Combining resonant and tail-based anomaly detection

Gerrit Bickendorf[,1] Manuel Drees,[1] Gregor Kasieczka,[2] Claudius Krause[,3] and David Shih[4]

[1]*Bethe Center for Theoretical Physics and Physikalisches Institut, Universität Bonn,*
Nussallee 12, 53115 Bonn, Germany
[2]*Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany*
[3]*Institut für Theoretische Physik, Universität Heidelberg,*
Philosophenweg 12, 69120 Heidelberg, Germany
[4]*New High Energy Theory Center, Rutgers University Piscataway, New Jersey 08854-8019, USA*

In many well-motivated models of the electroweak scale, cascade decays of new particles can result in highly boosted hadronic resonances (e.g., $Z/W/h$). This can make these models rich and promising targets for recently developed resonant anomaly detection methods powered by modern machine learning. We demonstrate this using the state-of-the-art classifying anomalies through outer density estimation (CATHODE) method applied to supersymmetry scenarios with gluino pair production. We show that CATHODE, despite being model agnostic, is nevertheless competitive with dedicated cut-based searches, while simultaneously covering a much wider region of parameter space. The gluino events also populate the tails of the missing energy and H_T distributions, making this a novel combination of resonant and tail-based anomaly detection.

I. INTRODUCTION

The absence of new physics at the LHC is an enduring mystery. Many well-motivated theoretical frameworks such as supersymmetry, extra dimensions, and composite Higgs have predicted signatures of new particles at the weak scale, yet countless searches for these new particles have not found any significant evidence for them to date.

Nearly all of these searches for physics beyond the Standard Model (BSM) are model specific to some degree, optimized for specific signal scenarios, often using simulations. It is highly likely that these searches have not thoroughly covered the full phase space at the LHC, leaving a real possibility of new physics simply hiding in the data at the LHC, undiscovered because we have not searched for it.

Recently there has been considerable interest in developing more model-agnostic search strategies for the LHC [1–3]. In particular, a lot of activity has focused on "resonant anomaly detection" methods [4–22]. In these approaches, one singles out a specific kinematic feature (e.g., the invariant mass of something in the event) in which new physics is postulated to be localized (resonant) to a window. This window serves as the signal region (SR) of

the anomaly search. Then one uses the sidebands and modern machine learning techniques to learn a multivariate, data-driven background template in additional features x. Finally, one employs further techniques (such as a classifier) to learn the difference between the background template and the data itself in the SR, in the form of an anomaly score

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

If $p_{bg}(x)$ is the true background density and the classifier is optimal, this is the Neyman-Pearson optimal (idealized) anomaly detector in the SR. By cutting on $R(x)$, one can greatly enhance the significance of any resonant new physics in the SR.

So far this activity has almost exclusively focused on new physics that is fully localized—both in the SR and in the features x—and using a global resonant feature such as the invariant mass of a dijet system. Here we point out that the resonant anomaly detection technique is more general and both assumptions can be easily relaxed.

First, resonant anomaly detection methods can be applied to any resonant feature in the event, as long as the background satisfies the assumption of smoothness in that feature. One strong motivation for considering this broader perspective is that in many well-motivated models, such as those for the electroweak hierarchy, highly boosted resonances (either $Z/W/h$ from the SM or additional BSM
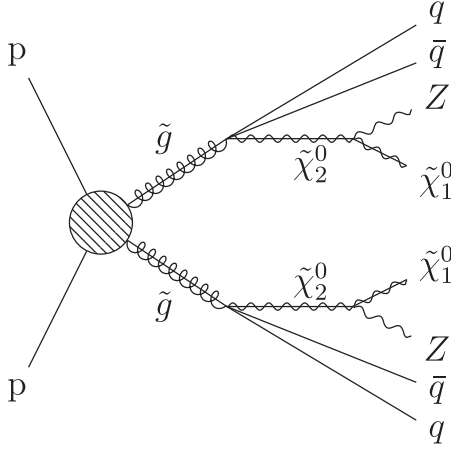
FIG. 1. Diagram of the signal process $pp \to \tilde{g}\tilde{g}$ with $\tilde{g} \to q\bar{q}\tilde{\chi}_2^0, \tilde{\chi}_2^0 \to Z\tilde{\chi}_1^0$.

particles such as a heavier Higgs boson) can be quite common in the decays of heavier particles.

Additionally, we seek to broaden the scope of resonant anomaly detection in this work by pointing out that the signal need not be localized in all (or any) of the features x; it can also appear on the tails of the x distribution (although of course the signal needs to be distinguishable in some of the features). This is a feature of resonant anomaly detection that has not been utilized so far. Anomalies on the tails of distributions such as p_T^{miss}, H_T, and M_{eff} are quite common and plausible in models of TeV-scale new physics.

In this work, we illustrate this broader application of resonant anomaly detection using a supersymmetric (SUSY) scenario as a well-motivated example. This SUSY scenario consists of gluino pair production, with the gluinos decaying to neutralinos plus a pair of (light quark) jets, and the neutralino decaying to another neutralino (LSP) through an on-shell Z boson, as shown in Fig. 1. The LSP neutralino is much lighter than the second neutralino, meaning the Zs are highly boosted. Therefore every event has two boosted Zs, jets, and missing energy. CMS previously searched for this signal with a cut-based analysis [23]. It defined a series of SRs requiring leading and subleading AK8 jets within $m \in [70, 100]$ GeV and considering p_T^{miss} in different exclusive bins. The background was estimated in two steps. First, the total number of events B_{norm} in the SR was determined using sidebands in the leading AK8 jet mass (the subleading AK8 jet was required to be in the SR). Then the distribution in p_T^{miss} bins (shape) is determined using the p_T^{miss} distribution in control regions defined by requiring both AK8 jets to be outside the SR, renormalized to B_{norm}.

Here we point out that we can use one of the Zs to define a SR for resonant anomaly detection, and then we can use the rest of the kinematic variables (m_{jet} of the subleading AK8 jet, p_T^{miss}, H_T, etc.) to play the role of x in resonant anomaly detection. We show that this allows for a potentially more expansive and model-agnostic search, while not

sacrificing much in sensitivity to the original SUSY signal. We illustrate this with additional SUSY-motivated scenarios (different decay branching ratios to h and Z), as well as hypothetical nonminimal scenarios involving non-SM resonances.

Notably, all of these scenarios have p_T^{miss}, and in fact the p_T^{miss} is essential to suppress the resonant backgrounds from SM $Z/W +$ jets with hadronically decaying Z/W. This leads to a novel combination of a resonant and nonresonant anomaly detection strategy.[1] This is also the first application of model-agnostic strategies to the SUSY domain and opens up the potential for many more new avenues in the search for SUSY and other well-motivated top-down scenarios. Our method should be contrasted with existing machine learning-based approaches to SUSY in the literature which are fully supervised (see, e.g., [25–29]).

The outline of our paper is as follows: Section II describes how the signal and background processes are simulated. In Sec. III we summarize the steps involved in classifying anomalies through outer density estimation (CATHODE). We show the results of applying CATHODE to different signal processes in Sec. IV. We conclude in Sec. V. Finally, in two Appendices we describe our recasting of the LHC analysis, and the CATHODE receiver operating characteristic (ROC) curves for various signal models.

II. DATA

Since all the methods described here (both the CMS search and CATHODE) fully rely on data for estimating backgrounds (aka are "fully data driven"), the simulation data we generate here is meant to play the role of real data, and all background estimates and significances etc. we derive are meant to illustrate the result one would get applying these methods to collider data. There will be no events generated here that play the role of simulations at the LHC.

For Standard Model (SM) background data, we take into account the three largest contributions of background events to the CMS search, arising from $Z +$ jets, $W +$ jets, and $t\bar{t}+$ jets. W and Z events were generated with one to four additional final state partons while $t\bar{t}$ were generated with up to three additional partons.

