Published for SISSA by Springer

Received: October 25, 2021
Accepted: February 17, 2022
Published: March 9, 2022

Challenges for unsupervised anomaly detection in particle physics

Katherine Fraser, Samuel Homiller, Rashmish K. Mishra, Bryan Ostdiek and Matthew D. Schwartz

Department of Physics, Harvard University,
Cambridge, MA 02138, U.S.A.
NSF AI Institute for Artificial Intelligence and Fundamental Interactions
E-mail: kfraser@g.harvard.edu, shomiller@g.harvard.edu, rashmishmishra@fas.harvard.edu, bostdiek@g.harvard.edu, schwartz@g.harvard.edu

Abstract: Anomaly detection relies on designing a score to determine whether a particular event is uncharacteristic of a given background distribution. One way to define a score is to use autoencoders, which rely on the ability to reconstruct certain types of data (background) but not others (signals). In this paper, we study some challenges associated with variational autoencoders, such as the dependence on hyperparameters and the metric used, in the context of anomalous signal (top and W) jets in a QCD background. We find that the hyperparameter choices strongly affect the network performance and that the optimal parameters for one signal are non-optimal for another. In exploring the networks, we uncover a connection between the latent space of a variational autoencoder trained using mean-squared-error and the optimal transport distances within the dataset. We then show that optimal transport distances to representative events in the background dataset can be used directly for anomaly detection, with performance comparable to the autoencoders. Whether using autoencoders or optimal transport distances for anomaly detection, we find that the choices that best represent the background are not necessarily best for signal identification. These challenges with unsupervised anomaly detection bolster the case for additional exploration of semi-supervised or alternative approaches.

Keywords: Jet substructure, Beyond Standard Model, Hadron-Hadron Scattering , Jets

ArXiv ePrint: 2110.06948

https://doi.org/10.1007/JHEP03(2022)066

JHEP03(2022)066

Contents

JHEP03(2022)066

1 Introduction

While many searches for physics beyond the Standard Model have been carried out at the Large Hadron Collider, new physics remains elusive. This may be due to a lack of new physics in the data, but it could also be due to us looking in the wrong place. Trying to design searches that are more robust to unexpected new physics has inspired a lot of work on anomaly detection using unsupervised methods including community wide challenges such as the LHC Olympics [1] and the Dark Machines Anomaly Score Challenge [2]. The goal of anomaly detection is to search for events which are "different" than what is expected. When used for anomaly detection, unsupervised methods attempt to characterize the space of background events in some way, independent of signal. The hope is then that signal events will stand out as being uncharacteristic.

Anomaly detection techniques can be broadly split into two categories. For some signals, the signal events look similar to typical background events and one must exploit information about the expected probability distribution of the background to find the signal. Many anomaly detection techniques have been developed to find signals of this type [1, 3–24]. Alternatively, some signals are qualitatively different from prototypical background and then methods that try to characterize an individual event as anomalous can be used [2, 25–54]. Here, we restrict to the latter type of anomaly detection, where an

anomaly score for individual events can be determined from the background ensemble and used for discrimination, without needing to characterize the full probability distribution of the signal ensemble. With an effective method, events with a small score are likely to be a part of the background, while events with a larger score are not. There are many different ways of defining an anomaly score. Some rely on traditional high-level observables, like mass or N-subjettiness [55], (see e.g. [7, 13, 29] which use traditional variables in anomaly detection). Others attempt to directly learn how likely a given event or object is using low-level information, like individual particle momenta (see e.g., ref. [50]). Some methods that search for outliers rely on abstract representations to try to characterize the event space, such as the latent space of an autoencoder [20, 48]. Others give the event space itself a geometric interpretation in terms of distances [56–58]. Given the complexity and high-dimensionality of data at the LHC, many anomaly detection techniques employ machine learning.

In this paper, we begin by exploring the use of autoencoders for anomaly detection on individual fat jets within events. Autoencoders were initially introduced for dimensionality reduction, similar to principal component analysis, to learn the important information in data while ignoring insignificant information and noise [59]. Autoencoders contain an encoder, which reduces the dimensionality of the input to give some latent representation, and a decoder, which transforms the latent space back to the original space. In particle physics, autoencoders were first used for anomaly detection in refs. [26–28], where they are meant to reconstruct certain types of data (background) but not others (signals). In order to work as an anomaly detector, an autoencoder should have a small reconstruction error for background events and a large reconstruction error for signal events. To do so, the autoencoder must establish a delicate balance in achieving a reconstruction fidelity which is accurate, but not too accurate. There are several cases where training a network with adequate discriminating power is especially difficult, such as when the signal looks very similar to the background, when the dataset has certain topological properties [18], or when innate characteristics of the samples make the signal sample simpler than the background sample to reconstruct [46, 48].

