

Weakly supervised anomaly detection with event-level variables

Liam Brennan^{1,*}, Tamas Almos Vami^{1,†}, Oz Amram^{2,‡}, Sanjana Sekhar³, Yuta Takahashi⁴, Louis Moureaux⁵,
Manuel Sommerhalder⁵, Petar Maksimovic³, Tianji Cai^{6,7} and Nathaniel Craig^{1,8}

¹*University of California Santa Barbara, Santa Barbara, California, USA*

²*Fermilab, Batavia, Illinois, USA*

³*Johns Hopkins University, Baltimore, Maryland, USA*

⁴*University of Florida, Gainesville, Florida, USA*

⁵*University of Hamburg, Hamburg, Germany*

⁶*School of Physical Science and Engineering, Tongji University, Shanghai, China*

⁷*State Key Laboratory of Autonomous Intelligent Unmanned Systems, MOE Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai, China*

⁸*Kavli Institute for Theoretical Physics, Santa Barbara, California, USA*



(Received 17 April 2025; accepted 25 August 2025; published 26 September 2025)

We introduce a new topology for weakly supervised anomaly detection searches, diobject plus X. In this topology, one looks for a resonance decaying to two standard model particles produced in association with other anomalous event activity (X). This additional activity is used for classification. We demonstrate how anomaly detection techniques which have been developed for dijet searches focusing on jet substructure anomalies can be applied to event-level anomaly detection in this topology. To robustly capture event-level features of multiparticle kinematics, we employ new physically motivated variables derived from the geometric structure of a collision's phase space manifold. As a proof of concept, we explore the application of this approach to several benchmark signals in the di- τ and di- μ plus X final states. We demonstrate that our anomaly detection approach can reach discovery-level significances for signals that would be missed in a conventional bump-hunt approach.

DOI: [10.1103/rk29-518p](https://doi.org/10.1103/rk29-518p)

I. INTRODUCTION

Despite many searches for new particles at the Large Hadron Collider (LHC), no clear evidence of beyond standard model (BSM) physics has been found. However, the space of possible signals which could be hiding in LHC data far exceeds the coverage of the dedicated searches performed thus far. In recent years, anomaly detection has been proposed as a new paradigm for conducting searches at the LHC [1–3]. Rather than designing a search to target a specific signal model, anomaly detection analyses take a model-agnostic approach. They use novel machine-learning-based strategies to identify events which differ from typical backgrounds without reference to a specific signal. Both the ATLAS and CMS collaborations have now released searches involving these new anomaly detection approaches [4–8] and

additional studies have been performed using CMS open data [9,10].

The majority of proposed anomaly detection algorithms can generally be split into two separate categories: unsupervised and weakly supervised algorithms. Unsupervised algorithms, first proposed in Refs. [11,12], are trained solely on background events and are then used to define an anomaly score which quantifies how similar an input event is to typical background events. Weakly supervised methods build on the classification without labels paradigm [13] to learn to distinguish between signal and background. This is accomplished by training a classifier to distinguish events in the signal region, which may contain some small fraction of signal, from a carefully constructed sample of background-only events [14,15]. Weakly supervised methods have the advantage that they are asymptotically optimal: the optimal weakly supervised classifier will learn the likelihood ratio, which is the ideal signal versus background classifier [13].

The training procedure of weakly supervised anomaly detection methods requires a high quality estimate of the background in the high-dimensional feature space, allowing the classifier to identify the anomalous data events in the signal region which are distinct from background. Defects in the background estimate will cause the classifier to learn the

* Contact author: Liam.Robert.Brennan@cern.ch

† Contact author: Tamas.Almos.Vami@cern.ch

‡ Contact author: Oz.Amram@cern.ch

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

difference between the true background events in the signal region and the estimate, rather than identifying the signal. This requirement means that simulations, which are known to have minor mismodelings, are generally not suitable for such a background estimate and data-driven methods must be used. Most weakly supervised anomaly detection methods have therefore focused on resonant signals, so the background composition in a signal mass window can be learned through an interpolation from sideband regions. Several approaches have been proposed to achieve the necessary data-driven high-quality background estimate [14–23]. Many of these first anomaly detection approaches have focused on heavy resonances decaying to two boosted large radius jets, where well-studied jet substructure observables, with minimal correlations with the resonance mass, can be used as the classification features. There is also a large space of signal models which could produce many phenomenological varieties of jets, making this topology ideal for anomaly detection. The weakly supervised searches from at the LHC [4,7,8] have also targeted this dijet topology. Now that these first searches have been performed, it is timely to explore other signal topologies which are amenable to a weakly supervised resonant anomaly detection approach.

In this paper, we explore the application of weakly supervised anomaly detection in a new diobject + X topology using the CATHODE methodology [18] with event-level features. The CATHODE method follows the general strategy described earlier and leverages a generative model trained on sideband data to characterize background distributions, subsequently interpolating these learned representations into the signal region to produce realistic background samples for weakly supervised anomaly detection. We consider signals which produce a two-body resonance in addition to other particles (X). Such a signature can easily arise if the resonance is produced through an extended decay chain or from associated production with additional particles. These production mechanisms can lead to striking experimental signatures, with many objects in the event, potentially with rich kinematic correlations, and/or high missing energy. These distinct signatures can be leveraged to significantly reduce backgrounds; however, given the large phenomenological space of possibilities it is infeasible to design analysis strategies for every possible scenario. These two factors make for an ideal application of anomaly detection methods. This work develops a general framework for weakly supervised searches in this topology using event-level classification. The usage of weakly supervised, rather than unsupervised, anomaly detection may also offer larger discovery potential for anomalies due to improved classification performance [24].

As an exploratory study, we examine the specific case where the resonance decays into charged lepton pairs (μ or τ). Since we expect similar behavior between electrons and muons, we focus our analysis on μ pairs instead of electrons. Searches featuring τ leptons are strongly motivated by

various BSM theories, which often preferentially couple to third-generation fermions, but are less explored at the LHC due to their more complex experimental signatures. Examples of searches include Z' bosons [25], leptoquarks [26–28], and additional Higgs bosons [29,30]. These new particles are motivated by the observed anomalies in B -hadron decays to $D^{(*)}\tau\nu$, which show enhanced rates compared to the corresponding decays involving lighter leptons [31–39]. If these anomalies are caused by heavy BSM particles, it is likely they also influence the di- τ production at high invariant masses. Searches for μ lepton resonances are also strongly motivated by various BSM theories. While there are a plethora of experimental searches looking for BSM resonances decaying to second generation fermions [40–46], it is likely there are still many unexplored exotic signatures.

Though we focus on the di- τ and di- μ final states in this work, the methodology we develop is entirely general and is applicable to other resonance + X searches, such as dielectron, diphoton, or dijet resonance + X. Although these other final states may have somewhat better existing experimental coverage, there are almost certainly still unexplored “X” signatures which could be covered by our approach. These other final states may also offer the advantage of larger samples available for the training of the background model of the CATHODE method which can improve performance.

Several other works have studied the application of anomaly detection methods to topologies beyond dijet resonances. In Refs. [47,48], weak supervision was applied to topologies with high missing energy and a boosted jet while in Refs. [49,50] the application of CATHODE to supersymmetry scenarios producing pairs of gluinos was examined. Recently, Ref. [10] applied the CATHODE method to CMS open data to demonstrate the identification of Υ particles decaying to muon pairs. This study used only features related to the dimuon resonance for classification rather than focusing on the other event activity we exploit here. A previous ATLAS search [6] applied an event-level unsupervised autoencoder-based anomaly detection algorithm to search for resonances for several diobject combinations, where at least one of the objects was required to be a jet. These analyses used specific characteristics of these signal topologies to design the data-driven background estimate necessary for weak supervision, making generalization to other topologies unclear. The resonance + X topology considered in this paper complements these prior works by offering a framework to apply weakly supervised anomaly detection to searches in many different final states.

The simulated data used in this paper can be found under Refs. [51–53] and the analysis code is available on GitHub [54].

II. SETUP

A. Final states

Unlike dijet searches where the QCD multijet background is dominant, di- τ and di- μ searches receive

considerable backgrounds from multiple physics processes. In the di- τ study, we focus on the final state where both τ leptons decay hadronically, which is the decay mode with the largest branching ratio. The dominant backgrounds in the fully hadronic di- τ and the di- μ final states are Drell-Yan and $t\bar{t}$. QCD and $W + \text{jets}$ also contribute in this channel, with the fraction dependent on the purity of the τ and μ identification criteria being employed and preselection requirements. As this is an exploratory work, and these effects are not modeled well with simulation, we focus here on the Drell-Yan and $t\bar{t}$ backgrounds only. The methodology we employ should generally account for any background which varies smoothly in di- τ or di- μ mass. Because of the multiple decay modes of the τ and limited reconstruction efficiencies, di- τ searches feature significantly smaller event samples than the dijet or di- μ searches of prior weakly supervised searches. This presents an interesting challenge for the application of weakly supervised methods which require a large sample of background events in order to properly learn the probability distribution of the background events.