For the benchmark signal (to be used to compare the performance of the CMS search vs the CATHODE method), we follow the CMS search and generate gluino pair production (with zero to two additional partons), with subsequent cascade decay $pp \to \tilde{g}\tilde{g}, \tilde{g} \to q\bar{q}\tilde{\chi}_2^0, \tilde{\chi}_2^0 \to Z\tilde{\chi}_1^0$

---

[1]See [24] for a different, fully nonresonant application of weakly supervised anomaly detection to the jet constituents of the monojet $+ p_T^{\mathrm{miss}}$ final state. Motivated by (nonresonant) dark showers, they did not obtain their background templates from sidebands in the jet mass; instead they considered an idealized (perfect) background template from simulated $Z(\nu\nu) +$ jets events.

where the neutralino $\tilde{\chi}_2^0$ is the next to lightest supersymmetric particle (NLSP) and $\tilde{\chi}_1^0$ is the lightest supersymmetric particle (LSP). The mass splitting between the gluinos and NLSP is set to 50 GeV while the LSP mass is 1 GeV. This results in soft jets from the first step of the decay and a highly boosted Z boson. The LSP escapes the detector and contributes large amounts of missing energy.

Later we will also consider decays of $\tilde{\chi}_2^0$ to $X\tilde{\chi}_1^0$ where the X is either a Standard Model Higgs boson or a new Higgs boson with mass besides 125 GeV like the new Higgs bosons in supersymmetric extensions of the Standard Model. The Standard Model Higgs boson decays in $\sim 58\%$ of cases to $b\bar{b}$ while for the latter case we set the branching ratio to 100%.

All events are generated with MadGraph5_AMC@NLO 3.2.0 with $\sqrt{s} = 13$ TeV. The NNPDF3.1LO PDF set [30] is used throughout. At the generator level a minimum H_T cut of 250 GeV is imposed. Gluinos are decayed spin uncorrelated with MADSPIN [31] to $q\bar{q}\tilde{\chi}_2^0$ via an off-shell squark and subsequently $\tilde{\chi}_2^0 \to X\tilde{\chi}_1^0$. Showering is done using PYTHIA 8.306 [32] with MLM merging. PYTHIA-Tune CP5 was used for background events while CP2 [33] was used for the signal samples. The number of background events in each channel is scaled to match their respective next-to-leading-order cross sections [34]. Detector effects are simulated using DELPHES 3.5.0 [35] with the delphes_card_CMS.tcl detector card modified to account for the lepton isolation criterion. Particles are clustered into jets using the anti-k_T clustering algorithm with cone-radius parameter $R = 0.4$ for AK4 jets and $R = 0.8$ for AK8 jets. To be considered, jets have to have $p_T > 30$ and $|\eta| < 2.4$.

The following selection criteria are imposed for both the classical CMS-recast and the dataset for CATHODE:

(1) $N_{\text{AK4 jet}} \geq 2$
(2) $p_T^{\text{miss}} > 300$ GeV
(3) $H_T > 400$ GeV, where $H_T = \sum_{\text{AK4 jets}} |\vec{p}_T|$
(4) $|\Delta\phi_j, \vec{H}_T^{\text{miss}}| > 0.5(0.3)$ for the first two (up to next two) AK4 jets, where $\vec{H}_T^{\text{miss}} = -\sum_{\text{AK4 jets}} \vec{p}_T$
(5) no isolated photon, electron, or muon candidate with $p_T > 10$ GeV with isolation variables $I < 0.1, 0.2$, and 1.3 GeV$/p_T + 0.005$ for isolated electron, muon, and photon, respectively
(6) no isolated track with $m_T = \sqrt{2p_T^{\text{track}} p_T^{\text{miss}} (1 - \cos(\phi^{\text{miss}} - \phi^{\text{track}})} < 100$ GeV and $p_T > 5$ GeV for tracks identified as an electron/muon or else 10 GeV
(7) at least two AK8 jets with $p_T > 200$ GeV.

The number of background events that pass this baseline selection is shown in the first line of Table I. In total, the dataset is composed of 107,421 background events corresponding to $\mathcal{L}_{\text{int}} = 300$ fb^{-1} after cuts 1–7. Signal events are injected according to the gluino-pair production cross section.

TABLE I. Number of events passing each selection requirement for $\mathcal{L}_{\text{int}} = 300$ fb^{-1}.

Selection	W	Z	$t\bar{t}$
Baseline selection	73790	25725	7906
$m_{J_1} \in [70$ GeV, 100 GeV$]$	5936	2401	1320
CMS-SUS-19-013 [23] signal region	420	237	153

Figure 2 shows that the feature m_{J_1} is smooth for the background while it is resonant for the signal. (Hadronically decaying Ws and Zs are eliminated by the requirements on p_T^{miss}.) This is a necessary feature for the application of the CATHODE method employed in Sec. III. Figure 3 shows that the signal of new physics is found on the tail of the p_T^{miss} distribution, while the background peaks at lower p_T^{miss}. We will show that the powerful discriminator p_T^{miss} can be leveraged by CATHODE even though the signal is found on the tail of the distribution.

III. CATHODE

Here we recap the main points of the inner workings of classifying anomalies through outer density estimation (CATHODE) (for more details, see [14]). In very broad strokes, CATHODE aims to learn the density of background events in a signal-depleted region and estimates the density inside the signal enriched region by interpolation. Then, artificial samples are generated in that region, which should follow a signal-depleted distribution. Using a classifier, which is trained to distinguish between the artificial and real events, we can approximate the likelihood ratio (1). This would be the ideal (optimal) model-agnostic anomaly detector, as it is monotonic with $p_{\text{signal}}(x)/p_{\text{bg}}(x)$ for any signal [since $p_{\text{data}}(x)$ is an admixture of $p_{\text{signal}}(x)$ and $p_{\text{bg}}(x)$] [36]. This allows CATHODE to classify data events as background-like or signal-like. The whole method works
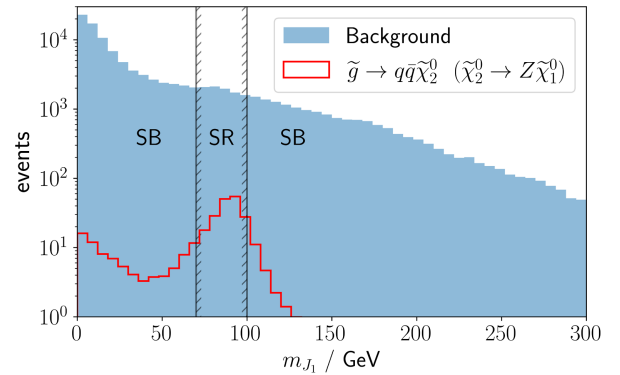

