Using Machine Learning for Model-Independent New Physics Discovery at the Large Hadron Collider

Luc Le Pottier

Department of Physics

University of Michigan

Supervisor

Professor Christine Aidala

In partial fulfillment of the requirements for the degree of

Bachelor of Science in Physics, Honors

May 1, 2021

Acknowledgements

This thesis was only possible with the help of many advisors, fellow students, friends, and colleagues. Thanks to Wolfgang Lorenzon, Tancredi Carli, Florencia Canelli, Annapaola de Cosa, Maurizio Pierini, Christine Aidala, and Benjamin Nachman for the supervision, sponsorship, and opportunities for collaboration they offered during my time in their research groups. Thanks also to fellow undergraduates Kees Benkendorfer, Alexis Mulski, Murali Saravanan, Alex Mieth, and Alexa Zielinski – their collaboration, advice, and friendship has helped immensely in completing this work. The funding for this research was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI), the University of Michigan Weiser Center for Emerging Democracies under the Summer Fellowship program, the University of Michigan International Institute under the II Student Fellowship program, and the Swiss Office of Science, Technology, and Higher Education at the Embassy of Switzerland in Washington, D.C. under the *ThinkSwiss* summer research fellowship.

In loving memory of Thomas Edens

Abstract

Despite compelling theoretical and observational evidence for new physics phenomena, direct searches at the Large Hadron Collider (LHC) have not yet discovered any physics Beyond the Standard Model (BSM). While the sensitivity and reach of direct searches for BSM models at the LHC has been steadily increasing, there is a growing need for search methods capable of discovery in unexpected scenarios. A promising avenue of recent research to this end has used machine learning to construct new model-independent search methods. In this thesis, I present an analysis of several such methods ranging over a variety of model dependencies, finding that a measured combination of approaches has the potential to broadly increase the reach of the new physics search program at the LHC.

Contents

Chapter 1

Introduction

1.1 Background

In 2012, both general purpose detector groups at the Large Hadron Collider (LHC) – the Compact Muon Solenoid (CMS) and A Toroidal LHC ApparatuS (ATLAS) groups – announced that they had discovered a new particle "consistent with the Higgs Boson" [1, 2]. In the subsequent months both groups refined their findings, providing the first complete experimental confirmation of the Standard Model (SM). But despite this "completeness," there exists a wide variety of evidence for the existence of new physics phenomena. In particular, many of the theorized resolutions to both the Hierarchy problem and the nature and origin of Dark Matter (DM) predict new physics visible at the energy scale of the LHC [3].

Many LHC experiment analyses have shifted their focus towards probing for such new physics, with major search efforts being undertaken at all LHC experiments [4–10]. With improvements in analysis techniques such as machine learning (ML), these efforts are continuing to see improved sensitivity [11–14], though most of them still target a particular class of Beyond the Standard Model (BSM) physics models. While such searches are well-motivated, their sensitivity is largely constrained to the small subset of all BSM physics models which appear like the one they're searching for. While it is crucial to continue searching for specific BSM typologies, there is a growing need for the development of new search strategies with sensitivity to multiple unexpected scenarios.

1.2 Anomaly Detection

A variety of new model-agnostic search routines have been developed to search for unexpected BSM physics, leveraging recent developments in ML to construct new physics-specific ML methods [15–43]. Such methods are generally referred to as *anomaly detection* (AD) methods. In general, a viable search for new physics must be both **(1)** sensitive to new physics and **(2)** able to estimate the Standard Model background [31]. These criteria together determine a search method's degree of model dependence, or reliance on signal/background models.

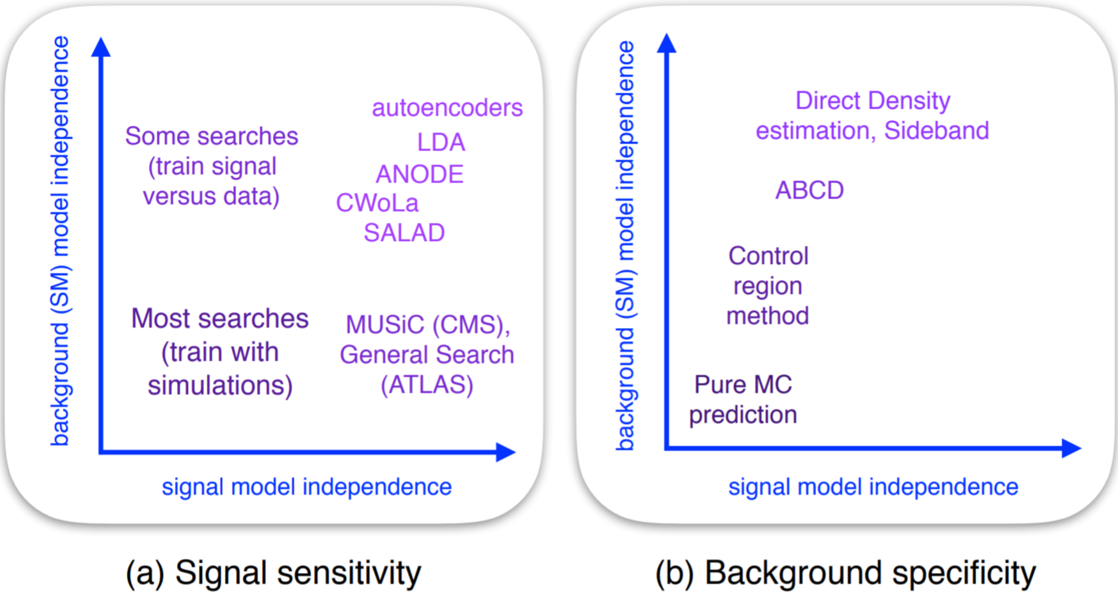


Figure 1.1 Signal sensitivity and background specificity plots for a number of new physics searches, from Ref. [31]. The Model Unspecific Search for New Physics (MUSiC) and General Search strategies are CMS and ATLAS approaches to model-independent searches, respectively. Latent Dirichlet Allocation (LDA), ANOmaly detection with Density Estimation (ANODE), Classification Without Labels (CWoLa), Simulation Assisted Likelihood-free Anomaly Detection (SALAD), and Direct Density Estimation are all general AD techniques.

In terms of signal sensitivity, the majority of searches at the LHC simulate both the signal and background, indicated by the lower left hand corner of Figure 1.1a. Some use unlabeled data and simulated signals, shown in the upper left hand corner of Figure 1.1a, and are called *background model agnostic* methods. A large number of methods additionally have been developed which are signal model agnostic and background model dependent, shown in the lower right hand corner of of Figure 1.1a. These methods generally involve the comparison of data to SM background simulations, and have been completed by a number of experimental collaborations, from the D0 collaboration [44] to the CMS and ATLAS groups [45, 46]. Lastly, a variety of recent approaches have attempted to reach both signal and background model independence, with varying degrees of success. Such methods are the focus of this thesis.

A variety of background estimation methods – the second component of AD routines – are shown in Figure 1.1b. In the lower left hand corner are pure Monte Carlo predictions and the Control Region method, where searches are complemented by a measurement to constrain the simulation. Both methods have very low background model independence. The *ABCD* method (see e.g. Refs. [47, 48]) has slightly higher model independence than those methods more directly using Monte Carlo data. Lastly, the *Sideband* method requires only that the background be smooth in the region where a potential signal exists, such that some function (usually parametric) can be fit to the sidebands and compared to the observed data in the signal region. To split data into signal and sideband regions, the sideband method requires that the new physics signal be *resonant*, i.e localized in some feature such as jet mass. The sideband method, otherwise known as the "bump hunt," is a common discovery method for particle physics – see e.g. Refs. [49, 50].

This thesis is organized as follows. Chapter 2 provides an analysis using neural Autoencoders (AEs) to search for a particular signal topology with unconstrained BSM model parameters, done in collaboration with ETH Zurich Professor Annapaola de Cosa and CERN Research Physicist Maurizio Pierini. Chapter 3 presents a direct comparison of several weakly supervised approaches to AD, quantifying their signal sensitivity, background estimation, and robustness to decorrelation. This section, completed in collaboration with LBNL Staff Scientist Benjamin Nachman, also presents a new AD method, improving on previous approaches with a loss function modification. Lastly, chapter 4 provides a broad outlook on the results of the thesis and avenues of possible future work.

Chapter 2

Autoencoders

2.1 Overview

2.1.1 Background

The usage of neural AEs as an AD method is rapidly gaining popularity in physics [20, 23, 24, 39] after having been a useful tool in computing systems analysis for some time [51]. Despite broad claims of signal sensitivity based on BSM benchmark processes, there have been few full analyses of the ability of AEs to detect a specific signal topology with unknown parameters. This chapter provides such an analysis for the case of Semi Visible jet detection. It also incorporates the EnergyFlow variables introduced by Ref. [52], which prove to be highly useful in discriminating between the QCD background and our signal.

This chapter begins with an overview of Semi-Visible jets in section 2.1.2, followed by a technical description of the data generation, selection, and preprocessing in section 2.2. Our results are presented in section 2.4, along with a discussion of the implication and possible further work in section 2.5.

2.1.2 Semi-Visible Jets

The existence of dark matter provides one of the strongest motivations for increasing BSM search sensitivity at the LHC. Assuming that dark matter takes on a particle form and that the particle is neutral and stable, it is generally thought to escape detection in LHC experiments and manifest as missing transverse energy $\not\!\!E_T$. While a number of searches at the LHC have aimed to expose this dark sector by tagging events with significant $\not\!\!E_T$ [53], relaxing the assumption that the dark sector is weakly coupled allows for an entirely new class of dark matter signatures which appear similar to hadronic jets. Such signals, dubbed *semi-visible* jets by Ref. [54], have a relatively straightforward discovery strategy and have been the subject of a number of direct LHC searches [54–56].

Under this model, semi-visible jets would be one visible signature at the LHC for a large number of dark-sector BSM theories. In general these theories may be simply described as a model for a $U(1)'$ gauge boson coupled to the SM baryon current J^{μ}_{SM}, as per [54]:

$$\mathcal{L} \supset -\frac{1}{4}Z'^{\mu\nu}Z'_{\mu\nu} - \frac{1}{2}M^2_{Z'}Z'_{\mu}Z'^{\mu} - g^{\text{SM}}_{Z'}Z'_{\mu}J^{\mu}_{\text{SM}} \tag{2.1}$$

A number of decays are possible for the Z' in this formulation, as detailed in [54], but only a few have a direct impact on the resultant observables. Here we consider two primary variables: the mass of the Z' boson $M_{Z'}$ and the average fraction of stable hadrons produced in the jet shower r_{inv}, defined as

$$r_{\text{inv}} \equiv \left\langle \frac{\#\text{ of stable hadrons}}{\#\text{ of hadrons}} \right\rangle \tag{2.2}$$

The subscript "inv" is shorthand for "invisible," as r_{inv} characterizes the average fraction of stable dark sector particles and dark sector particles are considered functionally "invisible" to detectors at the LHC by Ref. [54]. This variable roughly parameterizes the missing transverse energy $\not\!\!E_T$.

The parameter space of possible semi-visible jet models is 2-dimensional, with $r_{\text{inv}} \in [0, 1]$ and $M_{Z'} \approx 1$ TeV. Since the semi-visible jet signature might show up for a number of BSM dark sector models, there are few methods of further constraining the parameters – meaning that there is a large parameter space to cover for direct searches. This characteristic of the semi-visible jet DM model makes it an excellent candidate for a model-independent analysis, which I present here.

2.2 Data

2.2.1 Generation

Simulated QCD multijet production data are generated using Pythia8 [57] for event generation and Delphes [58] for an approximation of the detector response, modeled after the CMS upgrade design. Collisions are simulated with two proton beams, each with an energy of 6.5 TeV, which generate approximately 20 proton-proton collisions per crossing. Data is generated in batches of 10^4 events, with 350 total batches, for a total of $3.5 \cdot 10^6$ events before analysis cuts. Additionally, 30 unique semi visible jet signals were generated by the same event generator and detector simulation, for all combinations of the constraining parameters $M_{Z'} \in [1.5,\ 2,\ 2.5,\ 3,\ 3.5,\ 4]$ TeV and $r_{\text{inv}} \in [0.15,\ 0.30,\ 0.45,\ 0.60,\ 0.75]$.

To plot the reconstruction and error distributions of our models, we add portions of all signal samples together in equal proportion to make a single "signal" dataset, labeled as such in the plots which follow. In all other cases we analyze model efficacy on a per-signal basis, occasionally combining all signals with the same stable hadron fraction r_{inv}.

2.2.2 Event Selection

The training sample is prepared by applying a set of selection cuts aimed at isolating semi-visible jet signals, consistent with the cuts outlined in Ref. [54]. These cuts are made at the event level and focus on dijets with high-p_T constituent jets, a high transverse mass M_T, a small distance in pseudorapidity η between leading and subleading jets, and a high ratio of $\not{E}_T/M_T$ as a missing energy requirement. Transverse mass is chosen because it incorporates the missing transverse energy; it is defined as

$$M_T^2 = m_{JJ}^2 + 2\left(\sqrt{m_{JJ}^2 + p_{T_{JJ}}^2}\,\not{E}_T - \vec{p}_{T_{JJ}} \cdot \vec{\not{E}}_T\right) \tag{2.3}$$

The cuts are as follows:

- Require the event to have no isolated leptons. In particular, we cut events having leptons with $p_T > 10$ GeV, $|\eta| < 2.4$, and isolation $I < 0.4$, suppressing events from vector Bosons + jets and leptonic $\mathrm{t\bar{t}}$.

