Residual anomaly detection with density estimation

Ranit Das,[1,*] Gregor Kasieczka,[2,†] and David Shih[1,‡]

[1]*NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA*
[2]*Institut für Experimentalphysik, Universität Hamburg, 22761 Hamburg, Germany*

(Received 12 March 2025; accepted 18 November 2025; published 31 December 2025)

We present R-ANODE, a new method for data-driven, model-agnostic resonant anomaly detection that raises the bar for both performance and interpretability. The key to R-ANODE is to enhance the inductive bias of the anomaly detection task by fitting a normalizing flow directly to the small and unknown signal component, while holding fixed a background model (also a normalizing flow) learned from sidebands. In doing so, R-ANODE is able to outperform all classifier-based, weakly supervised approaches as well as the previous ANODE method, which fit a density estimator to all of the data in the signal region instead of just the signal. We show that the method works equally well whether the unknown signal fraction is learned or fixed and is even robust to signal fraction misspecification. Finally, with the learned signal model, we can sample and gain qualitative insights into the underlying anomaly, which greatly enhances the interpretability of resonant anomaly detection and offers the possibility of simultaneously discovering and characterizing the new physics that could be hiding in the data.

DOI: 10.1103/6hgk-q52g

I. INTRODUCTION

Despite countless searches at the LHC [1–7], so far none has turned up any definitive evidence for new physics beyond the Standard Model yet. Since the vast majority of these searches has been model specific, there has been increasing interest in developing new, model-agnostic search strategies powered by modern machine learning in recent years (for reviews and original references, see Refs. [8–11]). Among the model-agnostic search strategies proposed so far, methods for *resonant anomaly detection*—or using additional features x to enhance the sensitivity of a bump hunt in a primary resonant variable m—have received a lot of attention. All of these methods have focused on learning good approximations to the Neyman-Pearson optimal anomaly detector:

$$R_{\text{optimal}}(x) = \frac{p_{\text{data}}(x)}{p_{\text{bg}}(x)}. \tag{1}$$

Here, x are the additional features, and $p_{\text{data}}(x)$ and $p_{\text{bg}}(x)$ are the data and background probability densities, respectively, in the signal region (SR), defined as a window in m. As discussed in Refs. [12,13], $R_{\text{optimal}}(x)$ is monotonic with the signal-to-background likelihood ratio for any unknown signal. Therefore, cutting on it can potentially enhance the significance of any anomaly in the signal region by a large amount.

The main strategy for learning R_{optimal} has been based on weak supervision: training a classifier between the data and a sample of events constructed to resemble the background as closely as possible. One exception to this has been ANODE [12], which is technically an *unsupervised* approach. With ANODE, one trains conditional density estimators (normalizing flows in practice) on the signal region and sideband events, interpolates the latter into the SR, and constructs R_{optimal} by taking their ratio directly. Since this approach is solely based on unsupervised density estimation and does not involve any classifiers, it is technically an unsupervised approach to resonant anomaly detection.

It has been recognized [12,13] that, since density estimation is much more difficult than classification, ANODE suffers in sensitivity relative to classifier-based approaches. However, weak supervision and classifier-based approaches also have their drawbacks. In particular, they do not perform well when the number of signal events is too small, and they can be confused by noisy or uninformative features. Finally, the classifier by itself is not interpretable in that it does not tell us where the signal is, only where it is more overdense relative to the background.

In this paper, we present a new method for resonant anomaly detection that remedies the shortcomings of ANODE. We call our new method Residual ANODE or

*Contact author: ranit@physics.rutgers.edu
†Contact author: gregor.kasieczka@uni-hamburg.de
‡Contact author: shih@physics.rutgers.edu

R-ANODE for short. The idea of R-ANODE is to explicitly fit to the signal component in the SR while holding the background fixed. Assuming data to be a mixture of the signal and background distributions

$$p_{\text{data}}(x,m) = w p_{\text{sig}}(x,m) + (1-w) p_{\text{bg}}(x,m), \quad (2)$$

with w as the fraction of signal in the data and $p_{\text{bg}}(x,m)$ as the fixed background template, we attempt to model $p_{\text{sig}}(x,m)$.

Fitting for the signal density directly, instead of the full data density, not only improves on the performance of ANODE, but it even surpasses the performance of the Idealized Anomaly Detector (IAD), which sets an upper bound to classifier-based approaches. The IAD was introduced in Ref. [13] and corresponds to training a classifier on data vs perfectly simulated background (i.e., the same background model that the background in the data was drawn from). The IAD is limited by finite training statistics, noisy features, and finite model capacity. So, it cannot fully approach the optimal anomaly detector or (by extension) the fully supervised classifier. R-ANODE is able to surpass the IAD since it assumes more about the signal vs background mixture.

In R-ANODE, one has the option to fix the signal fraction w during training or to let it be a learnable parameter along with the signal density. We explore both options in this work, finding that for fixed w the method is quite robust to w-misspecification, retaining excellent sensitivity to the signal even when w is larger or smaller than the true w by nearly an order of magnitude. For learnable w, we find that the method is robust and there is only a slight drop in overall signal sensitivity; furthermore, the learned w tracks the true w well down to a nominal significance of $\sim 0.6\sigma$ (corresponding to $\sim$200 injected signal events). Thus, R-ANODE could potentially be used to place measure or place limits on the signal fraction directly.

Finally, with the learned signal density, one can also draw potentially unbiased signal samples and directly learn about the properties of the new physics model hiding in the data. In this way, R-ANODE simultaneously offers significantly improved performance over other methods and also much greater interpretability.

