Robust Anomaly Detection for Particle Physics Using Multi-Background Representation Learning

Abhijith Gandrakota[1,+], **Lily Zhang**[2,+], **Aahlad Puli**[2], **Kyle Cranmer**[3], **Jennifer Ngadiuba**[1], **Rajesh Ranganath**[2], **and Nhan Tran**[1]

[1]Fermi National Accelerator Laboratory, Batavia, IL, USA 60510
[2]New York University, New York, NY, USA, 10012
[3]University of Wisconsin-Madison, Madison, WI, USA, 53706

ABSTRACT

Anomaly, or out-of-distribution, detection is a promising tool for aiding discoveries of new particles or processes in particle physics. In this work, we identify and address two overlooked opportunities to improve anomaly detection for high-energy physics. First, rather than train a generative model on the single most dominant background process, we build detection algorithms using representation learning from multiple background types, thus taking advantage of more information to improve estimation of what is relevant for detection. Second, we generalize decorrelation to the multi-background setting, thus directly enforcing a more complete definition of robustness for anomaly detection. We demonstrate the benefit of the proposed robust multi-background anomaly detection algorithms on a high-dimensional dataset of particle decays at the Large Hadron Collider.

1 Introduction

The Standard Model (SM) of particle physics has achieved remarkable success in predicting a wide range of phenomena, culminating in the recent groundbreaking discovery of the Higgs boson[1–3]. However, despite these triumphs, many phenomena in particle physics still remain unexplained[4–6]. Among the most important and exciting questions in physics is the discovery of phenomena beyond the Standard Model (BSM). This quest lies at the heart of the efforts at the Large Hadron Collider (LHC), dedicated to advancing humanity's understanding of the universe.

There are two main approaches for the discovery of new physics. The first, more traditional approach looks for a particular alternate hypothesis which is typically a family of parametrized distributions[7]. However, this approach seeks specific kinds of deviation and could be blind to other evidence of BSM physics not explicitly considered under the alternate hypothesis posed.

To address this shortcoming, a second, more recent approach uses machine learning-based anomaly detection (AD) algorithms to search for deviations from the standard model more broadly, without focusing on a specific BSM alternative hypothesis. There has been a growing interest in this data-driven approach in recent years, with the development of AD algorithms specific to high-energy physics[8–13] as well as their direct use in the analysis of data collected by the LHC experiements[14,15].

AD algorithms are employed to isolate collimated particle showers (jets) coming from the decays of new BSM particles produced by proton proton collisions at the LHC. The hope is to discover a bump or a deviation in the smooth kinematic distributions such as the invariant mass of the selected jets (relative to what is otherwise expected from SM backgrounds), in order to provide evidence of a new BSM particle[1,3,16]. In other words, the discovery of a new particle can be viewed as hypothesis testing based on these kinematic distributions. To increase the power of this statistical test, the primary goal of the anomaly detection algorithm is to filter out SM backgrounds such that any deviations in these kinematic distributions resulting from a potentially new BSM process can be discerned. At the same time, this selection must be performed without creating deviations from the filtering process itself. This secondary objective is often known as decorrelation or the robustness of the anomaly detection algorithm[17]. For further details please see Appendix A.

In this work, we build upon existing data-driven approaches along two dimensions. The first is to take advantage of extra information and assumptions via representation learning fom multiple background processes. Existing approaches use deep generative models or density estimators such as variational autoencoders (VAEs)[18,19] and normalizing flows (NF)[20,21] to model the most dominant SM process, quantum chromodynamics (QCD), as well as possible[22]; however, AD methods based on generative modeling have been found to lead to worse than random chance failures in benchmark machine learning tasks[23], likely due to the sensitivity of such approaches to estimation error[24]. One way to address this limitation of generative models

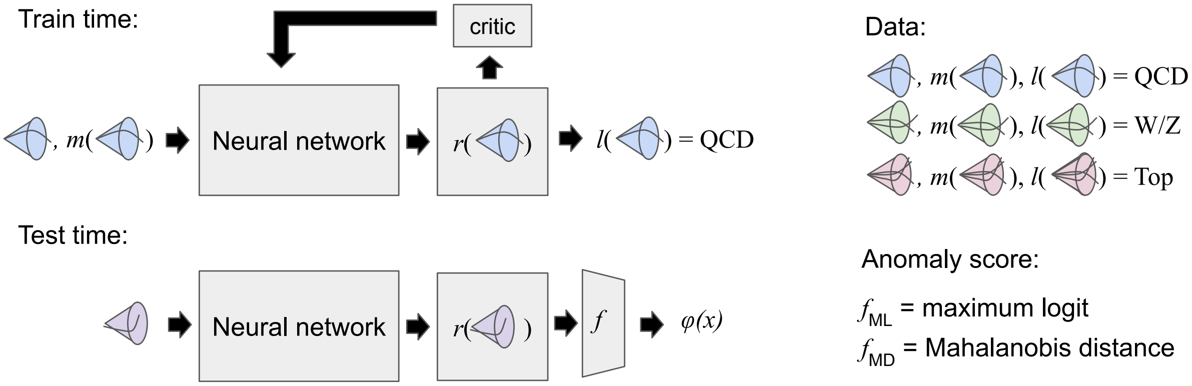


Figure 1. Overview of robust anomaly detection with multi-background representation learning. During training, jets (depicted by cones) as well as their mass m and label l are used to learn robust multi-background representations r. The data used for training includes jets of different background processes, not just QCD. Then, these learned representations are used to derive an anomaly score $\phi = f \circ r$ used at test time.

and yield better anomaly detection is to transform the data inputs into representations that are easier to model and/or focus on information important for the task of discovery. In fact, in existing anomaly detection benchmarks on high-dimensional inputs, methods that build on the feature representations obtained from a model trained to distinguish non-anomalous data have been empirically shown to outperform detection based on generative models that directly estimate the non-anomalous input distribution[25]. These empirical results do not guarantee that the former will always perform better than the latter, but they do point to the potential benefits of incorporating additional information or assumptions into anomaly detection. To incorporate additional information, we propose *multi-background representation learning* for anomaly detection.

Our second contribution is developing a solution for robustness in the multi-background representation learning setting of anomaly detection. An important consideration for AD methods in jet physics is ensuring that anomaly scores do not depend on kinematic variables like the jet mass. Otherwise, the probability of a false positive discovery increases. Concretely, discovery in high-energy physics often involves looking for statistically significant deviations or "bumps" in the overall sample of flagged anomalies along certain search variables, and such bumps are more likely when anomaly scores depend on these kinematic variables across jets within a known process. To minimize such false positive discoveries, we introduce a decorrelation objective and algorithm in the presence of multiple backgrounds, taking inspiration from recent techniques in the robust representation learning and anomaly detection literature in machine learning[26,27].