- 2 or more jets, with the leading dijet pair of component jets i,j having
  - $|\eta_{i,j}| < 2.4$
  - $|\Delta\eta| < 1.5$ for the dijet
  - $p_{T_{i,j}} > 200$
  - $M_T > 1500$ for the dijet
  - $\not{E}_T/M_T > 0.25$ for the dijet

where η is the pseudorapidity, $\Delta\eta_{ij}$ is the absolute dijet η difference, p_T is the transverse momentum for each jet, $\not{E}_T$ is the missing momentum in the transverse plane for the event, and M_T is the transverse mass of the leading dijet. Lepton isolation is computed for each lepton ℓ as

$$\text{Isolation} = \frac{\sum_{p\neq\ell} p_T^p}{p_T^\ell} \tag{2.4}$$

where the index p is all photons, charged particles, and neutral hadrons in a cone of size $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$ from ℓ. The event selection is implemented in the ROOT [59] data analysis framework, and yielded approximately 52k QCD background (1.49% selection efficiency) and an average of 16k signal events per set of parameters (1.70% avg. selection efficiency). The QCD background events which pass event selection cuts are then an estimate of the QCD background which will end up in our signal sample after the analysis cuts, meaning they are the background we will train on when searching for Semi-Visible jets.

2.2.3 Feature Selection

After event selections, a subset of characteristic features are extracted from the full data set on which to train our models. Since we want to find anomalies on the basis of tagging individual anomalous jets rather than anomalous events, we choose to save features specific to each jet in the dijet pair of selected events.

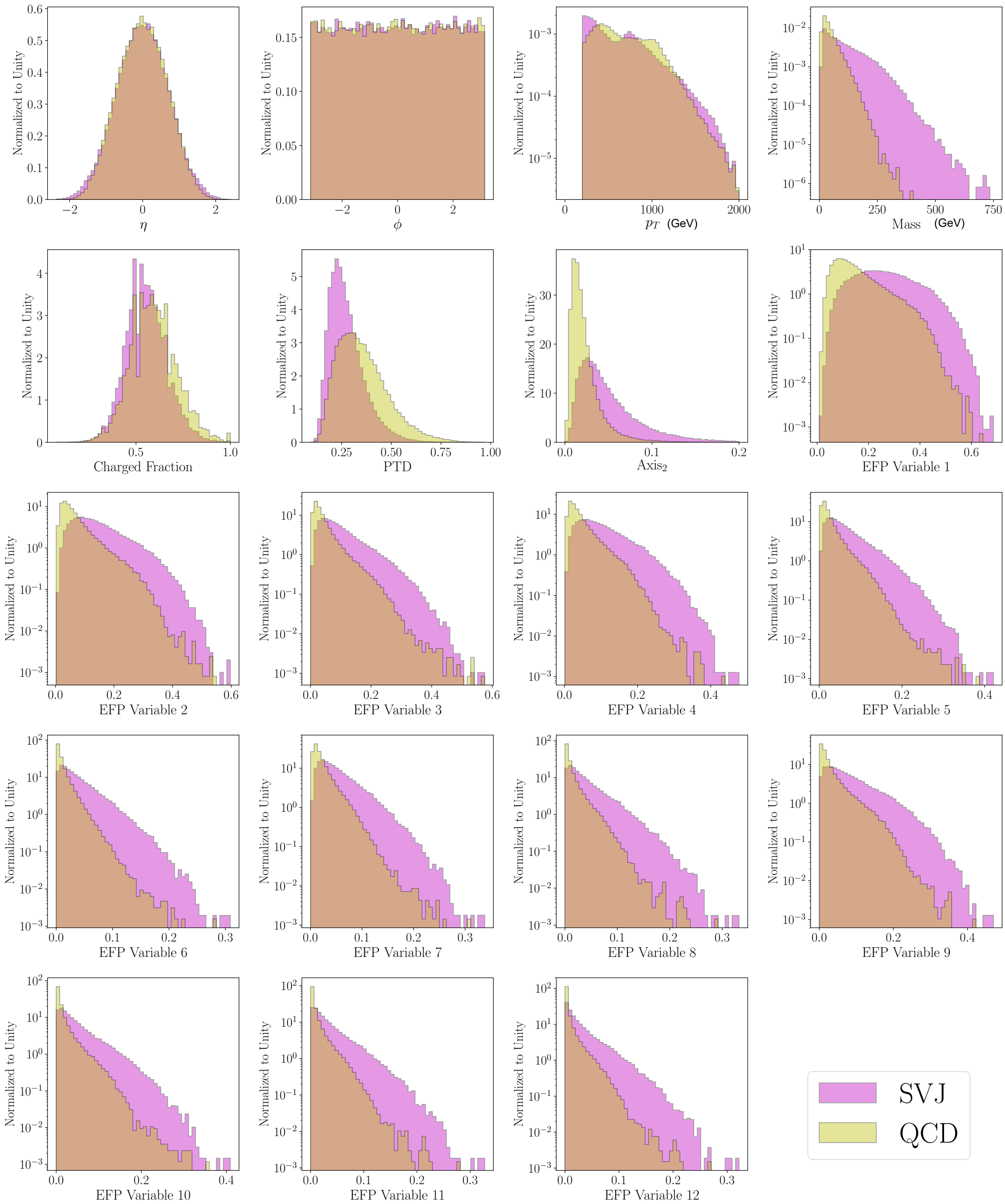


Figure 2.1 Input feature distributions for semi-visible jet signal (SVJ) and QCD background (QCD) jets. Here SVJ shows all signal models combined and normalized.

For this purpose we decided to use a number of the high-level features reconstructed by DELPHES, rather than using lower-level features (i.e. the momentum and position of jet constituents). This decision was primarily motivated by the dramatic decrease in training set size which results from distilling jet constituents into jet observables. In our case, this choice reduced the number of feature inputs by about an order of magnitude. In addition to allowing quicker prototyping, training, and testing times, doing this reduces hadronization model dependence by reducing our reliance on the detailed distributions of hadrons in our jets provided by the DELPHES generator. We chose a number of common jet features, listed below:

- Psuedorapidity η
- Azmuithal angle ϕ
- Transverse momentum p_T
- Invariant mass M
- Fraction of charged particles among jet constituents, $\frac{N_{charged}}{N_{charged}+N_{neutral}}$
- Jet fragmentation function $p_T D$ [60]
- Jet ellipse minor axis axis$_2$
- 12 significant features resulting from the Energy Flow Polynomial (EFP) basis, calculated for all multigraphs with degree $d \leq 3$ [52]

The first seven of these features are provided directly by reconstructions from DELPHES. The EFP basis is calculated using a list of jet constituents, which itself is obtained by collecting the 4-momenta for all particles and photons in the event which lie within a cone of radius $R = 0.5$ about the jet's axis. EFPs, presented in Ref. [52], provide a linear basis for characterizing jet constituent substructure. We did not do any significant optimization of the input features, except to verify that including the EFPs yielded improved model performance. All high level features for the combined signal and background datasets are shown in figure 2.1, labeled as "SVJ" and "QCD," respectively.

In theory, all of the selected high level features contain some discrimination capability, with the exception of positional variables η and ϕ. These are included to allow us to test the autoencoder's response to simulated detector defects. In principle, an AE continually trained on live data would be able to recognize and ignore anomalous jets in background data by "memorizing" the coordinates of problem spots in the detectors.

2.3 Models

2.3.1 Principle

A neural autoencoder is a neural network which approximates the identity transformation under the influence of some constraint. The idea is that the constraint forces the neural network to learn relationships between features in the dataset. A wide variety of constraints are available depending on the intended purpose of the AE, e.g. limiting the number of active neurons per layer, limiting the types of internal connections, etc. In the most basic neural autoencoder, this constraint takes the form of a *bottleneck* in the network architecture such that the input vector $x \in \mathbb{R}^N$ is fed through a hidden layer of $M < N$ nodes, and then reconstructed in the output vector $\hat{x} \in \mathbb{R}^N$.

This architecture is usually described as the composition of two functions; an *encoding* function $f : \mathbb{R}^N \to \mathbb{R}^M$ which compresses the inputs to an M dimensional vector, and a *decoding* function $g : \mathbb{R}^M \to \mathbb{R}^N$ which reconstructs the input vector using the compression to the best of its ability. Autoencoders are generally trained by minimizing some function computing the similarity of the inputs and the outputs, usually the Mean Squared Error (MSE):

$$L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \tag{2.5}$$

With the correct balancing of training parameters, an autoencoder trained mostly on "normal" data will learn to reconstruct this data (and anything similar to it) quite well. When evaluated on data unlike what it was trained on, however, the reconstruction error will be anomalously high. By cutting on the reconstruction error of unlabeled data, one can in principle select the fraction of the "most anomalous" events, as defined by the autoencoder.

2.3.2 Architecture

After some exploration of the model architectures and parameters, we settled on two hidden layers of 30 nodes each for both the encoder and the decoder, with a bottleneck layer of 8 nodes. The activation function for each layer is the Rectified Linear Unit (ReLu), defined as

$$\textsc{ReLu}(x) = \max(0, x) \tag{2.6}$$

We also explored a number of other autoencoder flavors such as Variational AEs and Sparse AEs, but found that none consistently provided the same performance and ease of training as neural AEs for our case. The architecture we used for our models is shown in figure 2.2

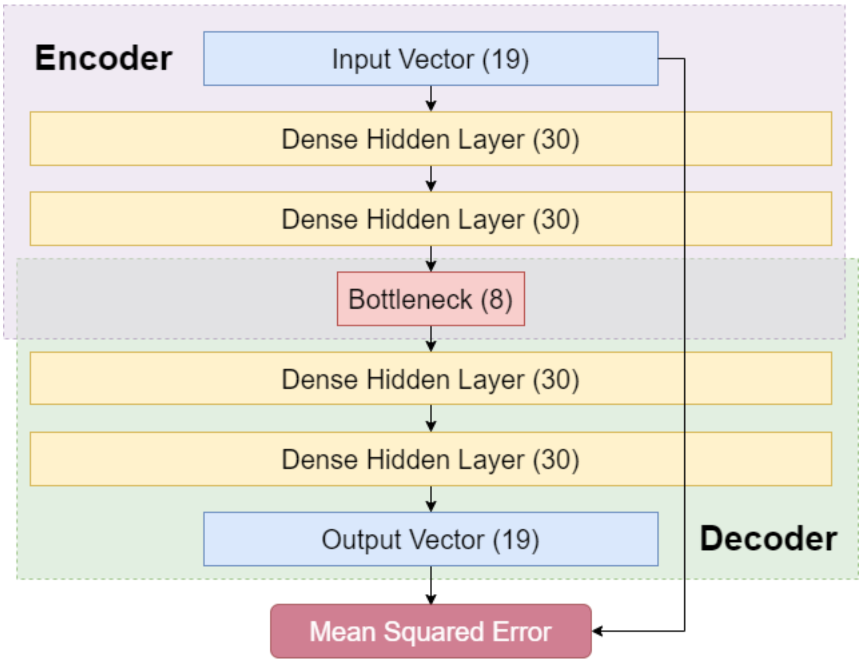


Figure 2.2 Neural network architecture for the neural AEs used in this work. The values in parentheses here indicate the number of nodes in each network layer. The input/output dimensions of 19 nodes each allow us to feed our selected features, and apply the Mean Squared Error between them as our loss function.

2.3.3 Training

For training, we took 80% of the QCD data and calculated a normalization function Norm(x), such that the 25th and 75th percentiles of each feature were scaled to 0 and 1, respectively. This same normalization was applied to all subsequent test and signal data. We found that this normalization routine (1) is robust to outliers in the data and (2) minimizes the effects of random initialization across trials. While we experimented with other normalization schemes, e.g. Min-Max norm and Mu-Sigma scaling, we found this method of percentile scaling to consistently perform best.

To train our models, we developed a training and testing framework using Keras [61] with the Tensorflow [62] backend. We used this framework to perform a Bayesian parameter search across the number of epochs, bottleneck dimension, and optimal learning rate, using the total autoencoder reconstruction error on a test dataset as the minimization metric.

Here the *number of epochs* refers to the number of times that the algorithm will iterate over the entire training dataset, updating the AE's internal model parameters with each sample in the training dataset. In general, model performance increases with increasing epochs, however it may overfit the dataset if trained for too long, "memorizing" the training dataset and performing poorly on any test datasets. Similarly, the *learning rate* is a parameter of the optimization routine Adam, controlling the rate of convergence for the neural network. For too low a learning rate training will be very slow, and for too high a learning rate the loss function will often have trouble converging. The *batch size* specifies the number of training samples to use for each update of the model parameters, averaging the gradients across each batch to provide some regularization and speed up training.

The training procedure which resulted from this and other optimization searches is as follows: models are trained using the Adam [63] optimizer for 100 epochs, using early stopping with a patience of 12 epochs. The learning rate is set to $5.1 \cdot 10^{-4}$, with a batch size of 64. Since optimization is not always deterministic, 200 models are trained and the 10% with the lowest total reconstruction error are selected for evaluation.

