly described in the following. The background process is dijets and the signal is $W' \to X(\to qq)Y(\to qq)$ with $m_{W'} = 3.5$ TeV, $m_X = 500$ GeV, and $m_Y = 100$ GeV. All events are generated using Pythia8 [97] and Delphes3.4.1 [98–100]. The jets are clustered using FastJet [101] with the anti-k_T algorithm [102] using $R = 1$. Finally, all events are required to have at least one jet with $p_T > 1.2$ TeV.

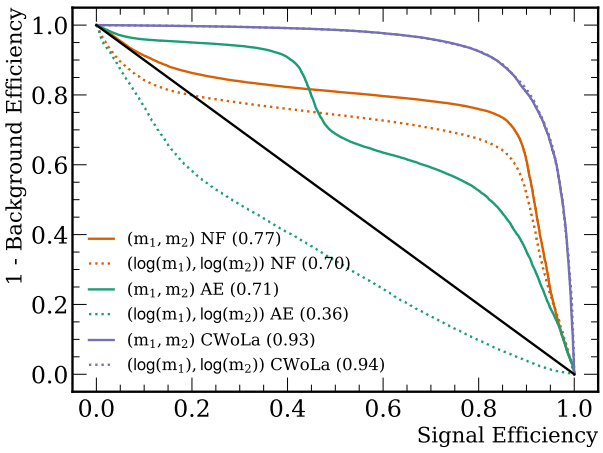
Some important discriminating features in the LHC Olympics dataset are the masses of the leading and subleading jets. In particular, the masses of the leading (m_1) and subleading (m_2) jets should approximately correspond to the masses of the X and Y particles for the signal. Since the masses have a large kinematic range, they are often preprocessed by taking the natural logarithm, $m \mapsto \log(m/\text{TeV})$ (henceforth, the units are suppressed). Other natural examples include the n-subjettiness observables τ_1 and τ_2 [103, 104]. These observables quantify the extent to which a jet is more consistent with having one or two prongs. The variable τ_1 captures similar properties of the jet radiation pattern as the jet mass. A researcher aiming to pre-process as minimally as possible might attempt to do anomaly detection with (τ_1, τ_2) directly, while someone wanting to use standard pre-processing might use instead $(\tau_1, \tau_2/\tau_1)$. The n-subjettiness ratio $\tau_{21} = \tau_2/\tau_1$ is one of the most widely used taggers for identifying two-prong substructure. This is characteristic of Lorentz-boosted W/Z boson decays, but it is also the case for our BSM particles X and Y. We show results for m and $\log(m)$, but we found similar, although less dramatic, results for n-subjettiness.

If (m_1, m_2) is described by probability density p, then the transformed coordinates are described by density $\tilde{p}(\log(m_1), \log(m_2)) = p(m_1, m_2)m_1m_2$. This shows that the ordering by anomaly score can be reversed depending on the relative sizes of p, m_1 and m_2.

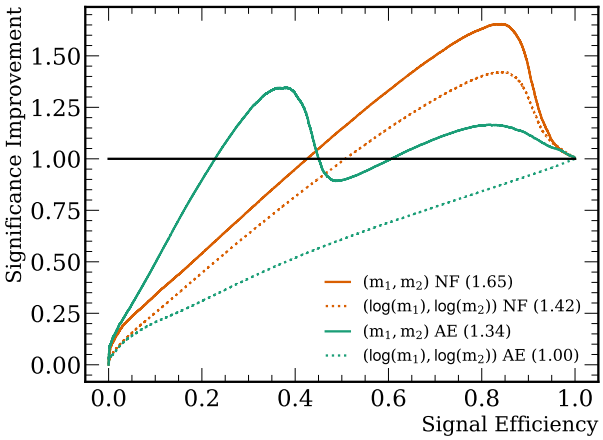
Unlike in the Gaussian case, for the LHC Olympics dataset we do not know the probability densities analytically and so we can only estimate the densities numerically. A comparison of the selected anomalies in the background-only case with a NF are presented in Fig. 5. The shaded regions in Fig. 5 indicate the selected anomalies using a 1% criteria. Interestingly, the two selections agree on only about 20% of events. This means that **even though we have the same events and the same input features, we have different anomaly selections depending on the coordinates we use to represent the events.**

Analogously to the previous section, we compare anomaly detection strategies in Fig. 6. We employ the same neural network models and hyperparameters as the ones used in the Gaussian example. Once again, the performance of the weakly-supervised training is independent from the choice of coordinates, while all other algo-

rithms show differences in performance based on the initial choice of coordinates. The AE and NF have a similar performance, reinforcing the claim that the approaches are targeting similar regions of phase space. However, the change in performance after the coordinate transformation is more pronounced for the AE, which may have other contributions aside from the indirect density estimation.