II. RESIDUAL ANODE METHOD

In R-ANODE, we model the signal distribution $p_{\text{sig}}(x,m)$ for $m \in \text{SR}$ directly with a normalizing flow and use it to fit to data from Eq. (2). We minimize the negative log likelihood averaged over the SR data,

$$L = -\mathbb{E}_{x,m\sim \text{SR data}} \log p_{\text{data}}(x,m), \quad (3)$$

with respect to the parameters of $p_{\text{sig}}(x,m)$ while keeping $p_{\text{bg}}(x,m)$ fixed during training.

To obtain the joint background density in the SR from the sidebands, we break it up into two factors:

$$p_{\text{bg}}(x,m) = p_{\text{bg}}(x|m) p_{\text{bg}}(m). \quad (4)$$

The first factor, the conditional density $p_{\text{bg}}(x|m \in \text{SR})$, is obtained by interpolating from the sidebands similarly as in Refs. [12,13]. The second factor, the background mass distribution $p_{\text{bg}}(m \in \text{SR})$, can be similarly obtained by interpolating from sidebands or (as we do in this work) approximating it with the data mass distribution under the assumption that there is no statistically significant anomaly in the inclusive bump hunt. This allows us to get $p_{\text{bg}}(x,m)$ for $m \in \text{SR}$.

For the signal fraction w, we explore two options:

(1) Hold w fixed during training. We then scan over different values of w, exploring the effect of fixing w to be larger or smaller than the true w.
(2) Keep w as a learnable parameter during training, with the same optimizer and hyperparameters used for $p_{\text{sig}}(x,m)$.[1]

We note that the true signal fraction w_{true} is related to the number of signal and background events in the SR:

$$w_{\text{true}} = \frac{N_{\text{sig,SR}}}{N_{\text{sig,SR}} + N_{\text{bg,SR}}}. \quad (5)$$

For small amounts of signal that we assume throughout this work, this relation is approximately linear, i.e., $w_{\text{true}} \approx N_{\text{sig,SR}}/N_{\text{bg,SR}}$.

Finally, the anomaly score is constructed as

$$R(x,m) = \frac{p_{\text{sig}}(x,m)}{p_{\text{bg}}(x,m)}. \quad (6)$$

III. SETUP

A. Dataset

We use the LHCO R&D Dataset [8,14] for our studies with dataset and train-val-test splits similarly as in Refs. [12,13]. In the following, a brief summary of the dataset is given.

QCD dijet events form the Standard Model (SM) background and $W' \to X(\to qq)Y(\to qq)$ events with $m_{W'} = 3.5$ TeV, $m_X = 500$ GeV and $m_Y = 100$ GeV are used as signal. These are simulated using Pythia [15,16] and Delphes3.4.1 [17–19]. The reconstructed particles are clustered into jets using the anti-k_T algorithm [20,21] with $R = 1$ using FastJet [22]. Events are required to satisfy the $p_T > 1.2$ TeV jet trigger.

[1]Using a different optimizer and hyperparameter to learn w might be interesting, but it could also necessitate further hyperparameter tuning, so we save this for future work.

The training features used are

$$m = m_{JJ}, \qquad x = [m_{J_1}, \Delta m_J, \tau_{21}{}^{J_1}, \tau_{21}{}^{J_2}], \tag{7}$$

where invariant masses of the subjets satisfy $m_{J_1} < m_{J_2}$, and $\Delta m_J = m_{J_2} - m_{J_1}$. The n-subjettiness ratios are defined as $\tau_{ij} = \tau_i/\tau_j$ [23,24].[2] The resonant variable is chosen as the dijet invariant mass m_{JJ}, with the signal region (SR) defined as $m \in [3.3, 3.7]$ and its complement $m \notin [3.3, 3.7]$ forming the sideband (SB) regions.

For the primary dataset (meant to represent actual unlabeled data from the experiment), we use all 1 million SM background events from the original R&D dataset and inject different amounts of signal (1000 or lower) from the first 70,000 signal events of the R&D dataset. The amount of signal in the SR is approximately 76% of the total injected signal, i.e., $N_{\text{sig,SR}} \approx 0.76 \times N_{\text{sig}}$. Meanwhile, the amount of background in the SR is $N_{\text{bg,SR}} \approx 120,000$. The highest signal injection of $N_{\text{sig}} = 1000$ corresponds to $\approx 2\sigma$ significance for the inclusive bump hunt in m_{JJ} (so, below discovery threshold).

All methods split this primary dataset into training and validation sets, with 80/20 split for R-ANODE and 50/50 split for IAD and ANODE.[3] For IAD, following Ref. [13], we used an additional 272,000 QCD dijet events in the SR [26] (with the same split as for data) to train the classifier. For evaluating Significance Improvement Characteristic (SIC) curves, we use the remaining 30,000 signal events from the original R&D dataset, along with the additional 340,000 QCD dijet events in the SR from Ref. [26].

For each N_{sig}, 10 different datasets are produced by injecting different randomly selected signal events from the R&D dataset. These are used to derive error bars on the performance curves in the figures below.

B. Model architectures and selection

Here, we briefly describe the model architectures and model selection procedures used in this work. More details can be found in Appendix A.

In Refs. [12,13], the conditional normalizing flow models used to fit the data in the SR and SB regions were masked autoregressive flows with affine transformations [27,28]. Here, we use the same model for learning $p_{\text{bg}}(x|m)$ in the SB, but for learning $p_{\text{data}}(x|m)$ and $p_{\text{sig}}(x, m)$ in the SR (for ANODE and R-ANODE, respectively), we switch to rational-quadratic spline (RQS) transformations [29] instead. Using the more expressive RQS transformations for the data and signal distributions was found to improve the performance of the methods. Interestingly, using RQS transformations for fitting the background in the SB led to *worse* results. Since RQS transformations are more expressive, it possibly leads to overfitting of the simpler background and/or interpolation.

