Anomaly detection with density estimation

Benjamin Nachman[1,*] and David Shih[1,2,3,†]

[1]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[2]*NHETC, Department of Physics and Astronomy, Rutgers, Piscataway, New Jersey 08854, USA*
[3]*Berkeley Center for Theoretical Physics, University of California, Berkeley, California 94720, USA*

(Received 6 February 2020; accepted 6 April 2020; published 23 April 2020)

We leverage recent breakthroughs in neural density estimation to propose a new unsupervised ANOmaly detection with Density Estimation (ANODE) technique. By estimating the conditional probability density of the data in a signal region and in sidebands, and interpolating the latter into the signal region, a fully data-driven likelihood ratio of data versus background can be constructed. This likelihood ratio is broadly sensitive to overdensities in the data that could be due to localized anomalies. In addition, a unique potential benefit of the ANODE method is that the background can be directly estimated using the learned densities. Finally, ANODE is robust against systematic differences between signal region and sidebands, giving it broader applicability than other methods. We demonstrate the power of this new approach using the LHC Olympics 2020 R&D dataset. We show how ANODE can enhance the significance of a dijet bump hunt by up to a factor of 7 with a 10% accuracy on the background prediction. While the LHC is used as the recurring example, the methods developed here have a much broader applicability to anomaly detection in physics and beyond.

DOI: 10.1103/PhysRevD.101.075042

I. INTRODUCTION

Despite an impressive and extensive search program from ATLAS [1–3], CMS [4–6], and LHCb [7] for new particles and forces of nature, there is no convincing evidence for new phenomena at the Large Hadron Collider (LHC). However, there remain compelling theoretical (e.g., naturalness) and experimental (e.g., dark matter) reasons for fundamental structure to be observable with current LHC sensitivity. The vast majority of LHC searches are designed with specific signal models motivated by one of these reasons (e.g., gluino pair production from supersymmetry) in mind, and these searches are optimized with a heavy reliance on simulations, for both the signal and the Standard Model (SM) background. Given that it is impossible to cover every model with a specially optimized search (see e.g., [8,9] for comprehensive lists of currently uncovered models), and given that there are vast regions of unexplored LHC phase space, it is critical to consider extending the search program to include more model-agnostic methods.

A variety of model-agnostic approaches have been proposed to search for physics beyond the Standard Model (BSM) at colliders. These approaches are designed to be broadly sensitive to anomalies in data without focusing on specific models. Yet, they have varying degrees of both *signal model* and *background model independence*, as there is often a tradeoff between the broadness of a search and how sensitive it is to particular classes of signal scenarios. Existing and proposed model-agnostic searches range from fully signal model independent but fully background model dependent [10–26] (because they compare data to SM simulation), to varying degrees of partial signal model and background model independence [27–40]. A comprehensive overview of existing model-agnostic approaches and how they are classified in terms of signal and background model independence will be given in Sec. II.

This paper introduces a new approach called ANOmaly detection with Density Estimation (ANODE) that is complementary to existing methods and aims to be largely background and signal model agnostic. Density estimation, especially in high dimensions, has traditionally been a difficult problem in unsupervised machine learning. The objective of density estimation is to learn the underlying probability density from which a set of independent and identically distributed examples were drawn. In the past few years, there have been a number of breakthroughs in density estimation using neural networks and the performance of high-dimensional density estimation has greatly improved. The idea of ANODE is to make use of these

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

recent breakthroughs in order to directly estimate the probability density of the data. Assuming the signal is localized somewhere, one can attempt to use sideband methods and interpolation to estimate the probability density of the background. Then, one can use this to construct a likelihood ratio generally sensitive to new physics.

As with any search for BSM, it is not enough to have a discriminant that is sensitive to signals, one must also have a valid method of background estimation, otherwise it will be impossible to claim a discovery of new physics. The method of background estimation can further introduce possible sources of signal and background model dependence, and it is important to avail oneself of data-driven background methods in any truly model-agnostic search. This paper will explore two methods of data-driven background estimation, one based on importance sampling, and the other based on directly integrating the background density estimate obtained in the ANODE procedure.

Other neural network approaches to density estimation have been studied in high energy physics. Such methods include generative adversarial networks (GANs) [41–67], autoencoders [56,68], physically inspired networks [69,70], and flows [71,72]. GANs are efficient for sampling from a density and are thus promising for accelerating slow simulations, but they do not provide an explicit representation of the density itself. For this reason, ANODE is built using normalizing flows [71] and in particular the recently proposed masked autoregressive flow (MAF) [73]. These methods estimate densities by using a succession of neural networks to gradually map the original data to a transformed dataset that follows a simple distribution (e.g., normal or uniform).

The ANODE method is demonstrated using a simulated large-radius dijet search based on the LHC Olympics 2020 R&D dataset [74]. In particular, properties of hadronic jets are used as discriminating features to enhance a bump hunt in the invariant mass of pairs of jets. ANODE learns a parametrized density of the features using a sideband and this is combined with a density estimation of the same features in the signal region. The resulting likelihood ratio is able to enhance the sensitivity of a traditional bump hunt from $S/\sqrt{B} \sim 1$ to $S/\sqrt{B} \gg 5$. There is currently no dedicated search for generic dijet signatures where each of the jets can also originate from a BSM resonance [8,75–78]. Therefore, this particular application could be directly useful for extending the LHC physics search program. Many other applications to resonant new physics searches involving jets and other final states are also possible.

In order to benchmark the performance of ANODE, it is compared with the compared with classification without labels (CWoLa) hunting method [33,34]. The CWoLa approach is also a neural network-based resonance search, but does not involve density estimation. Instead, CWoLa hunting uses neural networks to identify differences between signal regions and neighboring sideband regions. By turning the problem into a supervised learning task [79], CWoLa is able to effectively find rare resonant signals. However, CWoLa hunting has certain requirements on the independence of the discriminating features and the resonant feature. ANODE does not have this requirement, and the potential for exploiting correlated features is studied by introducing correlations.

This paper is organized as follows. Section II reviews the landscape of model-independent searches at the LHC to provide context for the ANODE method. Section III introduces the details of the ANODE approach and provides a brief introduction to normalizing flows. The reminder of the paper illustrates ANODE through an example based on a dijet search using jet substructure. Details of the simulated samples are provided in Sec. IV, and the results for the signal sensitivity and background specificity are presented in Sec. V A and V B, respectively. A study of correlations between the discriminating features and the resonant feature is in Sec. V C. The paper ends with conclusions and outlook in Sec. VI.

II. AN OVERVIEW OF MODEL-(IN)DEPENDENT SEARCHES

A viable search for new physics generally must have two essential components: it must be sensitive to new phenomena and it must also be able to estimate the background under the null hypothesis (Standard Model only). The categorization of a search's degree of model (in)dependence requires consideration of both of these components. Figure 1 illustrates how to characterize model independence for both BSM sensitivity and SM background specificity. We will now consider each in turn.

A. BSM sensitivity

For BSM sensitivity, the various types of searches are categorized as follows:

(i) Almost all searches at the LHC are optimized (with or without machine learning) using simulations of both the SM and particular signal models. This is represented as the lower-left corner of Fig. 1(a).

(ii) A handful of searches use signal simulation and unlabeled data to optimize the event selection. These are background model agnostic and are depicted in the upper-left corner of Fig. 1(a). For example, this was used in the $\gamma\gamma$ channel of the recent $t\bar{t}h$ observation, using events with inverted selection criteria to define the background data sample for optimization [81,82].

(iii) A series of signal model agnostic, but background model-dependent searches have been performed by D0 [10–13], H1 [14,15], ALEPH [16], CDF [17–19], CMS [20,21], and ATLAS [22–24]. All of these searches share essentially the same approach: they

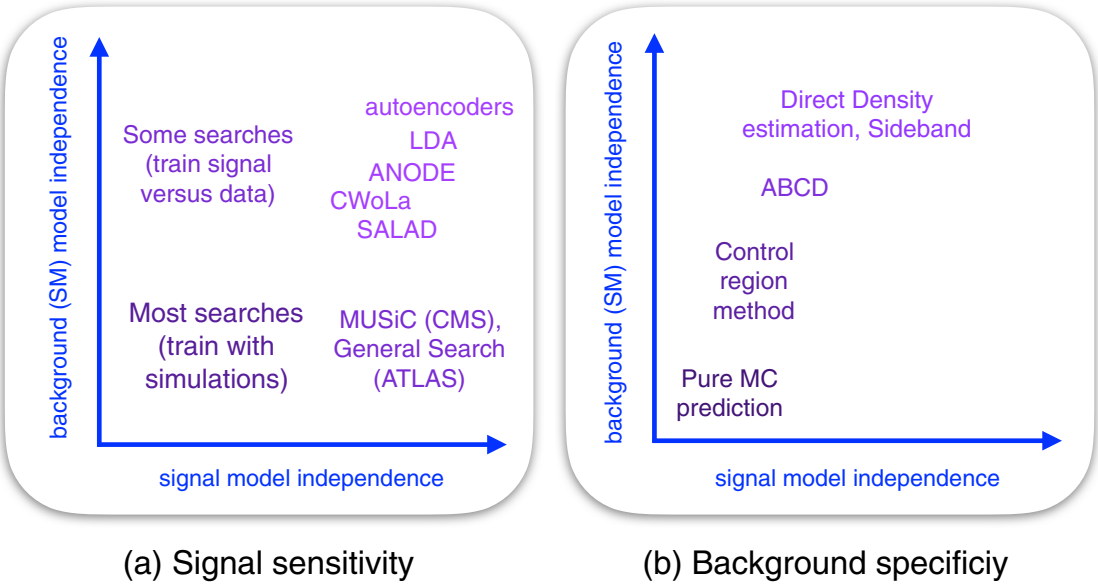

FIG. 1. A graphical representation of searches for new particles in terms of the background and signal model dependence for achieving signal sensitivity (a) and background specificity (b). The Model Unspecific Search for New Physics (MUSiC) [20,21] and general search [22–24] strategies are from CMS and ATLAS, respectively. LDA stands for latent dirichlet allocation [38,80], ANOmaly detection with Density Estimation (ANODE) is the method presented in this paper, CWoLa stands for classification without labels [33,34,79], and SALAD stands for simulation assisted likelihood-free anomaly detection [39]. Direct density estimation is a form of sideband where the multidimensional feature space density is learned conditional on the resonant feature (see Sec. III B).

compared histograms of data to histograms of SM simulations and looked for discrepancies. Such searches are represented in the lower-right part of Fig. 1(a). Recently, there have been proposals to extend these searches with deep learning [25,26].