B. The CATHODE method

We apply the CATHODE method for resonant anomaly detection [18]. In CATHODE, one defines a resonant variable, in our case the visible mass of the di- τ candidate (m_{vis}) or invariant mass of the di- μ candidate ($m_{\mu\mu}$), and a set of features to be used for classification, $\vec{x}$. Signals are assumed to manifest in a localized region of the resonant variable and backgrounds are assumed to vary smoothly as a function of the resonant variable. These assumptions allow the probability distribution of the background to be learned from neighboring sideband regions. The method proceeds in two steps. First, a generative model is trained on the sideband regions to learn the probability density of $\vec{x}$ conditioned on m for the background, $P_b(\vec{x}|m)$. The trained model is then interpolated into the signal region and used to generate synthetic background events. A weakly supervised classifier is then trained to distinguish between the real data events $p_{\text{data}}(\vec{x})$ in the signal region and the synthetic background events $p_{\text{bkg}}(\vec{x})$. If there is a true signal in the signal region abundant in sufficient quantity, then the classifier will learn to identify its events, which will look anomalous with respect to the interpolated background.

The application of weak supervision to event-level anomaly detection raises interesting choices in feature selection. Weak supervision is typically performed with a moderately sized set of high-level features, and its performance is known to deteriorate with the use of many uninformative features [55]. Because of the noise inherent to the weakly supervised training objective, the use of lower-level, higher-dimensional representations of the event can struggle to achieve good classification performance for moderate strength signals [56]. Though this degradation may be mitigated in the classification step

through the use of boosted decision trees as opposed to neural networks [55,57], using a very high-dimensional feature set may cause difficulties in learning an accurate background estimate given the limited size of the training set available in final states such as di- τ .

For these reasons, we explore the usage of the recently introduced [58,59] phase space variables to encode the relevant kinematic features of the event in a compact representation.

C. Phase space variables

In Refs. [58,59] a covariant description of the phase space of collider events involving a chosen number (N) of massless final-state particles was presented. The manifold of this N -body phase space, Π_N , was shown to be isomorphic to the product of an $(N-1)$ -simplex (Δ_{N-1}) and a $(2N-3)$ -hypersphere (S_{2N-3}), i.e., $\Pi_N \cong \Delta_{N-1} \times S_{2N-3}$. A set of explicit global coordinates on the phase space manifold were introduced: $(N-1)$ coordinates ρ on the simplex Δ_{N-1} and $(2N-3)$ angular coordinates ξ, χ on the hypersphere S_{2N-3} . For $N=2$, the simplex (hypersphere) coordinates lie on a line (circle), while for $N=3$ they span a triangle (sphere). The authors of Ref. [59] used this representation to develop a metric between events produced at a hadron collider in terms of their phase space coordinates and demonstrated that the distance between events could be used effectively for event classification. Although the phase space coordinates are simply functions of the final-state particles' four-momenta, they are the natural variables for distinguishing different physics processes since they define the space on which S -matrix elements live. In this work, we do not use the phase space metric itself, but rather directly use the phase space coordinates as an efficient representation of the kinematics of all of the final state objects in the event. We exploit the simplex coordinates only, since we found that they significantly enhance sensitivity while the generative model used for the background model, described in Sec. IV, has difficulties with the nontrivial topology of the hypersphere coordinates detailed in Appendix A 2.

At a hadron collider, in the final state's center of mass, the simplex coordinates take the simple form,

$$\rho_i = \frac{1}{Q}(E_i \pm p_{z,i}), \quad (1)$$

where Q is the total energy of the event in the center of mass frame, the index $i = 1, \dots, N$ runs over the final-state particles, and the choice of sign is a matter of convention. The constraint $\sum_i \rho_i = 1$ implies that the ρ_i are coordinates on an $N-1$ dimensional simplex. These simplex variables can be expressed in terms of rapidity y and transverse momentum p_T as

$$\rho_i = \frac{p_{T,i}}{Q} e^{\pm y_i}, \quad (2)$$

whereas the azimuthal angles of final-state particles are contained entirely in the sphere coordinates; we refer to Ref. [59] for further details. The simplex coordinates are a useful compact combination of p_T and y insofar as the simplex contribution to the phase space distance between events is just the Euclidean distance between the simplex coordinates. Indeed, we find that the simplex coordinates enhance sensitivity compared to a feature set of the same dimension using p_T ; see Appendix B for a detailed study.

It should be noted that the phase space coordinates encode only the kinematics of the event and, therefore, do not capture all event features that may be useful for event-level anomaly detection. For example, they do not encode which type of object in the event (jet, lepton, photon, etc.) is producing said kinematics. It therefore may be useful to supplement phase space variables with additional variables which encode object-type information. The phase space variables are also normalized such that they depend only on the relative fraction of energies carried by each object and its orientation. They do not contain information about the overall energy scale of the event; i.e., an event with 100 GeV of energy split evenly among ten jets in a certain topology is encoded the same as an event with 1 TeV split among those jets in the same topology. We therefore supplement such scale information to our feature set used for classification.

III. SIMULATED SAMPLES

We produced simulated samples with MADGRAPH_a MC@NLO [60] v3.5.4 using Universal FeynRules Outputs (UFOs) [61] from the FeynRules database. We set the center of mass energy to 13 TeV in order to match the run-2 conditions of the LHC. Hadronization was done with PYTHIA8 [62] v8.306. All samples are generated at leading order using the NNPDF2.3 LO parton distribution function [63,64], with default dynamic scale choices in MADGRAPH.

For the SM background samples we used the SM UFO [65] with Cabibbo–Kobayashi–Maskawa matrix mixing, and with the four-flavor scheme. We simulated extra jets (initial/final state radiation) up to two jets, i.e., 0, 1, or 2 jets besides the hard scatter. Matrix element and parton shower matching was performed using the M. L. Mangano scheme [66] with a matching scale of 30 GeV. For the $t\bar{t}$ sample, the top quarks were decayed in MADSPIN [67] to W bosons that decay to τ or μ leptons, denoted commonly as ℓ . In order to enhance the generation efficiency, we applied a generator

level cut on the di- ℓ invariant mass of 120 GeV. The cross section of the Drell-Yan sample is 17.37 pb and the cross section of the $t\bar{t}$ sample is 13.49 pb. Sufficient events were generated to have an equivalent luminosity of 138 fb $^{-1}$, to match that of the proton-proton collisions data taken during run-2 (2016–2018) of the LHC.

We explored a set of several BSM models as benchmarks to evaluate the performance of our approach. The details of the generation are summarized in Table I, where the number of events is after the generator cuts are applied. Some example Feynman diagrams of the signal processes are shown in Fig. 1. For simplicity, we set the masses of the di- ℓ resonance in all models to 250 GeV.

In the 2HDM model [68–70] we use the heavy Higgs particle, denoted with Φ , and let it decay to an ℓ lepton pair. In the eVLQ model [71,72], we let the vectorlike quark, T' , decay to a SM top and a BSM scalar S^0 . One of the S^0 then decays to an ℓ lepton pair while the other decays inclusively according to branching ratios similar to a heavy Higgs boson. In the NMSSM model [73–77] we simulate cascade processes, where the heavy resonance X decays to a Higgs boson (H) and an unknown particle Y . Y further decays to the ℓ leptons and the H decays invisibly, leading to missing transverse momentum. We explored nonresonant signals such as vectorlike antileptons [78], leptoquarks [79], and heavy neutral leptons [80]. We did not achieve significant improvements with our method on these signals, as expected since these models are not well localized to the signal region, making them difficult to identify with the CATHODE approach. We therefore did not study them further.

We used DELPHES [81] to simulate detector and reconstruction effects, using the CMS detector settings. Jets were clustered using the anti- k_T algorithm with $R = 0.4$ as implemented in FASTJET [82] via DELPHES. τ jets were reconstructed using the default CMS DELPHES τ tagging module with $p_T > 20$ GeV and $|\eta| < 2.5$. A 1% rate of mistagging QCD jets as τ and a 60% τ tagging efficiency were included [83,84]. No dedicated trigger simulation was included.