In this paper, we motivate and describe our robust multi-background representation learning approach to AD and demonstrate empirically that it rivals the current state-of-the-art method in jet anomaly detection, specifically in detecting jets from top quarks as out-of-distribution samples given jets from QCD and W/Z processes as in-distribution examples. We hope that this work inspires future work to consider underexplored ideas around multi-background representation learning for AD in high-energy physics.

2 Robust Multi-Background Anomaly Detection

Below, we motivate and describe our robust multi-background representation learning setup (Section 2.1, Section 2.2) and overall anomaly detection algorithm based on such representations (Section 2.3). A visual overview of the proposed method can be found in Figure 1.

2.1 Multi-Background Representation Learning

The primary purpose of representation learning in anomaly detection is to guide the information used to assign an anomaly score. For instance, image anomaly detection algorithms built from representations that can distinguish known classes have been shown to outperform anomaly detection algorithms based on representations that do not take into account known class labels[25]. In fact, among the latter, anomaly detection algorithms based on densities estimated from deep generative models have been shown to yield worse than random chance performance detecting anomalies that otherwise seem obvious based on human perception[23,24]. These failures in some ways represent a worst-case scenario for purely data-driven approaches: without guidance on what types of deviations matter most, generative models can end up focusing modeling efforts on unimportant details at the expense of poorly estimating details that matter for detecting practical anomalies. This ability to learn from data while also incorporating knowledge about what information is most important is where representation learning[28] shines.

We build representations using multiple known background process types to better utilize the data already available from known physics. Denote the input as $\mathbf{x}$, the searchvariable we wish to be robust to (e.g., jet mass) as $\mathbf{z}$, and the background process as $\mathbf{y}$. Define a representation function $r : \mathscr{X} \to \mathbb{R}^d$. A good representation retains relevant information for detection while removing information that simply yields an additional modeling burden to a downstream detection algorithm. We hypothesize that the information that is relevant for detecting new particles will generally vary across known particles. To capture this assumption, we learn a representation that can distinguish between existing processes by training representations using the labels of the known particle classes as supervision. In other words, we encourage representations that are useful for classifying between processes via $\arg\max_r p(\mathbf{y} \,|\, r(\mathbf{x}))$.

2.2 Decorrelation under Multi-Background Representation Learning

Second, we develop an approach for decorrelation under multi-background representations. Namely, we directly enforce decorrelation within each background process by enforcing an independence constraint on the representations we learn described below. We use $\perp\!\!\!\perp_p$ to denote independence of two random variables under distribution p. When there is no subscript specified, the independence is enforced under the original training distribution.

First, we introduce a distribution $p_{\perp\!\!\!\perp}$ such that the search variable (e.g., jet mass) and label are independent: $p_{\perp\!\!\!\perp}(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x}\,|\,\mathbf{y},\mathbf{z})p(\mathbf{y})p(\mathbf{z})$. In other words, the distribution $p_{\perp\!\!\!\perp}$ corresponds to the resulting distribution obtained after breaking dependence between $\mathbf{y}$ and $\mathbf{z}$ occuring in the training distribution.

Under this distribution, we first force representation $r(\mathbf{x})$ to be independent of $\mathbf{z}$, i.e., $r(\mathbf{x}) \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}$. This constraint ensures that the representation cannot predict jet mass under a distribution where jet mass and label do not provide information about each other. It is important that this independence constraint be enforced under $p_{\perp\!\!\!\perp}$, since enforcing marginal independence under the training distribution instead (i.e., $r(\mathbf{x}) \perp\!\!\!\perp \mathbf{z}$) would disallow representations to contain information important for distinguishing particles, just because such information is also correlated with mass.

Under $p_{\perp\!\!\!\perp}$, we also enforce the representations to be independent of mass conditioned on the label, i.e., $r(\mathbf{x}) \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}\,|\,\mathbf{y}$. This constraint ensures that jets belonging to a given background process cannot be differentially flagged as anomalies based on jet mass. Consequently, within a background process the distribution of jet mass will be the same for flagged jets as it is for all jets, an important condition for downstream analyses based on $\mathbf{z}$. For a more detailed discussion about downstream analyses and their assumptions, see Appendix A.

Combining all the above objectives together, we have the following objective for the representations:

$$\arg\max_r p_{\perp\!\!\!\perp}(\mathbf{y}\,|\,r(\mathbf{x})) \text{ s.t. } r(\mathbf{x}),\mathbf{y} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}. \tag{1}$$

The term $\arg\max_r p_{\perp\!\!\!\perp}(\mathbf{y}\,|\,r(\mathbf{x}))$ encourages informative representations, while the constraint $r(\mathbf{x}),\mathbf{y} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}$ enforces decorrelation in the multi-background setting. Note also that $r(\mathbf{x}),\mathbf{y} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}$ is the combination of $r(\mathbf{x}) \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}$ and $r(\mathbf{x}) \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}\,|\,\mathbf{y}$; moreover, for a given r, the predictor $p_{\perp\!\!\!\perp}(\mathbf{y}\,|\,r(\mathbf{x}))$ s.t. $r(\mathbf{x}),\mathbf{y} \perp\!\!\!\perp_{p_{\perp\!\!\!\perp}} \mathbf{z}$ in Equation (1) is equivalent to $p(\mathbf{y}\,|\,r(\mathbf{x}))$ s.t. $r(\mathbf{x}) \perp\!\!\!\perp \mathbf{z}|\mathbf{y}$ (see[27] for proof).

To learn a representation r that achieves Equation (1), we use Nuisance[1] Randomized Distillation (NuRD)[26]. First, to approximate the distribution $p_{\perp\!\!\!\perp}$ from the training distribution, we perform importance weighting:

$$p_{\perp\!\!\!\perp}(\mathbf{x},\mathbf{y},\mathbf{z}) = w_{\mathbf{y},\mathbf{z}}p(\mathbf{x},\mathbf{y},\mathbf{z}), \quad w_{\mathbf{y},\mathbf{z}} = p(\mathbf{y})/p(\mathbf{y}|\mathbf{z}). \tag{2}$$

Both $p(\mathbf{y})$ and $p(\mathbf{y}\,|\,\mathbf{z})$ can be estimated from the training data, assuming access to both the jet mass and label.