For a proof of concept, $p_{\text{bg}}(m \in \text{SR})$ was estimated using the mass histograms of each full dataset using SciPy RV_HISTOGRAM. Even though the full data contain signal, this should be an excellent approximation to $p_{\text{bg}}(m)$ under our assumption that there is no significant excess in the inclusive bump hunt. In future work, it would be interesting to also compare this against a proper interpolation of the mass histogram from the sidebands.

Finally, for the IAD and fully supervised classifier, we utilize boosted decision trees (specifically a `HistGradientBoostingClassifier` from `scikit-learn` [30,31]), which were shown in Ref. [25] to be robust under uninformative features.

For ANODE and R-ANODE, the method is trained on 20 random train/val splits of a given dataset, and for each split, the 10 epochs with lowest validation losses are selected. Probabilities from these 200 models (or fewer in the case of learnable w; see Sec. IV C for details) are ensembled to obtain the final predictions of the methods. For the IAD, the model is retrained on 50 random train/val splits of a given dataset, and the models with lowest validation loss for each retraining are chosen to ensemble, similarly as in Ref. [25].

C. Performance metrics

The main performance metric we use in this paper is the SIC. In terms of the signal efficiency (ϵ_S or true positive rate) and background efficiency [ϵ_B or false positive rate (FPR)], one has

$$\text{SIC} = \frac{\epsilon_S}{\sqrt{\epsilon_B}}. \tag{8}$$

The SIC characterizes the improvement to the nominal significance, using a cut on the anomaly score from a given method. We will also quote this nominal significance below, which is given by

$$\text{significance} = \text{SIC} \times \frac{N_{\text{sig,SR}}}{\sqrt{N_{\text{bg,SR}}}}. \tag{9}$$

The latter factor is the nominal significance from the inclusive bump hunt (i.e., prior to any enhancement from a resonant anomaly detection method).

In general, the SIC depends on the working point, and we will exhibit this dependence by plotting SIC vs the signal efficiency. At lower signal efficiencies, there are large statistical uncertainties in the SIC values. Hence, we introduce a cutoff in the SIC curves when the relative statistical error on the background efficiency exceeds 20%.

[2]These features correspond to the baseline features defined in Ref. [25], along with the invariant mass m_{JJ}. We leave exploring the effects of noisy features and extended set of features for a future study.

[3]Following previous work, we use the 50/50 split for ANODE and IAD. For R-ANODE, we found that the 80/20 split leads to a better performance as compared to the 50/50 split. We think that this is because R-ANODE has to explicitly learn the small amount of signal present in data, and this benefits from a higher amount of training data.

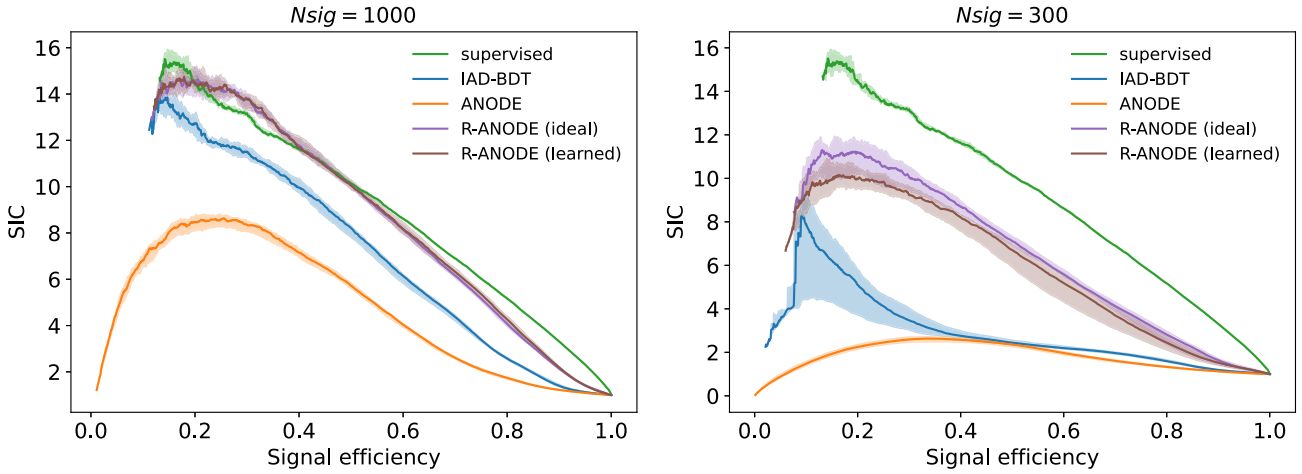

FIG. 1. SIC curves for $N_{\text{sig}} = 1000$ (left) and $N_{\text{sig}} = 300$ (right). R-ANODE with a learnable w almost saturates the performance of an idealized R-ANODE where w is fixed to its true value during training. Both these methods outperform the IAD and ANODE. The supervised classifier sets the upper limit for performance of all these methods, and at larger signal strengths, R-ANODE saturates this upper limit before the IAD does.

It is also useful to quantify the performance at particular working points. Maximum SIC is defined as the maximum of all the SIC values above the cutoff for statistical stability. We also use SIC at FPR $= 0.001$ to represent a typical, fixed working point that an experiment might choose in practice.