FIG. 2. Distribution of the resonant feature m_{J_1} for background and signal events in the sideband (SB) and the signal region (SR). The signal corresponds to $m_{\tilde{g}} = 1700$ GeV. The distributions are scaled to $\mathcal{L}_{\text{int}} = 300$ fb^{-1}.
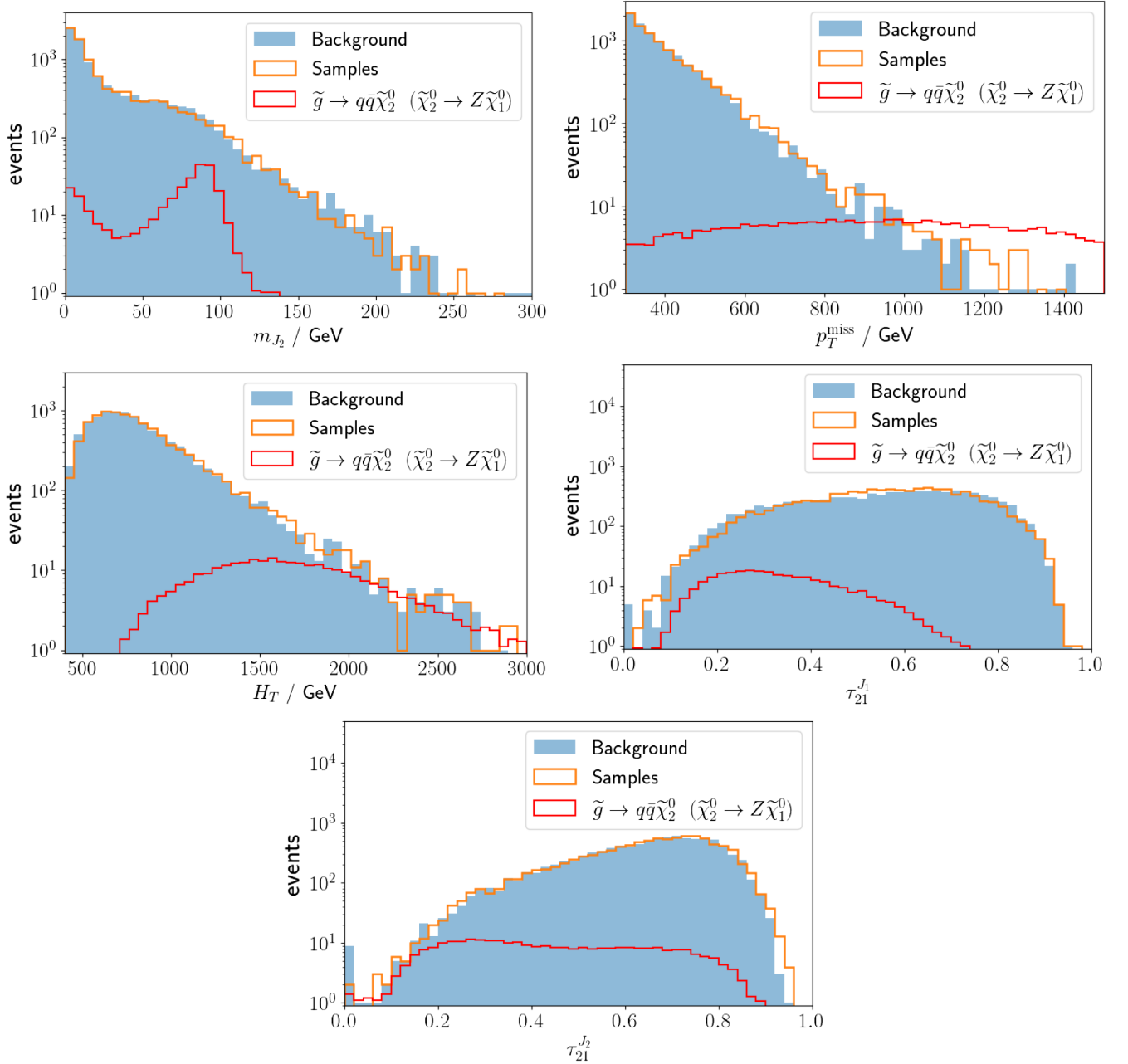
FIG. 3. Comparison of the signal and background distribution inside the signal region and the artificial samples. The artificial samples will be discussed in the next section. The signal corresponds to $m_{\tilde{g}} = 1700$ GeV. The distributions are scaled to $\mathcal{L}_{\text{int}} = 300$ fb^{-1}.

by learning directly from data. The training and model selection of both the density estimation and classification is completely agnostic of any signal truth label.

In this study, the events are represented as the tuple m_{J_1} and x with

$$x = \left(m_{J_2}, p_T^{\text{miss}}, H_T, \tau_{21}^{J_1}, \tau_{21}^{J_2} \right), \tag{2}$$

where J_1, J_2 are the leading/subleading AK8 jets and $\tau_{21} = \tau_2/\tau_1$ is the ratio of n-subjettiness variables [37]. To

compare the technique to the classical search more directly, we also consider the reduced set of features

$$x = \left(m_{J_2}, p_T^{\text{miss}}, H_T \right) \tag{3}$$

so that CATHODE only gets to use the same information. We use a slightly modified version of the original repository[2] to allow for any dimension for x.

---

[2]https://github.com/HEPML-AnomalyDetection/CATHODE.

A. Data preparation and density estimation

First, one defines the signal region (SR) as an interval in m_{J_1} where the signal is expected to be concentrated similar to a classical bump hunt. The complement of the SR defines the sideband (SB). As in any bump hunt, the SR window has to account for the position and the width of the signal bump. Because the reconstructed jet mass is not distributed symmetrically around the mass m of the mother particle (which is the Z, the Higgs, or a BSM Higgs in this paper), we chose parametrization

$$m_{J_1} \in \left[m \left(1 - \frac{4}{3} \sigma_m \right), m \left(1 + \frac{2}{3} \sigma_m \right) \right]. \qquad (4)$$

We estimate the mass resolution to $\sigma_{\mathrm{m}} = 15\%$ and round the window to the closest GeV. The lower sideband extends to $m_{J_1} = 0$ while the upper sideband is only bound by the phase space.

Events in the SB are partitioned into a training set (75%) used for the actual training and in a validation set (25%), used to select the models used in the next steps. To address the finite number of real SB events we use leave-one-out cross validation such that we get four datasets with non-overlapping validation sets. The data is transformed (pre-processed) for easier learning by shifting and scaling the observables in x to fit the interval $(0,1)$, then applying a logit tranformation,[3] and again shifting and scaling to unit standard deviation and zero mean.

For density estimation, a masked autoregressive flow (MAF) is used with affine transformations [38]. The MAF constructs invertible transformations with tractable Jacobians that map a simple multidimensional distribution (e.g., multiple Gaussians as is considered here) to the target density, in this case the conditional probability $p_{\mathrm{data}}(x|m_{J_1} \in \mathrm{SB})$. The MAF uses 15 blocks of masked autoencoder for distribution estimation (MADE) [39] to learn the transformations. The number of events it is trained on depends on the signal region but is typically of the order of 10^5.[4] Training is done with the hyperparameters listed in Table II.

After training the ten epochs with the lowest validation loss are selected for the sampling step.

B. Sampling SR events

The next step aims to sample synthetic events inside the SR using the four density estimates of the last step. Kernel density estimation with Gaussian kernel and bandwidth of 0.01 is used to model the m_{J_1} distribution inside the SR. This is then used to sample $N = 1,000$ events from each of the ten density estimator models which are combined,

---

[3]$\mathrm{logit}(x) = \ln \frac{x}{1-x}$.
[4]We emphasize that the number of events we are using for training was carefully tuned to match the actual number of events in data expected in $\mathcal{L} = 300/\mathrm{fb}$.

TABLE II. Parameters of the density estimator.

Hyperparameter	Value
optimizer	Adam
epochs	100
learning_rate	10^{-4}
batch_norm	true
batch_norm_momentum	1
batch_size	256

shuffled, and split between the training set (60%) and validation set (40%) for the next step. The training and validation sets of all four density estimators are combined respectively to form the synthetic dataset with a total of 40,000 events. Compared to the roughly 10,000 real events in the SR (see Table I second line) this is intentionally over-sampled to improve the classification performance [14]. Setting N even higher did not improve results systematically. The synthetic background events and the real SR events are then standardized in the SR without the logit transformation.

The distributions of the synthetic events are shown in orange in Fig. 3. In all of our models the signal is located in a resonance in m_{J_2} and in the tail of the p_T^{miss} distribution. The density estimation has to model the shape reasonably well so this powerful classification feature can be leveraged. This is accomplished successfully as shown in Fig. 3.

C. Classifier and anomaly detection

Now a classifier is trained on both the synthetic and real SR dataset to distinguish the sampled events, which should follow the background distribution, from the real events, which additionally might contain events following the signal distribution.

The classifier consists of three hidden layers with 64 nodes and ReLU activation each and it is optimized using the hyperparameters given in Table III. Because the datasets are imbalanced, a weight is assigned such that both classes contribute equally to the loss.

Since in a realistic example the number of events to train and validate on is limited, we employ an additional step of leave-one-out cross validation. The real SR data is partitioned into four subsets of equal size. In each subset, one quarter of the real events are held back as a test set for the anomaly detection while the remaining 75% are split

TABLE III. Parameters of the classifier.

Hyperparameter	Value
optimizer	Adam
epochs	100
learning_rate	10^{-3}
batch_size	128

between the training set (60%) and the validation set (40%). (The synthetic background events are also split into train/val sets with the same proportions.) After training, the ten model states with the lowest validation loss are selected and evaluated on the test set. The predicted labels are then averaged over the models and assigned as anomaly scores to the events. This is repeated for the next quarter of the SR data, and so on, until every event in the SR is assigned an anomaly score.