Figure 2 shows the reconstructed m_{vis} and $m_{\mu\mu}$ distributions of the di- τ and di- μ systems, respectively. While the generator level mass cut is at 120 GeV, because some of the di- τ momentum is lost to the neutrinos, the reconstructed visible mass is smoothly falling starting from 100 GeV. Histograms are normalized to unit area.

TABLE I. List of samples used, and the details of the production. The ℓ symbol denotes the τ or μ leptons.

UFO	Name	Extra jets	Parameters	Number of events
SM	Drell-Yan $\ell\bar{\ell}$	Up to 2	N/A	3M
SM	$t\bar{t} \rightarrow W^-W^+b\bar{b} \rightarrow \ell\bar{\nu} + \bar{\ell}\nu$	Up to 2	N/A	2M
2HDM	$t\bar{t}\Phi \rightarrow \ell\bar{\ell}$	0	$m_\Phi = 250$ GeV	100k
eVLQ	$T'\bar{T}' \rightarrow tS^0\bar{t}S^0 \rightarrow \ell\bar{\ell} + \text{incl}$	0	$m_{T'} = 1000$ GeV $m_S = 250$ GeV	100k
NMSSM	$X \rightarrow YH \rightarrow \ell\bar{\ell} + \text{inv}$	0	$m_X = 800$ GeV $m_H = 400$ GeV $m_Y = 250$ GeV	100k

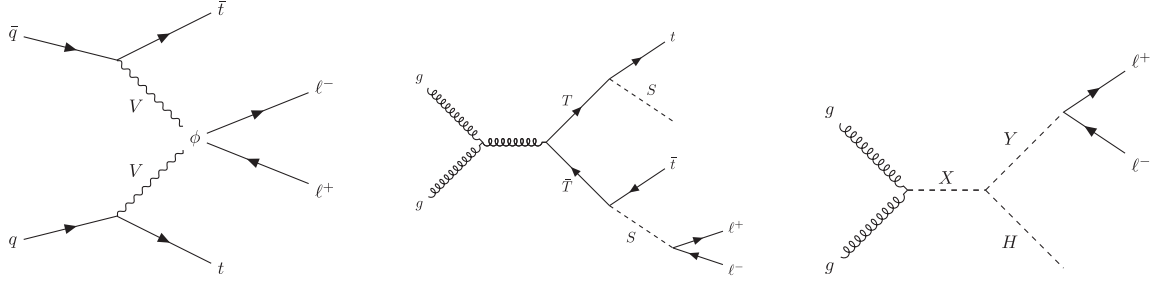


FIG. 1. Feynman diagrams of the $t\bar{t}\phi$ (left), $T\bar{T}$ (middle), and $X \rightarrow YH$ (right) signal models. The ℓ symbol denotes the τ or μ leptons. The S^0 decays inclusively according to its branching ratios. The H boson in the $X \rightarrow YH$ model decays invisibly.

IV. METHODS

For the di- τ channel, we select the two highest p_T τ -tagged jets with $|\eta| < 2.4$ as our resonance candidate. The leading τ jet is required to have $p_T > 40$ GeV. For the di- μ channel, we select the two highest p_T , oppositely charged, muons with $|\eta| < 2.4$ and $p_T > 20$ GeV. We use the m_{vis} and $m_{\mu\mu}$ as the resonance variable in the CATHODE approach in each channel. We require $m_{\text{vis}} > 100$ GeV ($m_{\mu\mu} > 120$ GeV) in the di- τ (di- μ) channel. For simplicity, we consider a SR with a mass window of 175–300 GeV for di- τ and 230–270 for di- μ , both of which contain the majority of our generated signals, as visible in Fig. 2. In a real search, a scan could be performed repeating the procedure for different signal window choices to cover the full range as has been done in previous anomaly detection searches. We expect the analysis to remain effective because the phase space variables should continue to capture the kinematic differences between signal and background for arbitrary values of the mass scale. When resonant SM backgrounds are present, their mass windows are masked so the flow learns the background density from uncontaminated sideband regions.

Within the SR, we reserve 15% of background events as a test set for performance evaluation and an equal number

of signal events. The remaining 85% of SR background events were divided equally between training (42.5%) and validation (42.5%) sets. In a true search, a cross-validation approach would be employed to ensure data events are not reused for training and signal extraction, as had been done in prior weakly supervised searches [4,7,8]. Signals are injected into the SR with varying strengths to test performance, as discussed in Sec. V.

A. Features

To establish a consistent representation of the phase space manifold, we require fixed dimensionality of objects across all events. We define events in terms of constituent objects (jets, electrons, muons, or photons), with each event represented by exactly 15 objects. For events with fewer than 15 physical objects, we perform zero padding with four-vectors as $(E, p_x, p_y, p_z) = (0.1, 0.1, 0, 0)$. We avoid four-vectors with exactly zero momentum because this leads to division by zero in the phase space coordinate calculation, causing numerical instabilities. Objects are ordered by decreasing transverse momentum to ensure consistent event representation.

The 15 objects yield 15 simplex coordinates per event, following the manifold construction described in Sec. II.

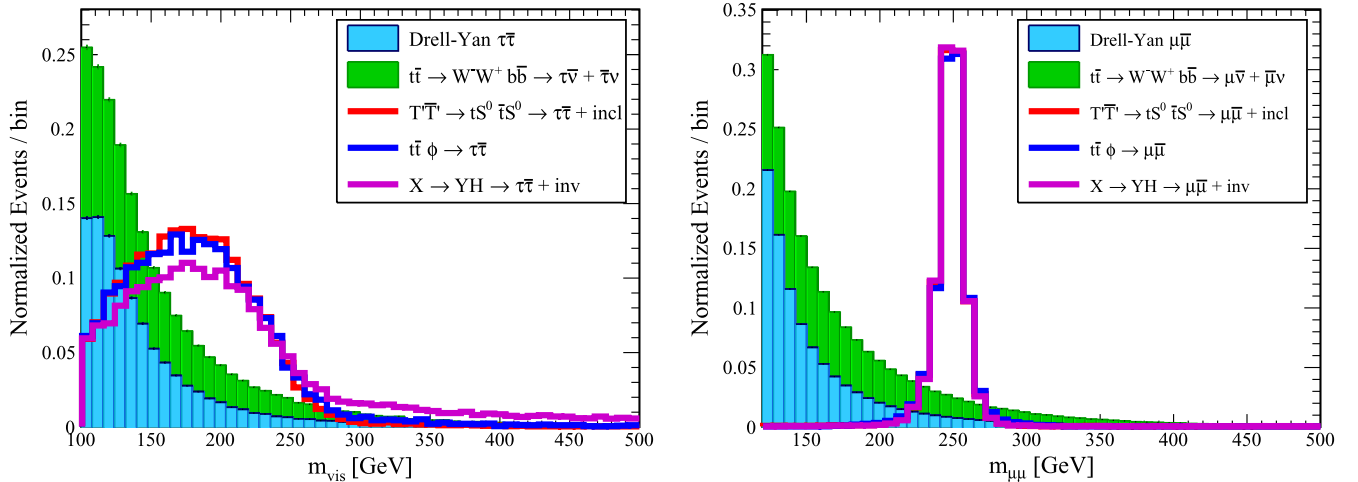


FIG. 2. Invariant visible mass distribution of the di- τ system (left) and the di- μ system (right).

Our analysis reveals that the first ten simplex coordinates, which corresponds to the ten highest p_T objects, capture the majority of the relevant phase space topology, with the remaining coordinates primarily corresponding to the zero-padded objects in typical events.

To complement the phase space variables, we incorporate event-level kinematic observables: the scalar sum of non- τ jet p_T , the missing transverse momentum (MET), and the angular separation [$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, where η is the pseudorapidity and ϕ is the azimuthal angle] between the two leptons. These features encode global event characteristics that are not fully captured by the phase space coordinates alone. We focus on these continuous kinematic features, rather than discrete quantities like jet multiplicity, because normalizing flows requires differentiable and invertible transformations. The scalar sum of jet p_T encodes valuable information on jet kinematics and overall event structure, while preserving the continuity required for a normalizing flow. Our final feature vector consists of the first ten simplex coordinates, MET, non- τ jet p_T sum, and lepton ΔR .

B. Classification and background evaluation

Our generative model uses conditional flow matching [85–87], and is trained in the sidebands. For simplicity, the training is performed on a background-only sideband sample, assuming the signal contamination to be negligible. A kernel density estimate is used to learn the m_{vis} or $m_{\mu\mu}$ distribution for use in synthetic background generation. Details of the model architecture and exact generation procedure are given in Appendix A 2.