(iv) More recently, a variety of approaches have been proposed, often relying on sophisticated deep learning techniques, that attempt to be both signal and background model agnostic, to varying degrees. These include approaches based on autoencoders [27–32], weak supervision [33,34], nearest neighbor algorithms [35–37], probabilistic modeling [38], reweighted simulation [39], and others [40]. These are indicated in the upper-right corner of Fig. 1(a).

In the upper-right corner of Fig. 1(a), we have also attempted to illustrate in finer detail the differences between some recent model-agnostic approaches. For example, the autoencoder is in the farthest corner since it assumes almost nothing about the signal or the background but can be run directly on the data, as long as the signal is sufficiently rare [27,28]. The tradeoff is that there is no optimality guarantee for the autoencoder—any signal that it does find will be found in a rather uncontrolled manner. Meanwhile, CWoLa hunting [33,34] is somewhat more signal and background model dependent than autoencoders, since this approach assumes that the signal is localized in a particular feature, and that there is an uncorrelated set of additional features on which one can train a classifier to distinguish signal region and sideband. In return, one obtains a guarantee of asymptotic optimality—the classifier approaches the likelihood ratio [83] in the limit of infinite statistics.[1]

The ANODE method introduced in this paper complements the other recently proposed techniques and is asymptotically optimal. To do this, ANODE estimates the density of the background-only scenario using sidebands and compares that with the density estimated in a signal-sensitive region (details are in Sec. III). Like the CWoLa hunting method, the new approach is broadly sensitive to resonant new physics and thus it is placed in the upper-right part of Fig. 1(a). The reason that ANODE is further right and above CWoLa hunting is that it is less sensitive to correlations, a feature that is discussed more below.

B. Background estimation

A variety of methods are commonly used for background estimation and are highlighted in Fig. 1(b). Generally, background estimation is less dependent on the signal model than achieving signal sensitivity and therefore the x-axis range of Fig. 1(b) is more compressed than Fig. 1(a).

(i) In some cases, the simulation is used to directly estimate the background. This is often the case for well-understood backgrounds such as electroweak phenomena or very rare processes that are difficult to constrain with data.

[1]See the Appendix for more details about "optimality."

TABLE I. A table with the common pairings of search strategy for signal sensitivity (left column), the background estimation (middle column), and an example search (right column).

Search	Typical background strategy	Recent examples
MUSiC and the general search	Pure MC prediction	[20,22]
Pure electroweak processes	Pure MC prediction	[85]
SUSY with top quarks and W bosons	Control region method	[86,87]
All-hadronic searches	ABCD method	[88,89]
Long-lived particle searches	ABCD method	[90,91]
BSM resonance searches	Sideband method	[92,93]
CWoLa hunting	Sideband method	[33,34]
ANODE	Sideband or direct density	This paper

(ii) Most searches use data in some way to constrain the background prediction. One common approach is the *control region method*, where a search is complemented by an auxiliary measurement to constrain the simulation. Knowledge of the signal is used to ensure that the auxiliary measurement is not biased by the presence of signal.

(iii) The two most common methods for background estimates that do not directly use simulation are the *ABCD method* and the *sideband method* (bump hunt). The ABCD method operates by identifying two independent features, each of which is sensitive to the presence of signal. Four regions, labeled A, B, C, and D are constructed by (anti)requiring a threshold on the two features. The background rate in the most signal sensitive region is estimated from the other three regions. Background simulations are required to verify independence of the two features.

(iv) Finally, the sideband fit only requires that the background be smooth in the region of a potential signal so that a parametric (or not [84]) function can be fit to sidebands and interpolated. However, this method only works for resonant new physics.

While strategies from Fig. 1(a) can often be matched with any approach in Fig. 1(b), there is often one combination that is used in practice. Table I provides examples of various searches and the background estimation technique that typically is associated with that search. Searches with a complex background may use multiple background estimation procedures.

ANODE can be combined with any background estimation technique, but it can also be used directly since the background density is already estimated to construct a signal-sensitive classifier. Even though directly providing an accurate background estimation puts stringent requirements on the accuracy of the density estimation, it also reduces the need for a full decorrelation between classification features and the resonant feature. A variety of decorrelation techniques exist [94–104], but ultimately decorrelating removes information available for classification.

III. THE ANODE METHOD

This section will describe the ANODE proposal for an unsupervised method to search for resonant new physics using density estimation.

Let m be a feature in which a signal (if it exists) is known to be localized around some m_0. The value of m_0 will be scanned for broad sensitivity and the following procedure will be repeated for each window in m. It is often the case that the width of the signal in m is fixed by detector properties and is signal model independent. A region $m_0 \pm \delta$ is called the signal region (SR) and $m \notin [m_0 - \delta, m_0 + \delta]$ is defined as the sideband region (SB). A traditional, unsupervised, model-agnostic search is to perform a bump hunt in m, using the SB to interpolate into the SR in order to estimate the background.

Let $x \in \mathbb{R}^d$ be some additional discriminating features in which the signal density is different than the background density. If we could find the region(s) where the signal differs from the background and then cut on x to select these regions, we could improve the sensitivity of the original bump hunt in m. The goal of ANODE is to accomplish this in an unsupervised and model-agnostic way, via density estimation in the feature space x.

More specifically, ANODE attempts to learn two densities $p_{\text{data}}(x|m)$ and $p_{\text{background}}(x|m)$ for $m \in \text{SR}$. Then, classification is performed with the likelihood ratio,

$$R(x|m) = \frac{p_{\text{data}}(x|m)}{p_{\text{background}}(x|m)}. \tag{3.1}$$

In the ideal case that $p_{\text{data}}(x|m) = \alpha p_{\text{background}}(x|m) + (1-\alpha) p_{\text{signal}}(x|m)$ for $0 \leq \alpha \leq 1$ and $m \in \text{SR}$, Eq. (3.1) is the optimal test statistic for identifying the presence of signal. In the absence of signal, $R(x|m) = 1$, so as long as $p_{\text{signal}}(x|m) \neq p_{\text{background}}(x|m)$, $R_{\text{data}}(x|m)$ has a nonzero density away from 1 in a region with no predicted background.

In practice, both $p_{\text{data}}(x|m)$ and $p_{\text{background}}(x|m)$ are approximations and so $R(x|m)$ is not unity in the absence of signal. The densities $p(x|m)$ are estimated using conditional neural density estimation as described in Sec. III A.

The function $p_{\text{data}}(x|m)$ is estimated in the signal region and the function $p_{\text{background}}(x|m)$ is estimated using the sideband region and then interpolated into the signal region. The interpolation is done automatically by the neural conditional density estimator. Effective density estimation will result in $R(x|m)$ in the SR that is localized near unity and then one can enhance the presence of signal by applying a threshold $R(x|m) > R_{\text{cut}}$ for $R_{\text{cut}} > 1$. The interpolated $p_{\text{background}}(x|m)$ can then also be used to estimate the background, as described in Sec. III B.

A. Neural density estimation

The ANODE procedure as described in the previous subsection is completely general with regards to the method of density estimation. In this work, we will demonstrate a proof of concept using normalizing flow models for density estimation. Since normalizing flows were proposed in Ref. [71], they have generated much activity and excitement in the machine learning community, achieving state-of-the-art performance on a variety of benchmark density estimation tasks.

The core idea behind a normalizing flow is to apply a change of variables from a random variable with a simple density (e.g., Gaussian or uniform) to one with a complex density that matches some training dataset. The transformation from one density describing random variable X to another density describing random variable Y follows the usual change of variables formula using the Jacobian,

$$p_Y(y) = p_X(x)\left|\det\left(\frac{\partial f}{\partial x}\right)\right|^{-1}, \tag{3.2}$$

where x and y are realizations of X and Y, respectively, X and Y have the same dimension, and $Y = f(X)$ is an invertible function. The process in Eq. (3.2) can be repeated to build a normalizing flow,

$$p_Y(y) = p_X(x)\prod_{i=1}^{N}\left|\det\left(\frac{\partial f_i}{\partial y_{i-1}}\right)\right|^{-1}, \tag{3.3}$$

where $Y_i = f_i(Y_{i-1})$, $Y_0 = X$, and $Y = f_N(Y_{N-1})$. The first neural density estimation with normalizing flows had the following form for $x \in \mathbb{R}^n$:

$$f(x) = x + \bar{x}\sigma(w \cdot x + b), \tag{3.4}$$

where σ is an elementwise nonlinearity and $\bar{x} \in \mathbb{R}^n, w \in \mathbb{R}^n, b \in \mathbb{R}$ are trainable parameters. The benefit of Eq. (3.4) is that the Jacobian evaluation is simple from the chain rule. Since the first development of normalizing flows, there has been significant development in extending their expressivity. One innovation is to combine flows with autoregressive density estimation [105]. An autoregressive flow [106] modifies the change of variables so that for $Y_i = f(X_i)$, $Y_{i,\alpha} = f_{i,\alpha}(X_{i,1}, \ldots, X_{i,\alpha})$, where the indices α denote the dimension of X_i and Y_i for $\alpha = 1, \ldots, n$. Any f that satisfies this condition is amenable to neural density estimation because the Jacobian determinant evaluation is simple. In particular, the Jacobian is upper triangular and therefore the determinant is the product of the diagonal elements: $\prod_{\alpha=1}^{n} \partial f_{i,\alpha}/\partial x_\alpha$. ANODE is built on an MAF [73]. For an MAF,

$$Y_{i,\alpha} = \mu_{i,\alpha}(Y_{i,1}, \ldots, Y_{i,\alpha-1}) + \sigma_{i,\alpha}(Y_{i,1}, \ldots, Y_{i,\alpha-1})X_{i,\alpha}, \tag{3.5}$$

where $\sigma_{i,\alpha} > 0$ and $\mu_{i,\alpha}$ are arbitrary functions and $Y_{i,1} = \mu_{i,1} + \sigma_{i,1}X_{i,1}$ for arbitrary numbers $\sigma_{i,1} > 0$, $\mu_{i,1}$. As in Eq. (3.3), this procedure is repeated multiple times to build a deep autoregressive flow. The masking in MAF comes from its use of masked autoencoder for distribution estimation (MADE) [107] to evaluate $\mu_{i,\alpha}$ and $\sigma_{i,\alpha}$ for all α in one forward pass. This approach eliminates the need for the recursion in Eq. (3.5). MAF is nearly the same as inverse autoregressive flows (IAFs) [105], which also use Gaussian autoregressions and are built on MADE. The main difference is that MAF is very efficient for density estimation and slow for sampling, while IAF is slow for density estimation and fast for sampling. As ANODE only needs to estimate the density without producing new samples, MAF is selected as the method of choice.