To reduce the statistical effects of severely overperforming and underperforming models, each dataset is shuffled 5 times to allow different selections. Then the entire process of the preceding paragraph is repeated to produce five different anomaly scores. All five anomaly score assignments are averaged to produce a final, more robust score.

Finally, to even out the influence of signal-event selection, everything is repeated 10 times with differing independent sets of signal events. In all the results we report below, we will report the mean and standard deviation of these ten different trials.

The signal to background ratio is improved by cutting on the anomaly score above a critical value R_c. Figure 4 shows the distributions of the anomaly score R for the signal and background. No additional selections are performed. In a real application, one would perform statistical inference by means of a bump hunt on the R distribution which is beyond the scope of this work. Instead the performance is evaluated using the nominal significance $Z = S/\sqrt{B}$ with S (B) the number of signal (background) events after imposing this cut. This makes use of the truth labels which an experiment would have to replace by other means of background estimation. One still has to choose a strategy to set R_c. In the following we will show the signal significance with R_c set to maximize Z with at least five background events left to show the best performance one could hope for. Since a real application does not have access to the truth labels, this is not immediately applicable. To show a more realistic method we also show the performance where R_c is
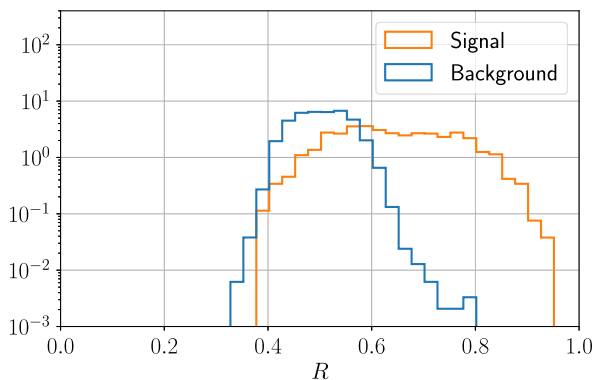
set so that 1% of SR events pass the cut while also containing at least five background events.

IV. RESULTS

A. Nominal signal model

We first turn our attention to the nominal signal model where $\tilde{\chi}_2^0 \to Z\tilde{\chi}_1^0$. This is the signal model the dedicated CMS search [23] was aimed at.

1. Three features

We start by using the limited feature set $x = (m_{J_2}, p_T^{\mathrm{miss}}, H_T)$ so CATHODE does not have access to more information than the classical search. To compare with CMS, we calculate the signal significance for events inside the signal region $m_{J_1/J_2} \in [70 \text{ GeV}, 100 \text{ GeV}]$ with the b-veto mentioned in Sec. A 1 applied. Since the search gets most of its sensitivity from the highest p_T^{miss} bins, we apply an additional cut $p_T^{\mathrm{miss}} > 800$ GeV.[5] This leads to roughly the same number of events as when only the top 1% of events are kept for CATHODE. For a gluino mass with sizable cross section like 1700 GeV, the classical search yields on average for ten independent signal injections $Z = 20$. Using CATHODE with three features the significance is on average $Z = 34 \pm 2$.

Evidently, CATHODE outperforms the classical approach, even though CATHODE is more model agnostic. The reason is that the classical approach, being cut based, misses correlations between the features that the multivariate classifier of CATHODE can pick up.

To confirm this, we also investigated the sensitivity of a fully supervised approach, using the same classifier architecture and hyperparameters as that of CATHODE. The training data for the fully supervised classifier consists of an additional 300 fb^{-1} background events and 10,000 signal events. 60% of this dataset is used in training while the remaining 40% is used as a validation set to select the best performing model. Evaluating this classifier again with selecting only the top 1% of anomaly scores results in a significance of on average $Z = 33 \pm 4$. We conclude that CATHODE is saturating the performance of the fully supervised classifier for this amount of signal (unsurprisingly, since this is a lot of signal), and that the deep neural networks of both CATHODE's classifier and the supervised classifier can leverage correlations to improve the signal significance significantly over the classical approach.

FIG. 4. Normalized distributions of the anomaly score R of the signal and background processes. The signal corresponds to the average distribution of ten independent injections with $m_{\tilde{g}} = 1900$ GeV.

---

[5]Technically, the original CMS search uses p_T^{miss} bins, and most of the sensitivity comes from the three highest bins, 800–1000 GeV, 1000–1200 GeV, and larger than 1200 GeV, where the background is comparable or subdominant to the signal hypothesis. To get a fair comparison with CATHODE we replace this with a single cut.

$$\widetilde{g} \to q\bar{q}\widetilde{\chi}_2^0 \quad (\widetilde{\chi}_2^0 \to Z\widetilde{\chi}_1^0)$$

$$\widetilde{g} \to q\bar{q}\widetilde{\chi}_2^0 \quad (\widetilde{\chi}_2^0 \to h\widetilde{\chi}_1^0)$$
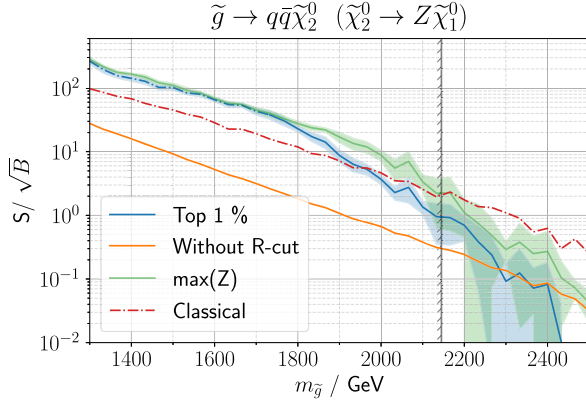
FIG. 5. Sensitivity of CATHODE and the classical strategy. The signal window is set as $m_{J_1} \in [70 \text{ GeV}, 100 \text{ GeV}]$. For the blue line R_c is set to allow 1% of events to pass this cut while the orange line omits the cut completely. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections. The dot-dashed part of the blue line represents parameter points where R_c has to be lowered to allow five background events. The vertical black line at 2145 GeV indicates gluino mass that is excluded at 95% confidence level by our 300/fb recreation of the dedicated search [23]. The red dot-dashed line is calculated using the classical strategy with $m_{J_1/J_2} \in [70 \text{ GeV}, 100 \text{ GeV}]$, $p_T^{\text{miss}} > 800$ GeV and the b-veto.

FIG. 6. CATHODE's performance for $\tilde{\chi}_2^0 \to h\tilde{\chi}_1^0$. The signal window is set as $m_{J_1} \in [100 \text{ GeV}, 140 \text{ GeV}]$. For the blue line R_c is set to allow 1% of events to pass this cut while the orange line omits the cut completely. The dot-dashed part of the blue line represents parameter points where R_c has to be lowered to allow 5 background events. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from 10 different signal injections. The vertical black line at 2355 GeV indicates gluino-mass that is expected to be excluded by rescaling the (expected) limit from a dedicated CMS search for this decay [41] from 137/fb to 300/fb integrated luminosity. There is no red line corresponding to the classical search (as in Fig. 5) because we did not perform a detailed recast of [41].

2. Five features

From now on, we will use the five features $(m_{J_2}, p_T^{\text{miss}}, H_T, \tau_{21}^{J_1}, \tau_{21}^{J_2})$ because the subjettiness variables τ_{21} are useful discriminants. Figure 5 shows CATHODE's performance compared to the classical strategy. We see that in the relevant region at high gluino masses the conservative cut on R (allowing only the top 1% to pass) reaches only slightly weaker results. We identify the mass where the signal significance is $Z = 1.645$ with the expected 95% limit on the mass in a real application [40]. The conservative cut on R alone excludes gluino masses up to $m_{\tilde{g}} = 2066$ GeV. This is only slightly weaker than the expected excluded mass of $m_{\tilde{g}} < 2145$ GeV for a dedicated search at this integrated luminosity. This is expected because a model specific search will be fine-tuned to the specific process while CATHODE is intentionally kept more general. CATHODE's strength lies in this generalization as it is able to detect different models without the need to tweak the approach as we will show in the following sections.