A generalization of autoencoders called variational autoencoders (VAEs) were introduced in ref. [60]. Unlike an ordinary autoencoder, where each input is mapped to an arbitrary point in the latent space, in a VAE, the latent space is a probability distribution which is sampled and mapped back to the original space by the decoder. In addition to the usual reconstruction error, the VAE loss also includes a Kullback-Leibler (KL) divergence component that pushes the latent space towards a Gaussian prior and regularizes network training. The latent space of the VAE encodes the probability distribution of the background training sample, which can be used in the anomaly score. VAEs were first used in anomaly detection in computer science in ref. [61], and first used for particle physics anomaly detection in refs. [26, 29]. They have been widely studied since then [20, 37, 41–43, 48, 53, 54, 62, 63].

The task of an autoencoder, variational or not, for unsupervised anomaly detection is to provide a strong universal signal/background discriminant for a variety of signals having access only to background for training. In principle, this approach is advantageous

JHEP03(2022)066

because it opens the possibility to bypass Monte Carlo simulations and work directly with experimental data, which is almost completely background.[1] The autoencoder paradigm is based on the vision that there is trade-off between efficacy and generality: the ideal discriminant for a given signal and given background would be ineffective for a different signal and different background while a general discriminant, like the autoencoder, would work decently on a broad class of signals and backgrounds. The ideal assumes, first, that such a general discriminant exists with an appropriate use case, and second that it can be found by training purely on one or more background samples without any direct information about the signal. However, one has reason to be suspicious: machine learning methods work great at optimizing a given loss, which is meant to correlate strongly with the problem one is trying to solve. For autoencoder anomaly detection, the optimization (background only) is not aligned with the ultimate problem of interest (signal discovery over background), so it should not be surprising if the autoencoder does poorly. In section 4, we explore the challenges induced by trying to optimize a VAE in a model agnostic way.

In order to understand what a VAE is learning, we study its latent space. In particular, we look at the distance between events in VAE latent space (see [48, 64] for other studies of VAE latent spaces in particle physics). Since we can think of the VAE anomaly score as a "distance" encoding how far any given event is from the background distribution, it is also natural to ask about the distances between individual events. We find there is a significant correlation between the Euclidean distance between events represented in the VAE latent space and the Wasserstein optimal transport distance between events represented as images. We study Wasserstein distances in particular because they were physically motivated in refs. [56–58].

The correlation we observe between distances in the VAE latent space and between the event images motivates us to explore using optimal transport distances between events to define an anomaly score in section 5. One method for using distances directly is to identify representative events in the background sample, and use an event-to-event distance between a given event and the representative event as the score. The advantages of this method we propose are that it does not require training a neural network and that it is easily adaptable to different background samples.

This paper is organized as follows. In section 2, we provide information about the dataset used in our study. In section 3, we provide relevant background information on the metrics used (section 3.1), and the details of the VAE architecture (section 3.2). In section 4, we explore the effectiveness of an image based convolutional autoencoder for anomaly detection, including its sensitivity to hyperparameters. We also explore correlations between Euclidean distances in the autoencoder's latent space and optimal-transport distances among the event images in section 4.2. This motivates the development of methods that directly use the optimal transport distances among events as an alternative to VAEs in section 5. We conclude in section 6.

[1] Of course, in the realistic situation where the training is performed on data, there could be signal events contaminating the background sample. Nevertheless, a number of studies have demonstrated that this has little effect on the autoencoder performance (e.g., refs. [27, 28]). We will ignore this complication when we refer to "background only" samples in what follows.

JHEP03(2022)066

2 Datasets

We begin by describing the datasets we use for our analysis. For concreteness, we focus on anomaly detection in simulated jet events at the LHC. We will consider QCD dijet events as the background, and consider both top and W jets as representatives of anomalous signal events. Although in practice anomaly detection techniques would not be used for top and W jets since there are dedicated experimental searches for these objects, these jets provide a simple benchmark for studying unsupervised methods. The authors of ref. [37] have provided a suite of jets for Standard Model and beyond the Standard Model particle resonances which are available on Zenodo [65]. A sample of QCD dijet background events are also provided on Zenodo using the same selection criteria, showering, and detector simulation parameters [66]. The datasets were generated with MadGraph [67] and Pythia8 [68] and used Delphes [69] for fast detector simulation. Jets were clustered using FastJet [70, 71] using the anti-k_T algorithm [72] with a cone size of $R = 1.0$. The event selection requires two hard jets, with leading jet having $p_T > 450\,\text{GeV}$ and the subleading jet having $p_T > 200\,\text{GeV}$. The QCD jets are created using the $pp \to jj$ process in MadGraph, while the top and W jets we examine are produced through a Z' which decays to $t\bar{t}$ or a W' which decays to a W and invisibly decaying Z. Samples are available in ref. [65] for a variety of top and W masses, but we use only those with the SM values. There are around 600,000 QCD dijet events and 100,000 events for the "anomalous" top and W events. We reserve 100,000 QCD events for testing and use 50,000 QCD events for validation when training the VAE.