The estimation of $p_{\text{background}}(x|m)$ for ANODE requires that the MAF provides a conditional density. This can be accomplished by adding m as an input to all functions μ_i and σ_i.

B. Estimating the background

An anomaly detection technique is only useful for finding new particles if the Standard Model background can be estimated. As mentioned earlier, one benefit of the direct density estimation in ANODE is that the background can be directly estimated with $p_{\text{background}}(x|m)$. This results in the following multiple possibilities for background estimation that are considered in this work:

(i) *Direct density estimation*: These methods use the interpolated $p_{\text{background}}(x|m)$ to directly compute the efficiency $\epsilon_{bg}(R_c|m)$ of the background after a threshold requirement on $R(x|m)$.

Density sampling. One could directly sample events from $p_{\text{background}}(x|m)$ using the stacked change of variables specified by Eq. (3.5). As mentioned in Sec. III A, this is less efficient for MAF compared with IAF. This sampling is not pursued in this paper.

Density integration. Another approach is to directly integrate $p_{\text{background}}(x|m)$ for events with $R(x|m) > R_c$,

$$\epsilon_{bg}(R_c|m) = \int dx\, p_{\text{background}}(x|m)\Theta(R(x|m) - R_c). \tag{3.6}$$

Importance sampling. Analytically integrating a function in high dimensions is impractical, so one can estimate the integral with importance sampling. An effective method to implement this sampling is to make the following observation:

$$\begin{aligned}\epsilon_{bg}(R_c|m) &= \int dx p_{\text{background}}(x|m)\Theta(R(x|m) - R_c) \\ &= \int dx p_{\text{full}}(x|m)\frac{1}{R(x|m)}\Theta(R(x|m) - R_c) \\ &= \int_{R_c}^{\infty} dR p_{\text{full}}(R|m)\frac{1}{R}. \end{aligned} \tag{3.7}$$

The last line in Eq. (3.7) can be estimated by computing the fraction of events in the SR (representing the full distribution) with $R > R_c$ and then weighting each event in the counting by $1/R$.

(ii) *Sideband in m*: As long as the requirement $R(x|m) > R_c$ does not sculpt a localized feature in m, one can estimate the background prediction by performing a fit in the m spectrum from the SB and interpolating to the SR. This is a standard approach, as discussed in Sec. I.

Further details about background estimation are presented in Sec. V B for the numerical example described in the next section.

C. Comparison with the CWoLa hunting method

The CWoLa hunting method [33,34] is a recently proposed model-agnostic sideband method that also uses machine learning and will serve as a benchmark for ANODE. In the CWoLa hunting approach, the signal sensitivity is achieved by training a classifier to distinguish the SR from the SB. This classifier will approach the likelihood ratio R_{CWoLa}, which is optimal under certain conditions,

$$\begin{aligned}R_{\text{CWoLa}}(x) &= \frac{p_{\text{data}}(x|\text{SR})}{p_{\text{data}}(x|\text{SB})} = \frac{p_{\text{data}}(x|\text{SR})}{p_{\text{background}}(x|\text{SB})} \\ &= \frac{p_{\text{data}}(x|\text{SR})}{p_{\text{background}}(x|\text{SR})}, \end{aligned} \tag{3.8}$$

where the second equality is true in the absence of signal in the sideband[2] and the third equality is true when x and m are independent. The background is estimated using a sideband fit after placing a selection based on the above classifier.

A key assumption of the CWoLa method is that x and m are independent. This condition is stronger than the requirement for the background fit, but is necessary for achieving signal sensitivity. In particular, in the presence of a dependence between x and m, the CWoLa classifier will learn the true differences between SB and SR. If these differences are larger than the differences between signal and background in the SR, the CWoLa classifier may not succeed in finding the signal.

In contrast, the ANODE method does not require any particular relationship between x and m to achieve signal sensitivity. In fact, the information about m could be fully contained within x, and ANODE could still succeed in principle. Therefore, ANODE can make use of features which are strongly correlated with m, thus extending the potential sensitivity to new signals. This is possible because of the two step density estimation, interpolating $p_{\text{background}}(x|m)$ from the sideband and then estimating $p_{\text{data}}(x|m)$ from the SR. Such an approach is not possible with CWoLa hunting, which directly learns the likelihood ratio. The only requirement for ANODE is that there are no nontrivial features in the SR that cannot be smoothly predicted from the SB. Section V C illustrates the ability of ANODE to cope with correlated features.

IV. DETAILS OF THE SAMPLE

A simulated resonance search using large-radius dijets is used to illustrate ANODE. The simulated datasets are from the LHC Olympics 2020 challenge research and development dataset [74]. For a background process, one million quantum chromodynamic (QCD) dijet events are simulated with PYTHIA 8 [108,109] without pileup or multiple parton interactions. The signal is a hypothetical W' boson ($m_{W'} = 3.5$ TeV) that decays into an X boson ($m_X = 500$ GeV) and a Y boson ($m_Y = 100$ GeV), with the same simulation setup as the QCD dijets. The X and Y bosons decay promptly into quarks and due to their large Lorentz boost in the laboratory frame, the resulting hadronic decay products are captured by a single large-radius jet. The detector simulation is performed with Delphes 3.4.1 [110–112] and particle flow objects are clustered into jets using the Fastjet [113,114] implementation of the anti-k_t algorithm [115] using $R = 1.0$ as the jet radius. Events are selected by requiring at least one such jet with $p_T > 1.3$ TeV. While there exist LHC searches for the case that X and Y are electroweak bosons [116,117], the generic case is currently uncovered by a dedicated search.

The resonant feature m will be the invariant mass of the leading two jets, m_{JJ}. These two jets are ordered by their mass m_J so that by construction, $m_{J_1} < m_{J_2}$. The discriminating features x are four-dimensional, consisting of the observables,

$$m_{J_1}, \qquad m_{J_2} - m_{J_1}, \qquad \tau_{21}^{J_1}, \tau_{21}^{J_2}, \tag{4.1}$$

where τ_{21} is the n-subjettiness ratio [118,119]. This observable is the most widely used single feature for identifying jets with a two-prong substructure. While the

[2]This is not strictly necessary—the classifier can still be optimal even if there is some signal in the sideband [79].

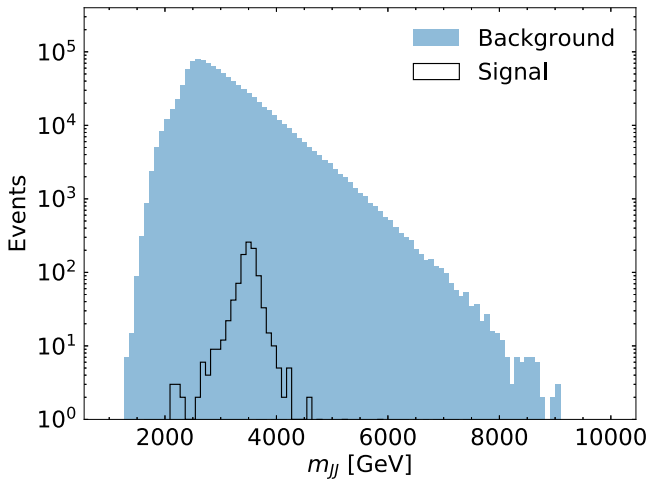

FIG. 2. Histograms for the invariant mass of the leading two jets for the Standard Model background as well as the injected signal. There are 1 million background events and 1000 signal events.

ultimate goal of ANODE is to perform density estimation on high-dimensional, low-level features, there is already utility in a search with high-level features from Eq. (4.1). Thus, to demonstrate how ANODE works, this will be the focus for the rest of this paper.

Simulated data are constructed by injecting 1000 signal events to the full background sample. A histogram of m_{JJ} is presented in Fig. 2. As expected, the signal peaks near $m_{W'}$. The signal region is defined by $m_{JJ} \in [3.3, 3.7]$ TeV and then the sideband is the rest of the spectrum. The simulated data are divided into two equal samples for training and testing; thus, we have $\approx 500,000$ background and ≈ 500 signal events in each sample. In the SR, we are left with $\approx 60,000$ background and ≈ 400 signal events in each sample. This corresponds to $S/\sqrt{B} = 1.6$ and $S/B = 0.6\%$ in the SR. This value of $S/\sqrt{B}$ would be the approximate significance from a sideband fit (ignoring the fit errors). Section V A will show how much this can be enhanced from ANODE.

The additional four features for classification are shown in Fig. 3. The lighter jet mass peaks near m_Y and the difference between masses peaks at about $m_X - m_Y = 400$ GeV. The τ_{21} observables are lower for the two-prong signal jets than for the mostly one-prong background jets. Jet mass and τ_{21} are negatively correlated for QCD jets [95] and so τ_{21} is higher for J_2 than for J_1.