After training, we oversample events from the background model to generate synthetic background events in the SR at 3 times the amount of data in the SR. This oversampling provides the classifier with additional background examples to learn from, improving the robustness of the weakly supervised classification. Class balance is maintained through an appropriate weighting in the loss function.

For the weakly supervised classification task, we employ an ensemble of 50 histogram-based gradient boosted decision trees (BDTs). Ensembles of BDTs have demonstrated superior performance over neural networks in weakly supervised training with high-dimensional feature spaces [55]. Our implementation utilizes the HistGradientBoosting classifier from SCIKIT-LEARN [88]. Hyperparameters for the BDT are kept to be the defaults as in the SK_CATHODE [89] repository.

Once we have trained the weakly supervised classifier, we define our final event selection based on a cut on the classifier score that is 1% efficient on the background. Though tighter cuts are seen to yield higher signal enhancements, this threshold is seen to give good classification performance for all signals, and is realistic to thresholds

used in previous weakly supervised anomaly detection searches [4,7,8].

C. Synthetic background quality validation

We mix two well-performing generated backgrounds in equal proportions, as each may capture different aspects of the background distribution. A comparison of the synthetic events generated by the background model and the true background events in the SR is shown in Fig. 3.

To validate the quality of our synthetic background, we applied our BDT classifier to distinguish real background events from synthetic ones in the signal region. The BDT classifier uses the same architecture and features as are used in the weakly supervised training. The training corresponds to what the weakly supervised classifier would see if there were only background events in the signal region. A near-random classification performance shows that such a classifier cannot tell the samples apart. This indicates that if there were instead signal events in the signal region the classifier would focus on correctly classifying them to improve the loss, rather than concentrating on the differences between the synthetic and true background. This test yields a classification area under curve of 0.52 for both channels, close to the random-chance value of 0.5.

Imperfect modeling of the background may dampen sensitivity in the CATHODE approach, as the weakly supervised classifier may focus on the differences between the synthetic and true backgrounds rather than identifying signals. However, even with our imperfect background modeling, we have seen strong anomaly detection performance with our chosen signals. We therefore conclude our background model is of sufficiently high quality for the anomaly detection procedure to work, and leave further improvements to future studies.

V. RESULTS

The performance of the signal classification is expected to vary as a function of the amount of signal present in the dataset. We run the training procedure for several different injection strengths for each signal and evaluate the resulting classification performance. For each injection the procedure is repeated 5 times to assess the stability of results. The injection strength is quantified by the approximate signal significance, $\sigma = \frac{S}{\sqrt{B}}$, prior to any additional selection from the anomaly detector. We test injection strengths of 0.25σ , 0.5σ , 1σ , 1.5σ , and 2σ . Significance improvement curves (SIC) are used to quantify classification performance. The curves are created by scanning across different classifier thresholds and plotting the signal efficiency divided by the square root of the background efficiency $\frac{\epsilon_s}{\sqrt{\epsilon_b}}$, as a function of the background efficiency. The SIC shows the approximate enhancement in signal statistical significance which can be achieved by cutting at a

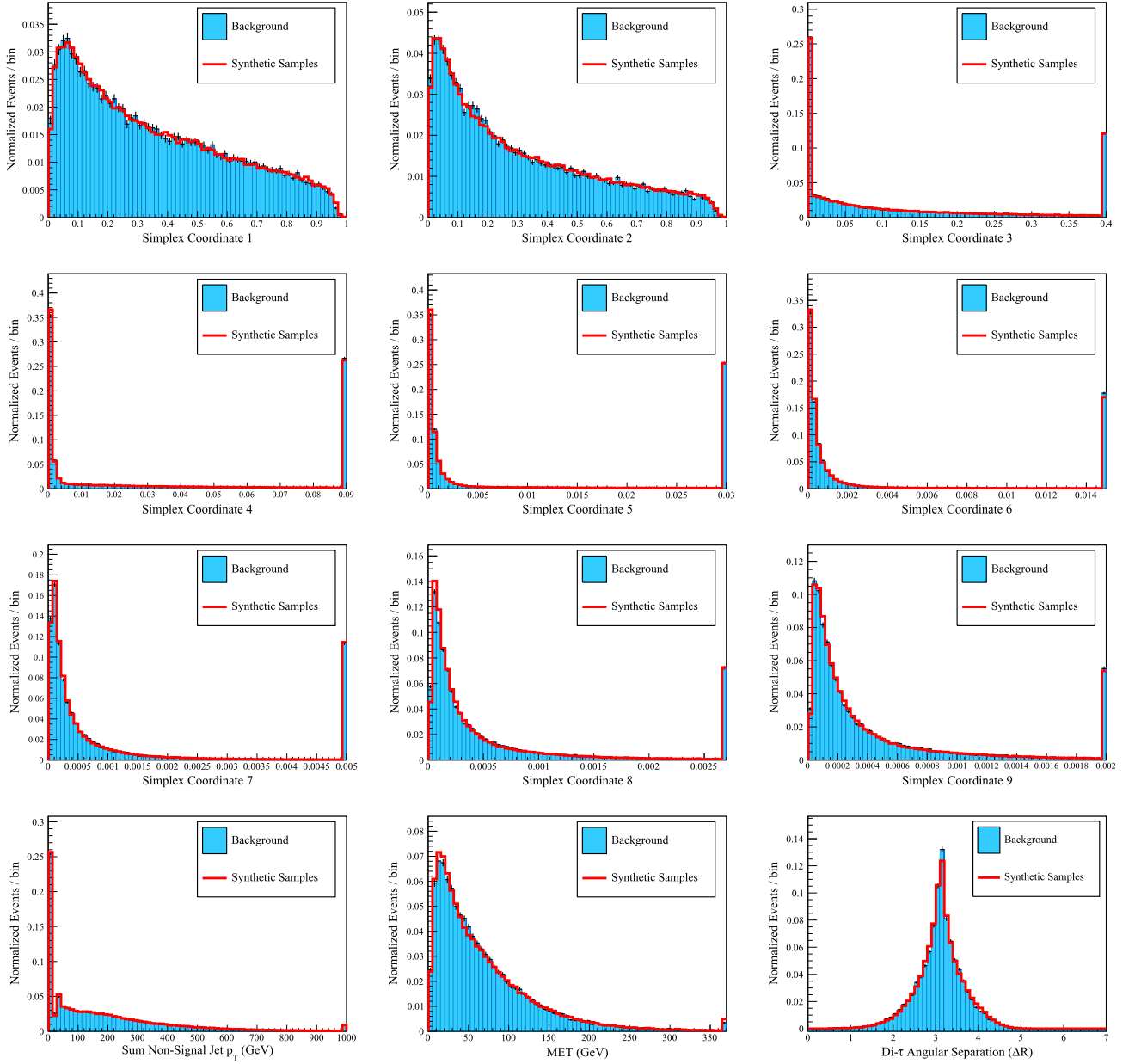


FIG. 3. A comparison of the distributions of true background events in the signal region and the synthetic ones generated by our normalizing flow for di- μ . Simplex coordinate 10 is omitted for brevity, but has similar distributions. The last bin in each histogram shows overflow entries beyond the range of the histogram axis. The flow is seen to model most of the distributions well. Some minor discrepancies are seen in the modeling close to variable boundaries at zero. The feature distributions and agreement between synthetic and true are similar in the di- τ channel.

given background efficiency. The achieved classification performance for different signal injections is shown in Fig. 4.

Figure 5 shows a summary of the anomaly detection performance for the different signal injections. The x axis denotes the strength of the signal injection and the y axis the approximate signal significance after the anomaly detection procedure has been performed. The final significance of a signal is estimated as the significance

improvement of the anomaly score cut, chosen to have a 1% background efficiency, times the original injection strength. For all three signal models considered, our procedure is able to enhance the signal significance above the discovery threshold (5σ) for moderately sized signal injections $\leq 2\sigma$. This demonstrates the potential of our anomaly detection approach to discover signals that may have been below the sensitivity threshold of a standard inclusive bump hunt search.