B. Alternate signal model: Decays to SM Higgs

Now we turn our attention to another model, where the neutralinos decay via $\tilde{\chi}_2^0 \to h\tilde{\chi}_1^0$ where h is the 125 GeV Standard Model Higgs boson. All that has to be done for CATHODE is select a new signal window around 125 GeV. A scan over the gluino mass is shown in Fig. 6. A b-jet selection criterion would be beneficial in this case, but we omit this to keep CATHODE as general as possible. Even
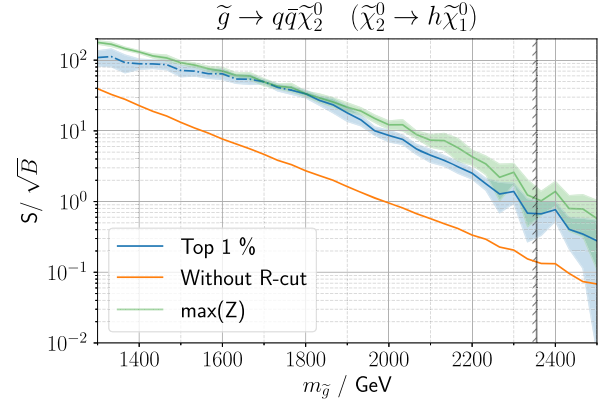
without the b-tag CATHODE still generates a sizable signal significance for gluino masses comparable to the expected excluded value. While the dedicated search is expected to exclude gluino masses below 2355 GeV, CATHODE with the 1% cut reaches $Z > 1.645$ for all masses up to 2233 GeV. With the best possible cut on R this can be pushed to 2300 GeV. As expected CATHODE results in slightly weaker bounds. The opportunity cost of this is significantly lower than a specialized search. The only change in the approach is the choice of the signal region. The intended use of CATHODE scans the signal region over the entire mass range, such that both the decay to Z and Higgs bosons would be included automatically in this strategy without any extra considerations.

C. Alternate signal model: Mixed Z/h decays

Setting the branching ratio of the $\tilde{\chi}_2^0 \to h\tilde{\chi}_1^0$ or $\tilde{\chi}_2^0 \to Z\tilde{\chi}_1^0$ decays to 100% is a rather unnatural choice. Therefore we also show CATHODE's performance for a model where both branching ratios are 50%. This time the anomaly detection has to find two bumps simultaneously. For this we chose the signal window to contain both resonances: $m_{J_1} \in [70 \text{ GeV}, 140 \text{ GeV}]$. The results of a scan over the gluino masses is shown in Fig. 7. This time CATHODE seems to outperform the extrapolated bound from the dedicated search [42]. The extrapolation from 35.9/fb to 300/fb integrated luminosity is quite far and should be taken with a grain of salt. The dedicated search classifies events in 0, 1,
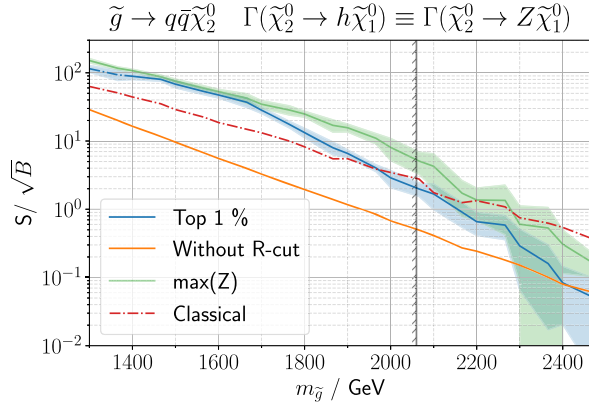
FIG. 7. Sensitivity of CATHODE and the classical strategy. The signal window is set as $m_{J_1} \in [70 \text{ GeV}, 140 \text{ GeV}]$. For the blue line R_c is set to allow 1% of events to pass this cut while the orange line omits the cut completely. The dot-dashed part of the blue line represents parameter points where R_c has to be lowered to allow five background events. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections. The vertical black line at 2060 GeV indicates gluino mass that is expected to be excluded by rescaling the expected excluded cross section obtained by the dedicated CMS search for this decay [42] from 35.9/fb to 300/fb integrated luminosity.
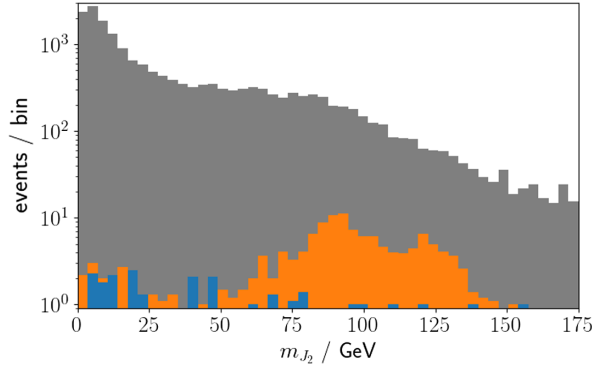
FIG. 8. The distribution of the data inside the signal region before the anomaly score cut is shown in gray. After selecting the top 1% of events in the SR the remaining signal events are shown in orange while the remaining background events are shown in blue. The signal corresponds to $m_{\tilde{g}} = 1700$ GeV.

and 2 Higgs categories using b-tags. The signal model populates all categories simultaneously. The approach using CATHODE only uses the single signal region without further thought to generate these results.

In Fig. 8 we show that CATHODE is indeed capable of recovering both bumps corresponding to the decay into Z and Higgs bosons, respectively.

Figure 9 shows that CATHODE is very robust against changes in branching ratios. We vary the branching ratio $\text{Br}(\tilde{\chi}_2^0 \rightarrow Z\tilde{\chi}_1^0)$ with $\text{Br}(\tilde{\chi}_2^0 \rightarrow h\tilde{\chi}_1^0) = 1 - \text{Br}(\tilde{\chi}_2^0 \rightarrow Z\tilde{\chi}_1^0)$ and calculate the significance. Regardless of branching
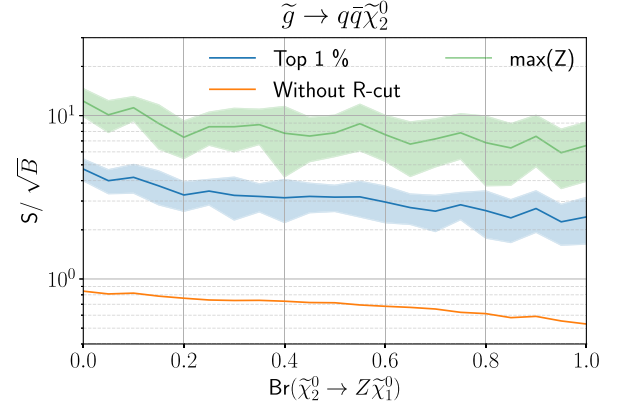

FIG. 9. Sensitivity of CATHODE for varying branching ratios to Z bosons for $m_{\tilde{g}} = 2000$ GeV. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections.

ratio, the multiplicative gain of significance by applying the technique is always between 5 and 6. This shows the real strength of the CATHODE approach over the dedicated searches [23,41,42]. With the enlarged SR that covers both decay modes, CATHODE only needs to be trained once, independent of the assumption on the BRs, compared to performing a dedicated analysis for each BR assumption.

D. Alternate signal model: Decays to BSM Higgs

Until now we applied CATHODE only to models where the position of the bump is known beforehand. But one strength of the technique is that we do not even need to know that. To discuss this further, we now focus on another model that induces the neutralino decay $\tilde{\chi}_2^0 \rightarrow H\tilde{\chi}_1^0$, where H is one of the additional Higgs bosons introduced by the (next to) minimal supersymmetric standard model [(N)MSSM] that has a mass different from 125 GeV. Because the decay of H depends on the specific implementation of SUSY-breaking parameters we set the branching ratio $\text{BR}(H \rightarrow b\bar{b}) = 100\%$. To find the signal, CATHODE is applied to different signal regions given by varying mass hypotheses m in Eq. (4), scanning the entire mass range in discrete steps and the signal significance is determined. To demonstrate this we chose $m_H = 100$ GeV and $m_{\tilde{g}} = 2000$ GeV and show the result in Fig. 10. Once the signal window has significant overlap with the signal bump, the signal significance gets sufficiently improved to show the presence of anomalous events. In a real application, this would then warrant further investigation with a dedicated search.