To encourage joint independence, we use a mutual information penalty. Namely, we estimate $\mathbf{I}_{p_{\perp\!\!\!\perp}}(r(\mathbf{x}),\mathbf{y};\mathbf{z})$, where any nonzero quantity suggests the lack of joint independence. To estimate this mutual information, we employ the density ratio trick[29], following Ref.[26]. Concretely, we train an additional neural network, referred to as the critic model, to estimate the relevant density ratio for mutual information. The critic model p_γ is trained to distinguish between inputs from $p_{\perp\!\!\!\perp}(r(\mathbf{x}),\mathbf{y},\mathbf{z})$ and inputs from $p_{\perp\!\!\!\perp}(r(\mathbf{x}),\mathbf{y})p_{\perp\!\!\!\perp}(\mathbf{z})$, the latter of which we represent by shuffling the jet masses within the batch. The critic model predicts unshuffled inputs with $c = 1$ and shuffled inputs with $c = 0$. Then, given well-calibrated critic model such that the probability $p_\gamma(c = 1\,|\,r(\mathbf{x}),\mathbf{y},\mathbf{z})$ corresponds to the probability of the input coming from distribution $p_{\perp\!\!\!\perp}(r(\mathbf{x}),\mathbf{y},\mathbf{z})$, we can estimate the mutual information as follows:

$$\mathbf{I}_{p_{\perp\!\!\!\perp}}(r(\mathbf{x}),\mathbf{y};\mathbf{z}) = \mathbb{E}_{p_{\perp\!\!\!\perp}(r(\mathbf{x}),\mathbf{y},\mathbf{z})}\left[\log p_\gamma(c = 1\,|\,r(\mathbf{x}),\mathbf{y},\mathbf{z}) - \log(1 - p_\gamma(c = 1\,|\,r(\mathbf{x}),\mathbf{y},\mathbf{z}))\right]. \tag{3}$$

To approximate the mutual information under $p_{\perp\!\!\!\perp}$, we weight the examples using the weights in Equation (2).

[1] In prior work[26,27], $\mathbf{z}$ was called a nuisance variable, not to be confused with nuisance parameter which is much more common in the high-energy physics context. To avoid confusion, in this work we refer to $\mathbf{z}$ as a search feature, given its role in the discovery process.

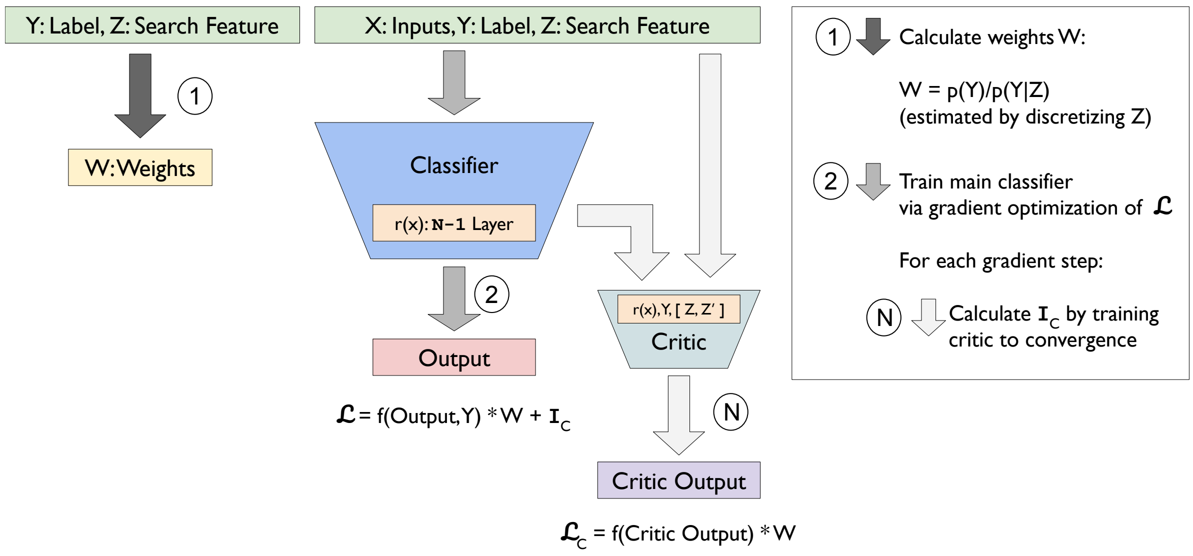

Y: Label, Z: Search Feature
1
W: Weights
X: Inputs, Y: Label, Z: Search Feature
Classifier
r(x): N-1 Layer
2
Output
$\mathcal{L}$ = f(Output, Y) * W + $\mathbf{I}_C$
r(x), Y, [Z, Z']
Critic
N
Critic Output
$\mathcal{L}_C$ = f(Critic Output) * W
1 Calculate weights W:
W = p(Y)/p(Y|Z)
(estimated by discretizing Z)
2 Train main classifier
via gradient optimization of $\mathcal{L}$
For each gradient step:
N Calculate $\mathbf{I}_C$ by training
critic to convergence

Figure 2. An overview of the Nuisance-Randomized Distillation algorithm[26] used in this work for learning robust multi-background representations for anomaly detection in high-energy physics. First, we calculate weights w for each input based on its label y and jet mass z. These weights are used to approximate the distribution $p_{\perp\perp}$ such that the y and z are marginally independent. Then, we train our classifier to optimize a reweighted objective. We additionally include a mutual information-based penalty term to the loss by training a critic model n gradient steps for every gradient step of the classifier.

Putting both ideas together, we learn a representation r by optimizing parameters θ via gradient descent under the following objective:

$$\max_{\theta} \quad \log p_{\perp\perp}(\mathbf{y} \,|\, r_\theta(\mathbf{x})) - \lambda \mathbf{I}_{p_{\perp\perp}}(r_\theta(\mathbf{x}), \mathbf{y}; \mathbf{z}). \tag{4}$$

In other words, our objective seeks a representation that distinguishes between known background process types as well as possible while penalizing representations that together with the background process label can predict the search feature. See Figure 2 for a schematic of the overall algorithm.

2.3 Anomaly Scores based on Robust Multi-Background Representations

Given a representation built using multiple backgrounds (i.e., more than one distinct value of $\mathbf{y}$) and the appropriate decorrelation of mass with respect to all of them, we define a robust multi-background detection score $\phi : \mathscr{X} \to \mathbb{R}$ as a function of such a representation, i.e., $\phi(\mathbf{x}) = f(r(\mathbf{x})),\ f : \mathbb{R}^d \to \mathbb{R}$. Here, we describe two anomaly scores, ϕ_{ML} and ϕ_{MD}, which utilize the aforementioned robust multi-background representations. In particular, we define $\phi_{ML} = f_{ML} \circ r$ and $\phi_{MD} = f_{MD} \circ r$, where functions f_{ML} and f_{MD} map the representation to some scalar anomaly score. Following the deep learning anomaly detection literature, we define two anomaly scores yielding high accuracy on image anomaly detection tasks; the first, the max logit (ML)[30], takes the maximum logit of the classifier trained on in-distribution inputs and labels. The second, the Mahalanobis distance (MD)[31], fits class-conditional Gaussians to the data distribution representations and uses the probability under this fitted distribution as the anomaly score. We describe both in more detail below.

The max logit anomaly score flags a jet as an anomaly if the maximum of the logits is small[30]. Since a good classifier on the data distribution of known classes should generally assign high maximum logits to jets from known classes, a jet that is assigned relatively small outputs relative to jets seen during training is likely different in some way (e.g. does not activate the same feature maps as strongly) and thus more likely to be anomalous. To implement the max logit score, we use the classifier that trains the representation described in the previous section.