(a)



(b)

FIG. 6. (a) Receiver Operating Characteristic (ROC) curves for the three anomaly detection algorithms evaluated using the LHC Olympics dataset for different choices of inputs. The black line denotes the expected ROC curve for a random algorithm. The number in parentheses represents the area under the curve. (b) Same as (a), but instead of the background efficiency, the dependent variable is the Significance Improvement Characteristic (SIC) defined as the signal efficiency divided by the square root of the background efficiency.

V. CONCLUSIONS AND OUTLOOK

In this paper, we have described the sensitivity of anomaly detection approaches to coordinate transformations. We have connected BSM hypotheses with ML strategies in order to make explicit what is being assumed and when coordinate transformations are potentially important. While it is not new, we have highlighted the coordinate sensitivity of unsupervised approaches (targeting 'rare' events). There is no optimal set of coordinates *a priori*, but for a given signal hypothesis, some set of coordinates will be optimal for a particular unsupervised learning algorithm. This does not mean that we should not use unsupervised algorithms – on the contrary, these approaches provide valuable complementarity to other less-than-supervised methods. However, we should be cautious about optimal claims, and it seems wise to explore multiple coordinate systems when determining the sensitivity. While weakly-supervised approaches are formally coordinate-independent, it could be that in practice some set of coordinates enables more efficient learning. These and other practical issues are critically important to explore as anomaly detection proposals become physics results in the near future.

CODE AVAILABILITY

The code for this paper can be found at https://github.com/ViniciusMikuni/Rareisnotuniversal.

ACKNOWLEDGMENTS

We are grateful to the participants, speakers, and co-organizers of the PhyStat-Anomaly workshop (https://indico.cern.ch/event/1138933/) for many useful discussions before, at, and after the meeting. We additionally thank the participants of Hammers and Nails 2022 for useful discussions, especially Kyle Cranmer. We also thank Sascha Caron, Bob Cousins, and Louis Lyons for feedback on the manuscript. RM, VM, BN, and MP are supported by the U.S. Department of Energy (DOE), Office of Science under contract DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award HEP-ERCAP0021099. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The work of DS was supported by DOE grant DOE-SC0010008. GK acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2121 "Quantum Universe" – 390833306.