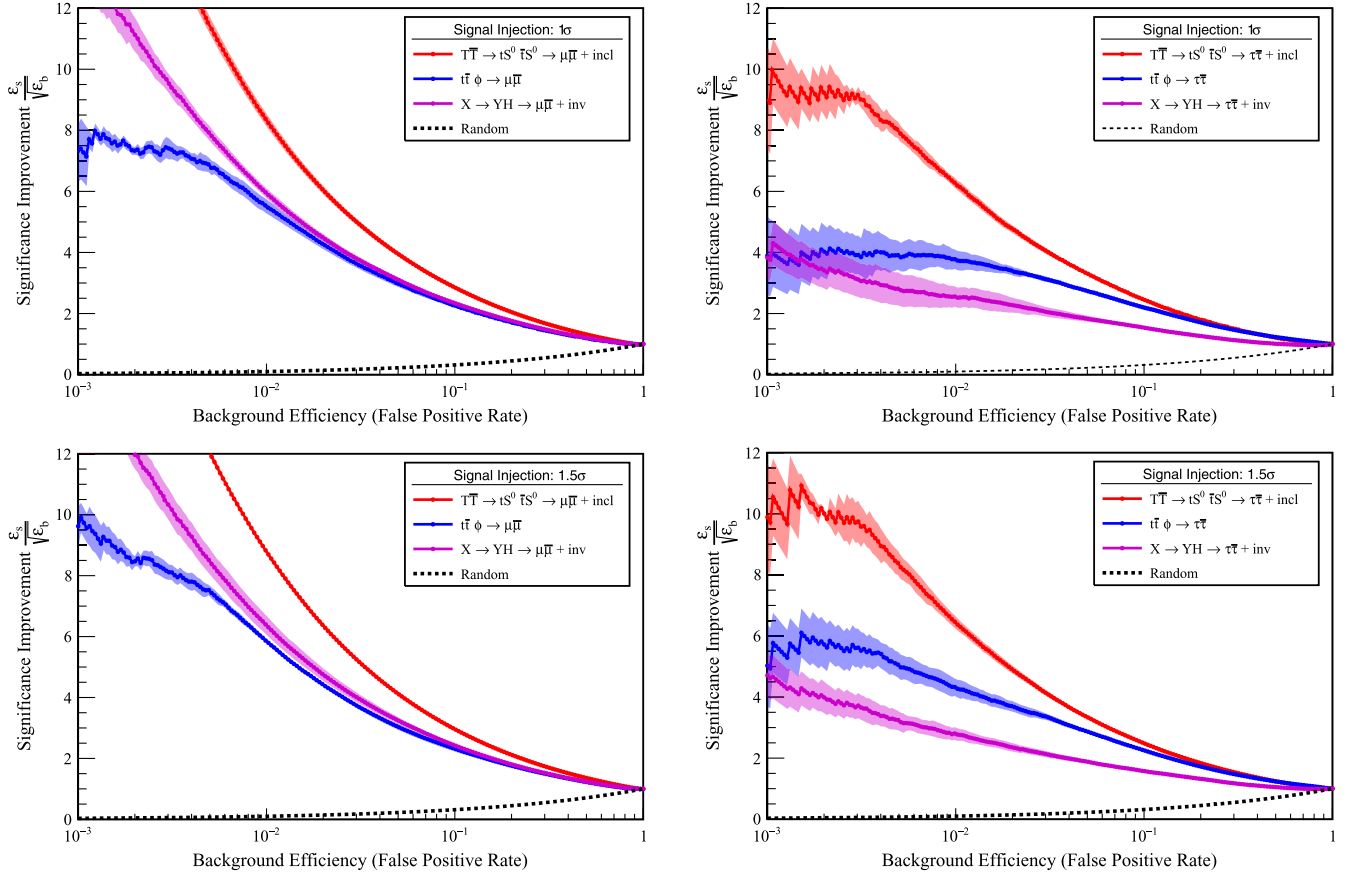


FIG. 4. Significance improvement curves showing classification performance across our different signal models for both di- μ (left) and di- τ (right) for signal injections of 1σ (top) and 1.5σ (bottom). The shaded bands represent the statistical uncertainty (± 1 standard deviation) in the measured average significance improvement from the five different runs.

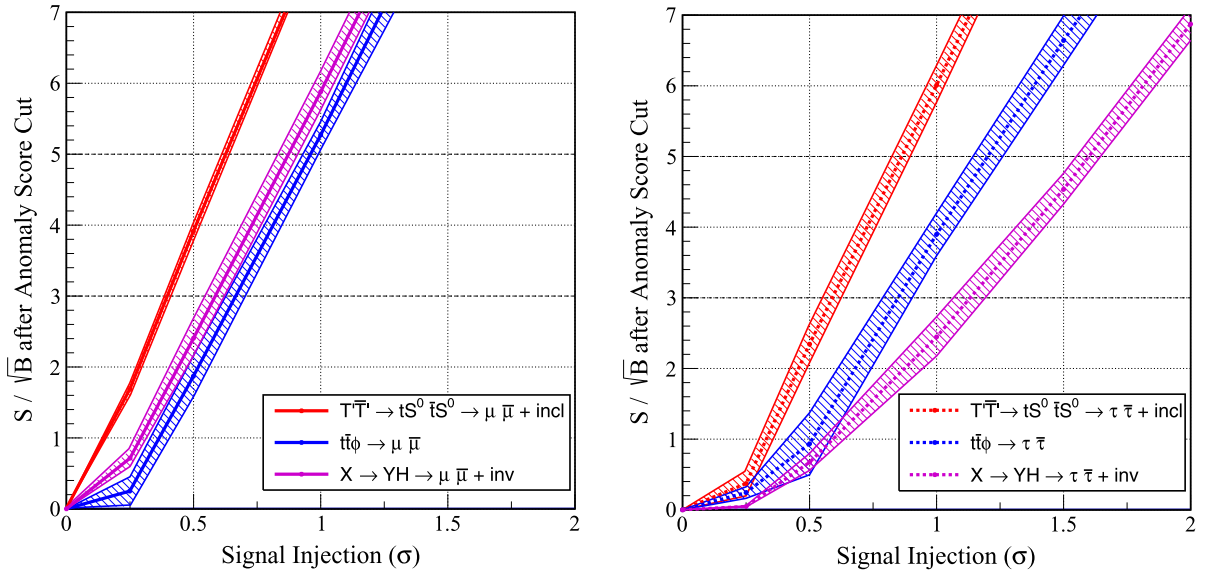


FIG. 5. A summary of the sensitivity enhancement of our anomaly detection approach for both di- μ (left) and di- τ (right). The x axis shows the size of the signal, quantified as $\frac{S}{\sqrt{B}}$ prior to any anomaly detection selection. The y axis shows the approximate signal significance achieved after the application of an anomaly score cut that is 1% efficient on the background. Our procedure is able to enhance the signal significance above the discovery threshold (5σ) for moderately sized signal injections $\leq 2\sigma$ for all considered signal models.

VI. CONCLUSION

We have demonstrated the potential for weakly supervised anomaly detection to be applied to a new class of resonance searches. Diobject plus X searches, in which one looks for a resonance decaying to two standard model particles, but produced in association with other anomalous event activity, represent a new class of searches ripe for the application of anomaly detection. In this exploratory work, we focused on di- τ and di- μ plus X final states. We have employed new physically motivated phase space variables to robustly capture salient features of multiparticle kinematics. We demonstrated that our anomaly detection approach can reach discovery-level significances for signals that would be missed in a conventional bump-hunt approach. Further work can be done to improve the generative modeling of topologically interesting variables like the hypersphere phase space coordinates. We could potentially gain sensitivity from exploring other event-level variables that quantify object type to complement the kinematic information of the phase space variables. It may be fruitful to explore non-Euclidean normalizing flows, which may better capture the topological features of the phase-space variables.

The framework introduced in this paper lays the foundation for a program of searches that can be performed at the LHC with significant discovery potential. Further exploration of topologies suitable for weakly supervised anomaly detection will be essential to deliver on the potential of anomaly detection to transform collider search strategies.

ACKNOWLEDGMENTS

We would like to thank Joe Incandela, Danyi Zhang, and Sanjit Masanam for useful discussions. We used the SK_CATHODE [89] library for implementation. We used the UCSB computational facilities administered by the Center for Scientific Computing at the California NanoSystems Institute and Materials Research Laboratory (an NSF MRSEC; DMR-1720256) and purchased through NSF CNS-1725797. UCSB is supported by the U.S. Department of Energy under Grant No. DE-SC0011702. Support for UCSB is also made possible by the Joe and Pat Yzurdiaga endowed chair in experimental science. L. B. was also supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. NSF 2139319. O. A. is supported by FermiForward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. L. M. and M. S. acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy-EXC2121 "Quantum Universe"-390833306 as well as by the Bundesministerium für Bildung und Forschung under Project No. 05H24GUA. S. S. and P. M. are supported by NSF Grant No. PHY2310072.

DATA AVAILABILITY

The data that support the findings of this article are openly available [51–53].