The Mahalanobis distance focuses instead on the representations of individual jets. Namely, the Mahalanobis distance models the feature representations of in-distribution data as k class-conditional Gaussians with means μ_k and covariances Σ_k. At test time, the anomaly score is the minimum Mahalanobis distance from a new input's feature representations to each of these class distributions, $\min_k \sqrt{(r(\mathbf{x}) - \mu_k)\Sigma_k^{-1}(r(\mathbf{x}) - \mu_k)^\top} = \max_k p(r(\mathbf{x}) \mid \mathbf{y} = k)$. Assuming minimal overlap in probability across each class-conditional Gaussian (e.g., representations for each class form tight clusters that are far apart), this method can approximate density estimation on the representations: $\max_k p(r(\mathbf{x}) \mid \mathbf{y} = k) \propto \max_k p(r(\mathbf{x}) \mid \mathbf{y} = k)p(\mathbf{y} = k) \approx \sum_k p(r(\mathbf{x}) \mid \mathbf{y} = k)p(\mathbf{y} = k) = p(r(\mathbf{x}))$.

While density estimation based on Gaussians for each known class may seem relatively simple, this method is especially powerful coupled with a good representation that clusters similar jets. Moreover, the speed of inference, especially relative to running a more complex deep generative model, makes this score especially practical given the computational burden of processing petabytes of collision data[32].

We call the above methods Nuisance Randomized Distillation with Max Logit (nurd-ml) and Nuisance Randomized Distillation with Mahalanobis Distance (nurd-md). These two anomaly detection methods based on robust representations have been shown to offer significant performance gains in out-of-distribution benchmarks relative to approaches without these robustness considerations[27].

Summarizing the full method, first we learn representations via the NuRD objective (Equation (4)) by training a classifier to predict the background particle classes under a reweighted objective (Equation (2)) with a mutual information penalty (Equation (3)), implemented with an inner submodule that estimates a mutual information term. Then, we derive two anomaly scores from these representations, one which involves fitting class-conditional Gaussians on the representations, and the other which takes the maximum logit score from the classifier derived from the representations. See Figure 1 for an overview of the entire detection algorithm pipeline.

3 Related Work

Our multi-background and representation learning-based approach is related to and inspired by work in the anomaly and out-of-distribution detection literature in high-energy physics and machine learning. The move away from purely generative-based approaches towards an approach that considers other techniques is motivated by existing work on the limitations of deep generative models for anomaly detection[23,24]. The use of class labels of in-distribution processes to improve anomaly detection is inspired by existing work in the image out-of-distribution detection literature which shows that the most performant methods take advantage of label information[33]. In particle physics applications, existing works have considered the idea of building classifiers for representation learning[34], as well as embedding data into lower-dimensional spaces while preserving different choices of metrics[35].

The need for robust anomaly detection in high-energy physics has been discussed in[13,36]. Several approaches have been proposed for encouraging detection that is decorrelated with kinematic variables such as jet mass[17,37,38]. Ref.[37] trains an autoencoder with reconstruction loss and an additional adversarial term which penalizes the loss if an adversary (a neural network) can predict the jet mass from the prediction residual. In this approach, the objective encourages reconstruction error to be independent of the jet mass, which is equivalent to enforcing the independence between the anomaly score and the jet mass. Given that only QCD is considered, this can also be trivially viewed as a form of conditional independence, i.e. $\phi(\mathbf{x}) \perp\!\!\!\perp \mathbf{z}|\mathbf{y}$. Another approach to ensure robustness is post-hoc decorrelation, which chooses different thresholds for classification depending on jet mass. A common approach is to bin the data based on jet mass and then choose thresholds per bin such that the false positive rate is the same for all bins[39]. This objective is the same as the equal opportunity constraint (i.e., $\phi(\mathbf{x}) \perp\!\!\!\perp \mathbf{z}|\mathbf{y} = 1$) in the fairness in machine learning literature, which is a relaxation of the full conditional independence constraint $\phi(\mathbf{x}) \perp\!\!\!\perp \mathbf{z}|\mathbf{y}$, also known as equalized odds in the fairness literature[40]. Practically, a post-processing method can be applied on top of any detection method, including the one we propose. Moreover, our predictor also satisfies conditional independence, which means that it generalizes the objectives of both aforementioned approaches.

There also exist various methods which incorporate decorrelation into the classification of jets. Ref.[36] employs an additional adversarial term to encourage the distribution of a classifier's predictions to be invariant to a nuisance parameter either conditional on the class $\mathbf{y}$ or with class marginalized out. Rather than using an adversarial objective, Ref.[38] incorporates a distance correlation penalty which is zero if and only if the relevant variables are independent. The goals of jet classification are distinct from those of anomaly detection, but this work demonstrates how some of these ideas can be ported over to anomaly detection when the latter makes use of supervised representation learning.

4 Experimental Setup

To demonstrate the practical benefit of our method, we train and test nurd-ml and nurd-md on simulated high-momentum jets at the LHC. We first describe the data used to benchmark our approach (Section 4.1). Then, we describe details of the model architecture and training (Section 4.2). We compare our results with a baseline VAE approach as defined in[37], described in related work in Section 3. To test the benefits of the approach, we keep architectural choices between our proposed approach and baseline as close as possible; namely, we design our classifier to match the architecture of the VAE encoder so that empirical benefits cannot be attributed to architectural choices.

4.1 Data

The model is trained on high momentum jets from the hls4ml LHC Jet dataset set given in Refs.[41,42]. The jets are represented as an image of 50×50 pixels, where each pixel corresponds to the sum of the energy of all particles in the corresponding cell. These images are normalized into a probability distribution over the grid (i.e. pixel values sum to one), as in this task we are uninterested in the total energy within the grid but rather the spatial distribution over the grid. We train a classifier on two different standard model jet classes, QCD and W/Z bosons. For testing the algorithm, we use jets produced by top quarks as the out-of-distribution sample. For constructing the in-distribution dataset, we used 600,000 jets with equal proportions of QCD and W/Z boson jets. We split this dataset into training, validation, and test sets with proportions 60%, 20%, and 20% respectively. We use the training and validation data to train the classifier to obtain a representation function for anomaly detection. For evaluating anomaly detection performance, we use a dataset consisting of the test QCD jets alongside out-of-distribution top jets.

4.2 Model Architecture and Training

Our main classifier takes as input the 50×50-size images as input and predicts the particle class (QCD or W/Z) as output. This classifier is a convolutional neural network whose architecture closely follows that of the encoder in Ref.[13] for a fair comparison of methods. We train our classifier on a dataset comprised of QCD and W/Z jets in equal proportions, in order to bypass the optimization difficulties of a highly imbalanced distribution that training on a dataset of natural proportions would introduce.

We train the NuRD algorithm as described in Section 2 which modifies traditional supervised learning by additionally reweighting the examples to approximate $p_{\perp\!\!\!\perp}$ and estimating the mutual information term $\mathrm{I}_{p_{\perp\!\!\!\perp}}(r(\mathbf{x}), \mathbf{y}; \mathbf{z})$ to enforce joint independence.