[1] R. T. D'Agnolo and A. Wulzer, Phys. Rev. **D99**, 015014 (2019), arXiv:1806.02350 [hep-ph].
[2] J. H. Collins, K. Howe, and B. Nachman, Phys. Rev. Lett. **121**, 241803 (2018), arXiv:1805.02664 [hep-ph].
[3] J. H. Collins, K. Howe, and B. Nachman, Phys. Rev. **D99**, 014038 (2019), arXiv:1902.02634 [hep-ph].
[4] R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, (2019), 10.1140/epjc/s10052-021-08853-y, arXiv:1912.12155 [hep-ph].
[5] M. Farina, Y. Nakai, and D. Shih, (2018), 10.1103/PhysRevD.101.075021, arXiv:1808.08992 [hep-ph].
[6] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, SciPost Phys. **6**, 030 (2019), arXiv:1808.08979 [hep-ph].
[7] T. S. Roy and A. H. Vijay, (2019), arXiv:1903.02032 [hep-ph].
[8] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J.-R. Vlimant, JHEP **05**, 036 (2019), arXiv:1811.10276 [hep-ex].
[9] A. Blance, M. Spannowsky, and P. Waite, JHEP **10**, 047 (2019), arXiv:1905.10384 [hep-ph].
[10] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, (2018), 10.1103/PhysRevD.101.076015, arXiv:1807.10261 [hep-ph].
[11] A. De Simone and T. Jacques, Eur. Phys. J. **C79**, 289 (2019), arXiv:1807.06038 [hep-ph].
[12] A. Mullin, H. Pacey, M. Parker, M. White, and S. Williams, (2019), 10.1007/JHEP02(2021)160, arXiv:1912.10625 [hep-ph].
[13] G. M. Alessandro Casa, (2019), arXiv:1809.02977 [hep-ex].
[14] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, Phys. Rev. **D100**, 056002 (2019), arXiv:1904.04200 [hep-ph].
[15] A. Andreassen, B. Nachman, and D. Shih, Phys. Rev. D **101**, 095004 (2020), arXiv:2001.05001 [hep-ph].
[16] B. Nachman and D. Shih, Phys. Rev. D **101**, 075042 (2020), arXiv:2001.04990 [hep-ph].
[17] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, JHEP **11**, 163 (2017), arXiv:1709.01087 [hep-ph].
[18] M. Romão Crispim, N. Castro, R. Pedro, and T. Vale, Phys. Rev. D **101**, 035042 (2020), arXiv:1912.04220 [hep-ph].
[19] M. C. Romao, N. Castro, J. Milhano, R. Pedro, and T. Vale, (2020), 10.1140/epjc/s10052-021-08891-6, arXiv:2004.09360 [hep-ph].
[20] O. Knapp, G. Dissertori, O. Cerri, T. Q. Nguyen, J.-R. Vlimant, and M. Pierini, (2020), 10.1140/epjp/s13360-021-01109-4, arXiv:2005.01598 [hep-ex].
[21] ATLAS Collaboration, (2020), 10.1103/PhysRevLett.125.131801, arXiv:2005.02983 [hep-ex].
[22] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc, (2020), 10.1007/JHEP10(2020)206, arXiv:2005.12319 [hep-ph].
[23] M. C. Romao, N. Castro, and R. Pedro, (2020), 10.1140/epjc/s10052-020-08807-w, arXiv:2006.05432 [hep-ph].
[24] O. Amram and C. M. Suarez, (2020), 10.1007/JHEP01(2021)153, arXiv:2002.12376 [hep-ph].