The leading jet in each event is used for the analysis. We pre-process the raw four-vectors into an image following the procedure presented in ref. [73]. Using the EnergyFlow package [74], we boost and rotate the jet along the beam direction so that the p_T weighted centroid is located at $(\eta, \phi) = (0, 0)$. Next, the jet is rotated about the centroid such that the p_T weighted major principal axis is vertical. After this, the jet is flipped along both the horizontal and vertical axes so that the maximum intensity is in the upper right quadrant. Only after the centering, rotations, and flipping do we pixelate the data [73]. We use 40×40 pixel images covering a range of $\Delta\eta = \Delta\phi = 3.2$. The final step of the pre-processing is to divide by the total p_T in the image. Note that we do *not* standardize each pixel by, e.g., subtracting the mean and dividing by the standard deviation for the entire training dataset, because optimal transport requires positive values in every pixel. It is important to note that the individual images are very sparse and do not resemble the average of the dataset. For instance, out of the 1600 pixels, only 10.4 ± 5.3, 13.5 ± 4.3, and 10.1 ± 3.3 pixels account for more than 1% of the total p_T of the image for the QCD, top, and W jets, respectively.

3 Defining the anomaly score

Anomaly detection, in general, requires an anomaly score: we want to determine if an event is anomalous by measuring how far away it is from a typical background event. This anomaly score can also be thought of as the "distance" between an event and an ensemble.

JHEP03(2022)066

In order to define an event-to-ensemble distance it is helpful first to explore event-to-event distance measures. For instance, given an event-to-event metric, one could compute the distance from an event to some fiducial background event, and use this as a proxy for the event-to-ensemble distance. To understand both types of distances, we need to review the metrics used to define the distance, which we will do in section 3.1. We can also use an autoencoder to generate an implicit construction of an approximate event-to-ensemble distance, in the form of an anomaly score. We will provide background and discuss the architecture of our autoencoder in section 3.2.

3.1 Metrics

First, we define the metrics that can be used to compute event-to-event distances. One of the simplest event representations is to treat an event as an image, with pixel intensities representing the particles' transverse momentum [75].[2] A simple event-to-event metric, the "mean power error" (MPE), can then be written as:

$$d^{(\alpha)}_{\text{MPE}}(\mathcal{I}_1, \mathcal{I}_2) = \frac{1}{N_{\text{pixels}}} \sum_{i \in \text{pixels}} |\mathcal{I}_{1,i} - \mathcal{I}_{2,i}|^\alpha \ . \tag{3.1}$$

where $\mathcal{I}_{1(2),i}$ is the pixel intensity (transverse momentum) in pixel i of the image 1(2), and α is a parameter that governs the relative importance of pixels with high/low intensity differences. This type of metric is often used for doing regression. Frequently, the choice $\alpha = 2$ is made, inspired by the χ^2 statistic, in which case $d^{(2)}_{\text{MPE}}$ is known as the mean-square error (MSE). The mean-absolute error (MAE) is another well-known choice, corresponding to $\alpha = 1$.

While $d^{(\alpha)}_{\text{MPE}}$ makes sense in regression, using it on images does not make much sense from a physics point of view.[3] For instance, let $\mathcal{I}_1$ be the image of a particle with energy E in a single pixel and $\mathcal{I}_2$ be the image of a particle with same energy E in the neighboring pixel. These events are nearly identical physically, but will have a very large MSE distance. Moreover, we will still get the same MSE distance if we move one of the two pixels much further way. Physically similar events do not necessarily result in small MSE distances.