IV. RESULTS

We first show the performance of an ideal version of R-ANODE with the correct value of $w = w_{\text{true}}$ fixed during training, for different amounts of injected signal. Then, we present the results for different fixed values of w for $N_{\text{sig}} = 1000$. And finally, we show the results for the case where we attempt to learn w.

We compare R-ANODE to ANODE, the IAD, and a fully supervised classifier. For all figures, unless otherwise mentioned, the curves show the median value and 68% confidence bands for the results obtained by retraining the methods on 10 different datasets described in Sec. III A corresponding to different randomly selected signal injections. Since the upper limit for performance of the classifier-based data-driven approaches like CWoLA [32,33], CATHODE [13], CURTAINS [34], etc. is the IAD, we omit the explicit comparison to these methods. The supervised classifier sets the upper bound to performance for all methods on this signal hypothesis. Interestingly, as observed in Refs. [13,25,34], for these signal injections, there is a difference in performance between the IAD and supervised classifier. This is because the IAD is not actually fully optimal—it is limited by finite training statistics and finite model capacity. The truly optimal anomaly detector given by Eq. (1) would be completely equivalent to the fully supervised classifier, since it would be monotonic with it [13].

A. Idealized version: Fixing $w = w_{\text{true}}$

We study the performance of R-ANODE for different amounts of signal injections, in the idealized scenario where $w = w_{\text{true}}$ is held fixed during training. In Fig. 1, we show the SIC curves for $N_{\text{sig}} = 1000$ and $N_{\text{sig}} = 300$ (which correspond to 2.2σ and 0.7σ nominal inclusive significance in the SR, respectively). In both cases, the R-ANODE (ideal) method outperforms ANODE and the IAD across a wide range of signal efficiencies. For $N_{\text{sig}} = 1000$, R-ANODE (ideal) nearly closes the gap between the IAD and the fully supervised classifier.[4] Meanwhile, for $N_{\text{sig}} = 300$, we see that the gap between R-ANODE and the fully supervised classifier widens, but the gap with the IAD widens even further. The IAD suffers relative to the fully supervised classifier due to limited training statistics, whereas R-ANODE benefits from not being classifier based and from its stronger inductive bias.

Next, we examine the SIC values and total achieved significances at a fixed FPR of 10^{-3}. These are shown in Fig. 2 as a function of the number of injected signal events, for the different methods. We see from both plots that R-ANODE (ideal) achieves a better sensitivity to signal at all signal strengths and allows us discover the signal at cross sections that are ∼25% lower than the IAD.

B. Scanning fixed w

Next, we study the more realistic case of scanning across different, fixed values of w during training. We focus on the benchmark case of $N_{\text{sig}} = 1000$ for simplicity.

[4]Interestingly, R-ANODE even seems to outperform the fully supervised classifier for a small range of signal efficiencies. This should not be possible, and it may not be statistically significant, or it could point to a deficiency of the model architecture chosen for the fully supervised classifier.

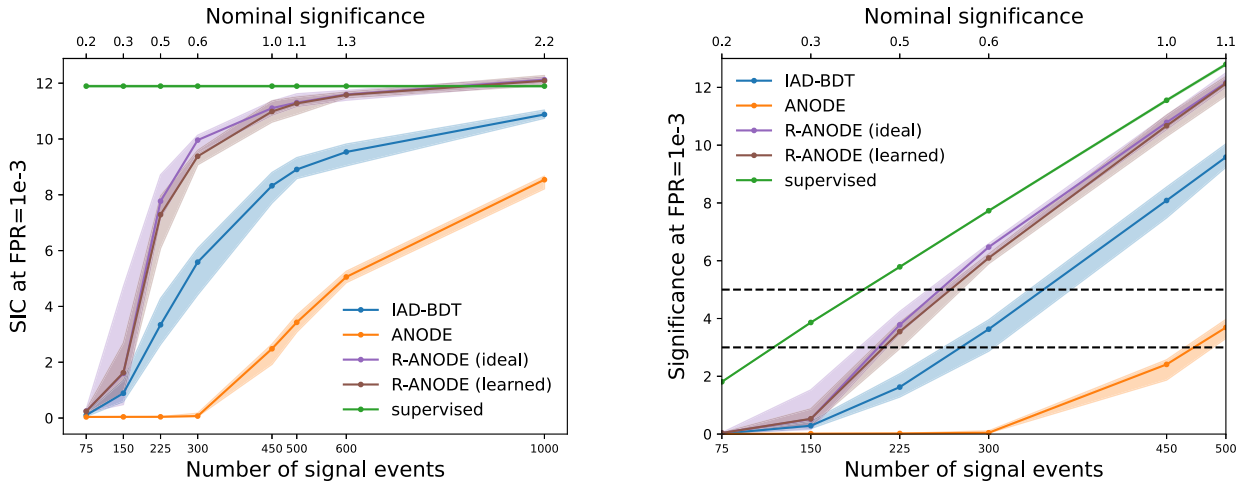

FIG. 2. Left: SIC (at FPR = 0.001) vs N_{sig} (amount of signal injected to data). Right: total significance achieved (at FPR = 0.001) vs N_{sig}. Again, we see that R-ANODE with learnable w matches the idealized R-ANODE with $w = w_{\text{true}}$, and both outperform IAD and ANODE, across a wide range of signal levels.

From Fig. 3, we see that the performance curves and SIC values are actually quite robust to incorrect choices of w. We see that only at w-values an order of magnitude different than w_{true} is there some noticeable difference in the max SIC. At $w < w_{\text{true}}$, the trained signal model could be overestimating regions in phase space where the signal density is higher than background. This might help it locate the region with higher signal density easier, which could explain why in Fig. 3 we see a higher max SIC than w_{true}. At $w \ll w_{\text{true}}$, R-ANODE loses the capability of modeling the signal, which leads to a sharp decline in performance. This could be possibly used to put a lower bound on w. Finally, for $w > w_{\text{true}}$, the SIC declines slowly but does not completely go away. Here, the method is learning to model the signal in addition to some amount of background. We expect that as $w \to 1$ the R-ANODE method smoothly interpolates to the original ANODE method.