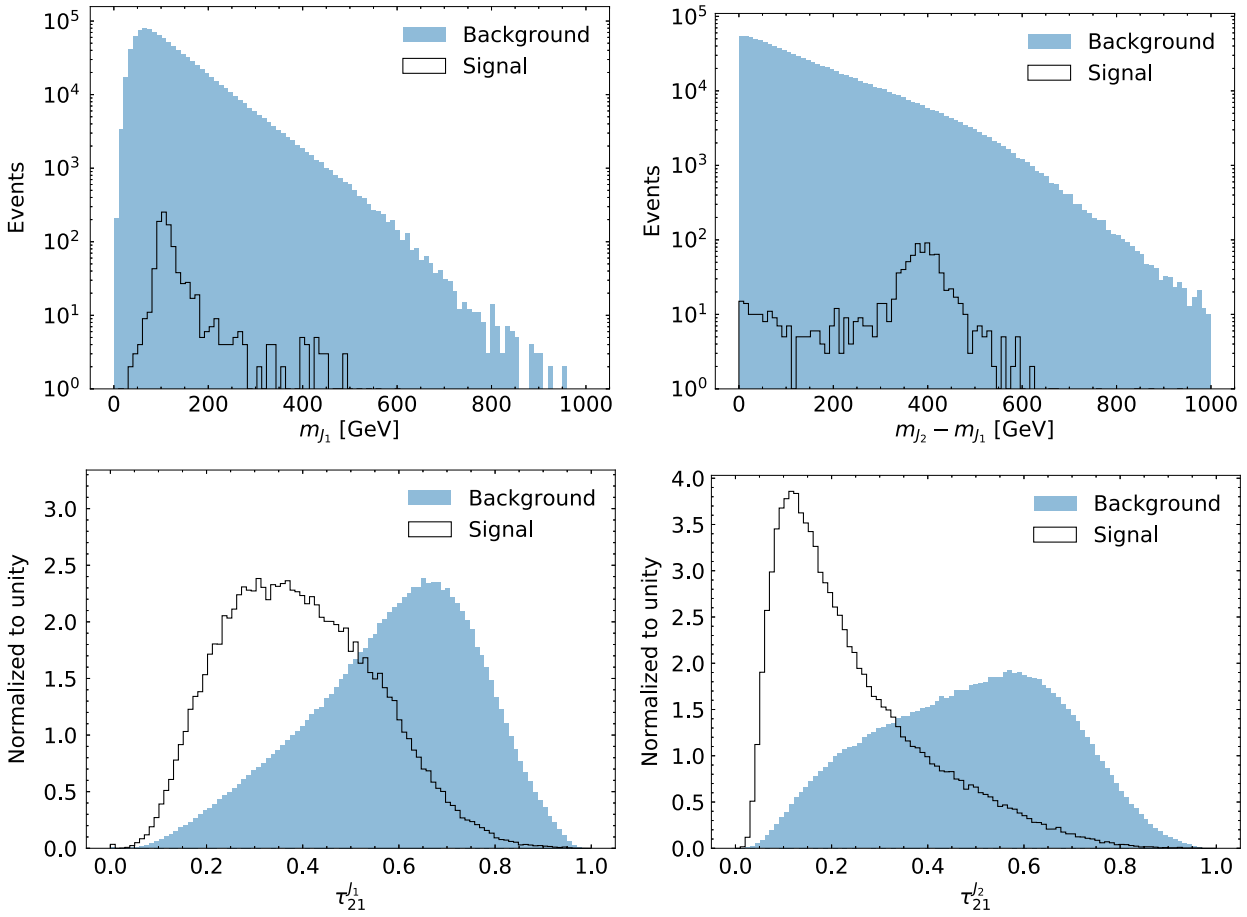

FIG. 3. The four features used for classification: m_{J_1} (top left), $m_{J_1} - m_{J_2}$ (top right), $\tau_{21}^{J_1}$ (bottom left), and $\tau_{21}^{J_2}$ (bottom right). These histograms are inclusive in m_{JJ}. There are 1 million background events and 1000 signal events for the mass histograms.

The conditional MAF (along with most methods of density estimation) has difficulty at sharp, discontinuous edges and boundaries, so we first transform the dataset before performing density estimation. First, all features are linearly scaled to be (feature) $\mapsto x \in [0,1]$. Then, the logit transformation $\log(x/(1-x))$ is applied to map the scaled features to be between $(-\infty,\infty)$. The Jacobian for this map is accounted for when computing probability densities for the original feature space. Even with this transformation, density estimation is difficult near the boundaries. Therefore, the scaled features are required to have $0.05 < x < 0.95$. This keeps 95% (72%) of the signal (background) in the SR. Below we will refer to this as the "fiducial region." All results below are computed with respect to the number of events after this truncation.

V. RESULTS

A. Sensitivity

The conditional MAF is optimized[3] using the log-likelihood loss function, $\log(p(x|m))$. All of the neural networks are written in PyTorch [120]. For the hyperparameters, there are 15 MADE blocks (one layer each) with 128 hidden units per block. Networks are optimized with Adam [121] using a learning rate 10^{-4} and weight decay of 10^{-6}. The SR and SB density estimators are each trained for 50 epochs. No systematic attempt was made to optimize these hyperparameters, and it is likely that better performance could be obtained with further optimization. For the SR density estimator, the last epoch is chosen for simplicity and it was verified that the results are robust against this choice. The SB density estimator significantly varies from epoch to epoch. Averaging the density estimates pointwise over 10 consecutive epochs results in a stable result. Averaging over more epochs does not further improve the stability. All results with ANODE present the SB density estimator with this averaging scheme for the last 10 epochs.

Figure 4 shows a scatter plot of $R(x|m)$ versus $\log p_{\text{background}}(x|m)$ for the test set in the SR. As desired, the background is mostly concentrated around $R(x|m) = 1$, while there is a long tail for signal events at higher values of $R(x|m)$ and between $-2 < \log p_{\text{background}}(x|m) < 2$. This is exactly what is expected for this signal: it is an overdensity ($R > 1$) in a region of phase space that is relatively rare for the background ($p_{\text{background}}(x|m) \ll 1$).

The background density in Fig. 4 also shows that the $R(x|m)$ is narrower around 1 when $p_{\text{background}}(x|m)$ is large and more spread out when $p_{\text{background}}(x|m) \ll 1$. This is evidence that the density estimation is more accurate when the densities are high and worse when the densities are low. This is also to be expected: if there are many data points close to one another, it should be easier to estimate their density than if the data points are very sparse.

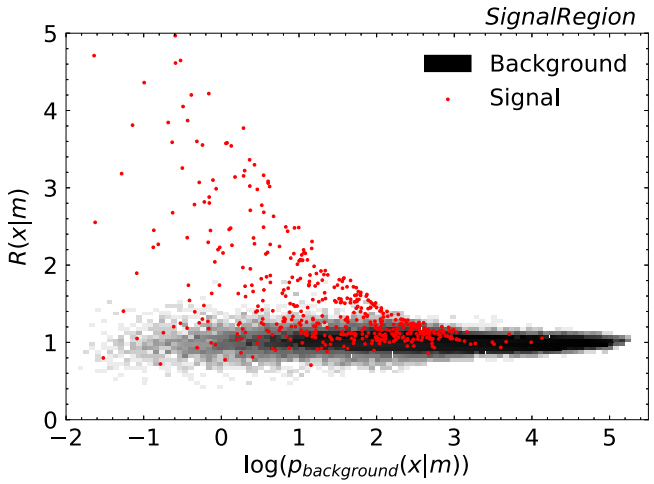

FIG. 4. Scatter plot of $R(x|m)$ versus $\log p_{\text{background}}(x|m)$ across the test set in the SR. Background events are shown (as a two-dimensional histogram) in gray scale and individual signal events are shown in red.

Another view of the results is presented in Fig. 5, with one-dimensional information about $R(x|m)$ in the SR. The left plot of Fig. 5 shows that the background is centered and approximately symmetric around $R = 1$ with a standard deviation of approximately 17%. This width is due to various sources, including the accuracy of the SR density, the accuracy of the SB density, and the quality of the interpolation from SB to SR. Each of these sources has contributions from the finite size of the datasets used for training, the neural network flexibility, and the training procedure. The right plot of Fig. 5 presents the number of background and signal events as a function of a threshold $R > R_c$. The starting point are the original numbers background (40,000) and signal (400) numbers in the SR window and the fiducial window. Starting from low S/B and $S/\sqrt{B}$ one can achieve $S/B > 1$ and a high $S/\sqrt{B}$ with a threshold requirement on R. Figure 6 shows that the signal is clearly visible in the x distribution after applying such a threshold requirement.

The performance of R as an anomaly detector is further quantified by the receiver operating characteristic (ROC) and significance improvement characteristic (SIC) curves in Fig. 7. These metrics are obtained by scanning R and computing the signal efficiency (true positive rate) and background efficiency (false positive rate) after a threshold requirement on R. The area under the curve for ANODE is 0.82. For comparison, the CWoLa hunting approach is also shown in the same plots. The CWoLa classifier is trained using sideband regions that are 200 GeV wide on either side of the SR. The sidebands are weighted to have the same number of events as each other and in total, the same

[3]Based on code from https://github.com/ikostrikov/pytorch-flows.

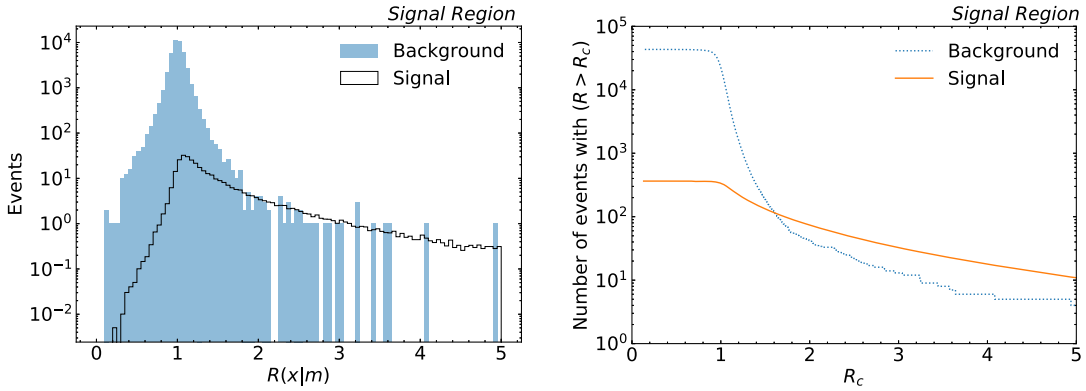

FIG. 5. Left: histogram of $R(x|m)$ evaluated on the test set; right: the integrated number of events that survive a threshold on $R(x|m)$. The two distributions are scaled to represent the rates for 500,000 total background events and 500 total signal events, as introduced in Sec. IV.