A completely different way to assign distance between two events is to compute the minimum "effort" needed to transform one image into the other, known as the optimal transport distance. There are many possible optimal transport algorithms (see ref. [76] for a broad review). Finding the minimum effort is an optimization problem: given a cost function c_{ij}, where i and j label elements (e.g. pixel labels) of the two events, we optimize

[2] In principle, it would be interesting to consider the complete set of four-vectors of the particles in an event as a representation, rather than the pixelated image, and define a metric on these. The p-Wasserstein distances described later in this section are well-suited for such a representation, but building an autoencoder architecture on the full set of four-vectors is more challenging. It is also important to comment that our image representation is dependent on its preprocessing. Although recent studies have shown that the processing done to events before anomaly detection is inherently model dependent [46], we work with the images as described.

[3] In contrast, if one designs a neural network with higher-level variables as the input data representation, using MPE as the metric is a sensible choice.

JHEP03(2022)066

over the transport plan, f_{ij}. The cost can be thought of as how much work it takes to transport a single unit of intensity a given distance, and the plan describes how much intensity to transport and where to transport it to. In terms of the cost and plan, the total optimal transport cost d_{OT} is then defined as

$$d_{\text{OT}} = \min_{f} \sum_{i,j} f_{ij}\, c_{ij} \,. \tag{3.2}$$

In some cases, the cost function c_{ij} is itself a positive definite distance, in which case d_{OT} is also a distance. One example is the set of p-Wasserstein distances:

$$d^{(p)}_{\text{Wass}} = \left(\min_{f} \sum_{i,j} f_{ij}\ (c_{ij})^p \right)^{1/p} , \tag{3.3}$$

Depending on the problem, the set of f_{ij} may have to satisfy additional constraints.

We define the underlying cost c_{ij} as the Euclidean distance in the (η, ϕ) plane between pixel i in image $\mathcal{I}_1$ and pixel j in image $\mathcal{I}_2$. The transport plan f_{ij} is defined by the amount of p_T that is moved from pixel i in image $\mathcal{I}_1$ to pixel j in image $\mathcal{I}_2$. The transport plan is constrained such that the amount of p_T moved from a pixel cannot be more than what was there, $\sum_j f_{ij} \leq p_{T,i}$. Similarly the amount of p_T moved into a pixel cannot exceed the amount in that pixel in $\mathcal{I}_2$: $\sum_i f_{ij} \leq p'_{T,j}$. Here, we consider normalized images, preprocessed such that the total intensity summed over all pixels is equal to unity, so that there is no extra cost of creating or destroying p_T. In mathematical language, we are considering "balanced optimal transport".

In particle physics applications, unbalanced optimal transport with the choice $p = 1$ is commonly referred to as the Energy Movers Distance (EMD) [56, 57], as it has the interpretation of work required to rearrange an energy pattern. This interpretation makes the EMD a natural choice for a metric on collider events. This has prompted further work on using the EMD to define event shape observables characterizing the event isotropy [77], which can be useful in searching for signals that are very non-QCD-like [78, 79]. Sometimes, $p > 1$ has been considered [57], while the case of $0 < p < 1$ has been less explored. Intuitively, $p < 1$ gives more importance to smaller distances. While the EMD includes an additional term to account for energy differences between jets, in our results, we will restrict to balanced optimal transport, since we normalize the images.

The p-Wasserstein optimal transport metrics are more aligned with what one expects for physical events than MPE. For example, two single-particle events where the particles are nearby will have a much smaller p-Wasserstein distance than when they are far from each other, in contrast to their MSE distance. We find that the 1-Wasserstein distance and MSE have mild correlations, as shown in figure 1.

We reiterate that both $d^{(\alpha)}_{\text{MPE}}$ and $d^{(p)}_{\text{Wass}}$ are used to compare the distance between two images (or events). However, for anomaly detection, we want to know how far an event is from the expected distribution. One way to do this is with an autoencoder, which we describe next.

JHEP03(2022)066

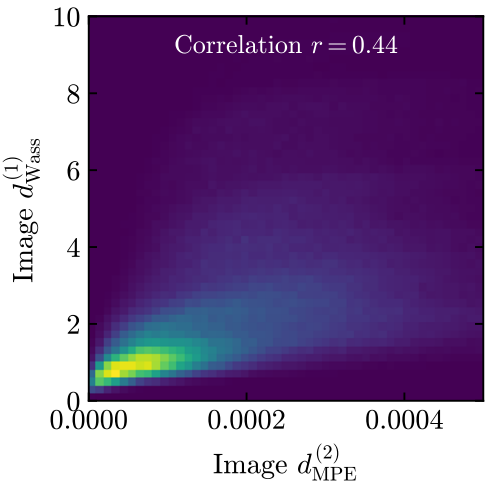

Figure 1. The pairwise (event-to-event) distances between event images for different metrics. The mean squared error ($d^{(2)}_{\text{MPE}}$) is displayed along the x-axis and the 1-Wasserstein distance ($d^{(1)}_{\text{Wass}}$) is along the y-axis.