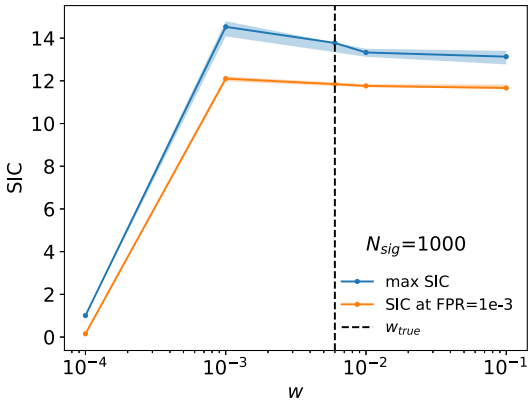

FIG. 3. For $N_{\text{sig}} = 1000$, we scan across different w-values, while holding it fixed during training. The significance improvement is robust for incorrect choices of w; however, the performance drops significantly for $w \ll w_{\text{true}}$.

C. Learnable w

Finally, we explore what happens if we allow w to be learned during training. For simplicity, we use the same optimizer for w as the rest of the signal model; i.e., we just include w in the list of learnable parameters during training.

For each dataset—as described in Sec. III A—we take w-values from 20 retrainings and 10 lowest validation losses from each training. For learnable w, we find that some trainings converge to extremely low values, below 10^{-6}. For 100,000 events in SR, such low w values correspond to $N_{\text{sig}} < 1$. We label these trainings as having "undetected signal" and choose to exclude them from the analysis. To do so, we form histograms of the 200 learned w-values as shown in Fig. 6 of Appendix B. These histograms show a clearly well separated, bimodal structure, with one mode corresponding to the undetected signal case. We therefore devise a simple by-eye cut to remove this mode for a given dataset. In Appendix B, we explore other ways of removing this mode.

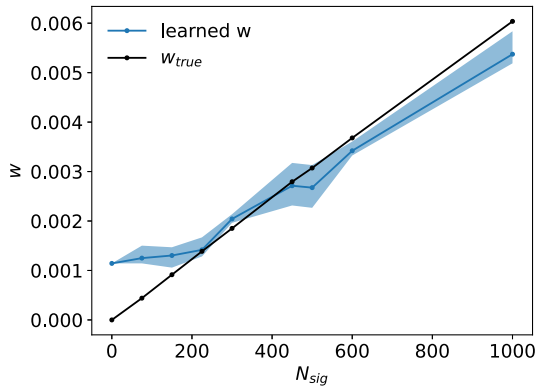

FIG. 4. A comparison between learned w and w_{true} for different N_{sig} values.