While model training should be optimized to minimize the loss rather than maximize

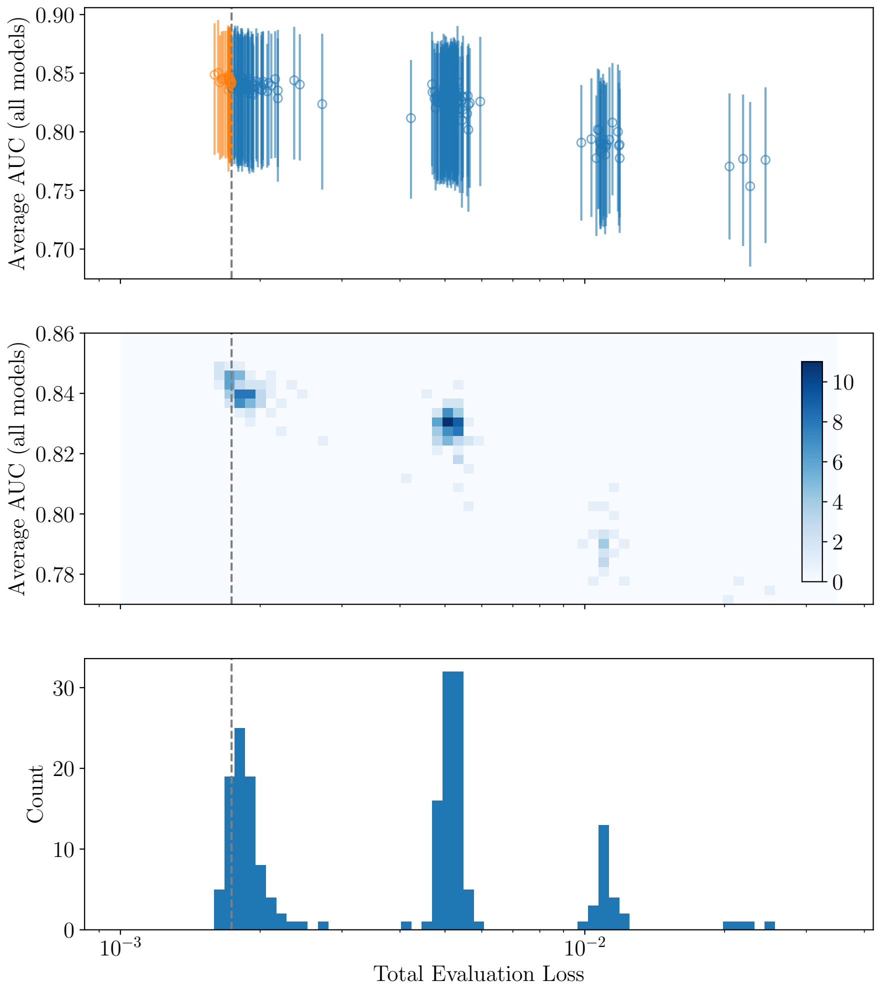


Figure 2.3 Here AE loss on the evaluation dataset is compared to the Area Under the Curve (AUC) values on all signal models. The AUC roughly parameterizes the signal model idenfication capability. In the top plot, the average and min/max AUC for all 30 signals is plotted against the loss for each of the 200 trials. In the center plot, average AUC across all 30 signals is plotted against the loss. In the bottom plot, a histogram of the loss alone is plotted. Grey dotted lines indicate the 10th percentile of evaluation loss.

signal identification, it is interesting to see the relationship between the two. Figure 2.3 shows this relationship for the 200 models, indicating that AEs with identical training parameters generally converge to one of about four evaluation loss outcomes. Additionally, it is clear that AEs with lower evaluation loss generally have better signal identification.

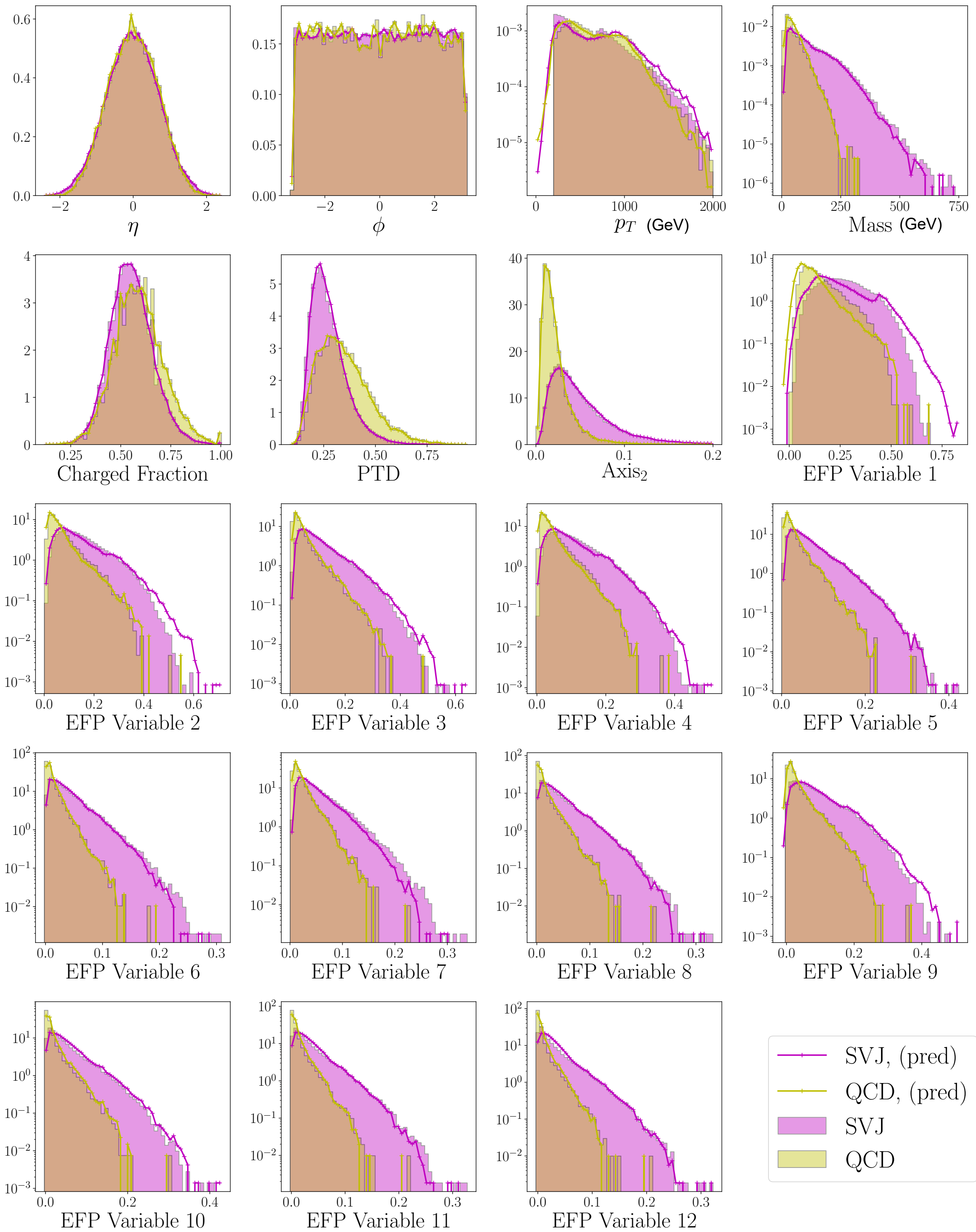


Figure 2.4 Input and reconstructed variable distributions across all jets for each of the input features in this analysis, for both semi-visible jet signal (SVJ) and QCD background (QCD). Here "SVJ" shows all signal models combined in equal proportion and normalized. Input distributions are indicated by filled histograms, and reconstructed distributions are differentiated by their solid line plots and predicate label "pred" in the plot legend.

2.3.4 Code and data availibility

The code for this work can be found at https://github.com/luclepot/autoencodeSVJ. In particular, the selection, data processing, and full final analysis are readily available. The data is currently internal to a CMS project group, however it may be recreated to high accuracy using the parameters given in section 2.2.

2.4 Results

Here we present specific results from one of the top 10% of AE trainings, as described in the previous section.

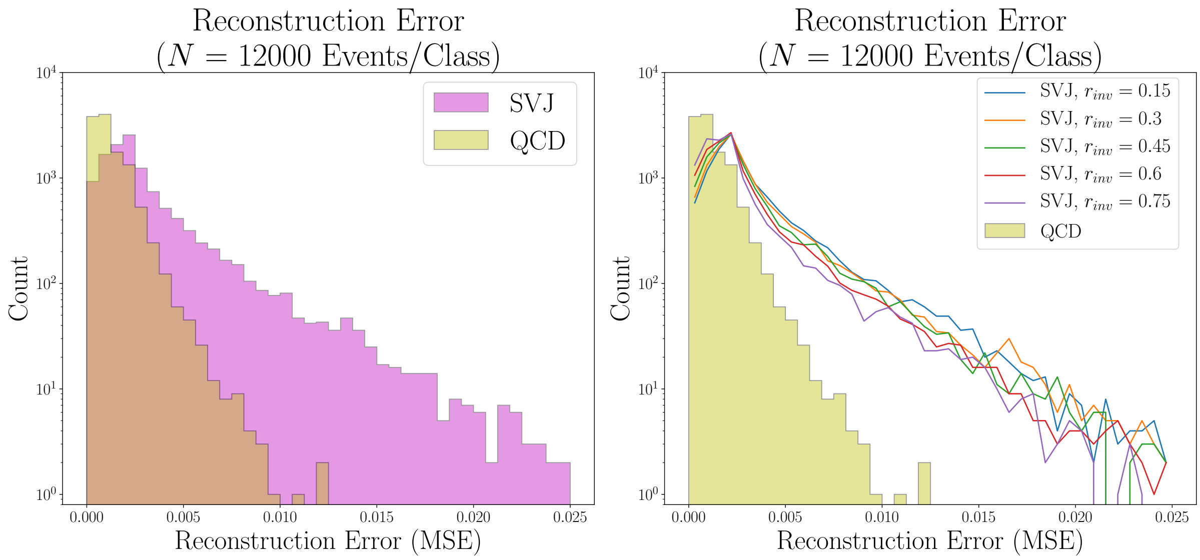
Input and reconstruction distributions are shown for both SVJ and QCD data in Figure 2.4. We can see from these reconstructions that the jet transverse momentum p_T, jet mass, jet PTD, and jet EFP variables have much worse reconstructions than the Charged Fraction and positional variables. We also see that the QCD variable distributions are generally reconstructed much better than the SVJ variable distributions, partially satisfying our intuitive expectations for the autoencoders.

Next we plot the sum of the error distributions across each feature for each sample. The result is shown in figures 2.5a and 2.5b; by inspection it is clear that the signal and background distributions are better separated by this metric than by any of the high level features alone. It is also clear from this plot that the reconstruction error tends to be lower across the entire distribution for higher values of r_{inv}, though this effect is fairly small.

We can also evaluate the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve for the signal for each set of parameters. A ROC curve is one which graphically shows both the True Positive Rate (TPR) and the False Positive Rate (FPR), over a range of cuts on the AE reconstruction error. These are defined as

$$\text{TPR} = \frac{TP}{TP + FN} \quad \text{and} \quad \text{FPR} = \frac{FP}{FP + TN} \tag{2.7}$$

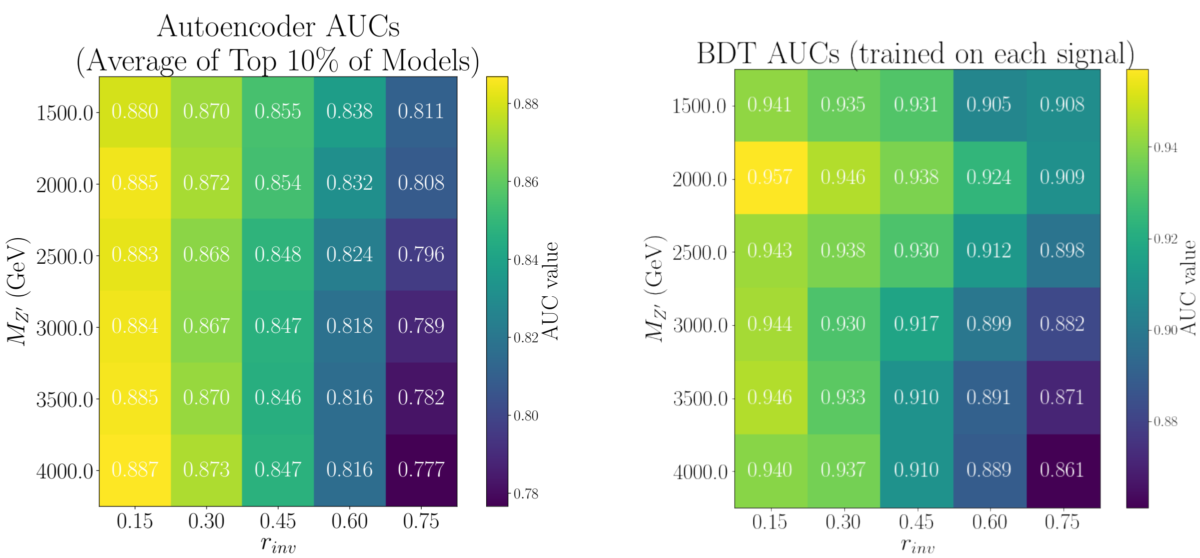
where TP is the number of true positive events (signal passing the cut), FN the number of false negative events (signal not passing the cut), FP the number of false positive events



(a) Combined reconstruction errors

(b) Reconstruction errors split by r_{inv}

Figure 2.5 Reconstruction error distributions, with $N = 12000$ events per class. Here both QCD background jets and Semi-Visible Jet (SVJ) signal jets are shown. At left, we see a plot of the equal proportion combined signal samples, and at right, the same distributions split by the fraction of dark hadrons r_{inv}.



(a) Autoencoder AUC values

(b) Boosted Decision Tree AUC values

Figure 2.6 Area Under Curve (AUC) metric for both Autoencoders and Boosted Decision Trees (BDTs) across all generated signal samples. Here the BDT AUC scores represent that of a BDT trained to recognize each signal, indicating the absolute upper bound on classification performance in this context.

(background passing the cut), and TN the number of true negative events (background not passing the cut). In the case of BSM physics searches, TPR is the signal efficiency, while FPR is background contamination. The AUC metric is the integral over this ROC curve, given formally as a function of the threshold value $T \in \mathbb{R}$ as

$$\mathrm{AUC} = \int_{-\infty}^{\infty} \mathrm{TPR}(T)\mathrm{FPR}'(T)dT \tag{2.8}$$

Figure 2.6a shows the AUC metrics for the average of the top 10% of models, across all signal topologies. The corresponding results for Boosted Decision Trees (BDTs), a strong baseline model for supervised discrimination, are shown for each signal instance in Figure 2.6b. In this case the BDT for each signal was trained directly on that signal, meaning the AUC value for the BDTs are an upper bound for discrimination in any case. In both plots it is clear that lower values of r_{inv} are easier to identify overall, in general both models find good signal identification across all signal topologies. The AE performance is notably good for every signal topology, and the fractional performance decrease from the optimal BDT case appears to be relatively consistent across the signal space.

Figure 2.7 plots the ROC curves for each signal, colored as a function of their stable hadron fraction r_{inv}. The relative separation between levels of r_{inv} is notable, again showing that signals with a lower r_{inv} are generally better detected by our AEs.

Since this method is applied specifically on a jet-by-jet basis, we do not approach a background fit of any event level parameter such as the invariant dijet mass m_{JJ}. Given that the autoencoder and other autoregressive models are generally robust to correlations in the training data [31], and that the signal topology we are studying is generally localized in the invariant dijet mass spectrum, we expect that a background fit on increasingly tight cuts of autoencoder reconstruction error would be appropriate. Such fits are well documented by other semi-visible jet analyses [55, 56]; the purpose of this analysis was to study the pure signal sensitivity of AEs.