To reweight the examples, we discretized the jet mass into bins of 5 GeV for both the QCD and W/Z samples to accommodate the resolution of the W/Z resonance peaks. Then, the weight given to a jet in the bin i is the inverse of the probability of that bin under the distribution of the jet mass for that class: more specifically, we use weights N/n_i, where N is the total number of jets in the class and n_i number of jets in bin i. This is equivalent to using the weight $\frac{p(\mathbf{y})}{p(\mathbf{y}|\mathbf{z})}$ described in the Section 2, except that the jet mass has been discretized. We utilize these weights to estimate the empirical risk under $p_{\perp\!\!\!\perp}$ when computing the loss over a batch.

To estimate the mutual information penalty, we initialize a critic model which takes in the 20-dimensional representation $r(\mathbf{x})$ from the penultimate activations of the main classifier, as well as a scalar label and jet mass. The network is a multilayer perceptron (MLP) with 3 hidden layers of 256, 128, and 64 neurons and predicts which distribution the input came from (i.e., $p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}, \mathbf{z})$ or $p_{\perp\!\!\!\perp}(r(\mathbf{x}), \mathbf{y}) p_{\perp\!\!\!\perp}(\mathbf{z})$).

For each step of the main classifier, we train and update the critic on a 10% fraction of the dataset and use the resulting critic model to approximate the mutual information per batch. This term is added as a penalty to the reweighted cross-entropy loss, which we optimize via gradient descent using the Adam optimizer with learning rate of 10^{-3}.

We train a baseline VAE on 300,000 QCD jets, using the network architecture and the hyperparameters as defined in Ref.[37]. We evaluate both proposed methods nurd-ml and nurd-md, and the baseline VAE method on a test set containing only QCD and out-of-distribution top-quark jets, following previous evaluations[37].

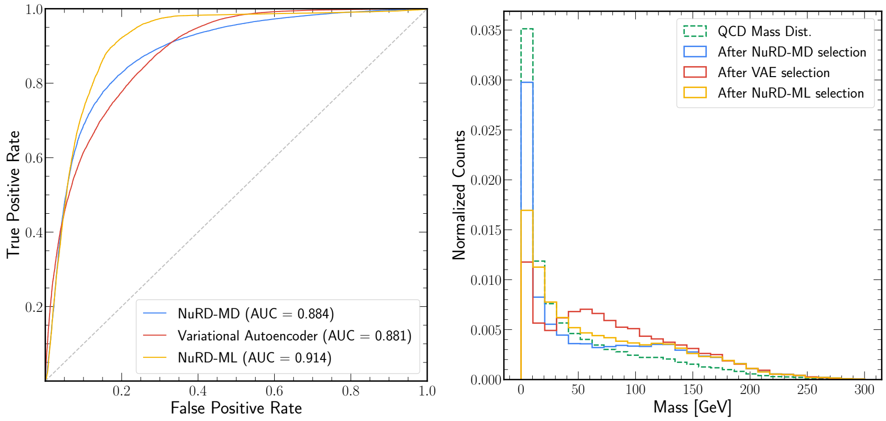
5 Results

Figure 3(a) shows that nurd-ml and nurd-md yield better detection performance than the baseline VAE method as measured via AUROC. Recall that both the baseline and nurd-md perform anomaly detection via density estimation, one directly on the inputs and the other on a representation of the inputs learned from multiple types of background processes. Given this parallelism, the superiority of nurd-md despite its use of a simpler generative model (i.e. class-conditional Gaussians rather than a deep generative model) suggests that a well-crafted representation can assist in anomaly detection by making the task of density estimation easier. The fact that nurd-ml also outperforms the baseline suggests that the learned representation can be useful even for anomaly scores that do not directly correspond to density estimation.

Moreover, Figure 3(b) shows the QCD mass distribution overall (green) and after subsetted to the 50% most anomalous events. An ideal algorithm would show no difference in this mass distribution before and after thresholding, such that downstream analyses could treat the flagged anomalies in aggregate using known physics about QCD. On the other hand, a poor algorithm results in a "sculpting" of the mass distribution into one unlike the original QCD mass distribution, increasing the potential for false positive discoveries. Both nurd-md and nurd-ml yield less sculpting of the jet mass distribution than the baseline VAE method. The degree of sculpting is quantified in Table 1 which reports the Jensen-Shannon divergence and L2 Wasserstein

Distance between the jet mass distributions before and after applying the threshold on the anomaly score. The fact that the proposed methods achieve better detection performance as well as less sculpting of the QCD mass distribution suggests they are a superior option to the baseline. The presence of some amount of sculpting from the proposed methods is likely due to the use of a penalty in the objective rather than a hard constraint.

Nurd-ml and nurd-md also exhibit an overall better significance improvement (Table 1). The significance improvement for a given true positive rate is defined as the signal efficiency (true positive rate) divided by the square root of the background misidentification (false positive rate). This metric is indicative of the discovery potential[43]. Table 1 shows the maximum significance improvement versus the signal efficiency. nurd-ml and nurd-md yield a better signal-to-background ratio, indicating they falsely flag a smaller proportion of background jets for a given proportion of anomalous jets flagged. Both proposed methods also have a higher maximum significance improvement, suggesting they can achieve a better optimal signal-to-background ratio with a high signal efficiency.



(a) ROC Curve

(b) Mass distribution

Figure 3. Our proposed robust multi-background detection methods outperform the baseline VAE implementation of Ref.[13] in both the overall detection performance (AUROC, left) as well as decorrelation with jet mass (less sculpting, right).

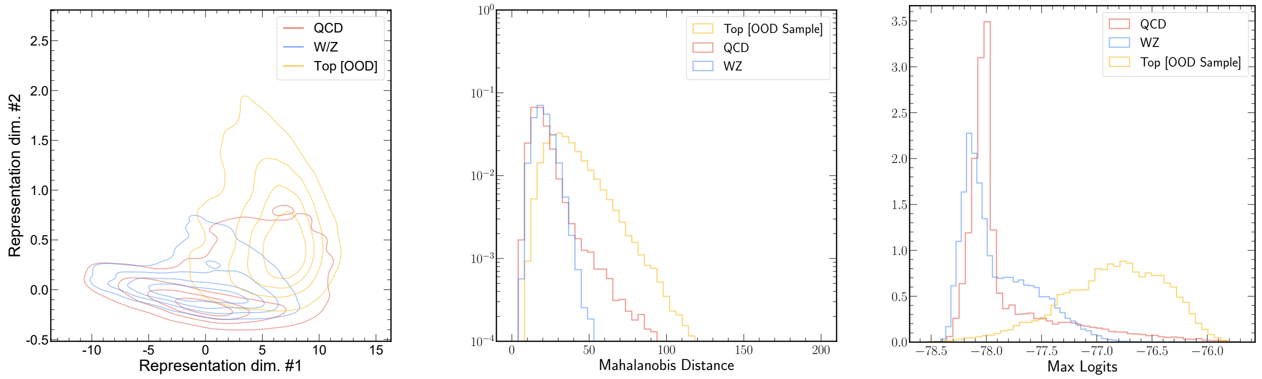
Method	AUROC (↑)	JSD (↓)	L2 WD (↓)	SI (↑)
VAE	0.881	0.255	34.3	2.03
nurd-ml	**0.914**	**0.168**	**24.4**	**2.32**
nurd-md	**0.884**	**0.118**	**19.1**	**2.23**

Table 1. Main results comparing the baseline with proposed methods. We look at the area under the ROC curve (AUROC) when classifying in-distribution from out-of-distribution samples; the Jensen-Shannon Divergence (JSD) and L2 Wasserstein distance (L2 WD) between the QCD mass distribution before and after filtering to the 50% most anomalous inputs; and the maximum significance improvement (SI). We bold all results better than baseline and underline the best performance.