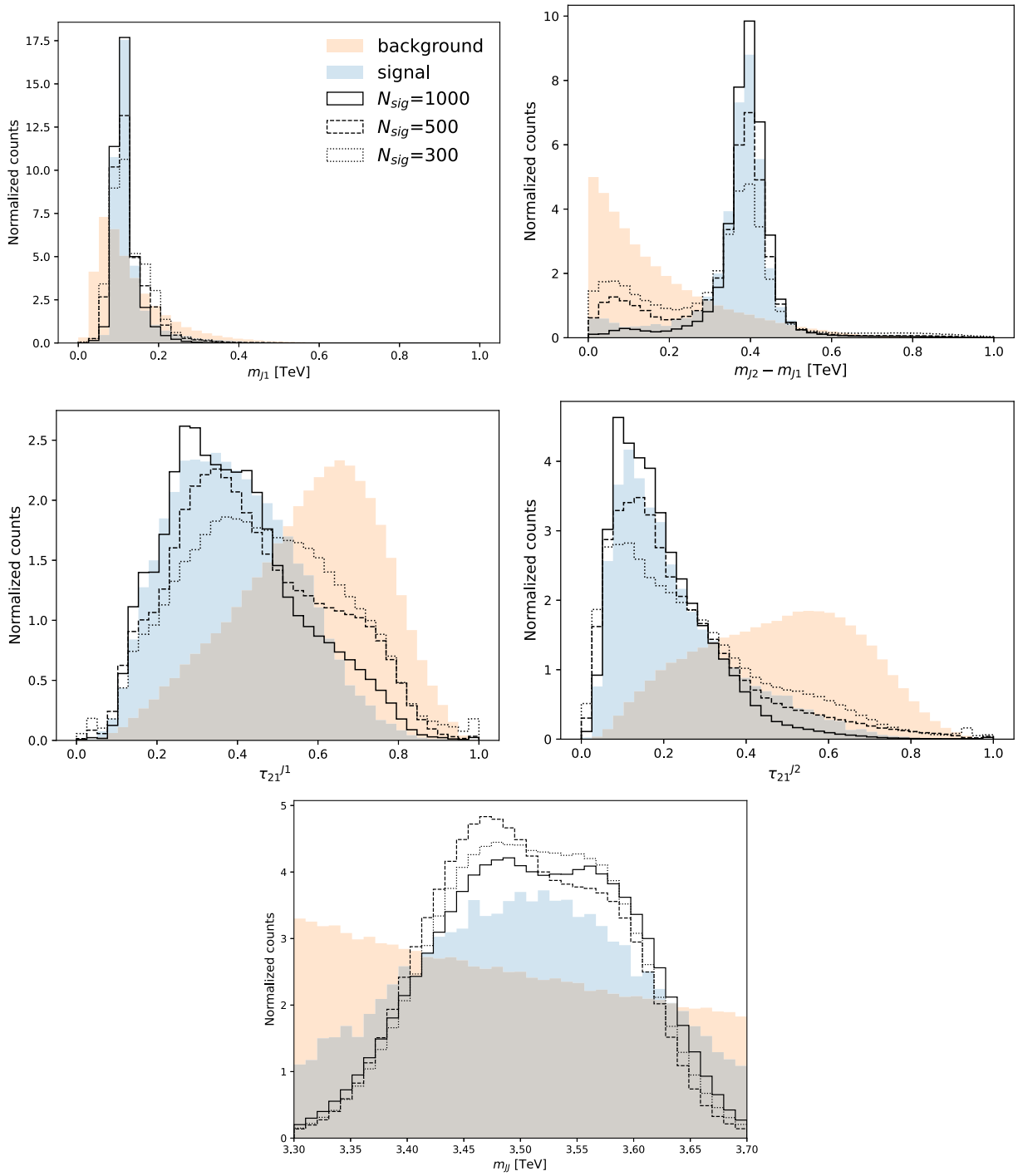

FIG. 5. Samples for R-ANODE with a learnable w during training, for $N_{\text{sig}} = [1000, 500, 300]$. Solid histograms show the true background (signal) distributions in orange (blue), while black lines indicate the signal distributions learned for different injected signal amounts. Overall, the qualitative agreement between the learned and actual signal distributions is decent, although the samples clearly get worse for lower signal injections due insufficiency of training data.

The remaining w-values are then averaged to produce a final output for the learned w for a given dataset. (The corresponding p_{sig} distributions are also averaged as described above, to produce the anomaly score, etc. in the learned w case.)

In Fig. 4, we show the final learned w vs true w, for a range of N_{sig} values. We also indicate the spread of outcomes across the 10 different datasets as a blue uncertainty band. We see that for the most part there is a good agreement between the learned w and true w values. At larger w, there is a slight downward bias in the learned w, while at smaller w, there is a pronounced upward bias. Indeed, for the no-signal fit, the learned w values have an

average value of $\approx 10^{-3}$. According to Fig. 4, our method with learnable w is unable to detect $\lesssim 300$ signal events, for the given amount of background events.

The undetectability of signal below some threshold may be an irreducible limitation of the R-ANODE method with learnable w, or it could be a byproduct of our decision to exclude the undetected signal trainings and keep trainings where a nonzero w was returned. Ideally, one could devise a data-driven criterion that would correctly switch between the undetected and detected signal modes as N_{sig} decreases. Alternatively, one could stick to the current procedure and devise a process to quantify the lower threshold on signal strength in a data-driven way. For example, one could imagine applying R-ANODE to synthetic background events sampled from the interpolated sideband model $p_{\text{bg}}(x, m)$, which might enable us to measure the "floor" for w. We save such investigations for future work.

As noted in Sec. IV B, R-ANODE as an anomaly detector is robust to incorrect choices of w. In Figs. 1 and 2, the R-ANODE performance curves with learnable w are shown in brown. Although a slight performance drop relative to the idealized $w = w_{\text{true}}$ case is seen for lower signal strengths, overall it is encouraging to see that the performance of the learnable w case is actually extremely robust and nearly saturates the ideal R-ANODE performance.

D. Samples from the signal distribution

Using the IAD, one could make cuts on data with the anomaly score $R(x)$, to obtain the resulting signal samples. This requires a choice for the cut, which is not necessarily obvious from data. It also might leave too few signal events for interpretable distributions. Finally, the distributions one obtains would in general be biased, since the cut on $R(x)$ preferentially selects signal events that are more overdense relative to background. In R-ANODE, however, we benefit from being able to directly (over)sample signal events in an (in principle) unbiased way.

In Fig. 5, we show a comparison of these signal densities obtained from R-ANODE (learned) for different signal-injections.[5] We see that the learned signal densities show good qualitative agreement with the true signal densities and allow us to get rough estimates on properties of the signal, like the invariant mass, subjet mass, pronginess of each subjet, etc. There is a clear degradation in the quality of the learned signal densities at lower signal strengths. Clearly, an important future direction here will be to devise data-driven methods to place uncertainties on these learned distributions.

V. CONCLUSIONS

R-ANODE is a novel method for model-agnostic, data-driven resonant anomaly detection that sets a new standard for the state of the art in both performance and interpretability. The key to R-ANODE is to build in more inductive bias to the task of learning the optimal anomaly score. It builds on the previous approach of ANODE [12], but instead of fitting a density estimator for the data, it aims to isolate the small contribution from the unknown signal. By fitting a density estimator for just the signal component, holding fixed the background density learned from the sidebands, R-ANODE is able to better approach the optimal anomaly detector, surpassing approaches based on classifiers and weak supervision.

By learning the signal density, R-ANODE allows us to model the signal distributions directly from data, and this could be used to simultaneously discover and fully characterize the signal.

One of the issues we have not addressed in this paper and which we save for a future study is background estimation. Combining R-ANODE with a bump hunt would require us to examine the issue of mass sculpting in more detail, as was done in Ref. [35] for classifier-based approaches. Alternatively, perhaps the technique of "direct background estimation" from Ref. [12] could be revisited, or the uncertainties on the learned w could be properly calibrated somehow from data. We also plan to investigate the efficacy of using the direct estimation of w for anomaly detection as an alternative to the bump hunt.