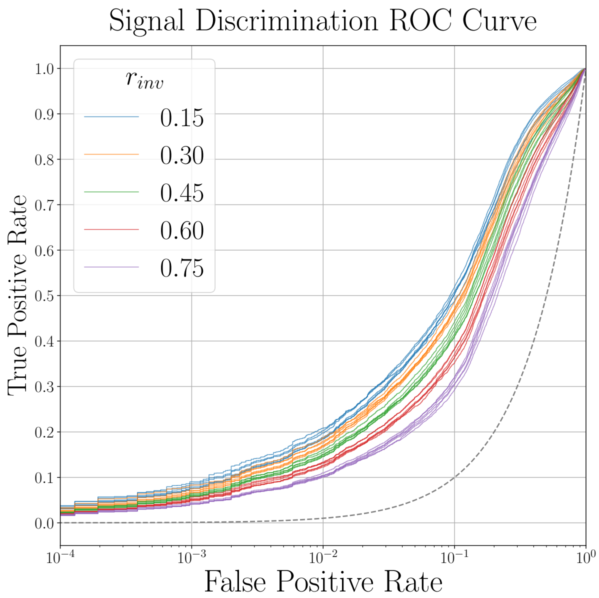


Figure 2.7 ROC Curves for all signals, colored by value of r_{inv}. Each individual ROC curve represents a signal with distinctly different dark hadron fraction r_{inv} and Z' boson mass $M_{Z'}$. The grey dotted line indicates random performance, i.e. the least informative classifier possible. Both high True Positive Rates and low False Positive Rates indicate better performance.

2.5 Outlook

This analysis shows that neural autoencoders can provide excellent signal detection performance when used to search for specific signal signatures, as in this search for semi-visible jets. Future analyses should consider studying the performance in comparison to other unsupervised anomaly detection methods, applying the invariant mass fit and calculating the exact expected significance increases by using AEs, and explore alternative model architectures.

Chapter 3

Weakly Supervised Methods

3.1 Overview

3.1.1 Background

While fully model-independent approaches such as the neural autoencoders discussed in Chapter 2 have been shown to be promising ways of acheiving general sensitivity to new physics signals, it is somewhat more difficult to control their training and anomaly detection qualities. Additionally, without manually constraining the signal topologies as done in chapter 2, the methods are largely limited to identifying individual anomalous events rather than finding consistent excesses in data. One class of model which aims to address this issue are AD methods based on *weakly-supervised* machine learning. This class of models usually has a somewhat higher signal model dependence than the unsupervised methods, as well as a higher degree of interpretability of their results.

This chapter begins with an overview of the need for decorrelation in weakly supervised models, discussed in section 3.1.2. It continues in chapter 3.2 by introducing three weakly supervised methods along with their background fitting method. Finally, it concludes by presenting the signal sensitivity and background specificity of all these methods under a variety of tests in section 3.3, before discussing future avenues of research in section 3.4.

3.1.2 Decorrelation

A promising subset of new weakly-supervised AD techniques targets *resonant new physics*. This is the case where sideband methods can be used to estimate the SM background directly from data, after applying a classifier to enhance signal presence in the data. A growing challenge with this methodology is that in the case where the classifiers are *dependent* on the resonant feature, artificial bumps in the distribution of this resonant feature are formed when applying a sideband fit. This phenomenon comes up often in dijet resonance searches, where the resonant variable is the invariant dijet mass m_{JJ}; in this case, the artificial bump sculpting is often called *mass sculpting*. While many automated decorrelation methods have been proposed to ensure that classifiers are decorrelated from particular features by construction [48, 64–74], they have trouble with performance over a broad range of applications. In particular, application of such decorrelation methods has been found to degrade the signal identification performance of weakly supervised approaches to AD. This is because many localized signals are expected to appear as a dependence between the resonant feature and other descriptive features; in this case, forcing independence would eliminate signal sensitivity. This chapter studies the behavior of multiple weakly supervised models, both alone and in the presence of correlations between the resonant feature and the descriptive features. In particular, two approaches are studied: Classification without Labels (CWoLa) [16–18, 75] and Simulation Assisted Likelihood-free Anomaly Detection (Salad) [30].

CWoLa is a method that does not depend on simulation and achieves signal sensitivity by comparing a signal region with nearby sideband regions in the resonance feature. As a result, CWoLa is particularly sensitive to dependencies between the classification features and the resonant feature. Salad uses a reweighted simulation to achieve signal sensitivity. Since it never directly uses the sideband region, Salad is expected to be more robust than CWoLa to dependencies.

In order to recover the performance of CWoLa in the presence of significant dependence between the classification features and the resonant feature, a new method called simulation augmented CWoLa (SA-CWoLa) is introduced. The SA-CWoLa approach augments the CWoLa loss function to penalize the classifier for learning differences between the signal region and the sideband region in simulation, which is signal-free by construction. All of these methods are investigated on both a gaussian dataset and an LHC physics example, in both the normal case and in the case where induced correlations are present.

3.2 Methods

Let m be a feature set which localizes a potential signal in a signal region (SR) and a sideband (SB) with higher and lower expected signal concentrations, respectively. Let x be an additional set of features useful for isolating this signal. In a prototypical LHC example, m might be the invariant dijet mass of the leading two jets in a proton-proton collision, and x a variety of other jet-level features such as mass, n-subjettiness, etc. We consider here the case where m is a single feature, as it is the simplest and most signal model-independent case.

For a single feature m and $x \in \mathbb{R}^n$, $(m, x) \in \mathbb{R}^{n+1}$, let $f : \mathbb{R}^n \to [0, 1]$ be a classifier parameterized by a neural network. All classifiers described here will ultimately aim to train f for maximum signal sensitivity and minimum response to correlations between m and x.

3.2.1 Classification Without Labels

The idea of CWoLa is to train a classifier $f(x)$ on two mixed samples rather than on direct signal and background samples. It can be proven [75] that a classifier trained to distinguish mixed samples is also optimal for distinguishing signal and background samples, in the asymptotic limit of data and training time. In practice the optimal classifier for both cases is not reached, however this still allows CWoLa excellent performance as it approaches that limit.

In our case, the mixed samples are the data in the SR and the SB, labeled as 1 and 0, respectively. The loss function used to train CWoLa is then given by

$$\mathcal{L}_{\text{CWoLa}}[f] = -\sum_{i \in \text{SR,data}} \log(f(x_i)) - \sum_{i \in \text{SB,data}} \log(1 - f(x_i)) \tag{3.1}$$

simply the binary cross entropy loss between the two classes.

3.2.2 Simulation-Assisted Likelihood-Free Anomaly Detection

The basic principle of the SALAD network is to compare data and simulation in the SB and learn a reweighting function which improves the simulation significantly. This reweighting function is then applied to SR simulation, and a classifier is trained to distinguish data from simulation in the SR. If the reweighting of the simulation is good enough, the classifier f trained to distinguish data from simulation instead learns to distinguish signal from background. This procedure proceeds (as described in [30]) as follows:

1. Let $g(m, x)$ be a function $g : \mathbb{R}^{N+1} \to [0, 1]$. g is then trained using binary cross entropy loss to distinguish data and simulation for $m \notin$ SR. More concretely, g is optimized using the loss function

$$\mathcal{L}_{\text{SALAD}}[g] = -\sum_{i \in \text{SB,data}} \log(g(m_i, x_i)) - \sum_{i \in \text{SB,sim.}} \log(1 - g(m_i, x_i)) \tag{3.2}$$

   It is well known [76–79] that an asymptotically trained g will yield a weight function defined by

$$w(x|m) \equiv \frac{g(m, x)}{1 - g(m, x)} = \frac{p(m, x|\text{data})}{p(m, x|\text{simulation})} \times \frac{p(\text{data})}{p(\text{simulation})} \tag{3.3}$$

2. Next, simulated SR events are reweighted using $w(x|m)$, interpolated automatically by the neural network. A second classifer f is introduced, and trained to distinguish the reweighted simulation from the data in the SR. This classifier is optimized again with binary cross-entropy loss, using the loss function

$$\mathcal{L}_{\text{SALAD}}[f] = -\sum_{i\in\text{SR,data}} \log(f(x_i)) - \sum_{i\in\text{SR,sim.}} w(x_i,m)\log(1-f(x_i)) \tag{3.4}$$

Trained in this way, $f(x)$ is asymptotically related by simple transformations to the optimal classifier:

$$\frac{f(x)}{1-f(x)} \propto \frac{p(x|\text{signal}+\text{background})}{p(x|\text{background})} \tag{3.5}$$

as desired.

3.2.3 Simulation Assisted CWoLa

While CWoLa has been effective for resonant anomaly detection [16, 17] and was recently employed in a low-dimensional analysis by the ATLAS experiment [18], there are in practice differences between the SR and the SB due to dependencies between m and x. This effectively means that CWoLa will only be able to find signals which introduce a bigger difference than is already present in the background.

To address this problem, we propose a modification of the usual CWoLa loss function to utilize simulation information, which we call the simulation-augmented (SA) CWoLa classifer:

$$\begin{aligned}\mathcal{L}_{\text{SA-CWola}}[f] = &-\sum_{i\in\text{SR,data}} \log(f(x_i)) - \sum_{i\in\text{SB,data}} \log(1-f(x_i)) \\ &- \lambda\left(\sum_{i\in\text{SR,sim.}} \log(1-f(x_i)) + \sum_{i\in\text{SB,sim.}} \log(f(x_i))\right)\,,\end{aligned} \tag{3.6}$$

where $\lambda > 0$ is a hyperparameter. In the limit $\lambda \to 0$, the regular CWoLa loss function is recovered, and for $\lambda > 0$ the loss function is *penalized* if it correctly identifies the SR and SB tags of the simulated samples. In the limit $\lambda \lesssim 1$, this is equivalent to penalizing the classifier for distinguishing the SR from the SB in the simulated samples. In the limit $\lambda \to \infty$, the method converges to tagging the simulation SB. With a balanced λ, the classifier is able to efficiently find signal without being too distracted by differences in the SR and SB. In particular, differences between the SR and SB that occur similarly in both the data and the simulation are ignored by this procedure; since signal is by design not present in the simulated sample, this heightens signal sensitivity overall.

3.2.4 Bump Hunt Analysis

In addition to quantifying performance with ROC curves, it is also useful to emulate a proper background estimation based on a bump hunt. To do this, we fit a histogram of the m_{jj} spectrum after applying a threshold on one of the classifiers described above to the following parametric function:

$$\frac{d\sigma}{dm_{jj}} = \frac{p_0\,(1-x)^{p_1}}{x^{p_x + p_3\,\log(x)}}\,, \tag{3.7}$$

where $x = m_{jj}/\sqrt{s}$ and p_i are fit parameters. This function has a long history and has also been recently used by the ATLAS and CMS collaborations (see e.g. [80, 81]). While non-parametric functions such as Gaussian processes [82] are also possible, these are not needed for the demonstration considered here. The SR is masked for the fit and then a p-value between the observed and fitted data is computed. This allows us to form a test statistic from the profile likelihood ratio:

$$\lambda_0 = \frac{\max_\theta p(n|\mu=0,\theta)}{\max_{\mu,\theta} p(n|\mu,\theta)}, \tag{3.8}$$

where n is the number of observed events in the SR and θ is a nuisance parameter from the sideband fit:

$$p(n|\mu,\theta) = \mathrm{Poisson}(n|b+\theta+\mu)e^{-\theta^2/2\sigma^2}, \tag{3.9}$$

where b and σ are the number of events and uncertainty from the sideband fit, respectively. The test statistic itself is $q_0 = -2\log(\lambda_0)$ when the extracted signal strength $(\mu,\theta) =$

$\text{argmax}_{\mu',\theta'} p(n|\mu',\theta')$ is $\mu > 0$ and 0 otherwise. Asymptotic formulae from Wald and Wilks then give the significance $Z = \sqrt{q_0}$ [83–85].

3.2.5 Code and data availability

The code for much of this work can be found at https://github.com/bnachman/DCTRHunting. Additionally, a PYTHON package providing versions of the models discussed here compatible with the popular Scikit-Learn PYTHON module [86] is available at https://pypi.org/project/anomaly-detection-models. The dataset used in section 3.3.2 is available on Zendo at https://zenodo.org/record/2629073#.YGD2I51Kh3g.

3.3 Results

All neural networks were implemented in KERAS [61] with the TENSORFLOW backend [62] and optimized with ADAM [87]. Each network is composed of three hidden layers with 32 nodes each and use the rectified linear unit (ReLU) activation function. The sigmoid function is used after the last layer to force the output to the range [0,1]. Training proceeds for 20 epochs with a batch size of 256.

3.3.1 Gaussian Example

We begin by applying these models to a test example, for which the likelihood ratio is much simpler than in any possible physics example. We generated a simple test dataset of events drawn from two-dimensional Gaussian distributions. The 'data' distributions in the SR and SB were drawn with mean and covariance matricies

$$\mu_{\text{SR}} = (0,0), \quad \mu_{\text{SB}} = (1,0), \quad \sigma_{\text{SR}} = \sigma_{\text{SB}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.10}$$

Signal events had mean and covariance

$$\mu = (1.5, 1.5), \quad \sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}. \tag{3.11}$$

To simulate an imperfect background simulation, 'simulation' events had mean and covariance

$$\mu_{\mathrm{SR}} = (0.1, -0.1), \quad \mu_{\mathrm{SB}} = (1.05, 0), \quad \sigma_{\mathrm{SR}} = \sigma_{\mathrm{SB}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.12}$$

These distributions are plotted in Fig. 3.1. To simulate a strong correlation between the resonant and descriptive features m and x, the signal events are more similar to the SB than the SR.