[25] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, and T. Golling, (2020), arXiv:2007.01850 [hep-ph].
[26] C. K. Khosa and V. Sanz, (2020), arXiv:2007.14462 [cs.LG].
[27] P. Thaprasop, K. Zhou, J. Steinheimer, and C. Herold, (2020), arXiv:2007.15830 [hep-ex].
[28] S. Alexander, S. Gleyzer, H. Parul, P. Reddy, M. W. Toomey, E. Usai, and R. Von Klar, (2020), arXiv:2008.12731 [astro-ph.CO].
[29] J. A. Aguilar-Saavedra, F. R. Joaquim, and J. F. Seabra, (2020), 10.1007/JHEP03(2021)012, arXiv:2008.12792 [hep-ph].
[30] K. Benkendorfer, L. L. Pottier, and B. Nachman, (2020), arXiv:2009.02205 [hep-ph].
[31] Adrian Alan Pol and Victor Berger and Gianluca Cerminara and Cecile Germain and Maurizio Pierini, (2020), arXiv:2010.05531 [cs.LG].
[32] V. Mikuni and F. Canelli, (2020), arXiv:2010.07106 [physics.data-an].
[33] M. van Beekveld, S. Caron, L. Hendriks, P. Jackson, A. Leinweber, S. Otten, R. Patrick, R. Ruiz de Austri, M. Santoni, and M. White, (2020), arXiv:2010.07940 [hep-ph].
[34] S. E. Park, D. Rankin, S.-M. Udrescu, M. Yunus, and P. Harris, (2020), arXiv:2011.03550 [hep-ph].
[35] D. A. Faroughy, (2020), arXiv:2012.08579 [hep-ph].
[36] G. Stein, U. Seljak, and B. Dai, (2020), arXiv:2012.11638 [cs.LG].
[37] G. Kasieczka *et al.*, (2021), arXiv:2101.08320 [hep-ph].
[38] P. Chakravarti, M. Kuusela, J. Lei, and L. Wasserman, (2021), arXiv:2102.07679 [stat.AP].
[39] J. Batson, C. G. Haaf, Y. Kahn, and D. A. Roberts, (2021), arXiv:2102.08380 [hep-ph].
[40] A. Blance and M. Spannowsky, (2021), arXiv:2103.03897 [hep-ph].
[41] B. Bortolato, B. M. Dillon, J. F. Kamenik, and A. Smolkovič, (2021), arXiv:2103.06595 [hep-ph].
[42] J. H. Collins, P. Martín-Ramiro, B. Nachman, and D. Shih, (2021), arXiv:2104.02092 [hep-ph].
[43] B. M. Dillon, T. Plehn, C. Sauer, and P. Sorrenson, (2021), arXiv:2104.08291 [hep-ph].
[44] T. Finke, M. Krämer, A. Morandini, A. Mück, and I. Oleksiyuk, (2021), arXiv:2104.09051 [physics.data-an].
[45] D. Shih, M. R. Buckley, L. Necib, and J. Tamanas, (2021), arXiv:2104.12789 [astro-ph.GA].
[46] O. Atkinson, A. Bhardwaj, C. Englert, V. S. Ngairangbam, and M. Spannowsky, (2021), arXiv:2105.07988 [hep-ph].
[47] A. Kahn, J. Gonski, I. Ochoa, D. Williams, and G. Brooijmans, (2021), arXiv:2105.09274 [hep-ph].
[48] T. Aarrestad *et al.*, (2021), arXiv:2105.14027 [hep-ph].
[49] T. Dorigo, M. Fumanelli, C. Maccani, M. Mojsovska, G. C. Strong, and B. Scarpa, (2021), arXiv:2106.05747 [physics.data-an].
[50] S. Caron, L. Hendriks, and R. Verheyen, (2021), arXiv:2106.10164 [hep-ph].
[51] E. Govorkova, E. Puljak, T. Aarrestad, M. Pierini, K. A. Woźniak, and J. Ngadiuba, (2021), arXiv:2107.02157 [physics.data-an].