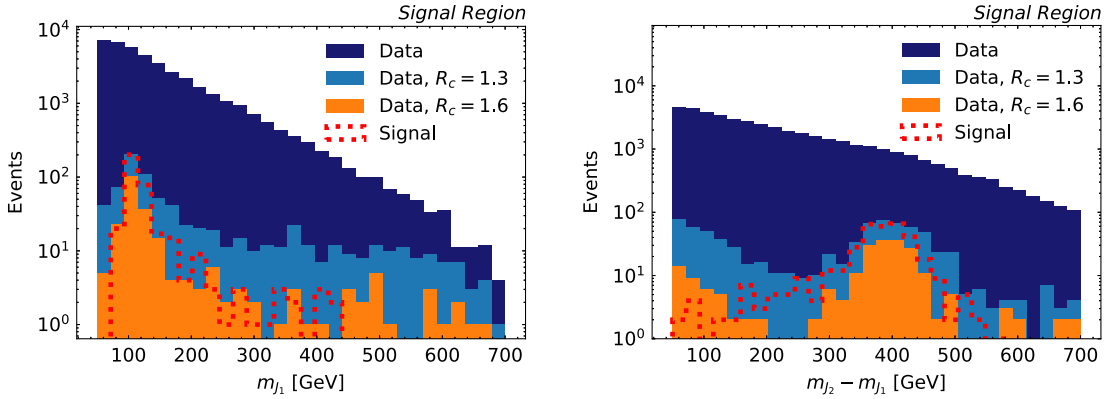

FIG. 6. Distributions of m_{J_1} (left) and $m_{J_2} - m_{J_1}$ (right) in the signal region after applying a threshold requirement on R.

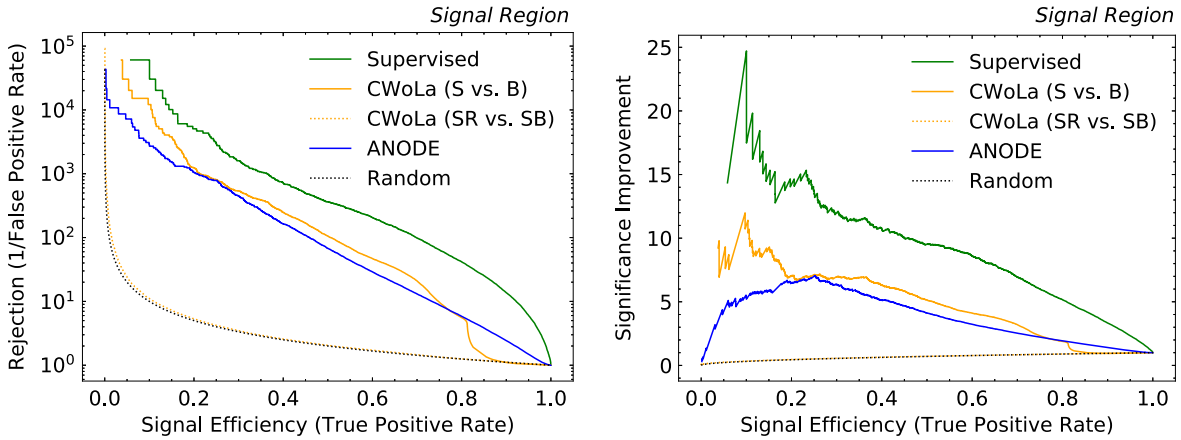

FIG. 7. Receiver operating characteristic (ROC) curve (left) and significance improvement characteristic (SIC) curve (right).

as the SR. A single NN with four hidden layers with 64 nodes each is trained using Keras [122] and TensorFlow [123]. Dropout [124] of 10% is used for each intermediate layer. Intermediate layers use rectified linear unit activation functions and the last layer uses a sigmoid. The classifier is optimized using binary cross entropy and is trained for

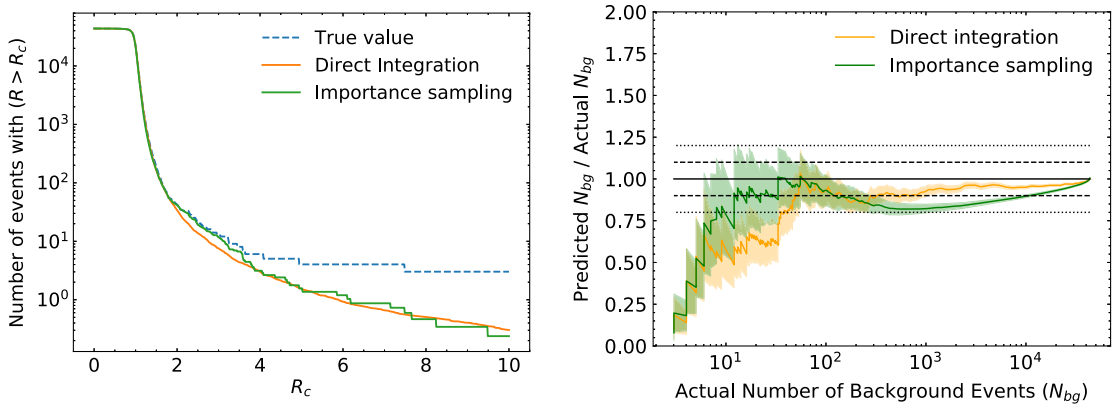

FIG. 8. Left: the number of events after a threshold requirement $R > R_c$ using the two integration methods described in Sec. III B, as well as the true background yield. Right: the ratio of the predicted and true background yields from the left plot, as a function of the actual number of events that survive the threshold requirement. The shaded bands around the central predictions are the 1σ statistical (Poisson) uncertainty derived from the observed background counts. The black dashed and dotted lines are 10% and 20% around a ratio of 1.

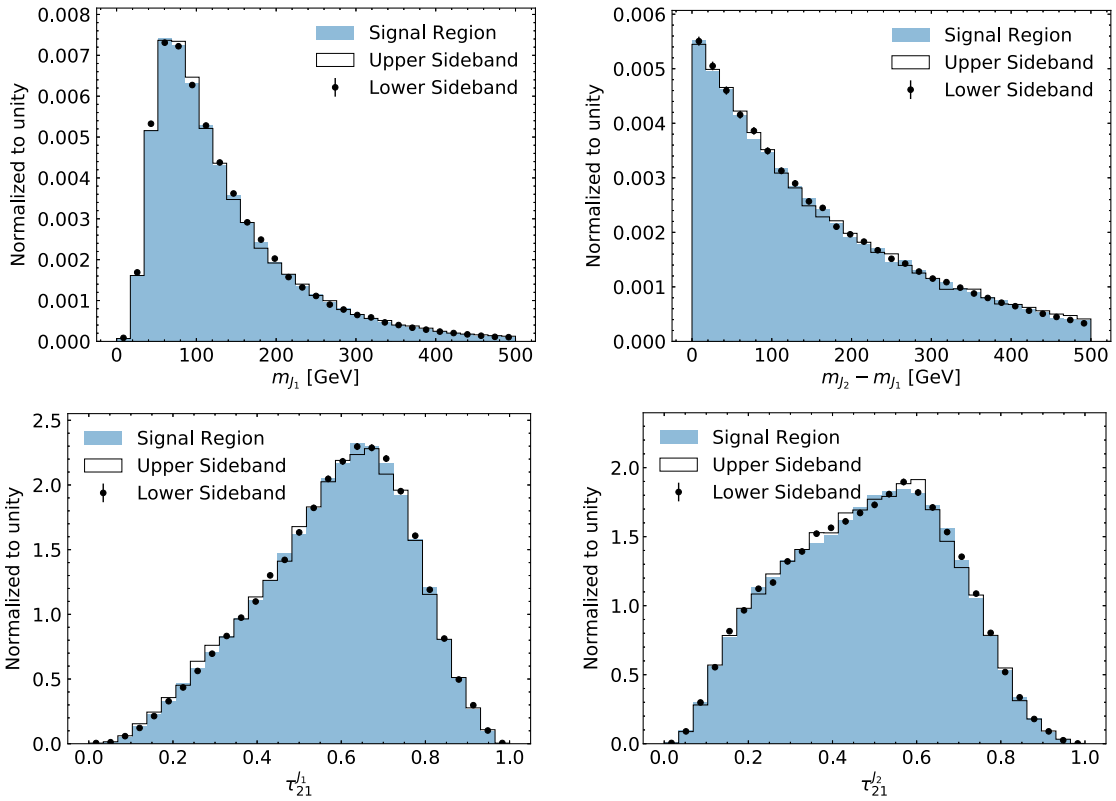

FIG. 9. A comparison of the four features x between the SR and two nearby sidebands defined by $m_{jj} \in [3.1, 3.3]$ TeV (lower sideband) and $m_{jj} \in [3.7, 3.9]$ TeV (upper sideband).

300 epochs. As with ANODE, ten epochs are averaged for the reported results.[4]

[4]A different regularization procedure was used in Refs. [33,34] based on the validation loss and k-folding. The averaging here is expected to serve a similar purpose.