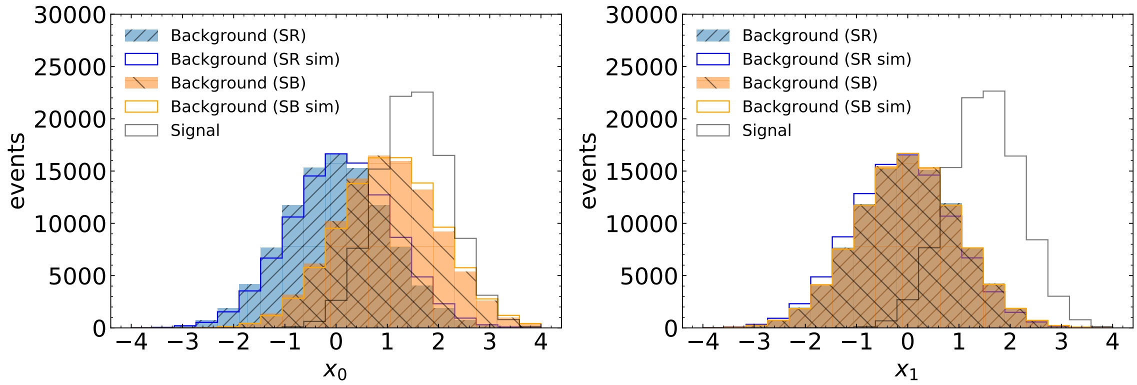


Figure 3.1 The Gaussian distributions generated to test the classification methods. Left: the first feature x_0 Right: the second feature x_1. Here the SR is shown in blue, the SB in orange, and the "data" is distinguished from the "simulation" (indicated here by the term "sim") by the filling and hatching of the former. Note how "data" and "simulation" are fairly similar for all cases, and the SR/SB differences are only significant in the x_1 variable. Also note how the signal is much more like the SB than the SR, in the case of the feature x_0.

To establish performance bounds, fully supervised classifiers were trained to distinguish between signal and data events using both available features and using only the second feature (x_1). For each other test, 500 signal events were injected into signal region samples of 50000 background events, corresponding to a signal significance of about 2.2σ. For non-supervised methods, 20 classifiers were trained, and their performance is exhibited in mean and standard deviation.

The results are displayed in Fig. 3.2. The solid purple line corresponds to the SALAD classifier, the solid orange line corresponds to the SA-CWoLa classifier with hyperparameter $\lambda = 1$, the dashed red line is standard CWoLa, the solid green line corresponds to the fully supervised classifier, and the dashed green line corresponds to the supervised classifier trained only on the second feature x_1. The dashed black line corresponds to a

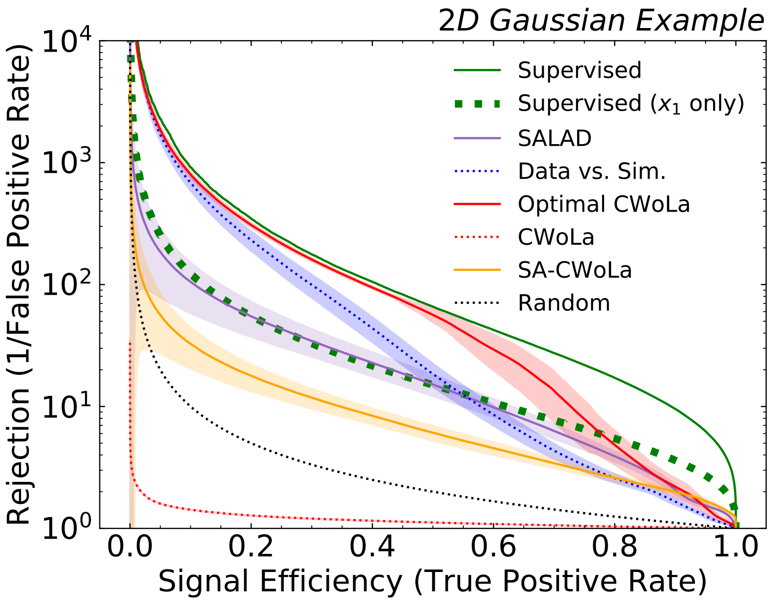


Figure 3.2 Receiver Operating Characteristic (ROC) curves for anomaly detection methods described in the text. Better performance is to the upper right. With the exception of the supervised classifiers, the solid/dashed lines correspond to the mean performance of 20 classifiers; the shaded bands display the standard deviation in performance. The SA-CWoLa hyperparameter was set to $\lambda = 1$.

classifier which randomly assigns values to inputs.

The solid red 'Optimal CWoLa' line displays the results of a classifier that was trained to distinguish a mixed sample of signal-region data and signal events from a pure sample of signal-region data events. It is optimal in the sense that the only difference between the two datasets is the presence of signal in one of them, meaning that the only way for the classifier to minimize its loss is to learn to tag signal events. In principle, no CWoLa-style classifier should be able to outperform the Optimal CWoLa classifier.

The dashed blue line labeled 'Data vs. Sim.' presents the the results of a classifier trained to distinguish the mixed sample of signal-region data and signal events from a pure sample of signal-region simulation events. Since the simulation is statistically similar, but not identical, to the data, the performance is slightly worse than that of Optimal CWoLa. Though it outperforms the SALAD, CWoLa, and SA-CWoLa methods in this case, the sensitivity of the direct data-simulation comparison to the accuracy of the simulation can be a hindrance in higher-dimensional applications, as will be shown in the physics example.

A couple of observations are worth mentioning. First, while the strong similarity between the sideband region and the signal causes CWoLa to consistently anti-tag signal, the SA-CWoLa classifier is able to recover significant signal-tagging ability. Additional tests showed that varying the SA-CWoLa hyperparameter λ had a large impact on performance; in some cases, the SA-CWoLa was able to surpass the signal-tagging performance of the x_1-only supervised classifier. The trade-off is that higher values of λ make it more likely that the classifier will anti-tag signal region background events while tagging sideband background events as signal-like, thereby carving a large deficit in the signal region during a bump hunt. In our case, setting $\lambda = 1$ balances the competing priorities of the classifier, though this may not be desirable in other applications. Note also that, in this example, the SALAD classifier performs strictly better than the SA-CWoLa classifier. This is not always true, as will be seen in the physics example which follows.

3.3.2 Resonant Physics Example

This study took the simulations produced for the 2020 LHC Olympics community challenge [88] as data. This dataset is composed of generic background dijet events, with at least one jet having $p_T > 1.3$ TeV. Signal events are $W' \to XY$ for $m_{W'} = 3.5$ TeV and two hypothetical particles X and Y with masses of 500 and 100 GeV, respectively, which decay into pairs of quarks. The final state of the W' boson here is characterized by two large-radius jets with characteristically two-pronged substructure. Pythia 8 [89, 90] is used to simulate both the background and signal events, with an additional background dataset provided by simulation from Herwig++ [91]. A detector simulation provided by Delphes 3.4.1 [92–94] is additionally performed, using the default CMS detector card. Jet clustering is performed on Particle flow objects with FastJet and the anti-k_t algorithm [95–97], using a radius parameter of $R = 1$. Both background samples contain one million events, corresponding to an integrated luminosity of 100 fb^{-1}. The dataset is split in half, with one half used for training and the other for testing. For the purposes of testing and evaluating the performance of our models, we take the Pythia dataset to be ‘Data’ and the Herwig dataset to be ‘Simulation’ (Sim).

The resonant feature for this analysis is the dijet invariant mass m_{jj}. The additional characterizing features x are the masses of the two leading jets, m_1 and m_2, as well as their N-subjettiness ratios [98, 99], $\tau_{21}^{(J_1)}$ and $\tau_{21}^{(J_2)}$. Distribution plots of these four input features are shown in figure 3.3 below.

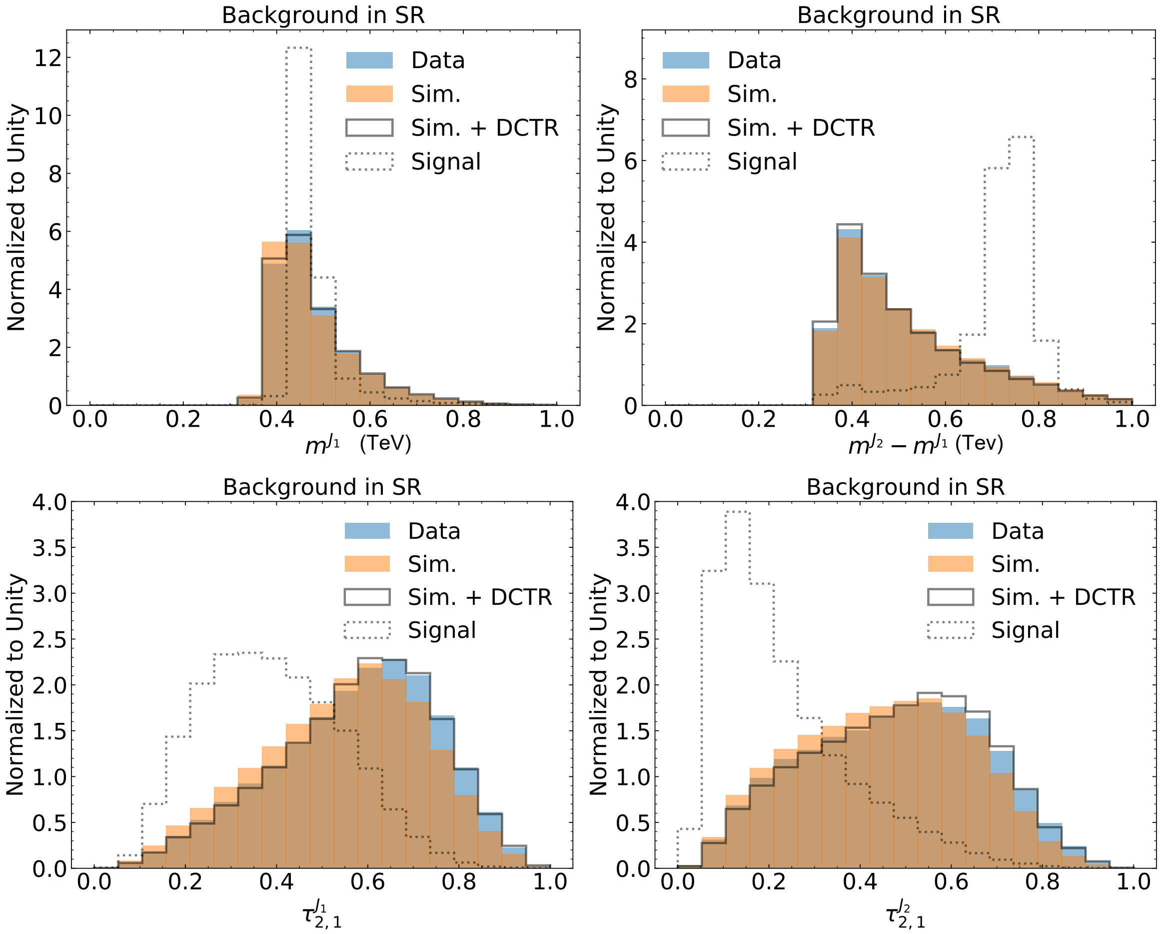


Figure 3.3 Left: the jet mass and τ_{21} of the jet with a smaller mass. Right: the difference between the heavier and lighter jet masses and τ_{21} of the heavier jet. In addition to showing the data, simulation, and signal, the histogram labeled 'Sim.+DCTR' is the simulation with weights derived from a parameterized reweighting function trained on long sidebands (the first step of the SALAD method).

For evaluation, 1000 signal events corresponding to a fitted significance of about 2σ are injected into the data for training. The entire signal sample is used for evaluation to improve statistics.

As a base case, a number of fully supervised models with all possible label information are trained; one on each input feature, and then an additional model using all four input

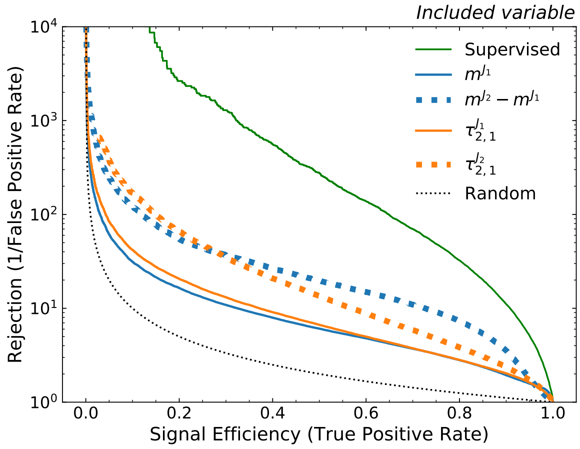


Figure 3.4 ROC curves for a variety of fully supervised models, used as baseline performance metrics on the training data. The solid green line labeled "Supervised" is a model trained with all four features, and represents the upper bound on any classifier for this task.

features. The results for this base case are shown in figure 3.4. In general, we see that the model employing all feature information yields the best results, as is expected. The green line labeled 'supervised' represents the upper bound on any classifier with this information, and is unachievable by any of the weakly supervised models in practice.

The solid red line labeled 'Optimal CWoLa' corresponds to a classifier trained using two mixed samples, one composed of pure background in the SR and the other composed of mostly background in the SR (independent from the first sample) with 500 injected signal events. The Optimal CWoLa line is far below the fully supervised classifier because the neural network needs to identify a small difference between the mixed samples over the natural statistical fluctuations in both sets. The actual CWoLa method is shown with a dotted red line. By construction, there is a significant difference between the phase space of the SR and SB and so the classifier is unable to identify the signal. At low efficiency, the CWoLa classifier actually anti-tags because the SR-SB differences are such that the signal is more SB-like than SR-like. Despite this drop in performance, the simulation augmenting modification (solid orange) with $\lambda = 1$ nearly recovers the full performance of CWoLa.