APPENDIX A: IMPLEMENTATION DETAILS

1. Preprocessing

The DELPHES files generated in Sec. III are further processed before being utilized with CATHODE. A luminosity under sampling is applied so the effective luminosity of the background samples is scaled to match the 138 fb^{-1} collected data in LHC run-2. A random selection of events with yields as detailed in the upper section of Table II is used for analysis. After under sampling, the two simulated background datasets are combined and preprocessed following the cuts discussed in Sec. IV, with the resulting event distributions shown in the lower section of Table II.

In the CATHODE method, the synthetic background samples generated by the normalizing flow should be indistinguishable from the true background events in the signal region. If this is not the case, then the weakly supervised classifier will fail to learn to identify the signal, instead learning to distinguish the real and synthetic background events.

We employ a sequence of transformations to optimize the feature space for our generative model. First, all features are shifted and scaled to the range $x \in (0, 1)$. Then a logit transformation [$\text{logit}(x) = \ln(\frac{x}{1-x})$] is applied to convert bounded domains into unbounded ones. The transformed features are then standardized by subtracting the mean and dividing by the standard deviation of the training set, respective to each feature. This preprocessing pipeline ensures that the normalizing flow operates on a well-conditioned feature space while preserving the essential structure of the phase space manifold.

Simulated and synthetic data in the signal and sideband regions are further split into training/validation/testing datasets following the methods described in Sec. IV. The corresponding event counts can be seen in Table III.

TABLE II. Number of events before and after luminosity under sampling (top). Event counts before and after preprocessing cuts (bottom).

Name	Number of events	Lumiscalled number of events
Drell-Yan	3M	2,397,067
$t\bar{t}$	2M	1,861,869
	Combined lumiscalled	After cuts
$\tau\bar{\tau}$ Drell-Yan + $t\bar{t}$	4,258,936	$\sim 158,000$
$\mu\bar{\mu}$ Drell-Yan + $t\bar{t}$	4,258,936	$\sim 1,325,000$

TABLE III. Numbers of events used for training, validation, and testing in the signal region and sidebands (rounded to the nearest thousand).

Name	Total number of events	Sideband	Signal region
$\tau\bar{\tau}$ Drell-Yan + $t\bar{t}$	$\sim 158,000$	$\sim 127,000$	$\sim 31,000$
$\mu\bar{\mu}$ Drell-Yan + $t\bar{t}$	$\sim 1,325,000$	$\sim 1,260,000$	$\sim 65,000$
$\tau\bar{\tau}$	Training	Validation	Testing
Sideband	89,000	38,000	...
Signal region	13,000 simulated 39,000 synthetic	13,000 simulated 39,000 synthetic	5000 background 5000 signal
$\mu\bar{\mu}$	Training	Validation	Testing
Sideband	882,000	378,000	...
Signal region	27,500 simulated 82,500 synthetic	27,500 simulated 82,500 synthetic	10,000 background 10,000 signal

2. Generative model

Generative models with conditional flow matching (CFM) use conditional normalizing flows by learning vector fields between conditional distributions. CFM avoids the computational intractability of traditional flow matching objectives [85] while offering improved efficiency and stability over diffusion models [85,90].

Our flow is parametrized by a multilayer perceptron with six hidden layers of 128 neurons each, using Exponential Linear Unit activation functions [91]. Time is encoded using sinusoidal embeddings with three frequencies, producing a six-dimensional time representation through concatenated sine and cosine components. The network takes as an input the concatenation of the features $\vec{x}$, time embeddings t , and conditioning variable m (di- τ or di- μ visible invariant mass).

Training employs the Adam optimizer [92] with a cosine annealing learning rate scheduler, starting at 1×10^{-3} and decaying to 1×10^{-6} . We used batch sizes of 1024 (di- μ) and 512 (di- τ). We trained the model over 2000 to 5000 epochs using A100 GPUs. We found that training beyond 5000 epochs did not improve the quality of the generated samples. We fine-tuned the hyperparameters of the generative model by minimizing the Receiver Operating Characteristic (ROC) between background and generated samples. Training for 5000 epochs takes 32 hours.

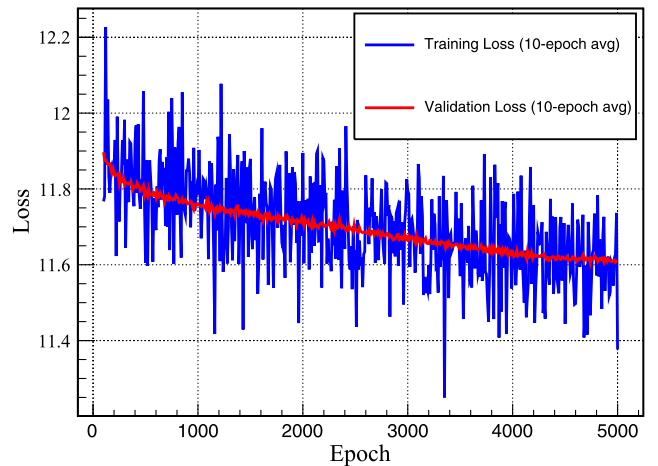
After training, we select the model state with the lowest validation loss for the subsequent interpolation and sampling steps. The training and validation losses of the di- μ background can be seen in Fig. 6.

To generate synthetic background events in the signal region, we employ a two-step process that requires the trained flow to extrapolate beyond its training domain: (1) We fit a kernel density estimate to the di- ℓ invariant mass distribution observed in the signal region in the training set using a Gaussian kernel with bandwidth 0.01. (2) We sample mass values m from this signal region mass distribution and use them to condition the learned normalizing flow that was trained on sideband data. This approach requires the learned normalizing flow to extrapolate from

sideband masses (training domain) to signal region masses (target domain), assuming the conditional dependence on m generalizes smoothly across this mass range.

Since the normalizing flow operates on transformed features, we apply the inverse of the standardization and logit transformation to convert the sampled events back to physical space.

We do not model the hypersphere coordinates in this analysis due to their non-Euclidean nature, where normalizing flows are designed for features in the d -dimensional Euclidean space $\mathcal{R}^d$ [93]. The flow struggled to learn how to model the connection between different coordinates needed for a good background distribution. The simplex still captures much of the event's topology, but the additional azimuthal information contained in the hypersphere could improve identification. There are limited works looking into flows on non-Euclidean manifolds [94–96]. However, implementing and improving upon their architectures for ten-dimensional hyperspheres is beyond the scope of this paper.

FIG. 6. Training and validation loss for the di- μ channel, as an example.

APPENDIX B: COMPARISON TO p_T -BASED FEATURES

To evaluate the effectiveness of the phase space representation, we conducted a comparative analysis replacing the ten simplex coordinates used with the p_T values of the ten highest p_T objects in each event, in descending order. We focus on the di- μ channel as a proof of concept for this comparison.

To maintain consistency with the phase space analysis, events with fewer than ten objects were zero padded using

randomized exponentially decreasing values from 0 to 25 GeV, corresponding to the minimum p_T threshold in DELPHES simulation. We retain the same complementary kinematic features used in the phase space analysis: the scalar sum of jet p_T , the missing transverse momentum (MET), and the angular separation ΔR between the two muons. These features provide essential event-level information not captured by the ranked p_T values alone.

The features can be seen in Fig. 7.

The same generative model architecture and training methodology employed for the phase space di- μ analysis