The performance of ANODE is comparable to CWoLa hunting in Fig. 7, which does slightly better at higher signal efficiencies and much better at lower signal efficiencies. This may be a reflection of the fact that CWoLa makes use of supervised learning and directly approaches the likelihood ratio, while ANODE is unsupervised and attempts to

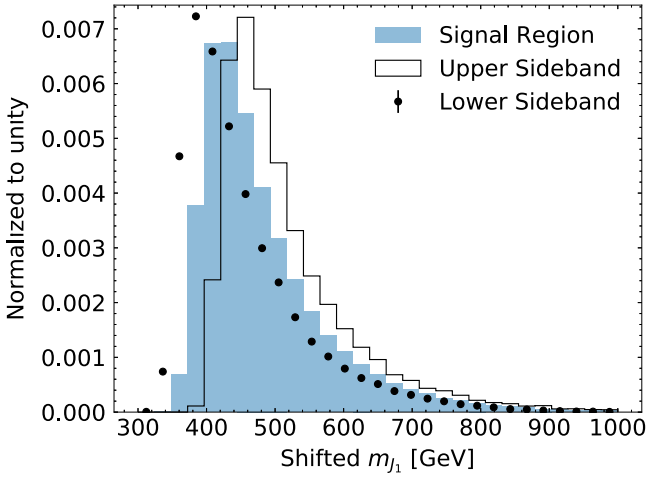

FIG. 10. The lighter jet mass for the SR and the lower and upper sideband regions after the shift defined by Eq. (5.1).

learn both the numerator and denominator of the likelihood ratio. With this dataset, ANODE is able to enhance the signal significance by about a factor of 7 and would therefore be able to achieve a local significance above 5σ given that the starting value of $S/\sqrt{B}$ is 1.6.

B. Background estimation

This section explores the possibility of using the estimate of $p_{\text{background}}(x|m)$ to directly determine the background efficiency in the SR after a requirement on $R > R_c$. Figure 8 presents a comparison between integration methods (direct integration and importance sampling) described in Sec. III B and the true background yields. Qualitatively, both methods are able to characterize the yield across several orders of magnitude in background efficiency. However, both methods diverge from the truth in the extreme tails of the R distribution. The right plot of Fig. 8 offers a quantitative comparison between methods. For efficiencies down to about 10^{-3}, both methods are accurate within about 25%. The direct integration method has a smaller bias of about 10%. This is consistent with Fig. 5, for which the standard deviation is between 10% and 20%.

C. Performance on a dataset with correlated features

The results presented in the previous sections have established that ANODE is able to identify the signal and estimate the corresponding SM backgrounds introduced in Sec. IV. One fortuitous aspect of the chosen features x introduced in Sec. IV is that they are all relatively independent of m_{jj}. This is illustrated in Fig. 9, using the SR and neighboring sideband regions. As a result of this independence, the CWoLa method is able to find the signal and presumably the ANODE interpolation from SB to SR is easier than if there was a strong dependence.

The purpose of this section is to study the sensitivity of the ANODE and CWoLa hunting methods to correlations in the features x with m_{jj}. Based on the assumptions of the two methods, it is expected that with strong correlations, CWoLa hunting will fail to find the signal while ANODE should still be able to identify the presence of signal in the SR as well as estimate the background. To study this sensitivity in a controlled fashion, correlations are introduced artificially. In practice, adding more features to x will inevitably result in some dependence with m_{jj}; the artificial example here illustrates the challenges already in low dimensions. New jet mass observables are created, which are linearly shifted,

$$m_{J_{1,2}} \to m_{J_{1,2}} + cm_{JJ}, \tag{5.1}$$

where $c = 0.1$ for this study. The resulting shifted lighter jet mass is presented in Fig. 10.

New ANODE and CWoLa models are trained using the shifted dataset and their performance is quantified in Fig. 11. As expected, the fully supervised classifier is

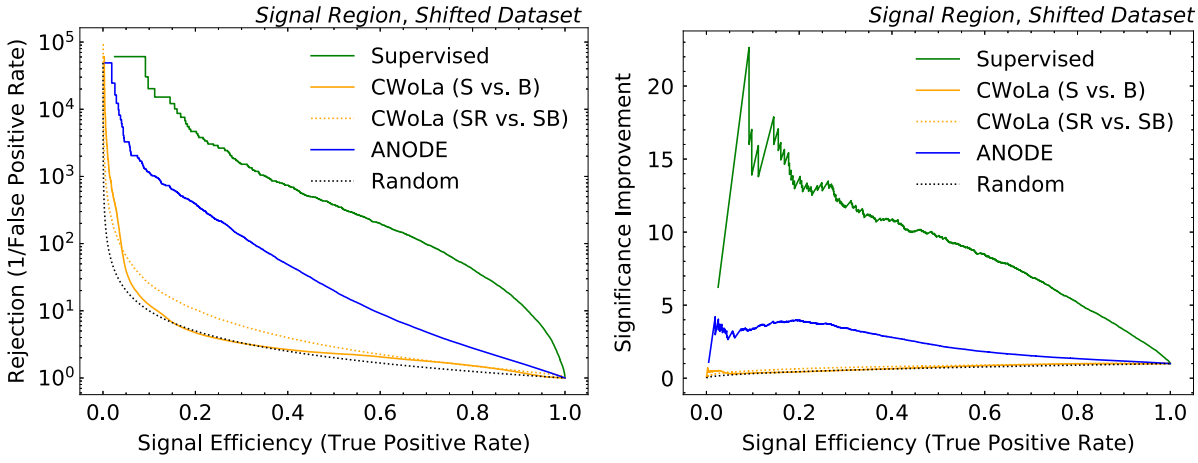

FIG. 11. ROC (left) and SIC (right) curves in the signal region using the shifted dataset specified by Eq. (5.1).

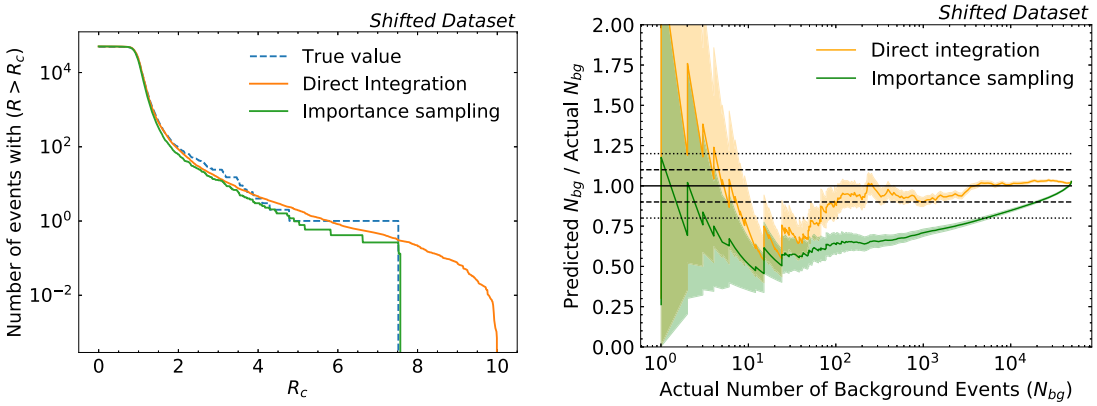

FIG. 12. The same as Fig. 8, but for the shifted dataset. In particular, these plots compare the background prediction from two direct density estimation techniques with the true background yield after a threshold requirement $R(x|m) > R_c$.

nearly the same as Fig. 7. ANODE is still able to significantly enhance the signal, with a maximum significance improvement near 4. While in principle ANODE could achieve the same classification accuracy on the shifted and nominal datasets, the performance on the shifted examples is not as strong as in Fig. 7. In practice, the interpolation of $p_{\text{background}}$ into the SR is more challenging now due to the linear correlations. This could possibly be overcome with improved training, better choices of hyperparameters, or more sophisticated density estimation techniques.

By construction, there are now bigger differences between the SR and SB than between the SR background and the SR signal. Therefore, the CWoLa hunting classifier is not able to find the signal. This is evident from the ROC curve in the left plot of Fig. 11, which shows that the signal-versus-background classifier is essentially random while the SR-versus-SB classifier has learned something non-trivial.

Last, Fig. 12 shows the performance of direct density estimation for the background prediction using the shifted dataset. The performance is comparable to the unshifted dataset (Fig. 8), meaning that ANODE could potentially be used as a complete anomaly detection method even in the presence of correlated feature spaces.

VI. CONCLUSIONS

This paper has presented a powerful new model-independent search method called ANODE, which is built on neural density estimation. Unlike other approaches, ANODE directly learns the background probability density and data probability density in a signal region. The ratio of these densities is a powerful classifier and the background density can be directly used to estimate the background efficiency from a threshold requirement on the classifier. Finally, ANODE is robust against correlations in the data, which tend to break other model-agnostic sideband methods such as CWoLa.

The results presented in this paper are meant to be a proof of concept of the general method, and there are many exciting future directions. For example, while this paper focused on collider searches for BSM, the ANODE method is completely general and could be applied to many areas beyond high energy physics, including astronomy and astrophysics. Similarly, while the demonstrations here were based on the innovative MAF density estimation technique, the ANODE method can be used in conjunction with any density estimation algorithm. Indeed, there are numerous other neural density estimation methods from the past few years that claim state-of-the-art performance, including neural autoregressive flows [125] and neural spline flows [126]; exploring these would be an obvious way to attempt to improve the results in this paper. In addition, it would be interesting to attempt the ANODE method on even higher-dimensional feature spaces, all the way up to the full low-level feature set of the four vectors of all the hadrons in the event. This might already be feasible with existing neural density estimators, at is it common to evaluate their performance on high-dimensional datasets ranging from UCI datasets [127] with up to ~50 features, to image datasets such as MNIST [128] and CIFAR-10 [129] which have hundreds and thousands of features, respectively. The prospects for the ANODE method are exciting: as the field of neural density estimation continues to grow within the machine learning community, ANODE will become more sensitive to resonant new physics in collider high energy physics and beyond.

ACKNOWLEDGMENTS

D. S. is grateful to Matt Buckley and John Tamanas for many fruitful discussions on neural density estimation. We are especially grateful to John Tamanas for help with the

conditional MAF code. Additionally, we would like to thank Kyle Cranmer and Uroš Seljak for helpful discussions and Nick Rodd and John Tamanas for helpful comments on the draft. This work was supported by the U.S. Department of Energy, Office of Science under Contract No. DE-AC02-05CH11231. D. S. is supported by DOE Grant No. DOE-SC0010008. D. S. thanks LBNL, BCTP, and BCCP for their generous support and hospitality during his sabbatical year.

APPENDIX: COMMENTS ON OPTIMALITY

The Neyman-Pearson lemma only applies to simple hypothesis tests. The lemma states that for a fixed probability of rejecting the null hypothesis when it is true (level), the probability for rejecting the null hypothesis when the alternative is true (power) is maximized with the likelihood ratio test statistic. For supervised searches with profiled nuisance parameters or for anomaly detection with a composite alternative hypothesis, there is no uniformly most powerful classifier. The goal of this brief section is to clarify what is meant by asymptotically optimal anomaly detection.