3.2 Autoencoders

A popular method for detecting anomalous data is with a neural-network autoencoder (AE). An autoencoder works by first encoding the data in a lower-dimensional *latent space*, and then decoding it back to the original higher-dimensional representation. The idea is that data similar to the training sample will be reconstructed well, whereas data that is not similar to the training sample may be reconstructed poorly. The reconstruction fidelity can then be used as an anomaly score. Often the data are represented as images, and the autoencoder uses the MSE metric (eq. (3.1) with $\alpha = 2$) to compare the input image to the reconstructed image.

In figure 2, we show an example of an autoencoder architecture that we will use, which we implement in pytorch [80]. The encoder is made up of some number of *downsampling blocks* (there are two in the figure, each marked by a dashed blue line). Each block contains two sets of 3×3 convolutional layers with a depth of five filters. The stride and padding are set to keep the image size the same and the ELU activation function [81] is applied after each layer. After the convolutional layers, the data is downsampled through a 2×2 average pooling layer. After the final downsampling block, the data is flattened and then followed by a dense layer with 100 nodes and an ELU activation. Finally the network is mapped to the latent space through another dense layer. We experiment with one, two, and three downsampling blocks, and use a fixed latent size of 64 dimensions. Our latent space is substantially larger than what is often used, for example ref. [28] uses a six dimensional latent representation and ref. [27] finds the optimal size to be around 20-34 for their top-tagging data. We chose the latent space size by preforming a small scan over latent dimension sizes [2, 4, 8, 16, 32, 64] on the "The Machine Learning Landscape of Top Taggers" data [82] and found that the larger latent space yielded better top tagging. We then changed to the current data set, as there are more signals to consider (i.e. the W).