[52] G. Kasieczka, B. Nachman, and D. Shih (2021) arXiv:2107.02821 [stat.ML].
[53] S. Volkovich, F. D. V. Halevy, and S. Bressler, (2021), arXiv:2107.11573 [hep-ex].
[54] E. Govorkova *et al.*, (2021), arXiv:2108.03986 [physics.ins-det].
[55] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, (2021), arXiv:2109.00546 [hep-ph].
[56] B. Ostdiek, (2021), arXiv:2109.01695 [hep-ph].
[57] K. Fraser, S. Homiller, R. K. Mishra, B. Ostdiek, and M. D. Schwartz, (2021), arXiv:2110.06948 [cs.LG].
[58] P. Jawahar, T. Aarrestad, M. Pierini, K. A. Wozniak, J. Ngadiuba, J. Duarte, and S. Tsan, (2021), arXiv:2110.08508 [hep-ph].
[59] J. Herrero-Garcia, R. Patrick, and A. Scaffidi, (2021), arXiv:2110.12248 [hep-ph].
[60] J. A. Aguilar-Saavedra, (2021), arXiv:2111.02647 [hep-ph].
[61] R. Tombs and C. G. Lester, (2021), arXiv:2111.05442 [hep-ph].
[62] C. G. Lester and R. Tombs, (2021), arXiv:2111.00616 [hep-ph].
[63] V. Mikuni, B. Nachman, and D. Shih, (2021), arXiv:2111.06417 [cs.LG].
[64] S. V. Chekanov and W. Hopkins, (2021), arXiv:2111.12119 [hep-ph].
[65] R. T. d'Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, (2021), arXiv:2111.13633 [hep-ph].
[66] F. Canelli, A. de Cosa, L. L. Pottier, J. Niedziela, K. Pedro, and M. Pierini, (2021), arXiv:2112.02864 [hep-ph].
[67] V. S. Ngairangbam, M. Spannowsky, and M. Takeuchi, (2021), arXiv:2112.04958 [hep-ph].
[68] L. Bradshaw, S. Chang, and B. Ostdiek, (2022), arXiv:2203.01343 [hep-ph].
[69] J. A. Aguilar-Saavedra, (2022), arXiv:2201.11143 [hep-ph].
[70] S. Alvi, C. Bauer, and B. Nachman, (2022), arXiv:2206.08391 [hep-ph].
[71] B. M. Dillon, R. Mastandrea, and B. Nachman, (2022), arXiv:2205.10380 [hep-ph].
[72] M. Birman, B. Nachman, R. Sebbah, G. Sela, O. Turetz, and S. Bressler, Eur. Phys. J. C **82**, 508 (2022), arXiv:2203.07529 [hep-ph].
[73] J. A. Raine, S. Klein, D. Sengupta, and T. Golling, (2022), arXiv:2203.09470 [hep-ph].
[74] M. Letizia, G. Losapio, M. Rando, G. Grosso, A. Wulzer, M. Pierini, M. Zanetti, and L. Rosasco, (2022), arXiv:2204.02317 [hep-ph].
[75] C. Fanelli, J. Giroux, and Z. Papandreou, (2022), arXiv:2204.08609 [cs.LG].
[76] R. Verheyen, (2022), arXiv:2205.01697 [hep-ph].
[77] S. Caron, R. R. de Austri, and Z. Zhang, (2022), arXiv:2207.07631 [hep-ph].
[78] B. M. Dillon, L. Favaro, T. Plehn, P. Sorrenson, and M. Krämer, (2022), arXiv:2206.14225 [hep-ph].
[79] T. Buss, B. M. Dillon, T. Finke, M. Krämer, A. Morandini, A. Mück, I. Oleksiyuk, and T. Plehn, (2022), arXiv:2202.00686 [hep-ph].
[80] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, (2021), arXiv:2112.03769 [hep-ph].
[81] C. Le Lan and L. Dinh, Entropy **23** (2021), 10.3390/e23121690.
[82] D. P. Kingma and M. Welling, (2014), arXiv:1312.6114 [stat.ML].
[83] D. P. Kingma and M. Welling, Foundations and Trends in Machine Learning **12**, 307 (2019).
[84] S. Yoon, Y.-K. Noh, and F. C. Park, "Autoencoding under normalization constraints," (2021).
[85] I. J. Goodfellow *et al.*, Conference on Neural Information Processing Systems **2**, 2672 (2014).
[86] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, IEEE Signal Processing Magazine **35**, 53 (2018).
[87] D. J. Rezende and S. Mohamed, International Conference on Machine Learning **37**, 1530 (2015).
[88] I. Kobyzev, S. Prince, and M. Brubaker, IEEE Transactions on Pattern Analysis and Machine Intelligence , 1 (2020).
[89] A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, T. Plehn, D. Shih, and R. Winterhalder, (2022), arXiv:2202.09375 [hep-ph].
[90] K. Krzyzanska and B. Nachman, (2022), arXiv:2203.09601 [hep-ph].
[91] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer New York Inc., New York, NY, USA, 2001).
[92] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning* (Cambridge University Press, 2012).
[93] E. M. Metodiev, B. Nachman, and J. Thaler, JHEP **10**, 174 (2017), arXiv:1708.02949 [hep-ph].
[94] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, arXiv preprint arXiv:1810.01367 (2018).
[95] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015), software available from tensorflow.org.
[96] D. P. Kingma and J. Ba, *3rd International Conference on Learning Representations, ICLR 2015*, (2015), arXiv:1412.6980 [cs.LG].
[97] T. Sjöstrand, S. Mrenna, and P. Skands, Computer Physics Communications **178**, 852–867 (2008).
[98] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, Journal of High Energy Physics **2014** (2014), 10.1007/jhep02(2014)057.
[99] A. Mertens, Journal of Physics: Conference Series **608**, 012045 (2015).
[100] M. Selvaggi, Journal of Physics: Conference Series **523**, 012033 (2014).
[101] M. Cacciari, G. P. Salam, and G. Soyez, The European Physical Journal C **72** (2012), 10.1140/epjc/s10052-012-1896-2.
[102] M. Cacciari, G. P. Salam, and G. Soyez, Journal of High Energy Physics **2008**, 063 (2008).
[103] J. Thaler and K. Van Tilburg, Journal of High Energy Physics **2012** (2012), 10.1007/jhep02(2012)093.

[104] J. Thaler and K. Van Tilburg, Journal of High Energy Physics **2011** (2011), 10.1007/jhep03(2011)015.