For any given BSM model, the procedures labeled asymptotically optimal are likely not optimal. The sense in which they are optimal is as follows. Let the null hypothesis H_0 be that the data are distributed according to $p_{\text{background}}$, a density describing the phase space of the background-only. Furthermore, let the alternative hypothesis H_A be that the data are distributed according to p_{data}, the learned density of the data. Distinguishing H_0 from H_A is a simple hypothesis test. Therefore, the test statistic $p_{\text{background}}/p_{\text{data}}$ has the property that for a fixed probability for rejecting H_0 given data $\sim p_{\text{background}}$, the probability for rejecting H_0 is as high as possible when H_A is true (which it is). If $p_{\text{background}} = p_{\text{data}}$, then power = level. So ANODE is asymptotically optimal for rejecting the data as background-only, but is not "optimal" for rejecting any particular BSM model.

[1] ATLAS Collaboration, Exotic physics searches, 2019, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults.
[2] ATLAS Collaboration, Supersymmetry searches, 2019, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults.
[3] ATLAS Collaboration, Higgs and Diboson searches, 2019, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/HDBSPublicResults.
[4] CMS Collaboration, CMS exotica public physics results, 2019, https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO.
[5] CMS Collaboration, CMS supersymmetry physics results, 2019, https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS.
[6] CMS Collaboration, CMS beyond-two-generations (B2G) public physics results, 2019, https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G.
[7] LHCb Collaboration, Publications of the QCD, Electroweak and Exotica Working Group, 2019, http://lhcbproject.web.cern.ch/lhcbproject/Publications/LHCbProjectPublic/Summary_QEE.html.
[8] J. H. Kim, K. Kong, B. Nachman, and D. Whiteson, The motivation and status of two-body resonance decays after the LHC Run 2 and beyond, J. High Energy Phys. 04 (2020) 030.
[9] N. Craig, P. Draper, K. Kong, Y. Ng, and D. Whiteson, The unexplored landscape of two-body resonances, Acta Phys. Pol. B **50**, 837 (2019).
[10] B. Knuteson. Ph. D. thesis, University of California at Berkeley, 2000.
[11] B. Abbott *et al.* (D0 Collaboration), Search for new physics in $e\mu X$ data at DØ using Sherlock: A quasi model independent search strategy for new physics, Phys. Rev. D **62**, 092004 (2000).
[12] V. M. Abazov *et al.* (D0 Collaboration), A quasi model independent search for new physics at large transverse momentum, Phys. Rev. D **64**, 012004 (2001).
[13] B. Abbott *et al.* (D0 Collaboration), A Quasi-Model-Independent Search for New High p_T Physics at DØ, Phys. Rev. Lett. **86**, 3712 (2001).
[14] F. D. Aaron *et al.* (H1 Collaboration), A general search for new phenomena at HERA, Phys. Lett. B **674**, 257 (2009).
[15] A. Aktas *et al.* (H1 Collaboration), A general search for new phenomena in ep scattering at HERA, Phys. Lett. B **602**, 14 (2004).
[16] K. S. Cranmer, Searching for new physics: Contributions to LEP and the LHC. PhD thesis, Wisconsin U., Madison, 2005.
[17] T. Aaltonen *et al.* (CDF Collaboration), Model-independent and quasi-model-independent search for new physics at CDF, Phys. Rev. D **78**, 012002 (2008).
[18] T. Aaltonen *et al.* (CDF Collaboration), Model-independent global search for new high-p(T) physics at CDF, arXiv:0712.2534.
[19] T. Aaltonen *et al.* (CDF Collaboration), Global search for new physics with 2.0 fb^{-1} at CDF, Phys. Rev. D **79** (2009) 011101.
[20] CMS Collaboration, MUSiC, a model unspecific search for new physics, in pp collisions at $\sqrt{s} = 8$ TeV, CERN Report No. CMS-PAS-EXO-14-016, 2017.
[21] CMS Collaboration, Model unspecific search for new physics in pp collisions at $\sqrt{s} = 7$ TeV, CERN Report No. CMS-PAS-EXO-10-021, 2011.

[22] M. Aaboud *et al.* (ATLAS Collaboration), A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment, Eur. Phys. J. C **79**, 120 (2019).
[23] ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV, CERN Report No. ATLAS-CONF-2014-006, 2014.
[24] ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV, CERN Report No. ATLAS-CONF-2012-107, 2012.
[25] R. T. D'Agnolo and A. Wulzer, Learning new physics from a machine, Phys. Rev. D **99**, 015014 (2019).
[26] R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, Learning multivariate new physics, arXiv:1912.12155.
[27] M. Farina, Y. Nakai, and D. Shih, Searching for new physics with deep autoencoders, arXiv:1808.08992.
[28] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or what?, SciPost Phys. **6**, 030 (2019).
[29] T. S. Roy and A. H. Vijay, A robust anomaly finder based on autoencoder, arXiv:1903.02032.
[30] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Variational autoencoders for new physics mining at the large Hadron collider, J. High Energy Phys. 05 (2019) 036.
[31] A. Blance, M. Spannowsky, and P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches, J. High Energy Phys. 10 (2019) 047.
[32] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, Novelty detection meets collider physics, arXiv:1807.10261.
[33] J. H. Collins, K. Howe, and B. Nachman, Anomaly Detection for Resonant New Physics with Machine Learning, Phys. Rev. Lett. **121**, 241803 (2018).
[34] J. H. Collins, K. Howe, and B. Nachman, Extending the search for new resonances with machine learning, Phys. Rev. D **99** (2019) 014038.
[35] A. De Simone and T. Jacques, Guiding new physics searches with unsupervised learning, Eur. Phys. J. C **79**, 289 (2019).
[36] A. Mullin, H. Pacey, M. Parker, M. White, and S. Williams, Does SUSY have friends? A new approach for LHC event analysis, arXiv:1912.10625.
[37] G. M. Alessandro Casa, Nonparametric semisupervised classification for signal detection in high energy physics, arXiv:1809.02977.
[38] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, Uncovering latent jet substructure, Phys. Rev. D **100**, 056002 (2019).
[39] A. Andreassen, B. Nachman, and D. Shih, Simulation assisted likelihood-free anomaly detection, arXiv:2001.05001.
[40] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, A generic anti-QCD jet tagger, J. High Energy Phys. 11 (2017) 163.
[41] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, arXiv:1406.2661.
[42] L. de Oliveira, M. Paganini, and B. Nachman, Learning particle physics by example: Location-aware generative adversarial networks for physics synthesis, Comput. Softw. Big Sci. **1**, 4 (2017).
[43] M. Paganini, L. de Oliveira, and B. Nachman, Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters, Phys. Rev. Lett. **120**, 042003 (2018).
[44] M. Paganini, L. de Oliveira, and B. Nachman, CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks, Phys. Rev. D **97**, 014021 (2018).
[45] A. Butter, T. Plehn, and R. Winterhalder, How to GAN event subtraction, arXiv:1912.08824.
[46] J. Arjona Martinez, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Particle generative adversarial networks for full-event simulation at the LHC and their application to pileup description, in *19th International Workshop on Advanced Computing and Analysis Techniques in Physics Research: Empowering the revolution: Bringing Machine Learning to High Performance Computing (ACAT 2019) Saas-Fee, Switzerland, 2019* (CERN, Geneva, Switzerland, 2019).
[47] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, and R. Winterhalder, How to GAN away detector effects, arXiv:1912.00477.
[48] S. Vallecorsa, F. Carminati, and G. Khattak, 3D convolutional GAN for fast simulation, EPJ Web Conf. **214**, 02010 (2019).
[49] C. Ahdida *et al.* (SHiP Collaboration), Fast simulation of muons produced at the SHiP experiment using generative adversarial networks, arXiv:1909.04451.
[50] S. Carrazza and F. A. Dreyer, Lund jet images from generative and cycle-consistent adversarial networks, Eur. Phys. J. C **79**, 979 (2019).
[51] A. Butter, T. Plehn, and R. Winterhalder, How to GAN LHC events, SciPost Phys. **7**, 075 (2019).
[52] J. Lin, W. Bhimji, and B. Nachman, Machine learning templates for QCD Factorization in the search for physics beyond the standard model, J. High Energy Phys. 05 (2019) 181.
[53] R. Di Sipio, M. F. Giannelli, S. K. Haghighat, and S. Palazzo, DijetGAN: A generative-adversarial network approach for the simulation of QCD dijet events at the LHC, J. High Energy Phys. 08 (2019) 110.
[54] B. Hashemi, N. Amin, K. Datta, D. Olivito, and M. Pierini, LHC analysis-specific datasets with generative adversarial networks, arXiv:1901.05282.
[55] V. Chekalina, E. Orlova, F. Ratnikov, D. Ulyanov, A. Ustyuzhanin, and E. Zakharov, Generative models for fast calorimeter simulation: The LHCb case, EPJ Web Conf. **214**, 02034 (2019).
[56] ATLAS Collaboration, Deep generative models for fast shower simulation in ATLAS, Report No. ATL-SOFT-PUB-2018-001, 2018.
[57] K. Zhou, G. Endrodi, L.-G. Pang, and H. Stocker, Regressive and generative neural networks for scalar field theory, Phys. Rev. D **100**, 011501 (2019).
[58] F. Carminati, A. Gheata, G. Khattak, P. M. Lorenzo, S. Sharan, and S. Vallecorsa, Three dimensional generative adversarial networks for fast simulation, J. Phys. Conf. Ser. **1085**, 032016 (2018).