JHEP03(2022)066

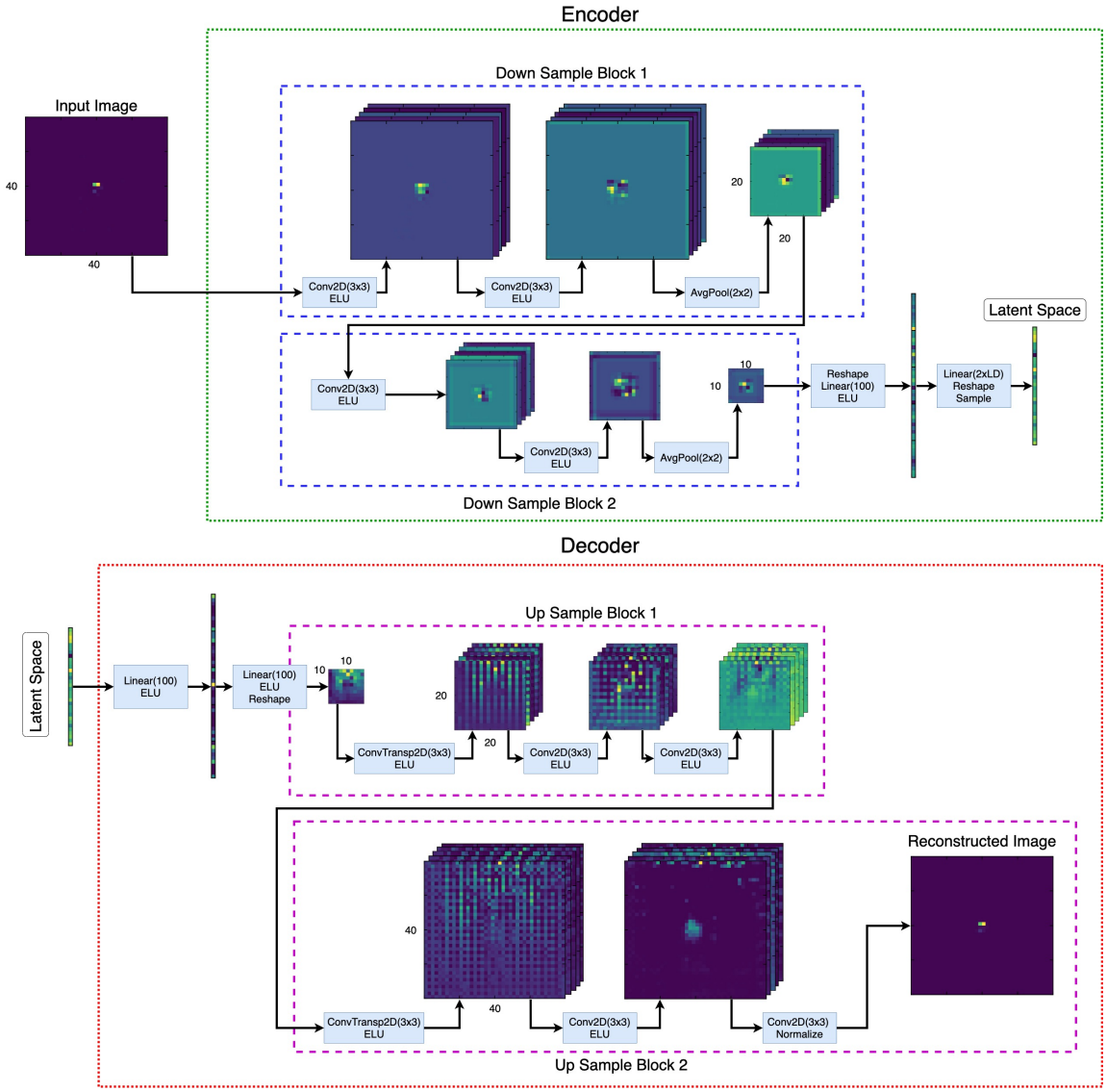

Figure 2. Example architecture of an autoencoder, as used for this study. The autoencoder is made of two networks, the encoder and the decoder, each with one, two, or three down(up)-sampling blocks.

The new data uses different transverse momentum cuts, so re-optimizing would in general be required. However, part of the point of this paper is to point out that one cannot optimize without a signal model in mind.[4]

The second part of the AE is the decoder, which maps the latent space back to the space of the input data. In our setup, the decoder is a mirror of the encoder. The first step is a dense layer with 100 nodes and ELU activation. From here, another dense layer is

[4]On the other hand, as discussed below, we can view the KL-divergence (and β) as regularizing the network. There will be a strong correlation with the optimal value of β with the size of the latent space. So making a different selection for the latent dimension would lead to a different value of β, but would not change the story.

JHEP03(2022)066

used, where the number of nodes is set to the number of pixels in the final downsampling block. The ELU activation function is used again, and then the data is reshaped into a square array. From here, the same number of *upsampling blocks* is applied as the number of downsampling blocks. In each upsampling block, the first operation is a 2D transposed convolution which doubles the shape of the image and contains a depth of five filters, followed by the ELU activation. After this, two 3×3 convolutional layers are used with the ELU activation with the stride and padding set to keep the image size the same. The final convolution operation reduces the depth to one channel.

During training and inference, the input image is compared with the reconstructed image via some choice of event-to-event metric. A common method is to use the MSE as the loss function, with the aim of reproducing the exact image. However, it is possible to use other metrics for the comparison. Furthermore, the metric used for training does not need to be the same as the metric used for the anomaly score (see for instance ref. [37]). We will refer to the difference between the input and reconstructed image as the *image distance*, also known as the reconstruction error.

A variational autoencoder enhances the basic autoencoder by adding stochasticity to the latent embedding. In a regular autoencoder, which is a deterministic function, very dissimilar events can be placed near each other in the latent space. Distances in the latent space of an ordinary AE therefore do not have a precise meaning. In a VAE, the stochastic element makes the network return a distribution in the latent space for each input event. Since the same input data can be mapped to several nearby points in a VAE, dissimilar events cannot be placed nearby. Returning a distribution in the latent space is therefore essential for making distances in the latent space meaningful. The stochasticity also connects the loss to the statistics method of variational inference [60, 83], as we summarize in appendix A (see also refs. [83, 84] for reviews). Specifically, we show that the autoencoder estimates a *lower bound* on the likelihood for any given event given the assumption that the event comes from the background distribution that the network is trained on.

To implement the stochasticity of a VAE, our networks are trained using the standard reparameterization trick [60, 85]. A single element of the input data now yields a distribution, and these distributions are treated as a set of D independent Gaussian distributions, where D is the dimension of the latent space. The output of the encoder is then doubled: instead of returning a single point in the latent space, it now outputs both the means μ and the variances σ^2 of the distribution in latent space. The loss function for the network also has to be modified: we want the background sample to be well modeled by a set of Gaussian distributions in latent space. This is done by introducing a Kullback-Leibler divergence (KLD) term (see appendix A for details), which is estimated as:[5]

$$\mathrm{KLD} = -\frac{1}{2}(1 + \log\sigma^2 - \mu^2 - \sigma^2). \tag{3.4}$$

[5]This estimation assumes Standard Normal priors for the likelihood of the latent data, as described in appendix A. There is a great deal of ongoing research into methods to improve the likelihood estimate by changing the latent space priors or improving the posterior approximations of the encoder [2, 48, 86, 87].