A classifier trained using simulation directly is also presented for comparison in Figure 3.5. The line labeled 'Data vs. Sim.' directly trains a classifier to distinguish the data and simulation in the SR, without reweighting. Since there are significant differences be-

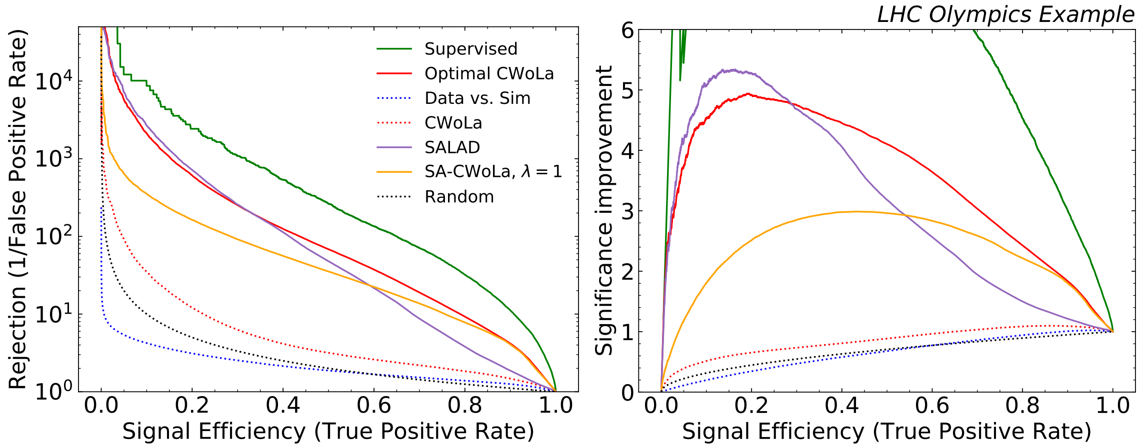


Figure 3.5 A Receiver Operating Characteristic (ROC) curve (left) and significance improvement curve (right) for various anomaly detection methods described in the text. Curves represent the mean of 20 model performances. The significance improvement is defined as the ratio of the signal efficiency to the square root of the background efficiency. A significance improvement of 2 means that the initial significance would be amplified by about a factor of two after employing the anomaly detection strategy. The supervised line is unachievable unless there is no mismodeling and one designed a search for the specific W' signal used in this paper. The curve labeled 'Random' corresponds to equal efficiency for signal and background.

tween the data and the simulated background, this classifier is not effective. In fact, the signal is more "similar" to the background simulation than the data background, so the classifier is worse than random (preferentially removes signal). The performance is significantly improved by adding in the parameterized reweighting, as described in Sec. 3.2. With this reweighting, the SALAD classifier is significantly better than random and is comparable to SA-CWoLa. The Optimal CWoLa line also serves as the upper bound in performance for SALAD because it corresponds to the case where the background simulation is statistically identical to the background in data. Full information on the means and standard deviations is also depicted in figure 3.6.

The loss function of the SA-CWoLa method has an additional free parameter, which can generally be tuned. Recall that $\lambda = 1$ corresponds to the case where signal sensitivity and decorrelation are approximately balanced, and $\lambda = 0$ is the normal CWoLa case. Figure 3.7 shows the performance of the SA-CWoLa classifier as a function of λ. In blue we see the mean and standard deviation of signal identification performance, and the dashed yellow lines show the classifier performance in the presence of *no* signal at

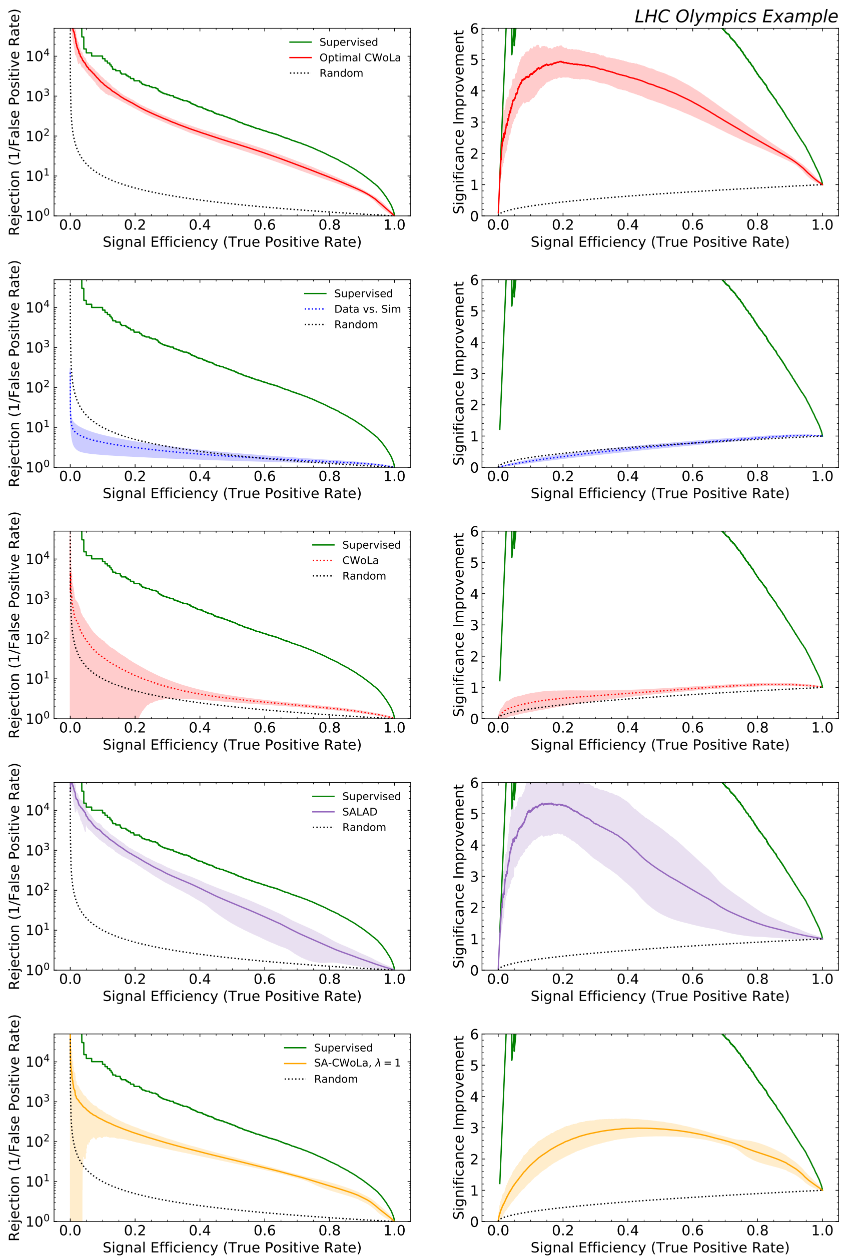


Figure 3.6 Spread of models on LHC Olympics datasets. Here the mean of 20 trials is indicated by the solid lines, and the standard deviation indicated by the shaded region. The means are identical to those displayed in figure 3.5. As before, the optimal supervised classifier performance is shown in green, with random performance shown by the black dotted line.

all. It is clear that the performance of the classifier is strong and relatively stable for the entire range of λ, and poor when not trained on any signal, as desired.

Next we implemented the background estimation given by the sideband fit and statistical test described in Section 3.2.4. These help to further communicate the performance

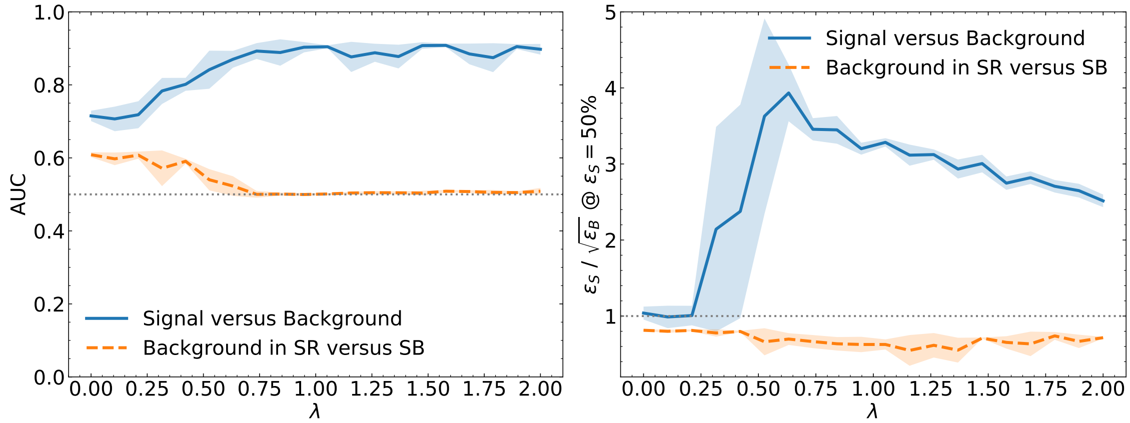


Figure 3.7 The Area Under the ROC Curve (AUC) (left) and significance improvement at 50% signal efficiency (right) using the SA-CWoLa method for a scan in the hyperparameter λ introduced in Eq. 3.6. Each curve is an average over 2 classifiers, with the standard deviation displayed in the shaded region. When $\lambda = 0$, SA-CWoLa is the same as the original CWoLa method. For comparison, the performance of the classifier for distinguishing signal and background is shown in blue and the performance for distinguishing the SR and SB is shown in orange. Ideally, the latter would have an AUC of 0.5, indicating random performance.

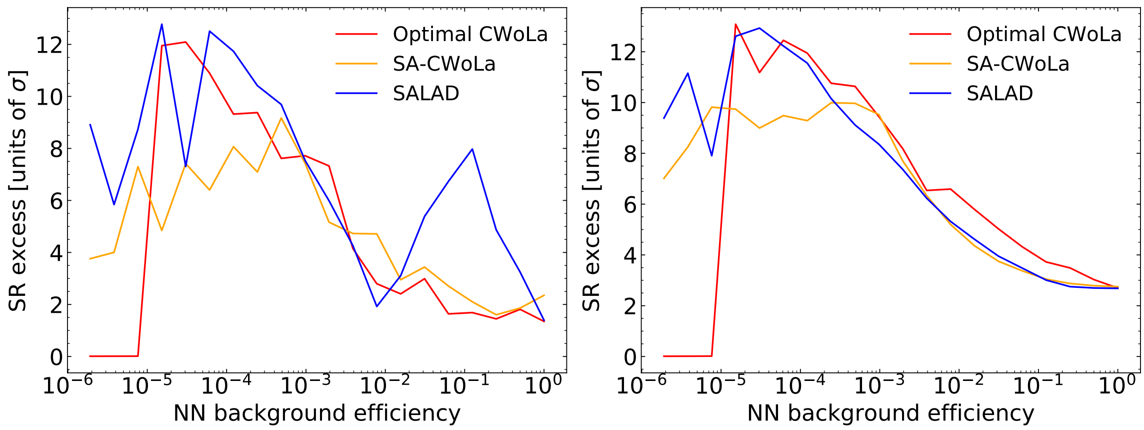


Figure 3.8 Fit excess with signal injected using the statistical procedure described in Sec. 3.2.4. There is a small local deficit in the simulation. The left plot shows the fitted excess without modifying the background while the right plot corrects for the initial deficit by subtracting the residuals of the background-only fit before performing the signal+background fit. In the latter case, the significances are still not $S/\sqrt{B}$ due to the uncertainty from the sideband fit. Since signal is injected, high SR excesses are desirable across the NN background efficiency range.

of these methods, especially in the context of practical searches for new physics in LHC experiments.

Figures 3.8 and 3.9 show the results of the sideband fit and statistical test (See

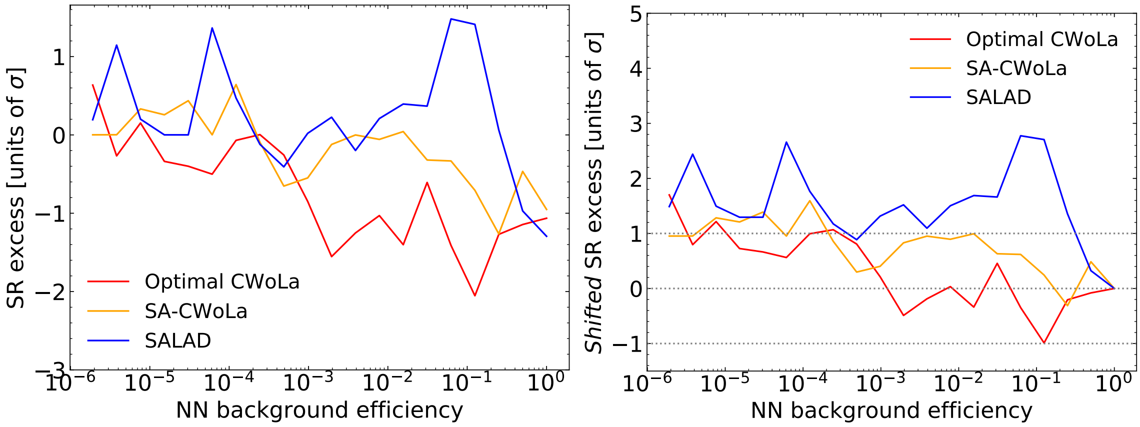


Figure 3.9 Fit excess without signal injected using the statistical procedure described in Sec. 3.2.4. Without any signal injected, there is a small ($\sim 1\sigma$) deficit in the simulation. The right plot shifts the curves so that the 100% efficiency point corresponds to 0σ. Since signal is not injected, SR excesses closer to 0 are ideal.

Sec. 3.2.4) in the case where the 2σ injected signal is included and excluded from the training dataset, respectively. When considering all bins, the fit quality is excellent (see Fig. 3.10), but there happens to be a small local deficit in the SR. The right plot of Fig. 3.8 removes this effect by subtracting the fitted residuals in the background-only case for each value of the NN background efficiency. The spectra after applying the CWoLa classifier cannot be fit to the same shape and are thus not included - see Fig. 3.11. This background sculpting is typical of methods with significant correlation.

As is expected of a good signal tagging algorithm, all of the models considered in figure 3.8 identify large excesses in the signal region. By inspection, one can see that at optimal thresholds, these models amplify the SR excesses from 2σ to $O(10\sigma)$, about a factor of 5. Similarly, in the case where no signal is injected shown in figure 3.9, no large SR excess is identified by any of the algorithms. The excesses shown in the background only case, on the order of 2-3σ, can be reduced by more advanced fitting techniques and optimization of the neural network training procedures.