Another potential issue which we look forward to exploring further in future work is performance and robustness under the inclusion of additional, noisy high-level features [25,36]. Perhaps the boosted decision trees-based density estimator used in Ref. [36] could be beneficial for this task, while speeding up the density estimation as well. It will also be important to examine how the performance of this method varies with different types of signal. Finally, it will be very interesting to go beyond high-level features and implement the R-ANODE method on the full phase space of jet constituents [37], where classifier-based approaches are especially challenged.

ACKNOWLEDGMENTS

We are grateful to Manuel Sommerhalder for assisting us with the scripts for training the background model in the SB. The work of R. D. and D. S. was supported by DOE Grant No. DE-SC0010008. G. K. acknowledges support by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy—EXC 2121 Quantum Universe—Grant No. 390833306. The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey [38] for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here.

DATA AVAILABILITY

The data that support the findings of this article are openly available [39].

[5] We found that fixing to w_{true} did not improve these distributions much, and hence we omit showing similar plots for R-ANODE (ideal).

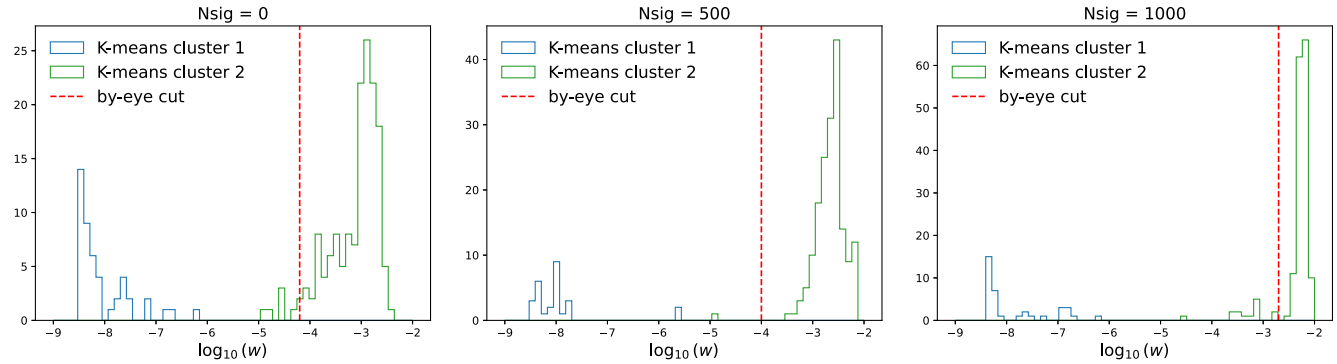

FIG. 6. Histograms of all 200 learned w values for three choices of N_{sig} (single dataset only). We see a clearly separated bimodal structure, with undetected signal trainings clustering at the low end (zero up to numerical precision). The red dashed lines are our thresholds selected by eye, for trainings with undetected signal; w values below these thresholds are rejected. Similarly, applying K-MEANS results in two clusters shown in blue and green; the w values in the blue cluster are discarded. The by-eye determined thresholds and K-MEANS clustering algorithm are applied independently to each dataset.

APPENDIX A: IMPLEMENTATION DETAILS

The background model $p_{\text{bg}}(x|m)$ is the same as in Refs. [12,13]: a masked autoregressive flow with affine transformations [27]. It contains 15 masked autoencoder for distribution estimation blocks, with each block consisting of one hidden layer of 128 nodes. We train the density estimator using PyTorch [40] in the SB region for 100 epochs with the ADAM optimizer [41], with a learning rate of 10^{-4}, batch size 256, and batch normalization with a momentum of 1.0. The base distribution chosen is unit normal. The SB data are split into a training-validation split of 50/50, and models with 10 lowest validation losses are selected. The probabilities obtained from these 10 models are averaged to obtain $p_{\text{bg}}(x|m)$.[6]

For $p_{\text{sig}}(x, m)$ (fit to data in the SR only), we use RQS transformations [29] with six masked autoencoder for distribution estimation blocks, with each block consisting of two hidden layers with 64 nodes, dropout 0.2, and batch normalization is applied in between layers. The same model is also used for $p_{\text{data}}(x|m)$ in the SR for ANODE. The RQS models for all cases are trained with a unit normal base distribution and a learning rate of 0.0003 with the ADAMW [42] optimizer, with a batch size of 256 for 300 epochs. The train/val split and ensembling of p_{sig} and p_{data} were described in Sec. III B. We select the lowest validation losses to avoid overfitting, and our ensembling avoids selecting the lowest loss from a single training as a downward fluctuation.

For the IAD and the fully supervised classifier, we train `HistGradientBoostingClassifier` with default hyperparameters for 200 epochs, similarly as in Ref. [25].

Similarly as in Ref. [13], the features x in the SB are first shifted and scaled so that $x \in (0, 1)$, logit transformed, and then standardized with mean 0 and variance 1. The same preprocessing parameters from the SB were reused for the SR, for all the methods. For R-ANODE and supervised, the mass is centered around zero, by subtracting 3.5 from all mass values.