JHEP03(2022)066

This KLD term acts to regularize the autoencoder by pushing the means in the latent space to zero and the variances to one. Depending on the metric used to determine the distance between the original and reconstructed data, more or less regularization may be needed. To account for this, we introduce another hyperparameter β, and define the loss function as

$$L = (1-\beta) \times \text{Image distance} + \beta \times \text{KLD} \ . \tag{3.5}$$

We scan over $\beta \in \{0, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$, typically finding the best results for small but nonzero β.

To minimize the loss given in eq. (3.5), we use the Adam optimizer [88] with the default parameters and an initial learning rate of 10^{-3}. The training data consists of around 550,000 QCD dijet events, and we reserve 50,000 QCD events for an independent validation set. After each epoch of training, the loss is evaluated on the validation set. When the loss has not improved on the validation set for five epochs, the learning rate is decreased by a factor of 10, with a minimum learning rate of 10^{-5}. Training concludes when the validation loss has not improved for 12 epochs. We then restore the weights of the network from the epoch with the best validation loss.

4 Autoencoder results

Here we present the results of our studies of variational autoencoders. We start by studying the metric dependence of VAE performance as anomaly detectors. Then we study the latent space to understand what the VAE is learning.

4.1 Autoencoder performance

Now we study the performance of variational autoencoders as anomaly detectors using different metrics. Anomaly detection with an autoencoder requires two metric choices. First, one must choose a *training metric*, used for computing the image distance during training. Next, one must choose an *anomaly metric* to compute an anomaly score which determines how similar an event is to the training sample. The training metric and anomaly metric can be the same, but do not have to be.

For the training metric, we consider MSE-type metrics $d^{(2)}_{\text{MPE}}$ and $d^{(1)}_{\text{MPE}}$ and p-Wasserstein metrics $d^{(1)}_{\text{Wass}}$ and $d^{(2)}_{\text{Wass}}$. Using a p-Wasserstein metric in the loss function to train an autoencoder is not standard, and requires a little bit of extra engineering.[6] The challenge is that the optimal-transport metrics are not well-suited for the back-propagation part of the training procedure of a neural network. To get around this, we used the Sinkhorn approximation within the GeomLoss package [89]. Even with this, training was slow and sometimes timed out after three days of training on GPU. In contrast, the MSE and MAE networks typically completed training in around 12 hours on the same platform.

For the anomaly metric, we consider either using the full loss (including both the training metric contribution and the KL-divergence part in the variational autoencoder), just the MSE error between the input and output images ($d^{(2)}_{\text{MPE}}$), the MAE ($d^{(1)}_{\text{MPE}}$), or the

[6] Ref. [64] also implements a VAE trained with a p-Wasserstein metric.

JHEP03(2022)066

p-Wasserstein distances ($d^{(p)}_{\text{Wass}}$) with $p = 0.5$, 1.0, and 2.0. The value of each of these is computed for the test samples for the QCD dijet events, the top-jet events, and the W-jet events.

To evaluate performance in anomaly detection, we train the autoencoder on a QCD background using the training metric. Then we evaluate the anomaly score using the anomaly metric for a boosted top jet signal sample and a boosted W-jet signal sample. For a figure of merit of performance we use the Area Under the receiver operating characteristic Curve (AUC). We also include the signal efficiency at a cut which allows only 10% of the QCD events to pass, which is denoted $\epsilon_S(\epsilon_B = 0.1)$.

Results are shown in table 1 for the training metric choices $d^{(2)}_{\text{MPE}}$ and $d^{(1)}_{\text{Wass}}$ and for different numbers of downsampling blocks in the network. For each number of down samplings, we trained the network with different values of the VAE parameter β, and in the table present the results for the value of β which achieved the smallest loss on the validation data. For the $d^{(2)}_{\text{MPE}}$ trained networks, the values of β which minimized the loss were 10^{-7}, 10^{-7}, and 10^{-8}, for the one, two, and three down sample block networks, respectively. The $d^{(1)}_{\text{Wass}}$ trained results are in the lower part of the table and had optimal values of β of 10^{-5}, 10^{-8}, and 10^{-7} for one, two, and three down sampling blocks, respectively. The entries highlighted in blue indicate the configuration with the best AUC and $\epsilon_S(\epsilon_B = 0.1)$ for top jets and W jets across all of our considered VAE architectures, training methods, and anomaly score methods. The top row in the table shows the results (in red) from a supervised approach, for comparison (see appendix B for details of the supervised algorithm).