Figure 3.10 shows one example of the result of our fitting procedure, as described in Section 3.2.4. In this case, no cut from any network has been applied to the data, and a fit is performed across the entire sideband. As we can see from this example, the fit quality to the invariant mass m_{jj} excellent in the case where there is no classifier cut. In general, the invariant mass distribution must roughly resemble that shown here for a parameterized

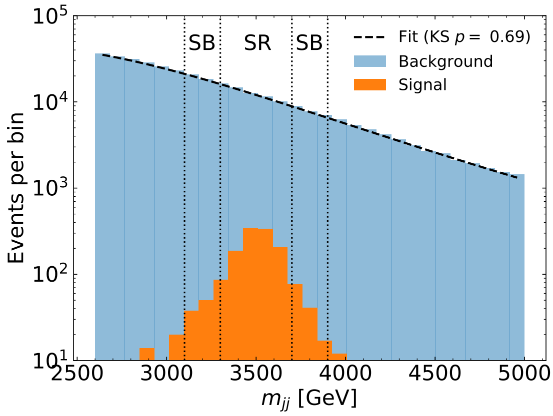


Figure 3.10 A fit to the m_{jj} distribution in the background-only case with no selection on any neural networks. The 500 signal events used for training is super-imposed for illustration. Vertical dashed lines indicate the SR and SB regions used for training. A Kolmogorov-Smirnov (KS) test using only bins outside of the SR yields a p-value of 0.69.

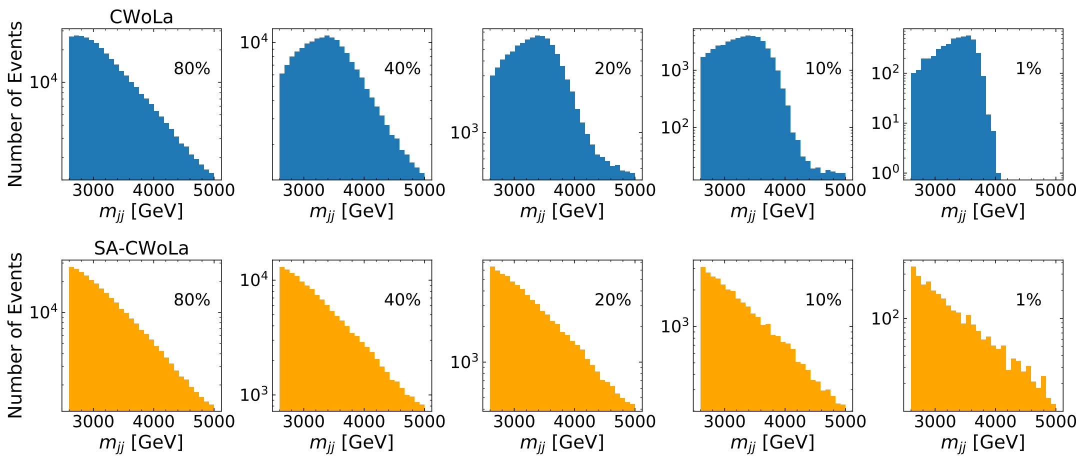


Figure 3.11 Histograms of m_{jj} for CWoLa (top) and SA-CWoLa (bottom) for various thresholds on the classifiers in the background-only case. From left to right, these thresholds are 80, 40, 20, 10, and 1%, and are indicated in the top right corner of each plot. Since signal is not injected here, more consistency in the distribution shapes as a function of threshold indicates better background estimation capability.

fit to be applied, since the fit can only use information from the sidebands. For any given classifier, the m_{jj} distribution filtered by a cut on the output of that classifier must not change so dramatically from the case shown here that it becomes impossible to apply a parametric fit. Furthermore, in the case of a classifier trained on a dataset without signal, m_{jj} distributions as a function of the threshold set on the classifier's output should show the same distribution shape as in figure 3.10, up to counting uncertainties.

Figure 3.11 shows how the CWoLa classifier fails to meet this criteria, whereas the

SA-CWoLa classifier succeeds. Due to the strong correlation between the invariant mass and the training features, CWoLa is easily able to tag the SR more frequently than the SB for tighter classifier thresholds, e.g. 1%. This creates a massive "sculpted bump," which incorrectly indicates the presence of signal in the context of the sideband fit method we are using. On the other hand, tightening thresholds on the SA-CWoLa output yields nearly identical m_{jj} distribution shapes, indicating that the classifier is able to successfully ignore the strong correlations between the invariant mass and the inputs.

3.4 Outlook

This chapter has investigated the impact of dependencies between localization features (i.e. m_{jj}) and classification features for resonant AD methods, particularly Salad, CWoLa, and the standard data vs. simulation comparison. It also proposes and evaluates a new iteration of CWoLa which makes use of simulation information to automatically decorrelate classification features from the localization feature. Both methods are able to exploit the physical priors of simulations without fully relying on simulations for background predictions. The demonstrated usefulness of simulations in both of these methods underscores their usefulness as a tool for mitigating the dependence of classifiers on resonant feature correlations.

In any case, each of these methods have advantages and weaknesses in the realm of AD as a whole. To achieve broad signal sensitivity, it is likely that a combination of multiple approaches will be required. This chapter puts forward a rigorous study of the sensitivity of each technique to dependencies and proposes one modification to improve robustness, representing a step forward in the decorrelation program for AD with machine learning. The most likely next step from here is a higher dimensional version of the existing ATLAS group weakly supervised anomaly detection search [18], as well as other related searches by other experimental collaborations in the future.

Chapter 4

Discussion

4.1 Conclusions

Here we have presented an analysis of multiple classes of Anomaly Detection techniques. First, Chapter 2 focused on studying the efficacy of autoencoders in the context of specific BSM analyses, showing that they can provide excellent performance when searching for broad classes of models with widely different parameters. As one of the most basic fully unsupervised AD methods, this autoencoder analysis provides a baseline performance for unsupervised Semi-Visible jet searches.

Chapter 3 focused on studying weakly-supervised anomaly detection techniques, with a specific focus on analysing the performance under the effects of strong correlations. Here we find that while some methods are automatically robust to correlations, others may require modification. We propose one such modified method and show its efficacy in the case of both a simple gaussian "toy" problem and a physics benchmark classification task from the 2020 LHC Olympics.

Both analyses contribute to the growing body of work which utilizes machine learning to increase the search program of model independent anomaly detection techniques at the frontier of high energy particle physics. A number of exciting future research paths are also clear from this work, to be discussed in the following section.

4.2 Further Work

The clearest next step for this work is to apply both the unsupervised autoencoder method and the weakly supervised methods to the same datasets to directly quantify their performances. While they both have different application methods, a direct comparison would be an excellent baseline comparison.

Future autoencoder analyses of the semi-visible jet detection capability should consider implementing other unsupervised anomaly detection methods, applying the invariant mass fit and calculating the exact expected significance increases by using AEs, and exploring alternative model architectures. Future progress could likewise be made with the weakly supervised methods discussed here by applying them to specific analyses, such as the semi-visible jet analysis done in chapter 2.

In both cases, work towards improving the understanding of the fitted signal sensitivity as done in chapter 3 is necessary, with a particular focus on using internal collaboration data and simulations in the lower energy regions. We have thus far completed several months of work focusing on the sensitivity of weakly supervised methods on ATLAS collaboration dijet data, using high fidelity ATLAS simulations and data in the high rapidity $|y^{j_1} - y^{j_2}| > 1.2$ region. Results on this work are still internal, though we hope to proceed with a collaboration wide analysis over the course of the next year. I personally plan to continue this project after graduation as a PhD student at the University of California, Berkeley.

References

[1] Georges Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Lett.* B716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[2] Serguei Chatrchyan et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Phys. Lett.* B716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[3] X. Cid Vidal et al. "Beyond the Standard Model Physics at the HL-LHC and HE-LHC". In: *arXiv:1812.07831 [hep-ex, physics:hep-ph]* (August 2019). arXiv: 1812.07831. URL: http://arxiv.org/abs/1812.07831 (visited on March 26, 2021).

[4] ATLAS Collaboration. *Exotic Physics Searches*. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults. 2019.

[5] ATLAS Collaboration. *Supersymmetry searches*. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults. 2019.

[6] ATLAS Collaboration. *Higgs and Diboson Searches*. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/HDBSPublicResults. 2019.

[7] CMS Collaboration. *CMS Exotica Public Physics Results*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO. 2019.

[8] CMS Collaboration. *CMS Supersymmetry Physics Results*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS. 2019.

[9] CMS Collaboration. *CMS Beyond-two-generations (B2G) Public Physics Results*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G. 2019.

[10] LHCb Collaboration. *Publications of the QCD, Electroweak and Exotica Working Group*. http://lhcbproject.web.cern.ch/lhcbproject/Publications/LHCbProjectPublic/Summary_QEE.html. 2019.

[11] Andrew J. Larkoski, Ian Moult, and Benjamin Nachman. "Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning". In: *Phys. Reports* (2019). DOI: 10.1016/j.physrep.2019.11.001. arXiv: 1709.04464 [hep-ph].

[12] Dan Guest, Kyle Cranmer, and Daniel Whiteson. "Deep Learning and its Application to LHC Physics". In: *Ann. Rev. Nucl. Part. Sci.* 68 (2018), p. 161. DOI: 10.1146/annurev-nucl-101917-021019. arXiv: 1806.11484 [hep-ex].

[13] Murat Abdughani et al. "Supervised deep learning in high energy phenomenology: a mini review". In: *Commun. Theor. Phys.* 71 (2019), p. 955. DOI: 10.1088/0253-6102/71/8/955. arXiv: 1905.06047 [hep-ph].

[14] Alexander Radovic et al. "Machine learning at the energy and intensity frontiers of particle physics". In: *Nature* 560 (2018), p. 41. DOI: 10.1038/s41586-018-0361-2.

[15] Raffaele Tito D'Agnolo and Andrea Wulzer. "Learning New Physics from a Machine". In: *Phys. Rev.* D99 (2019), p. 015014. DOI: 10.1103/PhysRevD.99.015014. arXiv: 1806.02350 [hep-ph].

[16] Jack H. Collins, Kiel Howe, and Benjamin Nachman. "Anomaly Detection for Resonant New Physics with Machine Learning". In: *Phys. Rev. Lett.* 121 (2018), p. 241803. DOI: 10.1103/PhysRevLett.121.241803. arXiv: 1805.02664 [hep-ph].

[17] Jack H. Collins, Kiel Howe, and Benjamin Nachman. "Extending the search for new resonances with machine learning". In: *Phys. Rev.* D99 (2019), p. 014038. DOI: 10.1103/PhysRevD.99.014038. arXiv: 1902.02634 [hep-ph].

[18] ATLAS Collaboration. "Dijet resonance search with weak supervision using 13 TeV pp collisions in the ATLAS detector". In: (2020). arXiv: 2005.02983 [hep-ex].

[19] Raffaele Tito D'Agnolo et al. "Learning Multivariate New Physics". In: (2019). arXiv: 1912.12155 [hep-ph].

[20] Marco Farina, Yuichiro Nakai, and David Shih. "Searching for New Physics with Deep Autoencoders". In: (2018). arXiv: 1808.08992 [hep-ph].

[21] Theo Heimel et al. "QCD or What?" In: *SciPost Phys.* 6 (2019), p. 030. DOI: 10.21468/SciPostPhys.6.3.030. arXiv: 1808.08979 [hep-ph].

[22] Tuhin S. Roy and Aravind H. Vijay. "A robust anomaly finder based on autoencoder". In: (2019). arXiv: 1903.02032 [hep-ph].

[23] Olmo Cerri et al. "Variational Autoencoders for New Physics Mining at the Large Hadron Collider". In: *JHEP* 05 (2019), p. 036. DOI: 10.1007/JHEP05(2019)036. arXiv: 1811.10276 [hep-ex].

[24] Andrew Blance, Michael Spannowsky, and Philip Waite. "Adversarially-trained autoencoders for robust unsupervised new physics searches". In: *JHEP* 10 (2019), p. 047. DOI: 10.1007/JHEP10(2019)047. arXiv: 1905.10384 [hep-ph].

[25] Jan Hajer et al. "Novelty Detection Meets Collider Physics". In: (2018). arXiv: 1807.10261 [hep-ph].

[26] Andrea De Simone and Thomas Jacques. "Guiding New Physics Searches with Unsupervised Learning". In: *Eur. Phys. J.* C79 (2019), p. 289. DOI: 10.1140/epjc/s10052-019-6787-3. arXiv: 1807.06038 [hep-ph].

[27] Anna Mullin et al. "Does SUSY have friends? A new approach for LHC event analysis". In: (2019). arXiv: 1912.10625 [hep-ph].

[28] Giovanna Menardi Alessandro Casa. "Nonparametric semisupervised classification for signal detection in high energy physics". In: (2019). arXiv: 1809.02977 [hep-ex].

[29] Barry M. Dillon, Darius A. Faroughy, and Jernej F. Kamenik. "Uncovering latent jet substructure". In: *Phys. Rev.* D100 (2019), p. 056002. DOI: 10.1103/PhysRevD.100.056002. arXiv: 1904.04200 [hep-ph].

[30] Anders Andreassen, Benjamin Nachman, and David Shih. "Simulation Assisted Likelihood-free Anomaly Detection". In: *Phys. Rev. D* 101.9 (2020), p. 095004. DOI: 10.1103/PhysRevD.101.095004. arXiv: 2001.05001 [hep-ph].

[31] Benjamin Nachman and David Shih. "Anomaly Detection with Density Estimation". In: *Phys. Rev. D* 101 (2020), p. 075042. DOI: 10.1103/PhysRevD.101.075042. arXiv: 2001.04990 [hep-ph].

[32] J. A. Aguilar-Saavedra, Jack H. Collins, and Rashmish K. Mishra. "A generic anti-QCD jet tagger". In: *JHEP* 11 (2017), p. 163. DOI: 10.1007/JHEP11(2017)163. arXiv: 1709.01087 [hep-ph].

[33] M. Romão Crispim et al. "Transferability of Deep Learning Models in Searches for New Physics at Colliders". In: *Phys. Rev. D* 101.3 (2020), p. 035042. DOI: 10.1103/PhysRevD.101.035042. arXiv: 1912.04220 [hep-ph].

[34] M. Crispim Romao et al. "Use of a Generalized Energy Mover's Distance in the Search for Rare Phenomena at Colliders". In: (April 2020). arXiv: 2004.09360 [hep-ph].