Finally, we show how wide the possible choice of m_H is that CATHODE can still help to find in our dataset with the given choice of features. For this we perform a parameter scan over m_H from 35 to 515 GeV in 10 GeV steps shown in Fig. 11. The method reaches reliably signal significances
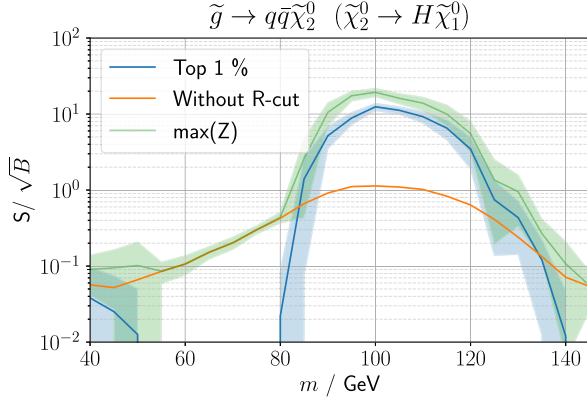
FIG. 10. Significance for a parameter scan over the mass hypothesis in 5 GeV steps, when the mass is not known *a priori*. The shaded region shows one standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections. Masses are chosen as $m_{\tilde{g}} = 2000$ GeV and $m_H = 100$ GeV.
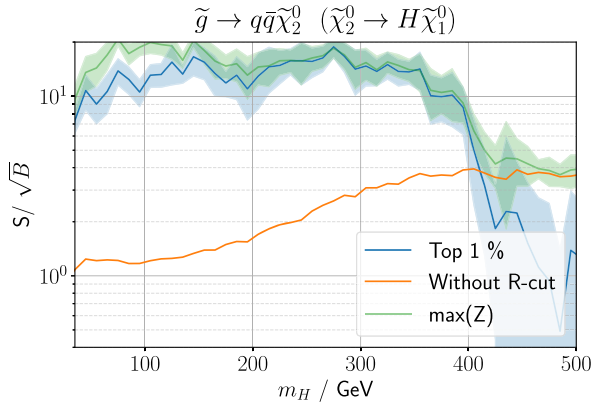

FIG. 11. Parameter scan of m_H with $m_{\tilde{g}} = 2000$ GeV to show which signals CATHODE can help find in the dataset. The shaded region shows 1 standard deviation around the mean $S/\sqrt{B}$ obtained from ten different signal injections.

of order 10 up to $m_H \sim 350$ GeV without using b-tags as otherwise powerful discriminators.

V. CONCLUSIONS

In this paper, we have shown how recently developed techniques for weakly supervised resonant anomaly detection can be easily extended to cover anomalies that also live on tails of distributions. This situation commonly arises in well-motivated weak-scale scenarios such as SUSY, where the cascade decays of heavier BSM particles can produce resonances such as Zs and Higgs bosons, while simultaneously populating the tails of features such as p_T^{miss} and H_T. As long as the signal is localized in one feature where the background is smooth, resonant anomaly detection can be brought to bear on these additional features in order to enhance the sensitivity to signal.

As a proof-of-concept demonstration, we applied the state-of-the-art anomaly detection method CATHODE [14] to the SUSY scenario $pp \to \tilde{g}\tilde{g}, \tilde{g} \to q\bar{q}\tilde{\chi}_2^0, \tilde{\chi}_2^0 \to X\tilde{\chi}_1^0$ where X is either a Z boson, Standard Model Higgs, or an additional (N)MSSM Higgs boson. Despite being model agnostic, we showed that the CATHODE method is competitive with existing, dedicated, cut-based searches [23,41,42], because—being inherently multivariate—it takes advantage of correlations between features. Moreover, whereas each decay scenario required a separate, optimized analysis, CATHODE—being model agnostic—is able to simultaneously target them all.

In this work we considered two different feature sets for the CATHODE algorithm, as shown in Eqs. (2) and (3). These were motivated by the SUSY scenarios we considered, and it would be interesting to generalize our study beyond these feature sets, both to increase the degree of model agnosticness of the method, and possibly to enhance the sensitivity to the SUSY signals considered here. For example, our benchmark signals all come with ∼4 additional jets from the gluino decay, and their detailed kinematic distributions (instead of just the aggregate feature H_T) may offer additional discriminating power versus the QCD background. Adding features related to additional jets in the event may also give us more sensitivity to spectra not explicitly considered here, for example where the NLSP mass is not so close to the gluino mass. As long as $m_{\mathrm{LSP}} + m_Z \ll m_{\tilde{g}}$, the Z will still be boosted, but the extra jets will get harder as m_{LSP} moves away from $m_{\tilde{g}}$.

All in all, using modern methods for resonant anomaly detection such as CATHODE allows for a broader and more efficient coverage of the parameter space of physics beyond the Standard Model. With much more data on the way, methods like these should prove indispensable for maximizing the discovery potential of the LHC.

APPENDIX A: RECASTING CMS

We describe how the background samples were simulated as closely as possible to an existing search and verified. We recreate the CMS SUSY search CMS-SUS-19-013 [23].

1. Recreating CMS-SUS-19-013

The recreation of CMS-SUS-19-013 [23] follows the most important analysis steps of the original publication.
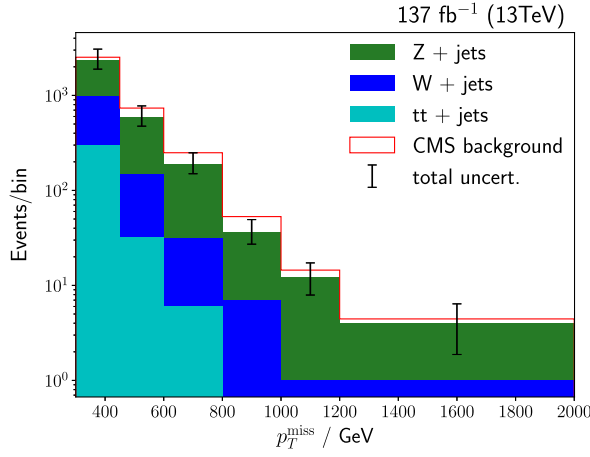
FIG. 12. p_T^{miss} spectrum of the three leading background processes. The background of the same three processes from the CMS publication is shown in red. The variation of cross section due to changing the energy scale by a factor of $1/2$ and 2 as computed by MadGraph is assigned as a systematic uncertainty and added to the statistic errors in quadrature and shown as the error bars.

The number of events is set to the integrated luminosity of $\mathcal{L}_{int} = 137$ fb^{-1}. First, a set of remaining cuts are applied to select Z candidates, then the background estimation is recreated before the statistical analysis is performed. The following cuts are applied to select hadronically decaying Z bosons:

(8) Softdropped $m_{jet} \in [40$ GeV, 140 GeV$]$ of the two highest p_T AK8 jets

(9) $\Delta R_{Z,b} > 0.8$ for the second highest p_T AK8 jet Z and any b-tagged jet where the angular separation is defined as $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$.

The resulting p_T^{miss} spectrum is shown in Fig. 12 which agrees with the spectrum shown in the original publication within uncertainties.

The background estimation consists of the normalization and the shape estimation. The signal region (SR) is defined as $m_{jet} \in [70$ GeV, 100 GeV$]$. First, one demands the subleading AK8-jet to be in the SR. Then a linear function is fitted to the m_{jet} spectrum of the leading AK8 jet outside its SR. The nominal yield $\mathcal{B}_{norm}$ is obtained by integrating the linear function in the SR. The statistical error of the yield is obtained from the spread of pseudo-experiments sampled from the fit. Additionally to the linear function Chebychev functions up to the fourth order are fitted. The largest deviation of the nominal yield is then assigned as an additional uncertainty.