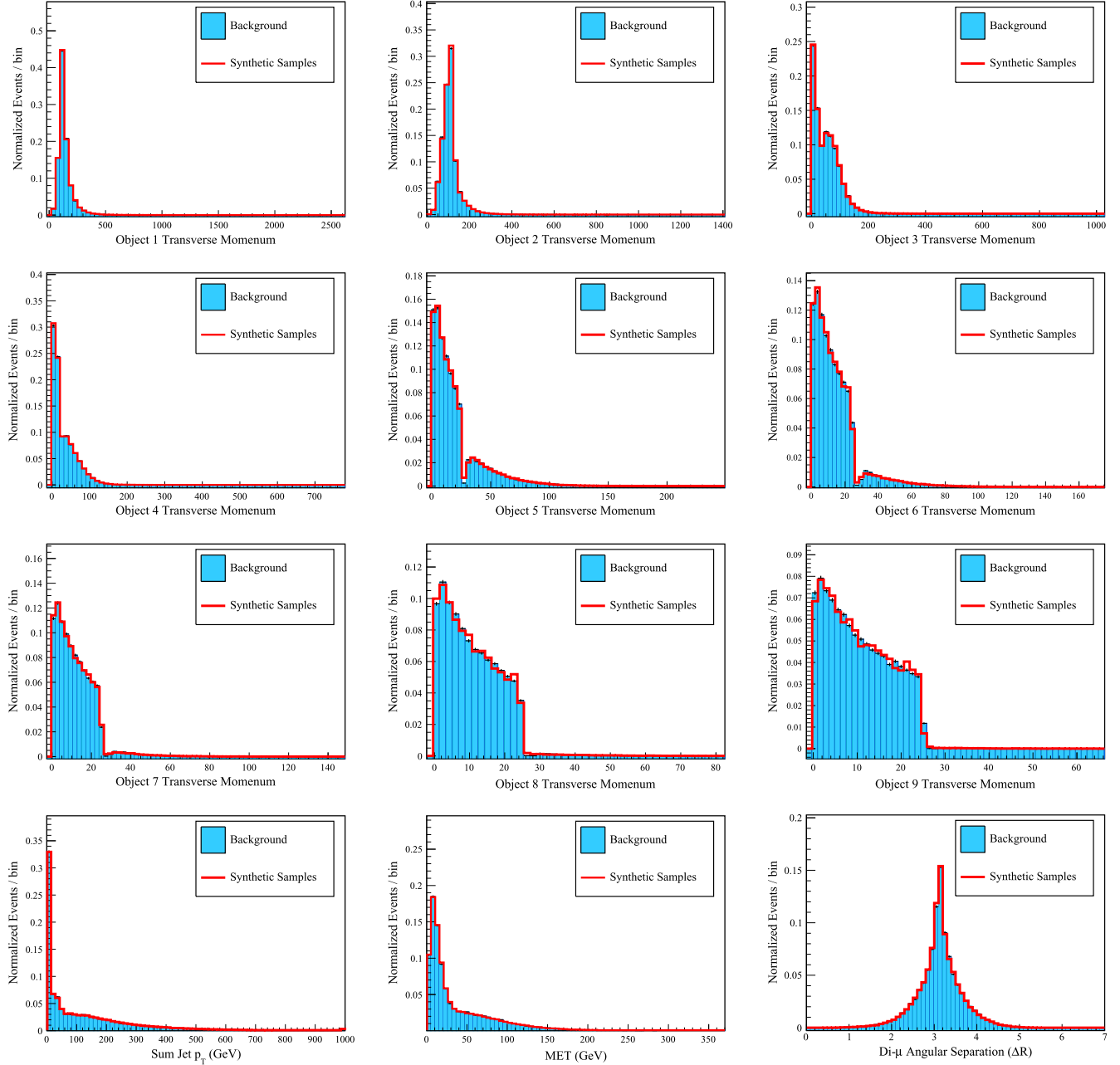


FIG. 7. A comparison of the distributions of true background events in the signal region and the synthetic ones generated by our normalizing flow for di- μ . Object 10 p_T is omitted for brevity, but has similar distributions.

were applied to the p_T -based features. The background generation achieved a ROC score of 0.53, comparable to the phase space background model performance (ROC = 0.52), indicating similar quality in modeling the background distribution.

Figure 8 shows the resulting significance improvement curves comparing the p_T and phase space approaches. The phase space coordinates consistently outperform the p_T -based features across all signal injection levels.

The superior performance of phase space coordinates is encouraging for several reasons. First, the simplex coordinates are specifically designed to capture the topological structure of multiparticle final states in a coordinate-invariant manner, whereas p_T rankings only preserve energy ordering without topological information. Second, our analysis utilizes only the simplex coordinates from the phase space decomposition and we have not yet incorporated the hypersphere coordinates, which encode additional angular correlations. This suggests significant untapped potential for further improvements in anomaly detection sensitivity using the complete phase space representation. The consistent performance advantage of phase space coordinates across different signal models demonstrates the robustness of this approach and supports its adoption for anomaly detection in collider physics applications.

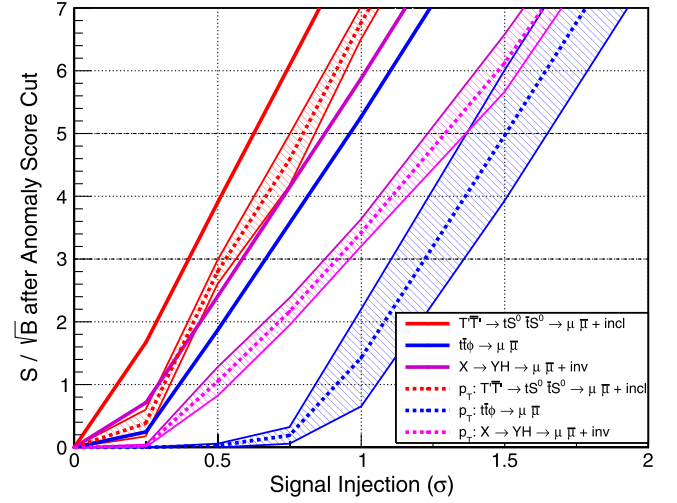


FIG. 8. A summary of the sensitivity enhancement of our anomaly detection approach for di- μ using phase space features (solid line) and p_T features (dashed line). The x axis shows the size of the signal, quantified as $\frac{S}{\sqrt{B}}$ prior to any anomaly detection selection. The y axis shows the approximate signal significance achieved after the application of an anomaly score cut that is 1% efficient on the background. Our procedure utilizing phase space features is superior to using the p_T of the objects alone in all signal cases.

-
- [1] G. Kasieczka *et al.*, *Rep. Prog. Phys.* **84**, 124201 (2021).
 - [2] T. Aarrestad *et al.*, *SciPost Phys.* **12**, 043 (2022).
 - [3] V. Belis, P. Odagiu, and T. K. Aarrestad, *Rev. Phys.* **12**, 100091 (2024).
 - [4] G. Aad *et al.* (ATLAS Collaboration), *Phys. Rev. Lett.* **125**, 131801 (2020).
 - [5] ATLAS Collaboration, *Phys. Rev. D* **108**, 052009 (2023).
 - [6] G. Aad *et al.* (ATLAS Collaboration), *Phys. Rev. Lett.* **132**, 081801 (2024).
 - [7] V. Chekhovsky *et al.* (CMS Collaboration), *Rep. Prog. Phys.* **88**, 067802 (2025).
 - [8] ATLAS Collaboration, *arXiv:2502.09770*.
 - [9] O. Knapp, O. Cerri, G. Dissertori, T. Q. Nguyen, M. Pierini, and J.-R. Vlimant, *Eur. Phys. J. Plus* **136**, 236 (2021).
 - [10] R. Gambhir, R. Mastandrea, B. Nachman, and J. Thaler, *Phys. Rev. Lett.* **135**, 021902 (2025).
 - [11] M. Farina, Y. Nakai, and D. Shih, *Phys. Rev. D* **101**, 075021 (2020).
 - [12] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, *SciPost Phys.* **6**, 030 (2019).
 - [13] E. M. Metodiev, B. Nachman, and J. Thaler, *J. High Energy Phys.* **10** (2017) 174.
 - [14] J. H. Collins, K. Howe, and B. Nachman, *Phys. Rev. Lett.* **121**, 241803 (2018).
 - [15] J. H. Collins, K. Howe, and B. Nachman, *Phys. Rev. D* **99**, 014038 (2019).
 - [16] O. Amram and C. M. Suarez, *J. High Energy Phys.* **01** (2021) 153.
 - [17] A. Andreassen, B. Nachman, and D. Shih, *Phys. Rev. D* **101**, 095004 (2020).
 - [18] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, *Phys. Rev. D* **106**, 055006 (2021).
 - [19] J. A. Raine, S. Klein, D. Sengupta, and T. Golling, *Front. Big Data* **6**, 899345 (2022).
 - [20] T. Golling, S. Klein, R. Mastandrea, and B. Nachman, *Phys. Rev. D* **107**, 096025 (2023).
 - [21] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, *Phys. Rev. D* **107**, 114012 (2023).
 - [22] D. Sengupta, S. Klein, J. A. Raine, and T. Golling, *SciPost Phys.* **17**, 046 (2024).
 - [23] R. Das, G. Kasieczka, and D. Shih, *arXiv:2312.11629*.
 - [24] J. H. Collins, P. Martín-Ramiro, B. Nachman, and D. Shih, *Eur. Phys. J. C* **81**, 617 (2021).
 - [25] G. Aad *et al.* (ATLAS Collaboration), *arXiv:2503.19836*.
 - [26] A. Hayrapetyan *et al.* (CMS Collaboration), *J. High Energy Phys.* **05** (2024) 311.
 - [27] G. Aad *et al.* (ATLAS Collaboration), *Eur. Phys. J. C* **83**, 1075 (2023).
 - [28] G. Aad *et al.* (ATLAS Collaboration), *Phys. Rev. D* **104**, 112005 (2021).