Finally, we visually assess the quality of our learned representation. We perform a Principle Component Analysis (PCA) on the 20-dimensional representation and visualize the results of the first two principal components in Figure 4(a). We observe that the representations of the out-of-distribution top-quark jets are well-separated from those of the in-distribution QCD and W/Z jets. The Mahalanobis distance computed from these representations also show a clear separation of OOD samples from the backgrounds (Figure 4(b)), as does the Max Logit scores (Figure 4(c)).

6 Summary

We present a new approach for anomaly detection in high energy physics via representation learning from multiple background distributions. Our approach takes advantage of more data than existing data-driven approaches by considering multiple jet types as well as classification labels distinguishing them. In addition, we take inspiration and motivation from the out-of-distribution detection and spurious correlations literature in machine learning to develop and motivate decorrelation in the multi-background setting. Our experiments illustrate empirically that the proposed robust multi-background representation learning approach yields better detection and decorrelation than state-of-the-art VAE-based detection[37] while keeping architectural decisions fixed.



(a) Representations under first two principal components.

(b) The Mahalanobis distance from the representations.

(c) Max Logit score for backgrounds and OOD

Figure 4. (a) PCA on the learned representations show that they yield good separation of the out-of-distribution top process and the in-distribution QCD and W/Z processes. This leads to good separation of the downstream anomaly detection scores, Mahalanobis distance (b) and Max Logit (c).

Based on these results, we encourage future work in anomaly detection for high-energy physics to consider multi-background approaches to improve the potential for discovery of new science.

Acknowledgements

AG, JN, and NT are supported by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the Department of Energy (DOE), Office of Science, Office of High Energy Physics. JN is also supported by the U.S. Department of Energy (DOE), Office of Science, Office of High Energy Physics “Designing efficient edge AI with physics phenomena” Project (DE-FOA-0002705). JN, NT are also supported by the DOE Office of Science, Office of Advanced Scientific Computing Research under the “Real-time Data Reduction Codesign at the Extreme Edge for Science” Project (DE-FOA-0002501). This work was supported in part by the AI2050 program at Schmidt Futures (Grant G-23-64934). NT is also supported by the DOE Early Career Research program under Award No. DE-0000247070. LZ, AP, and RR are partly supported by NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, NSF CAREER Award 2145542, ONR N00014-23-1-2634, and Google. AP is also partly supported by the Apple Scholars in AI/ML PhD fellowship.

Author contributions statement

References

1. Chatrchyan, S. *et al.* Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716**, 30, DOI: 10.1016/j.physletb.2012.08.021 (2012). 1207.7235.

2. Chatrchyan, S. *et al.* Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *JHEP* **06**, 081, DOI: 10.1007/JHEP06(2013)081 (2013). 1303.4571.

3. Aad, G. *et al.* Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716**, 1, DOI: 10.1016/j.physletb.2012.08.020 (2012). 1207.7214.

4. Fukuda, Y. *et al.* Evidence for oscillation of atmospheric neutrinos. *Phys. Rev. Lett.* **81**, 1562, DOI: 10.1103/physrevlett.81.1562 (1998).

5. Aguillard, D. P. *et al.* Measurement of the Positive Muon Anomalous Magnetic Moment to 0.20 ppm. *Phys. Rev. Lett.* **131**, 161802, DOI: 10.1103/PhysRevLett.131.161802 (2023). 2308.06230.

6. Barbier, R. *et al.* R-parity violating supersymmetry. *Phys. Rept.* **420**, 1–202, DOI: 10.1016/j.physrep.2005.08.006 (2005). hep-ph/0406039.

7. Sirunyan, A. M. *et al.* Search for pair-produced three-jet resonances in proton-proton collisions at $\sqrt{s}$ =13 TeV. *Phys. Rev. D* **99**, 012010, DOI: 10.1103/PhysRevD.99.012010 (2019). 1810.10092.

8. Park, S. E., Rankin, D., Udrescu, S.-M., Yunus, M. & Harris, P. Quasi anomalous knowledge: searching for new physics with embedded knowledge. *J. High Energy Phys.* **2021**, DOI: 10.1007/jhep06(2021)030 (2021).

9. Dillon, B. M., Mastandrea, R. & Nachman, B. Self-supervised anomaly detection for new physics. *Phys. Rev. D* **106**, DOI: 10.1103/physrevd.106.056005 (2022).

10. Canelli, F. *et al.* Autoencoders for semivisible jet detection. *J. High Energy Phys.* **2022**, DOI: 10.1007/jhep02(2022)074 (2022).

11. Hallin, A. *et al.* Classifying anomalies through outer density estimation. *Phys. Rev. D* **106**, DOI: 10.1103/physrevd.106.055006 (2022).

12. Farina, M., Nakai, Y. & Shih, D. Searching for new physics with deep autoencoders. *Phys. Rev. D* **101**, 075021, DOI: 10.1103/PhysRevD.101.075021 (2020).

13. Heimel, T., Kasieczka, G., Plehn, T. & Thompson, J. M. QCD or what? *SciPost Phys.* **6**, 030, DOI: 10.21468/SciPostPhys.6.3.030 (2019).

14. Aad, G. *et al.* Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector. *Phys. Rev. Lett.* **125**, 131801, DOI: 10.1103/PhysRevLett.125.131801 (2020). 2005.02983.

15. Aad, G. *et al.* Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector (2023). 2306.03637.

16. Button, J. *et al.* Pion-pion interaction in the reaction $\overline{p} + p \to 2\pi^+ + 2\pi^- + n\pi^0$. *Phys. Rev.* **126**, 1858–1863, DOI: 10.1103/PhysRev.126.1858 (1962).

17. Dolen, J., Harris, P., Marzani, S., Rappoccio, S. & Tran, N. Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure. *J. High Energy Phys.* **2016**, DOI: 10.1007/jhep05(2016)156 (2016).

18. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *Int. Conf. on Learn. Represent.* (2014).

19. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning* (2014).

20. Dinh, L., Krueger, D. & Bengio, Y. Nice: Non-linear independent components estimation. *ArXiv* (2015).

21. Dinh, L., Sohl-Dickstein, J. & Bengio, S. Density estimation using real nvp. *ICLR* (2017).

22. Workman, R. L. & Others. Review of Particle Physics. *PTEP* **2022**, 083C01, DOI: 10.1093/ptep/ptac097 (2022).

23. Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D. & Lakshminarayanan, B. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136* (2018).

24. Zhang, L., Goldstein, M. & Ranganath, R. Understanding failures in out-of-distribution detection with deep generative models. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 12427–12436 (PMLR, 2021).

25. Salehi, M. *et al.* A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051* (2021).

26. Puli, A., Zhang, L. H., Oermann, E. K. & Ranganath, R. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. *ICLR* (2022).

27. Zhang, L. H. & Ranganath, R. Robustness to spurious correlations improves semantic out-of-distribution detection. *AAAI* (2023).

28. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives (2014). 1206.5538.

29. Sugiyama, M., Suzuki, T. & Kanamori, T. *Density Ratio Estimation in Machine Learning* (Cambridge University Press, 2012).

30. Hendrycks, D. *et al.* Scaling out-of-distribution detection for real-world settings. *Int. Conf. on Mach. Learn.* (2022).

31. Lee, K., Lee, K., Lee, H. & Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS* (2018).

32. Duarte, J. *et al.* Fast inference of deep neural networks in FPGAs for particle physics. *JINST* **13**, P07027, DOI: 10.1088/1748-0221/13/07/P07027 (2018). 1804.06913.

33. Salehi, M. *et al.* A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges (2021). 2110.14051.

34. Cheng, T. & Courville, A. Invariant representation driven neural classifier for anti-QCD jet tagging. *JHEP* **10**, 152, DOI: 10.1007/JHEP10(2022)152 (2022). 2201.07199.

35. Park, S. E., Harris, P. & Ostdiek, B. Neural embedding: learning the embedding of the manifold of physics data. *JHEP* **07**, 108, DOI: 10.1007/JHEP07(2023)108 (2023). 2208.05484.

36. Louppe, G., Kagan, M. & Cranmer, K. Learning to pivot with adversarial networks. *NeurIPs* (2017).

37. Heimel, T., Kasieczka, G., Plehn, T. & Thompson, J. Qcd or what? *SciPost Phys.* (2018).

38. Kasieczka, G. & Shih, D. Robust jet classifiers through distance correlation. *Phys. Rev. Lett.* **125**, DOI: 10.1103/physrevlett.125.122001 (2020).

39. Dolen, J., Harris, P., Marzani, S., Rappoccio, S. & Tran, N. Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure. *JHEP* **05**, 156, DOI: 10.1007/JHEP05(2016)156 (2016). 1603.00027.

40. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *Adv. neural information processing systems* **29** (2016).

41. Moreno, E. A. *et al.* JEDI-net: a jet identification algorithm based on interaction networks. *The Eur. Phys. J. C* **80**, DOI: 10.1140/epjc/s10052-020-7608-4 (2020).

42. Pierini, M., Duarte, J. M., Tran, N. & Freytsis, M. Hls4ml lhc jet dataset (150 particles), DOI: 10.5281/zenodo.3602260 (2020).

43. Zyla, P. *et al.* Review of Particle Physics. *PTEP* **2020**, 083C01, DOI: 10.1093/ptep/ptaa104 (2020).

44. Tumasyan, A. *et al.* Search for resonant and nonresonant production of pairs of dijet resonances in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP* **07**, 161, DOI: 10.1007/JHEP07(2023)161 (2023). 2206.09997.

Appendix

A Background & Motivation

Here, we provide introductory background on anomaly detection in high-energy physics for non-physics readers. We first describe the common approach of bump hunting (Appendix A.1). Then we describe how existing approaches handle knowledge of multiple known processes (Appendix A.2). Finally, we motivate our proposed approach within the introductory framework described below (Appendix A.3).

A.1 Bump hunting

The existing paradigm of machine learning-based particle discovery typically proceeds as a bump hunt, which can be formalized as a hypothesis test based on the aggregate collection of jets in a given collision and a statistic such as jet mass. Concretely, the test asks whether the mass distribution of jets for a given experiment looks like the expected mass distribution of known background processes P_0, or whether one should reject this null hypothesis in favor of an alternative that additionally considers the presence of a new particle that creates a bump in the otherwise expected mass distribution[44]. This hypothesis test considers jets in aggregate because individual jets alone do not provide sufficient evidence for new physics in the presence of noise from external factors. The test focuses on mass rather than some other combination of statistics so that it can provide actionable information for theories priors which themselves consider mass (e.g., as a key property of a new particle).

So where does an anomaly detection algorithm operating on a per-jet basis come into play? First, considering all jets in aggregate makes it difficult to discern mass deviations from a novel particle that is only present in very small proportions; consequently, it is useful to first filter out as many jets from known processes as possible to allow jets from a potential new particle to make up a larger proportion of the overall distribution. This filtering process is performed via an anomaly detection algorithm: given an anomaly score $\phi : \mathscr{X} \to \mathbb{R}$, a detection score $d : \mathscr{X} \to \mathbb{R}$ is generally of the form $d(\mathbf{x}) = \mathbf{1}[\phi(\mathbf{x}) > \tau]$. The better the algorithm is at filtering out known processes (low false positive rate) while flagging novel processes to keep around in the hypothesis testing phase (high true positive rate), the more power the resulting hypothesis test has to find a deviation and thus a new particle. To maximize the accuracy of the detection algorithm, generally the high-dimensional information describing a jet is incorporated into this algorithm.

A key requirement for the anomaly detection algorithm is that the expected mass distribution under the null hypothesis must still be known for the jets remaining after the filtering step. To meet this requirement, one option is ensure that the distribution of mass does not change after filtering: i.e., $P_0(\mathbf{z}|d(\mathbf{x}) = 1) = P_0(\mathbf{z})$. One way to enforce this constraint is to define mass-based thresholds $\tau_{\mathbf{z}}$ for filtering. Alternatively, one can also force the upstream anomaly detection scoring function to be independent of mass for known jets: $\phi(\mathbf{x}) \perp\!\!\!\perp_{P_0} \mathbf{z}$, implying $d(\mathbf{x}) \perp\!\!\!\perp_{P_0} \mathbf{z}$ and thus $P_0(\mathbf{z}|d(\mathbf{x}) = 1) = P_0(\mathbf{z})$.