APPENDIX B: TRAININGS WITH UNDETECTED SIGNAL

As mentioned in Sec. IV C, Fig. 6 shows the full histograms of learned w values, for different N_{sig} values (single dataset only). The undetected signal trainings where w converged to zero (up to machine precision) are clearly seen in the bimodal histograms. The cuts shown in Fig. 6 (red dashed lines) are based on by-eye estimates, performed for each dataset, and all the results for R-ANODE (learned) are based on models corresponding to w-values surviving

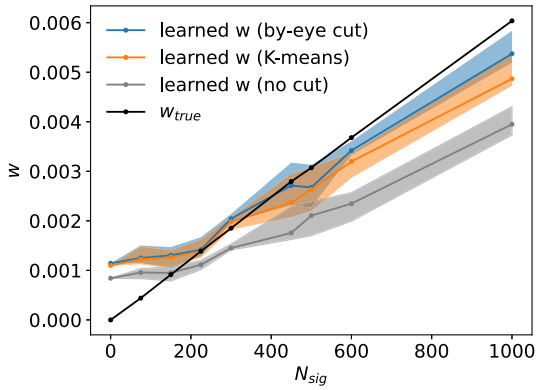

FIG. 7. A comparison between learned w and w_{true} for different N_{sig} values, after a by-eye cut, K-MEANS cut and without any cut.

[6]For simplicity, we use the same background model, obtained from a $N_{\text{sig}} = 1000$ case, as $p_{\text{bg}}(x|m)$ for all signal injections.

above these cuts. We also investigated clustering all 200w values for each dataset into two clusters using K-MEANS clustering [43]. Figure 6 shows the resulting clusters (in blue and green). We discard the w values in the cluster with lower mean as undetected signal. Figure 7 compares the w values obtained from K-MEANS, the bye-eye cut, and without any cut. We found that not imposing any cut underestimates the w compared to when unphysical w values are removed. Both the by-eye cut and K-MEANS clustering result in similar w values.

[1] ATLAS Collaboration, Exotic Physics Searches (2023), https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults.
[2] ATLAS Collaboration, Supersymmetry searches (2023), https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults.
[3] ATLAS Collaboration, Higgs and Diboson Searches (2023), https://twiki.cern.ch/twiki/bin/view/AtlasPublic/HDBSPublicResults.
[4] CMS Collaboration, CMS Exotica Public Physics Results (2023), https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO.
[5] CMS Collaboration, CMS Supersymmetry Physics Results (2023), https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS.
[6] CMS Collaboration, CMS Beyond-two-generations (B2G) Public Physics Results (2023), https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G.
[7] LHCb Collaboration, Publications of the QCD, Electroweak and Exotica Working Group (2023), http://lhcbproject.web.cern.ch/lhcbproject/Publications/LHCbProjectPublic/Summary_QEE.html.
[8] G. Kasieczka, B. Nachman, D. Shih *et al.*, Rep. Prog. Phys. **84**, 124201 (2021).
[9] T. Aarrestad *et al.*, SciPost Phys. **12**, 043 (2022).
[10] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, Nat. Rev. Phys. **4**, 399 (2022).
[11] V. Belis, P. Odagiu, and T. K. Aarrestad, Rev. Phys. **12**, 100091 (2024).
[12] B. Nachman and D. Shih, Phys. Rev. D **101**, 075042 (2020).
[13] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, Phys. Rev. D **106**, 055006 (2022).
[14] G. Kasieczka, B. Nachman, and D. Shih, R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge (2022).
[15] T. Sjöstrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026.
[16] T. Sjöstrand, S. Mrenna, and P. Skands, Comput. Phys. Commun. **178**, 852 (2008).
[17] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, J. High Energy Phys. 02 (2014) 057.
[18] A. Mertens, J. Phys. Conf. Ser. **608**, 012045 (2015).
[19] M. Selvaggi, J. Phys. Conf. Ser. **523**, 012033 (2014).
[20] M. Cacciari and G. P. Salam, Phys. Lett. B **641**, 57 (2006).
[21] M. Cacciari, G. P. Salam, and G. Soyez, J. High Energy Phys. 04 (2008) 063.
[22] M. Cacciari, G. P. Salam, and G. Soyez, Eur. Phys. J. C **72**, 1896 (2012).
[23] J. Thaler and K. Van Tilburg, J. High Energy Phys. 03 (2011) 015.
[24] J. Thaler and K. Van Tilburg, J. High Energy Phys. 02 (2012) 093.
[25] T. Finke, M. Hein, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, and M. Sommerhalder, Phys. Rev. D **109**, 034033 (2024).
[26] D. Shih, Additional QCD Background Events for LHCO2020 R&D (signal region only) (2021).
[27] G. Papamakarios, T. Pavlakou, and I. Murray, arXiv:1705.07057.
[28] L. Dinh, D. Krueger, and Y. Bengio, arXiv:1410.8516.
[29] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, arXiv:1906.04032.
[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, J. Mach. Learn. Res. **12**, 2825 (2011), https://www.jmlr.org/papers/v12/pedregosa11a.html.
[31] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Long Beach, CA, 2017).
[32] J. H. Collins, K. Howe, and B. Nachman, Phys. Rev. Lett. **121**, 241803 (2018).
[33] J. H. Collins, K. Howe, and B. Nachman, Phys. Rev. D **99**, 014038 (2019).
[34] J. A. Raine, S. Klein, D. Sengupta, and T. Golling, Front. Big Data **6**, 899345 (2023).
[35] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, Phys. Rev. D **107**, 114012 (2023).
[36] M. Freytsis, M. Perelstein, and Y. C. San, J. High Energy Phys. 02 (2024) 220.
[37] E. Buhmann, C. Ewen, G. Kasieczka, V. Mikuni, B. Nachman, and D. Shih, Phys. Rev. D **109**, 055015 (2024).
[38] https://it.rutgers.edu/oarc
[39] R. Das, Residual anode (2023), https://github.com/rd804/R-ANODE
[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, arXiv:1912.01703.
[41] D. P. Kingma and J. Ba, arXiv:1412.6980.
[42] I. Loshchilov and F. Hutter, arXiv:1711.05101.
[43] I. Loshchilov and F. Hutter, Biometrics **21**, 761 (1965).