- [29] G. Aad *et al.* (ATLAS Collaboration), *Phys. Rev. Lett.* **125**, 051801 (2020).
- [30] A. Tumasyan *et al.* (CMS Collaboration), *J. High Energy Phys.* **07** (2023) 073.
- [31] B. Capdevila, A. Crivellin, and J. Matias, *Eur. Phys. J. Special Topics* **1**, 20 (2023).
- [32] Y. S. Amhis *et al.* (HFLAV Collaboration), *Phys. Rev. D* **107**, 052008 (2023).
- [33] R. Aaij *et al.* (LHCb Collaboration), *Phys. Rev. Lett.* **115**, 111803 (2015); **115**, 159901(E) (2015).
- [34] J. P. Lees *et al.* (BABAR Collaboration), *Phys. Rev. D* **88**, 072012 (2013).
- [35] M. Huschle *et al.* (Belle Collaboration), *Phys. Rev. D* **92**, 072014 (2015).
- [36] S. Hirose *et al.* (Belle Collaboration), *Phys. Rev. D* **97**, 012004 (2018).
- [37] G. Caria *et al.* (Belle Collaboration), *Phys. Rev. Lett.* **124**, 161803 (2020).
- [38] R. Aaij *et al.* (LHCb Collaboration), *Phys. Rev. D* **97**, 072013 (2018).
- [39] R. Aaij *et al.* (LHCb Collaboration), *Phys. Rev. D* **108**, 012018 (2023); **109**, 119902(E) (2024).
- [40] C. Cesarotti, Y. Soreq, M. J. Strassler, J. Thaler, and W. Xue, *Phys. Rev. D* **100**, 015021 (2019).
- [41] CMS Collaboration, *J. High Energy Phys.* **10** (2023) 043.
- [42] CMS Collaboration, *J. High Energy Phys.* **06** (2018) 120.
- [43] M. Aaboud *et al.* (ATLAS Collaboration), *J. High Energy Phys.* **10** (2017) 182.
- [44] G. Aad *et al.* (ATLAS Collaboration), *Phys. Rev. D* **90**, 052005 (2014).
- [45] A. Policicchio and G. Crosetti, *Eur. Phys. J. C* **55**, 173 (2008).
- [46] M. Aaboud *et al.* (ATLAS Collaboration), *Phys. Rev. D* **99**, 012001 (2019).
- [47] T. Finke, M. Krämer, M. Lipp, and A. Mück, *J. High Energy Phys.* **08** (2022) 015.
- [48] G. Kasieczka, J. A. Raine, D. Shih, and A. Upadhyay, *arXiv:2404.07258*.
- [49] G. Bickendorf, M. Drees, G. Kasieczka, C. Krause, and D. Shih, *Phys. Rev. D* **109**, 096031 (2024).
- [50] D. Curtin, S. Dreyer, M. F. Costa, S. Heim, G. Kasieczka, L. Moureaux, D. Rouso, D. Shih, and M. Sommerhalder, *arXiv:2506.11192*.
- [51] T. A. Vami, DELPHES simulations of different signal models with di-tau final states at the LHC (2025), <https://zenodo.org/records/15203499>.
- [52] T. A. Vami, DELPHES simulations of different signal models with dimuon final states at the LHC (2025), <https://zenodo.org/records/16288044>.
- [53] T. A. Vami and L. Brennan, Datasets used in this paper (2025), <https://zenodo.org/records/16302772>.
- [54] L. Brennan *et al.*, Analysis github repository (2025), https://github.com/LiamBrennan-UCSB/Weakly_supervised_anomaly_detection_with_event_level_variables.
- [55] T. Finke, M. Hein, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, and M. Sommerhalder, *Phys. Rev. D* **109**, 034033 (2023).
- [56] E. Buhmann, C. Ewen, G. Kasieczka, V. Mikuni, B. Nachman, and D. Shih, *Phys. Rev. D* **109**, 055015 (2023).
- [57] M. Freytsis, M. Perelstein, and Y. C. San, *J. High Energy Phys.* **02** (2024) 220.
- [58] A. J. Larkoski and T. Melia, *Phys. Rev. D* **102**, 094014 (2020).
- [59] T. Cai, J. Cheng, N. Craig, G. Koszegi, and A. J. Larkoski, *J. High Energy Phys.* **09** (2024) 054.
- [60] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *J. High Energy Phys.* **07** (2014) 079.
- [61] L. Darmé *et al.*, *Eur. Phys. J. C* **83**, 631 (2023).
- [62] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *Comput. Phys. Commun.* **191**, 159 (2015).
- [63] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, *Nucl. Phys.* **B838**, 136 (2010).
- [64] R. D. Ball *et al.* (NNPDF Collaboration), *J. High Energy Phys.* **04** (2015) 040.
- [65] FeynRules Database, The Standard Model, <https://feynrules.irmp.ucl.ac.be/wiki/StandardModel> (2018).
- [66] J. Alwall *et al.*, *Eur. Phys. J. C* **53**, 473 (2008).
- [67] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk, *J. High Energy Phys.* **03** (2013) 015.
- [68] FeynRules Database, Two-Higgs-doublet model (2HDM), <https://cp3.irmp.ucl.ac.be/projects/feynrules/wiki/2HDM> (2018).
- [69] G. C. Branco, L. Lavoura, and J. P. Silva, *CP Violation* (Oxford University Press, New York, 1999), Vol. 103.
- [70] C. Degrande, *Comput. Phys. Commun.* **197**, 239 (2015).
- [71] FeynRules Database, Vector like quarks + exotic pNGBs, https://feynrules.irmp.ucl.ac.be/raw-attachment/wiki/NLO-Models/eVLQ_v8.1.tar.gz (2016).
- [72] A. Banerjee *et al.*, *arXiv:2203.07270*.
- [73] FeynRules Database, The next-to-minimal supersymmetric standard model, <https://cp3.irmp.ucl.ac.be/projects/feynrules/wiki/NMSSM> (2018).
- [74] U. Ellwanger, C. Hugonie, and A. M. Teixeira, *Phys. Rep.* **496**, 1 (2010).
- [75] A. Djouadi *et al.*, *J. High Energy Phys.* **07** (2008) 002.
- [76] P. Z. Skands *et al.*, *J. High Energy Phys.* **07** (2004) 036.
- [77] B. C. Allanach *et al.*, *Comput. Phys. Commun.* **180**, 8 (2009).
- [78] I. Bigaran, B. A. Dobrescu, and A. Russo, *Phys. Rev. D* **109**, 055033 (2024).
- [79] M. J. Baker, J. Fuentes-Martín, G. Isidori, and M. König, *Eur. Phys. J. C* **79**, 334 (2019).
- [80] B. Kayser and R. E. Shrock, *Phys. Lett.* **112B**, 137 (1982).
- [81] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), *J. High Energy Phys.* **02** (2014) 057.
- [82] M. Cacciari, G. P. Salam, and G. Soyez, *Eur. Phys. J. C* **72**, 1896 (2012).
- [83] A. M. Sirunyan *et al.* (CMS Collaboration), *J. Instrum.* **13**, P10005 (2018).
- [84] A. Tumasyan *et al.* (CMS Collaboration), *J. Instrum.* **17**, P07023 (2022).
- [85] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, *arXiv:2210.02747*.
- [86] A. Tong, K. Fatras, N. Malkin, G. Hugué, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio, *arXiv:2302.00482*.
- [87] P. Holderrieth and E. Erives, *arXiv:2506.02070*.
- [88] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg,

- J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [89] M. Sommerhalder *et al.*, SK CATHODE Github Repo (2025), https://github.com/uhh-pd-ml/sk_cathode/tree/main.
- [90] O. Amram *et al.*, [arXiv:2410.21611](#).
- [91] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, [arXiv:1511.07289](#).
- [92] D. P. Kingma and J. L. Ba, [arXiv:1412.6980](#).
- [93] Y. Lipman, M. Havasi, P. Holderrieth, N. Shaul, M. Le, B. Karrer, R. T. Q. Chen, D. Lopez-Paz, H. Ben-Hamu, and I. Gat, [arXiv:2412.06264](#).
- [94] Y. Liu, H. Liu, Y. Yin, Y. Wang, B. Chen, and H. Wang, [arXiv:2304.03937](#).
- [95] L. Falorsi, [arXiv:2104.14959](#).
- [96] D. J. Rezende, G. Papamakarios, S. Racanière, M. S. Albergo, G. Kanwar, P. E. Shanahan, and K. Cranmer, [arXiv:2002.02428](#).