In summary, we have the following procedure describing existing bump hunts. Let $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ denote the high-dimensional information, process label, and mass of a jet. Then, we have the following hypothesis test, where $d(\mathbf{x}) \perp\!\!\!\perp_{P_0} \mathbf{z}$:

$$
\begin{aligned}
&H_0 : \{\mathbf{z}|d(\mathbf{x}) = 1\} \sim P_0(\mathbf{z}) \\
&H_A : \{\mathbf{z}|d(\mathbf{x}) = 1\} \nsim P_0(\mathbf{z}).
\end{aligned}
$$

A.2 How existing approaches handle multiple known SM processes

Thus far we have abstracted out the details of this known reference distribution P_0. This distribution is in fact a mixture distribution of several processes, including QCD, W/Z, and Top quark processes; considering just these three processes, we have $P_0 = aP_{\text{QCD}} + bP_{\text{W/Z}} + cP_{\text{Top}}$ with known mixture coefficients. Generally, existing methods build anomaly detection algorithms on QCD jets only. Then, one of two procedures follows: either P_{QCD} is used as an approximation of P_0 in the above procedure, since $a >> b >> c$, or the other mixture components are accounted for in the hypothesis testing phase with a few assumptions, which we discuss next.

There exist very accurate approximations of $P_{\text{W/Z}}(\mathbf{z}), P_{\text{Top}}(\mathbf{z})$, and the distribution of kinematic variables in general for processes other than QCD. Given $P_{\text{W/Z}}(\mathbf{z}), P_{\text{Top}}(\mathbf{z})$ are known and $P_{\text{QCD}}(\mathbf{z})$ can be estimated from simulated data, we can estimate the filtered mass distribution $P_0(\mathbf{z}|d(\mathbf{x}) = 1)$ assuming the detection algorithm satisfies decorrelation within each process, i.e., $d(\mathbf{x}) \perp\!\!\!\perp \mathbf{z}|\mathbf{y}$ for each $\mathbf{y} \in \mathscr{Y}_{\text{known}}$:

$$P_0(\mathbf{z}|d(\mathbf{x})=1) = \sum_{\mathbf{y}\in\mathscr{Y}_{\text{known}}} P(\mathbf{z}, d(\mathbf{x})=1, \mathbf{y})/P(d(\mathbf{x}=1)) \tag{5}$$

$$= \sum_{\mathbf{y}\in\mathscr{Y}_{\text{known}}} P(\mathbf{z}|d(\mathbf{x})=1, \mathbf{y})P(d(\mathbf{x})=1|\mathbf{y})P(\mathbf{y})/P(d(\mathbf{x}=1)) \tag{6}$$

$$= \sum_{\mathbf{y}\in\mathscr{Y}_{\text{known}}} P(\mathbf{z}|\mathbf{y})P(d(\mathbf{x})=1|\mathbf{y})P(\mathbf{y})/P(d(\mathbf{x}=1)) \tag{7}$$

$$= \frac{aP_{\text{QCD}}(\mathbf{z})P(d(\mathbf{x})=1|\mathbf{y}=\text{QCD}) + bP_{\text{W/Z}}(\mathbf{z})P(d(\mathbf{x})=1|\mathbf{y}=\text{W/Z}) + cP_{\text{Top}}(\mathbf{z})P(d(\mathbf{x})=1|\mathbf{y}=\text{Top})}{\int P(\mathbf{z}, d(\mathbf{x}=1))d\mathbf{z}}. \tag{8}$$

$$= \frac{aP_{\text{QCD}}(\mathbf{z})P(d(\mathbf{x})=1|\mathbf{y}=\text{QCD}) + bP_{\text{W/Z}}(\mathbf{z})P(d(\mathbf{x})=1|\mathbf{y}=\text{W/Z}) + cP_{\text{Top}}(\mathbf{z})P(d(\mathbf{x})=1|\mathbf{y}=\text{Top})}{aP(d(\mathbf{x})=1|\mathbf{y}=\text{QCD}) + bP(d(\mathbf{x})=1|\mathbf{y}=\text{W/Z}) + cP(d(\mathbf{x})=1|\mathbf{y}=\text{Top})}. \tag{9}$$

In other words, one can accurately estimate the mass distribution of flagged jets under the null by knowing 1. the distribution of mass for each process, 2. the relative proportions of each process for the overall experiment, and 3. the detection algorithm's false positive rates for each known process type, assuming decorrelation is met. Note that all of 1, 2, 3 are either known through theory / accurate simulation or can be estimated, e.g. $P(d(\mathbf{x})=1|\mathbf{y})$ can be estimated by evaluating the detection algorithm on training data. Then, a key lever for improving the probability of successful scientific discovery is reducing the false positive rate of flagging known processes as anomalies (while successfully flagging true anomalies). Additionally, while existing works assume $d(\mathbf{x}) \perp\!\!\!\perp \mathbf{z}|\mathbf{y}$ for each $\mathbf{y} \in \mathscr{Y}_{\text{known}}$, they do not enforce it directly since the detection algorithms are trained on QCD jets only.

A.3 An alternative approach to increase power and reduce false positive discovery

As with any test, the goal of the above process is to have a test with high power and a low false discovery rate (FDR). The former goal of high power is achieved when the anomaly detection algorithm has a high true positive rate of flagging anomalies and a low false positive rate of flagging existing processes as anomalous. The result is then that novel jets make up a larger proportion of the flagged jets, making it easier for any test to detect their presence (in the form of a deviation from the expected mass distribution of known processes). The latter goal of low FDR is the purpose of decorrelation, to ensure that any deviations in the mass distribution are a result of new unknown processes rather systematic biases in the detection algorithm which "sculpt" the mass distribution away from its expected shape. In this work, we introduce robust multi-background anomaly detection and show how it can help both increase power and reduce false positive discoveries. Below, we provide a brief summary of the source of the gains.

First, rather than training a detection algorithm on QCD jets alone and assuming that other jets are consistently flagged as non-QCD, multi-background anomaly detection algorithms learn from data from all known jets to improve the rate at which all are filtered out. Representation learning guides what information is most important to learn, and in this work, we focus on the information that distinguishes known jets under the hypothesis that this information is most important to distinguish existing jets from new jets as well. If this is true, then such multi-background representation learning can increase overall power of the hypothesis test of interest.

Next, rather than assuming decorrelation is achieved across all jet types, we directly enforce decorrelation within each known background process, i.e. $d(\mathbf{x}) \perp\!\!\!\perp \mathbf{z}|\mathbf{y}$ for each $\mathbf{y} \in \mathscr{Y}_{\text{known}}$. Doing so ensures that the distribution of a kinematic variable such as jet mass is the same for a given background process whether we are considering flagged jets or all jets, i.e. $P(\mathbf{z}|\mathbf{y}, d(\mathbf{x})) = P(\mathbf{z}|\mathbf{y})$. This equivalence in the distribution of jet mass makes it possible to filter down the mass distribution of flagged jets further based on knowledge of $P(\mathbf{z}|\mathbf{y})$ obtained from accurate simulations. To achieve this decorrelation, we enforce an independence constraint on the representations we learn. Then, detection algorithms based on these representations are less susceptible to false discoveries than algorithms that do not enforce that decorrelation for all background processes.