[59] S. Vallecorsa, Generative models for fast simulation, J. Phys. Conf. Ser. **1085**, 022005 (2018).
[60] K. Datta, D. Kar, and D. Roy, Unfolding with generative adversarial networks, arXiv:1806.00433.
[61] P. Musella and F. Pandolfi, Fast and accurate simulation of particle detectors using generative adversarial networks, Comput. Softw. Big Sci. **2**, 8 (2018).
[62] M. Erdmann, L. Geiger, J. Glombitza, and D. Schmidt, Generating and refining particle detector simulations using the Wasserstein distance in adversarial networks, Comput. Softw. Big Sci. **2** (2018) 4.
[63] K. Deja, T. Trzcinski, and Ł. Graczykowski, Generative models for fast cluster simulations in the TPC for the ALICE experiment, EPJ Web Conf. **214**, 06003 (2019).
[64] D. Derkach, N. Kazeev, F. Ratnikov, A. Ustyuzhanin, and A. Volokhova, Cherenkov detectors fast simulation using neural networks, in *10th International Workshop on Ring Imaging Cherenkov Detectors (RICH 2018) Moscow, Russia, 2018* (CERN, Geneva, Switzerland, 2019).
[65] H. Erbin and S. Krippendorf, GANs for generating EFT models, arXiv:1809.02612.
[66] M. Erdmann, J. Glombitza, and T. Quast, Precise simulation of electromagnetic calorimeter showers using a Wasserstein generative adversarial network, Comput. Softw. Big Sci. **3**, 4 (2019).
[67] J. M. Urban and J. M. Pawlowski, Reducing autocorrelation times in lattice simulations with generative adversarial networks, arXiv:1811.03533.
[68] J. W. Monk, Deep learning as a parton shower, J. High Energy Phys. 12 (2018) 021.
[69] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, JUNIPR: A framework for unsupervised machine learning in particle physics, Eur. Phys. J. C **79**, 102 (2019).
[70] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, Binary JUNIPR: An Interpretable Probabilistic Model for Discrimination, Phys. Rev. Lett. **123**, 182001 (2019).
[71] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei (Proceedings of Machine Learning Research (PMLR), Lille, France, 2015), Vol. 3, pp. 1530–1538.
[72] M. S. Albergo, G. Kanwar, and P. E. Shanahan, Flow-based generative models for Markov chain Monte Carlo in lattice field theory, Phys. Rev. D **100**, 034515 (2019).
[73] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Long Beach, CA, 2017), Vol. 30, pp. 2338–2347.
[74] G. Kasieczka, B. Nachman, and D. Shih, R&D dataset for LHC Olympics 2020 anomaly detection challenge, 2019, https://doi.org/10.5281/zenodo.2629073.
[75] J. A. Aguilar-Saavedra, Stealth multiboson signals, Eur. Phys. J. C **77**, 703 (2017).
[76] J. A. Aguilar-Saavedra and F. R. Joaquim, The minimal stealth boson: Models and benchmarks, J. High Energy Phys. 10 (2019) 237.
[77] K. Agashe, J. H. Collins, P. Du, S. Hong, D. Kim, and R. K. Mishra, Detecting a boosted diboson resonance, J. High Energy Phys. 11 (2018) 027.
[78] K. Agashe, J. H. Collins, P. Du, S. Hong, D. Kim, and R. K. Mishra, Dedicated strategies for triboson signals from cascade decays of vector resonances, Phys. Rev. D **99**, 075016 (2019).
[79] E. M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.
[80] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. **3**, 993 (2003).
[81] M. Aaboud *et al.* (ATLAS Collaboration), Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector, Phys. Lett. B **784**, 173 (2018).
[82] A. M. Sirunyan *et al.* (CMS Collaboration), Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 11 (2018) 185.
[83] J. Neyman and E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, Phil. Trans. R. Soc. A **231**, 289 (1933), https://www.jstor.org/stable/91247.
[84] M. Frate, K. Cranmer, S. Kalia, A. Vandenberg-Rodes, and D. Whiteson, Modeling smooth backgrounds and generic localized signals with Gaussian processes, arXiv:1709.05681.
[85] M. Aaboud *et al.* (ATLAS Collaboration), Search for heavy ZZ resonances in the $\ell^+\ell^-\ell^+\ell^-$ and $\ell^+\ell^-\nu\bar{\nu}$ final states using proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Eur. Phys. J. C **78**, 293 (2018).
[86] M. Aaboud *et al.* (ATLAS Collaboration), Search for top-squark pair production in final states with one lepton, jets, and missing transverse momentum using 36 fb^{-1} of $\sqrt{s} = 13$ TeV pp collision data with the ATLAS detector, J. High Energy Phys. 06 (2018) 108.
[87] CMS Collaboration, Search for direct top squark pair production in events with one lepton, jets and missing transverse energy at 13 TeV, Technical Report No. CMS-PAS-SUS-19-009, CERN, Geneva, 2019.
[88] M. Aaboud *et al.* (ATLAS Collaboration), Search for new phenomena with large jet multiplicities and missing transverse momentum using large-radius jets and flavour-tagging at ATLAS in 13 TeV pp collisions, J. High Energy Phys. 12 (2017) 034.
[89] A. M. Sirunyan *et al.* (CMS Collaboration), Search for pair-produced resonances decaying to quark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. D **98**, 112014 (2018).
[90] M. Aaboud *et al.* (ATLAS Collaboration), Search for long-lived particles produced in pp collisions at $\sqrt{s} = 13$ TeV that decay into displaced hadronic jets in the ATLAS muon spectrometer, Phys. Rev. D **99**, 052005 (2019).
[91] A. M. Sirunyan *et al.* (CMS Collaboration), Search for long-lived particles decaying into displaced jets in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Rev. D **99**, 032011 (2019).
[92] G. Aad *et al.* (ATLAS Collaboration), Search for new resonances in mass distributions of jet pairs using 139 fb^{-1}

of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, arXiv:1910.08447.
[93] A. M. Sirunyan *et al.* (CMS Collaboration), Search for high mass dijet resonances with a new background prediction method in proton-proton collisions at $\sqrt{s} = 13$ TeV, arXiv:1911.03947.
[94] G. Louppe, M. Kagan, and K. Cranmer, Learning to pivot with adversarial networks, arXiv:1611.01046.
[95] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure, J. High Energy Phys. 05 (2016) 156.
[96] I. Moult, B. Nachman, and D. Neill, Convolved substructure: Analytically decorrelating jet substructure observables, J. High Energy Phys. 05 (2018) 002.
[97] J. Stevens and M. Williams, uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers, J. Instrum. **8**, P12013 (2013).
[98] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Søgaard, Decorrelated jet substructure tagging using adversarial neural networks, Phys. Rev. D **96**, 074034 (2017).
[99] L. Bradshaw, R. K. Mishra, A. Mitridate, and B. Ostdiek, Mass agnostic jet taggers, arXiv:1908.08959.
[100] ATLAS Collaboration, Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS, Report No. ATL-PHYS-PUB-2018-014 (2018).
[101] G. Kasieczka and D. Shih, DisCo fever: Robust networks through distance correlation, arXiv:2001.05310.
[102] L.-G. Xia, QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics, Nucl. Instrum. Methods Phys. Res., Sect. A **930**, 15 (2019).
[103] C. Englert, P. Galler, P. Harris, and M. Spannowsky, Machine learning uncertainties with adversarial neural networks, Eur. Phys. J. C **79**, 4 (2019).
[104] S. Wunsch, S. Jórger, R. Wolf, and G. Quast, Reducing the dependence of the neural network function to systematic uncertainties in the input space, arXiv:1907.11674.
[105] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, Improved variational inference with inverse autoregressive flow, in *Advances in Neural Information Processing Systems*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., Barcelona, Spain, 2016), Vol. 29, pp. 4743–4751.
[106] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, Neural autoregressive distribution estimation, J. Mach. Learn. Res. **17**, 1 (2016).
[107] M. Germain, K. Gregor, I. Murray, and H. Larochelle, Made: Masked autoencoder for distribution estimation, in *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei (Proceedings of Machine Learning Research (PMLR), Lille, France, 2015), Vol. 37, pp. 881–889.
[108] T. Sjöstrand, S. Mrenna, and P. Z. Skands, PYTHIA 6.4 physics and manual, J. High Energy Phys. 05 (2006) 026.
[109] T. Sjöstrand, S. Mrenna, and P. Z. Skands, A Brief Introduction to PYTHIA 8.1, Comput. Phys. Commun. **178**, 852 (2008).
[110] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaitre, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.
[111] A. Mertens, New features in Delphes 3, J. Phys. Conf. Ser. **608**, 012045 (2015).
[112] M. Selvaggi, DELPHES 3: A modular framework for fast-simulation of generic collider experiments, J. Phys. Conf. Ser. **523**, 012033 (2014).
[113] M. Cacciari, G. P. Salam, and G. Soyez, FastJet user manual, Eur. Phys. J. C **72**, 1896 (2012).
[114] M. Cacciari and G. P. Salam, Dispelling the N^3 myth for the k_t jet-finder, Phys. Lett. B **641**, 57 (2006).
[115] M. Cacciari, G. P. Salam, and G. Soyez, The anti-k_t jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.
[116] G. Aad *et al.* (ATLAS Collaboration), Search for diboson resonances in hadronic final states in 139 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, J. High Energy Phys. 09 (2019) 091.
[117] A. M. Sirunyan *et al.* (CMS Collaboration), A multi-dimensional search for new heavy resonances decaying to boosted WW, WZ, or ZZ boson pairs in the dijet final state at 13 TeV, Eur. Phys. J. C **80**, 237 (2020).
[118] J. Thaler and K. Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness, J. High Energy Phys. 02 (2012) 093.
[119] J. Thaler and K. Van Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.
[120] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Vancouver, Canada, 2019), Vol. 32, pp. 8024–8035.
[121] D. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.
[122] F. Chollet, Keras, https://github.com/fchollet/keras, 2017.
[123] 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* (CERN, Geneva, Switzerland, 2016), Vol. 16, pp. 265–283.
[124] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).
[125] C. Huang, D. Krueger, A. Lacoste, and A. C. Courville, Neural autoregressive flows, CoRR abs/1804.00779, 2018.
[126] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, arXiv:1906.04032.
[127] D. Dua and C. Graff, UCI machine learning repository, 2017.
[128] Y. LeCun, C. Cortes, and C. Burges, MNIST handwritten digit database.
[129] A. Krizhevsky, V. Nair, and G. Hinton, CIFAR-10 (Canadian Institute for Advanced Research).