[35] Oliver Knapp et al. "Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark". In: (2020). arXiv: 2005.01598 [hep-ex].

[36] Barry M. Dillon et al. "Learning the latent structure of collider events". In: (May 2020). arXiv: 2005.12319 [hep-ph].

[37] M. Crispim Romao, N.F. Castro, and R. Pedro. "Finding New Physics without learning about it: Anomaly Detection as a tool for Searches at Colliders". In: (June 2020). arXiv: 2006.05432 [hep-ph].

[38] Oz Amram and Cristina Mantilla Suarez. "Tag N' Train: A Technique to Train Improved Classifiers on Unlabeled Data". In: (February 2020). arXiv: 2002.12376 [hep-ph].

[39] Taoli Cheng et al. "Variational Autoencoders for Anomalous Jet Tagging". In: (July 2020). arXiv: 2007.01850 [hep-ph].

[40] Charanjit K. Khosa and Veronica Sanz. "Anomaly Awareness". In: (July 2020). arXiv: 2007.14462 [cs.LG].

[41] Punnathat Thaprasop et al. "Unsupervised Outlier Detection in Heavy-Ion Collisions". In: (July 2020). arXiv: 2007.15830 [hep-ex].

[42] J.A. Aguilar-Saavedra, F.R. Joaquim, and J.F. Seabra. "Mass Unspecific Supervised Tagging (MUST) for boosted jets". In: (August 2020). arXiv: 2008.12792 [hep-ph].

[43] J.A. Aguilar-Saavedra and B. Zaldıvar. "Jet tagging made easy". In: *Eur. Phys. J. C* 80.6 (2020), p. 530. DOI: 10.1140/epjc/s10052-020-8082-8. arXiv: 2002.12320 [hep-ph].

[44] B. Knuteson. *Quasi-Model-Independent Search for New Physics*. Ph.D. thesis, University of California at Berkeley (2000).

[45] Morad Aaboud et al. "A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment". In: *Eur. Phys. J.* C79 (2019), p. 120. DOI: 10.1140/epjc/s10052-019-6540-y. arXiv: 1807.07447 [hep-ex].

[46] "MUSiC, a Model Unspecific Search for New Physics, in pp Collisions at $\sqrt{s} = 8\,\mathrm{TeV}$". In: *CMS-PAS-EXO-14-016* (2017). URL: https://cds.cern.ch/record/2256653.

[47] W. Buttinger. *Background Estimation with the ABCD Method*. https://twiki.cern.ch/twiki/pub/Main/ABCDMethod/ABCDGuide_draft18Oct18.pdf. 2018.

[48] Gregor Kasieczka et al. "ABCDisCo: Automating the ABCD Method with Machine Learning". In: (July 2020). arXiv: 2007.14400 [hep-ph].

[49] J. J. Aubert et al. "Experimental Observation of a Heavy Particle J". In: *Phys. Rev. Lett.* 33 (23 December 1974), pp. 1404–1406. DOI: 10.1103/PhysRevLett.33.1404. URL: https://link.aps.org/doi/10.1103/PhysRevLett.33.1404.

[50] J. -E. Augustin et al. "Discovery of a Narrow Resonance in e^+e^- Annihilation". In: *Phys. Rev. Lett.* 33 (23 December 1974), pp. 1406–1408. DOI: 10.1103/PhysRevLett.33.1406. URL: https://link.aps.org/doi/10.1103/PhysRevLett.33.1406.

[51] Andrea Borghesi et al. "Anomaly Detection using Autoencoders in High Performance Computing Systems". In: *arXiv:1811.05269 [cs]* (November 2018). arXiv: 1811.05269. URL: http://arxiv.org/abs/1811.05269 (visited on July 15, 2019).

[52] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. "Energy flow polynomials: A complete linear basis for jet substructure". In: *JHEP* 04 (2018), p. 013. DOI: 10.1007/JHEP04(2018)013. arXiv: 1712.07124 [hep-ph].

[53] Vasiliki A. Mitsou. "SHEDDING LIGHT ON DARK MATTER AT COLLIDERS". en. In: *International Journal of Modern Physics A* 28.31 (December 2013), p. 1330052. ISSN: 0217-751X, 1793-656X. DOI: 10.1142/S0217751X13300524. URL: https://www.worldscientific.com/doi/abs/10.1142/S0217751X13300524 (visited on February 11, 2021).

[54] Timothy Cohen, Mariangela Lisanti, and Hou Keong Lou. "Semivisible Jets: Dark Matter Undercover at the LHC". en. In: *Physical Review Letters* 115.17 (October 2015), p. 171804. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.115.171804. URL: https://link.aps.org/doi/10.1103/PhysRevLett.115.171804 (visited on February 11, 2021).

[55] Pedro Schwaller, Daniel Stolarski, and Andreas Weiler. "Emerging jets". en. In: *Journal of High Energy Physics* 2015.5 (May 2015), p. 59. ISSN: 1029-8479. DOI: 10.1007/JHEP05(2015)059. URL: http://link.springer.com/10.1007/JHEP05(2015)059 (visited on February 11, 2021).

[56] Myeonghun Park and Mengchao Zhang. "Tagging a jet from a dark sector with jet substructures at colliders". en. In: *Physical Review D* 100.11 (December 2019), p. 115009. ISSN: 2470-0010, 2470-0029. DOI: 10.1103/PhysRevD.100.115009. URL: https://link.aps.org/doi/10.1103/PhysRevD.100.115009 (visited on February 11, 2021).

[57] Torbjörn Sjöstrand et al. "An Introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (June 2015). arXiv: 1410.3012, pp. 159–177. ISSN: 00104655. DOI: 10.1016/j.cpc.2015.01.024. URL: http://arxiv.org/abs/1410.3012 (visited on August 6, 2019).

[58] J. de Favereau et al. "DELPHES 3, A modular framework for fast simulation of a generic collider experiment". In: *Journal of High Energy Physics* 2014.2 (February 2014). arXiv: 1307.6346, p. 57. ISSN: 1029-8479. DOI: 10.1007/JHEP02(2014)057. URL: http://arxiv.org/abs/1307.6346 (visited on August 6, 2019).

[59] R. Brun and F. Rademakers. "ROOT: An object oriented data analysis framework". In: *Nucl. Instrum. Meth.* A389 (1997), pp. 81–86. DOI: 10.1016/S0168-9002(97)00048-X.

[60] Yang-Ting Chien et al. "Jet fragmentation functions in proton-proton collisions using soft-collinear effective theory". In: *Journal of High Energy Physics* 2016 (2016), pp. 1–26.

[61] François Chollet et al. *Keras*. https://github.com/fchollet/keras. 2015.

[62] Martın Abadi et. al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: http://tensorflow.org/.

[63] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014).

[64] Gilles Louppe, Michael Kagan, and Kyle Cranmer. "Learning to Pivot with Adversarial Networks". In: (2016). arXiv: 1611.01046 [stat.ME].

[65] James Dolen et al. "Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure". In: *JHEP* 05 (2016), p. 156. DOI: 10.1007/JHEP05(2016)156. arXiv: 1603.00027 [hep-ph].

[66] Ian Moult, Benjamin Nachman, and Duff Neill. "Convolved Substructure: Analytically Decorrelating Jet Substructure Observables". In: *JHEP* 05 (2018), p. 002. DOI: 10.1007/JHEP05(2018)002. arXiv: 1710.06859 [hep-ph].

[67] Justin Stevens and Mike Williams. "uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers". In: *JINST* 8 (2013), P12013. DOI: 10.1088/1748-0221/8/12/P12013. arXiv: 1305.7248 [nucl-ex].

[68] Chase Shimmin et al. "Decorrelated Jet Substructure Tagging using Adversarial Neural Networks". In: *Phys. Rev.* D96 (2017), p. 074034. DOI: 10.1103/PhysRevD.96.074034. arXiv: 1703.03507 [hep-ex].

[69] Layne Bradshaw et al. "Mass Agnostic Jet Taggers". In: (2019). arXiv: 1908.08959 [hep-ph].

[70] "Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS". In: *ATL-PHYS-PUB-2018-014* (2018). URL: http://cds.cern.ch/record/2630973.

[71] Li-Gang Xia. "QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics". In: *Nucl. Instrum. Meth.* A930 (2019), pp. 15–26. DOI: 10.1016/j.nima.2019.03.088. arXiv: 1810.08387 [physics.data-an].

[72] Christoph Englert et al. "Machine Learning Uncertainties with Adversarial Neural Networks". In: *Eur. Phys. J.* C79.1 (2019), p. 4. DOI: 10.1140/epjc/s10052-018-6511-8. arXiv: 1807.08763 [hep-ph].

[73] Stefan Wunsch et al. "Reducing the dependence of the neural network function to systematic uncertainties in the input space". In: (2019). arXiv: 1907.11674 [physics.data-an].

[74] Gregor Kasieczka and David Shih. "DisCo Fever: Robust Networks Through Distance Correlation". In: (January 2020). arXiv: 2001.05310 [hep-ph].

[75] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. "Classification without labels: Learning from mixed samples in high energy physics". In: *JHEP* 10 (2017), p. 174. DOI: 10.1007/JHEP10(2017)174. arXiv: 1708.02949 [hep-ph].

[76] Anders Andreassen and Benjamin Nachman. "Neural Networks for Full Phase-space Reweighting and Parameter Tuning". In: (2019). arXiv: 1907.08209 [hep-ph].

[77] Kyle Cranmer, Juan Pavez, and Gilles Louppe. "Approximating Likelihood Ratios with Calibrated Discriminative Classifiers". In: (2015). arXiv: 1506.02169 [stat.AP].

[78] Pierre Baldi et al. "Parameterized neural networks for high-energy physics". In: *Eur. Phys. J.* C76.5 (2016), p. 235. DOI: 10.1140/epjc/s10052-016-4099-4. arXiv: 1601.07913 [hep-ex].

[79] Johann Brehmer et al. "Mining gold from implicit models to improve likelihood-free inference". In: (2018). arXiv: 1805.12244 [stat.ML].

[80] Albert M Sirunyan et al. "Search for narrow and broad dijet resonances in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on dark matter mediators and other new particles". In: *JHEP* 08 (2018), p. 130. DOI: 10.1007/JHEP08(2018)130. arXiv: 1806.00843 [hep-ex].

[81] Georges Aad et al. "Search for new resonances in mass distributions of jet pairs using 139 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *JHEP* 03 (2020), p. 145. DOI: 10.1007/JHEP03(2020)145. arXiv: 1910.08447 [hep-ex].

[82] Meghan Frate et al. "Modeling Smooth Backgrounds and Generic Localized Signals with Gaussian Processes". In: (September 2017). arXiv: 1709.05681 [physics.data-an].

[83] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". In: *Ann. Math. Statist.* 9.1 (March 1938), pp. 60–62. DOI: 10.1214/aoms/1177732360. URL: https://doi.org/10.1214/aoms/1177732360.

[84] Abraham Wald. "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large". In: *Transactions of the American Mathematical Society* 54.3 (1943), pp. 426–482. ISSN: 00029947. URL: http://www.jstor.org/stable/1990256.

[85] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *Eur. Phys. J.* C71 (2011). [Erratum: Eur. Phys. J.C73,2501(2013)], p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0,10.1140/epjc/s10052-013-2501-z. arXiv: 1007.1727 [physics.data-an].

[86] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[87] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980 [cs]* (December 2014). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980 (visited on August 6, 2019).

[88] Gregor Kasieczka, Ben Nachman, and David Shih. *R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge*. Version v1. https://doi.org/10.5281/zenodo.2629073. Zenodo, April 2019. DOI: 10.5281/zenodo.2629073.

[89] Torbjorn Sjöstrand, Stephen Mrenna, and Peter Z. Skands. "PYTHIA 6.4 Physics and Manual". In: *JHEP* 05 (2006), p. 026. DOI: 10.1088/1126-6708/2006/05/026. arXiv: hep-ph/0603175 [hep-ph].

[90] Torbjorn Sjöstrand, Stephen Mrenna, and Peter Z. Skands. "A Brief Introduction to PYTHIA 8.1". In: *Comput. Phys. Commun.* 178 (2008), p. 852. DOI: 10.1016/j.cpc.2008.01.036. arXiv: 0710.3820 [hep-ph].

[91] M. Bahr et al. "Herwig++ Physics and Manual". In: *Eur. Phys. J.* C58 (2008), pp. 639–707. DOI: 10.1140/epjc/s10052-008-0798-9. arXiv: 0803.0883 [hep-ph].

[92] J. de Favereau et al. "DELPHES 3, A modular framework for fast simulation of a generic collider experiment". In: *JHEP* 02 (2014), p. 057. DOI: 10.1007/JHEP02(2014)057. arXiv: 1307.6346 [hep-ex].

[93] Alexandre Mertens. "New features in Delphes 3". In: *J. Phys. Conf. Ser.* 608 (2015), p. 012045. DOI: 10.1088/1742-6596/608/1/012045.

[94] Michele Selvaggi. "DELPHES 3: A modular framework for fast-simulation of generic collider experiments". In: *J. Phys. Conf. Ser.* 523 (2014), p. 012033. DOI: 10.1088/1742-6596/523/1/012033.

[95] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet User Manual". In: *Eur. Phys. J.* C72 (2012), p. 1896. DOI: 10.1140/epjc/s10052-012-1896-2. arXiv: 1111.6097 [hep-ph].

[96] Matteo Cacciari and Gavin P. Salam. "Dispelling the N^3 myth for the k_t jet-finder". In: *Phys. Lett.* B641 (2006), p. 57. DOI: 10.1016/j.physletb.2006.08.037. arXiv: hep-ph/0512210 [hep-ph].

[97] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[98] Jesse Thaler and Ken Van Tilburg. "Maximizing Boosted Top Identification by Minimizing N-subjettiness". In: *JHEP* 02 (2012), p. 093. DOI: 10.1007/JHEP02(2012)093. arXiv: 1108.2701 [hep-ph].

[99] Jesse Thaler and Ken Van Tilburg. "Identifying Boosted Objects with N-subjettiness". In: *JHEP* 03 (2011), p. 015. DOI: 10.1007/JHEP03(2011)015. arXiv: 1011.2268 [hep-ph].