The background p_T^{miss} shape is obtained by the sideband (SB) with both AK8 jets outside the SR. The content of the ith p_T^{miss} bin is denoted as N_i^{SB}. The transfer factor from the SB to the SR is then calculated as

$$\mathcal{T} \equiv \frac{\mathcal{B}_{norm}}{\sum_i N_i^{SB}} = 0.206 \pm 0.023, \qquad (A1)$$

which agrees with the original publication within uncertainties. The expected background in bin i is then

$$\mathcal{B}_i = \mathcal{T} N_i^{SB}. \qquad (A2)$$

ROOSTATS [43] is used for statistical modeling. It takes N_i^{SB} with statistical errors, $\mathcal{T}$ and $\Delta\mathcal{T}$ to model the background in the SR with uncertainties. The signal model contains signal events that pass all cuts and is rescaled to the approximate next-to-next-to leading order + next-to-next-to-leading logarithmic accuracy cross section [44]. The overall uncertainty of the cross section is applied to all signal bins. The resulting statistical model is then evaluated with the CL$_s$ approach and the asymptotic form of the onesided profile likelihood test statistic. This is used to obtain the 95% C.L. cross sections. The limits are shown in Fig. 13 for the integrated luminosity $\mathcal{L}_{int} = 137$ fb^{-1} and in Fig. 14 for $\mathcal{L}_{int} = 300$ fb^{-1}. We use the latter dataset for the application of the machine learning technique since the accuracy is greatly improved with more data points to learn on while in reach for the collider in the near future.
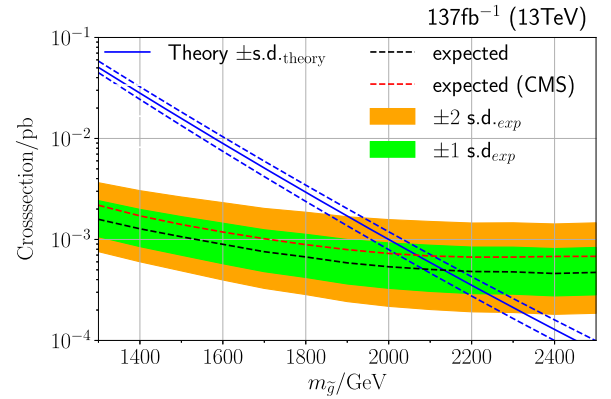

FIG. 13. Recreation of CMS-SUS-19-013 [23]. The red dashed line denotes the expected limits of original CMS search. The black dashed line shows the expected limits of the recreation.
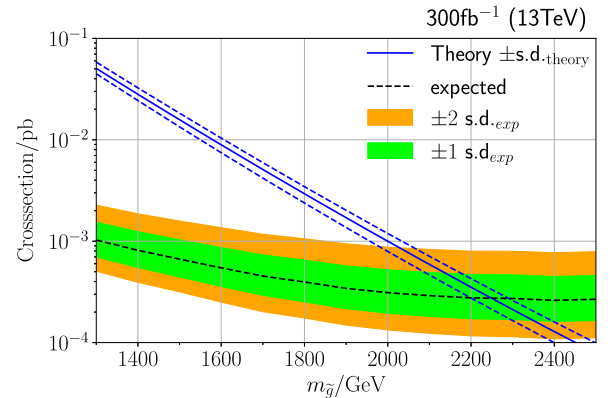

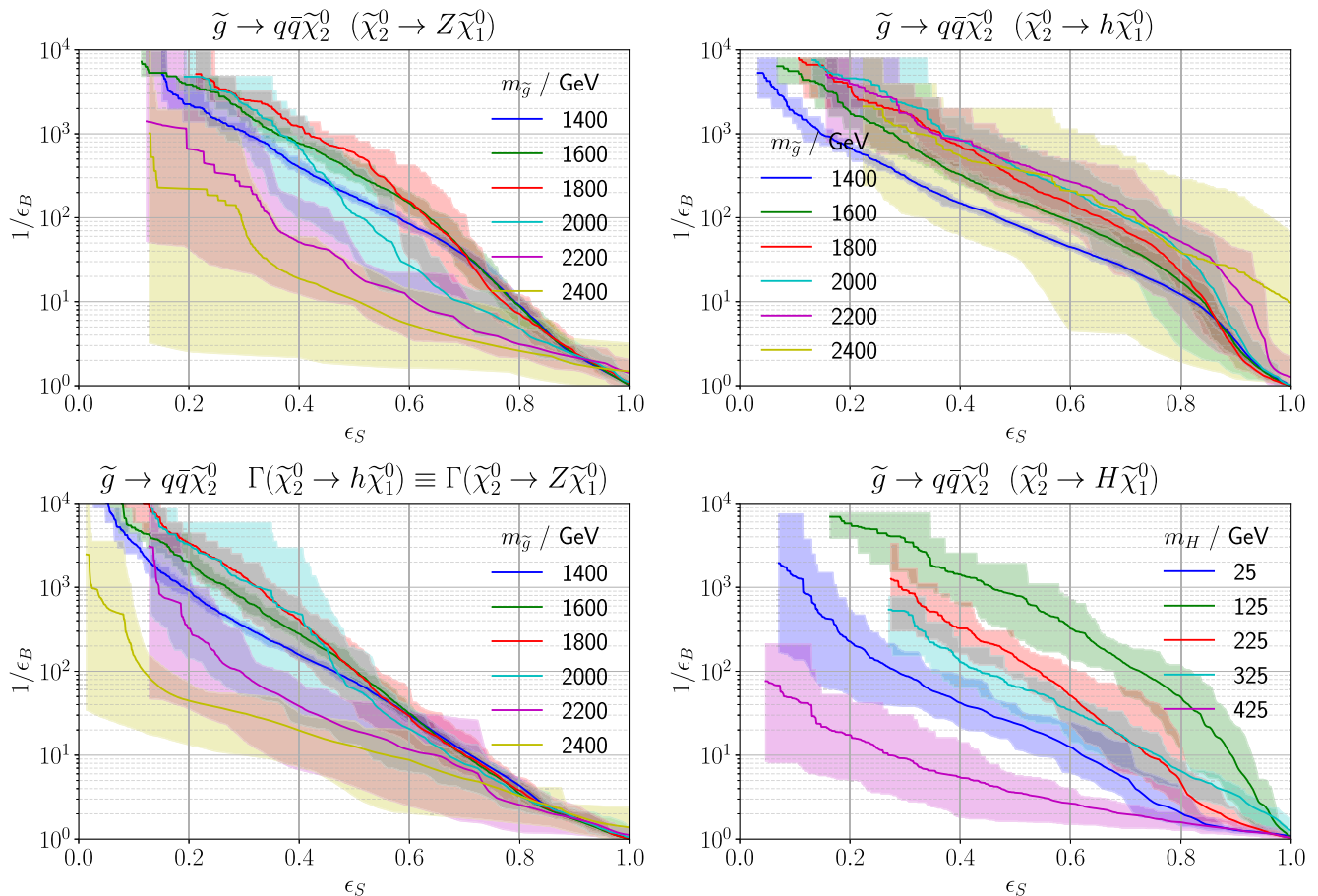FIG. 14. Results of the classical search for 300/fb integrated luminosity.

FIG. 15. ROC curves of all four benchmark models. Solid lines denote the mean value and shaded regions show the span between the minimum and maximum values obtained from ten different signal injections.

APPENDIX B: ROC CURVES

In Fig. 15 we show ROC curves, i.e., background suppression as a function of signal efficiency of our benchmark models. We also observe a common feature of anomaly detection techniques. With rising signal cross section the classifier learns to better separate background from signal-like events. At the same time, larger signal cross sections correspond to smaller gluino masses, which in turn leads to less expressive features. Both effects combined lead to intermediate gluino masses having the largest background suppression at the same signal efficiency compared to small masses with large cross sections or large masses with very obvious signatures, especially in the decay to Z and Standard Model Higgs bosons. We also observe in the bottom right figure that for low and high Higgs masses the background rejection is noticeably weaker than for intermediate masses. For light Higgs masses, the jets are too similar to background jets while high Higgs masses lead to wide jets that get reconstructed incorrectly.

[1] Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih, Machine learning in the search for new fundamental physics, arXiv:2112.03769.

[2] Gregor Kasieczka *et al.*, The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics, Rep. Prog. Phys. **84,** 124201 (2021).

[3] Thea Aarrestad *et al.*, The dark machines anomaly score challenge: Benchmark data and model independent event classification for the Large Hadron Collider, SciPost Phys. **12,** 043 (2022).

[4] Jack H. Collins, Kiel Howe, and Benjamin Nachman, Anomaly detection for resonant new physics with machine learning, Phys. Rev. Lett. **121,** 241803 (2018).

[5] Jack H. Collins, Kiel Howe, and Benjamin Nachman, Extending the search for new resonances with machine learning, Phys. Rev. D **99**, 014038 (2019).

[6] Benjamin Nachman and David Shih, Anomaly detection with density estimation, Phys. Rev. D **101**, 075042 (2020).

[7] Anders Andreassen, Benjamin Nachman, and David Shih, Simulation assisted likelihood-free anomaly detection, Phys. Rev. D **101**, 095004 (2020).

[8] Oz Amram and Cristina Mantilla Suarez, Tag N' Train: A technique to train improved classifiers on unlabeled data, J. High Energy Phys. 01 (2021) 153.

[9] Georges Aad *et al.*, Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector, Phys. Rev. Lett. **125**, 131801 (2020).

[10] Kees Benkendorfer, Luc Le Pottier, and Benjamin Nachman, Simulation-assisted decorrelation for resonant anomaly detection, Phys. Rev. D **104**, 035003 (2021).

[11] George Stein, Uros Seljak, and Biwei Dai, Unsupervised in-distribution anomaly detection of new physics through conditional density estimation, in *Proceedings of the 34th Conference on Neural Information Processing Systems* (2020), arXiv:2012.11638.

[12] Sang Eon Park, Dylan Rankin, Silviu-Marian Udrescu, Mikael Yunus, and Philip Harris, Quasi anomalous knowledge: Searching for new physics with embedded knowledge, J. High Energy Phys. 21 (2020) 030.

[13] Jack H. Collins, Pablo Martín-Ramiro, Benjamin Nachman, and David Shih, Comparing weak and unsupervised methods for resonant anomaly detection, Eur. Phys. J. C **81**, 617 (2021).

[14] Anna Hallin, Joshua Isaacson, Gregor Kasieczka, Claudius Krause, Benjamin Nachman, Tobias Quadfasel, Matthias Schlaffer, David Shih, and Manuel Sommerhalder, Classifying anomalies through outer density estimation, Phys. Rev. D **106**, 055006 (2022).

[15] Jernej F. Kamenik and Manuel Szewc, Null hypothesis test for anomaly detection, Phys. Lett. B **840**, 137836 (2023).

[16] John Andrew Raine, Samuel Klein, Debajyoti Sengupta, and Tobias Golling, CURTAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals, Front. Big Data **6**, 899345 (2023).

[17] Gregor Kasieczka, Radha Mastandrea, Vinicius Mikuni, Benjamin Nachman, Mariel Pettee, and David Shih, Anomaly detection under coordinate transformations, Phys. Rev. D **107**, 015009 (2023).

[18] Anna Hallin, Gregor Kasieczka, Tobias Quadfasel, David Shih, and Manuel Sommerhalder, Resonant anomaly detection without background sculpting, Phys. Rev. D **107**, 114012 (2023).

[19] Mayee F. Chen, Benjamin Nachman, and Frederic Sala, Resonant anomaly detection with multiple reference datasets, J. High Energy Phys. 07 (2023) 188,

[20] Tobias Golling, Samuel Klein, Radha Mastandrea, and Benjamin Nachman, Flow-enhanced transportation for anomaly detection, Phys. Rev. D **107**, 096025 (2023).

[21] Tobias Golling, Gregor Kasieczka, Claudius Krause, Radha Mastandrea, Benjamin Nachman, John Andrew Raine, Debajyoti Sengupta, David Shih, and Manuel Sommerhalder, The interplay of machine learning–based resonant anomaly detection methods, Eur. Phys. J. C **84**, 241 (2024).

[22] Debajyoti Sengupta, Samuel Klein, John Andrew Raine, and Tobias Golling, CURTAINs flows for flows: Constructing unobserved regions with maximum likelihood estimation, arXiv:2305.04646.

[23] Albert M Sirunyan *et al.*, Search for supersymmetry in proton-proton collisions at $\sqrt{s} = 13$ TeV in events with high-momentum Z bosons and missing transverse momentum, J. High Energy Phys. 09 (2020) 149.

[24] Thorben Finke, Michael Krämer, Maximilian Lipp, and Alexander Mück, Boosting mono-jet searches with model-agnostic machine learning, J. High Energy Phys. 08 (2022) 015.

[25] Anna Mullin, Stuart Nicholls, Holly Pacey, Michael Parker, Martin White, and Sarah Williams, Does SUSY have friends? A new approach for LHC event analysis, J. High Energy Phys. 02 (2021) 160.

[26] CMS Collaboration, Search for supersymmetry in final states with disappearing tracks in proton-proton collisions at 13 TeV, Technical Report No. CMS-PAS-SUS-21-006, CERN, Geneva, 2023, http://cds.cern.ch/record/2859611.

[27] Georges Aad *et al.*, Search for supersymmetry in final states with missing transverse momentum and three or more b-jets in 139 fb^{-1} of proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Eur. Phys. J. C **83**, 561 (2023).

[28] Armen Tumasyan *et al.*, Search for top squarks in the four-body decay mode with single lepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 06 (2023) 060.

[29] Arghya Choudhury, Arpita Mondal, Subhadeep Mondal, and Subhadeep Sarkar, Improving sensitivity of trilinear RPV SUSY searches using machine learning at the LHC, Phys. Rev. D **109**, 035001 (2024).

[30] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Patrick Groth-Merrild, Alberto Guffanti, Nathan P. Hartland, Zahari Kassabov, José I. Latorre, Emanuele R. Nocera, Juan Rojo, Luca Rottoli, Emma Slade, and Maria Ubiali, Parton distributions from high-precision collider data, Eur. Phys. J. C **77**, 663 (2017).

[31] Pierre Artoisenet, Rikkert Frederix, Olivier Mattelaer, and Robbert Rietkerk, Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations, J. High Energy Phys. 03 (2013) 015.

[32] Christian Bierlich, Smita Chakraborty, Nishita Desai, Leif Gellersen, Ilkka Helenius, Philip Ilten, Leif Lönnblad, Stephen Mrenna, Stefan Prestel, Christian T. Preuss, Torbjörn Sjöstrand, Peter Skands, Marius Utheim, and Rob Verheyen, A comprehensive guide to the physics and usage of PYTHIA 8.3, SciPost Phys. Codebases **2022**, 8 (2022).

[33] Albert M Sirunyan *et al.*, Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements, Eur. Phys. J. C **80**, 4 (2020).

[34] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.

[35] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, DELPHES 3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.

[36] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.

[37] Jesse Thaler and Ken Van Tilburg, Identifying boosted objects with N-subjettiness, J. High Energy Phys. 03 (2011) 015.

[38] George Papamakarios, Theo Pavlakou, and Iain Murray, Masked autoregressive flow for density estimation, arXiv:1705.07057.

[39] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle, Made: Masked autoencoder for distribution estimation, arXiv:1502.03509.

[40] Prudhvi N. Bhattiprolu, Stephen P. Martin, and James D. Wells, Criteria for projected discovery and exclusion sensitivities of counting experiments, Eur. Phys. J. C **81,** 123 (2021).

[41] Armen Tumasyan *et al.*, Search for Higgsinos decaying to two Higgs bosons and missing transverse momentum in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 05 (2022) 014.

[42] Albert M Sirunyan *et al.*, Search for physics beyond the standard model in events with high-momentum Higgs bosons and missing transverse momentum in proton-proton collisions at 13 TeV, Phys. Rev. Lett. **120,** 241801 (2018).

[43] Rene Brun *et al.*, root-project/root: v6.18/02, 10.5281/zenodo.3895860 (2019).

[44] Christoph Borschensky, Michael Krämer, Anna Kulesza, Michelangelo Mangano, Sanjay Padhi, Tilman Plehn, and Xavier Portell, Squark and gluino production cross sections in pp collisions at $\sqrt{s} = 13$, 14, 33 and 100 TeV, Eur. Phys. J. C **74,** 3174 (2014).