In general, we find the networks trained with $d^{(2)}_{\text{MPE}}$ as the training metric and using the full loss as the anomaly metric has the best AUC. The exception is when only a single down sample layer is used, in which case using $d^{(1)}_{\text{Wass}}$ as the anomaly metric does slightly better for the top-jet signal than using the full loss as the anomaly metric. When $d^{(1)}_{\text{Wass}}$ is used as the training metric, the best performance is with $d^{(2)}_{\text{MPE}}$ as the anomaly metric.

We can see at this stage the proliferation of choices one has to make when deciding what architecture, training metric, and anomaly metric to use. Making these choices is especially hard to do if one wants to remain model agnostic. For instance, figure 3 shows the results of the network trained with $d^{(2)}_{\text{MPE}}$ as the reconstruction loss. The left panel contains the loss on the QCD validation events. Using the idea that minimizing the loss is getting a better estimate of the probability of an event, one would expect that the network configuration (number of down samplings and value of β) which minimizes the loss will have learned the QCD distribution the best. However, the remaining panels show the ability of the networks to distinguish top and W jets from the QCD background in the upper and lower panels, respectively. The middle panels display the AUC and the right panels show the signal acceptance at a cut that allows only 10% of the QCD background events to pass. In particular, we see that the value of β which minimizes any of the loss curves does not yield the best signal separation. We also point out that the network with a single down sample block has the lowest loss, but is consistently the worst anomaly detector. This figure also highlights the danger of using the metrics of a particular signal to chose the

JHEP03(2022)066

Signal			Top jet		W jet	
Training Metric	Down Samplings	Anomaly Metric	AUC	$\epsilon_S(\epsilon_B = 0.1)$	AUC	$\epsilon_S(\epsilon_B = 0.1)$
Supervised	-	-	**0.94**	**0.81**	**0.96**	**0.91**
MSE	1 ($\beta = 10^{-7}$)	Loss	0.82	0.45	0.61	0.10
		MSE	0.82	0.45	0.60	0.10
		MAE	0.79	0.34	0.48	0.03
		Wass(0.5)	0.82	0.42	0.45	0.04
		Wass(1)	0.83	0.47	0.41	0.05
		Wass(2)	0.81	0.45	0.39	0.08
	2 ($\beta = 10^{-7}$)	Loss	0.83	0.48	**0.65**	0.14
		MSE	0.83	0.48	0.65	0.14
		MAE	0.80	0.37	0.53	0.04
		Wass(0.5)	0.82	0.43	0.51	0.04
		Wass(1)	0.82	0.44	0.51	0.04
		Wass(2)	0.81	0.44	0.54	0.06
	3 ($\beta = 10^{-8}$)	Loss	**0.84**	0.49	0.65	0.12
		MSE	0.84	0.48	0.65	0.12
		MAE	0.81	0.39	0.53	0.04
		Wass(0.5)	0.83	0.46	0.52	0.04
		Wass(1)	0.84	0.51	0.52	0.05
		Wass(2)	0.82	**0.51**	0.54	0.08
Wass(1)	1 ($\beta = 10^{-5}$)	Loss	0.78	0.35	0.44	0.04
		MSE	0.71	0.23	0.57	0.12
		MAE	0.72	0.20	0.49	0.03
		Wass(0.5)	0.75	0.26	0.47	0.03
		Wass(1)	0.78	0.35	0.44	0.04
		Wass(2)	0.76	0.37	0.39	0.05
	2 ($\beta = 10^{-8}$)	Loss	0.79	0.37	0.46	0.04
		MSE	0.76	0.33	0.61	**0.15**
		MAE	0.75	0.26	0.52	0.04
		Wass(0.5)	0.77	0.31	0.49	0.03
		Wass(1)	0.79	0.37	0.46	0.04
		Wass(2)	0.77	0.38	0.40	0.06
	3 ($\beta = 10^{-7}$)	Loss	0.79	0.36	0.41	0.03
		MSE	0.79	0.38	0.60	0.13
		MAE	0.77	0.31	0.51	0.03
		Wass(0.5)	0.79	0.33	0.47	0.03
		Wass(1)	0.79	0.36	0.41	0.03
		Wass(2)	0.7	0.32	0.36	0.06

Table 1. Results showing the ability of a VAE trained on QCD only samples to distinguish top and W jets from QCD jets. The Training Metric column shows which distance metric is used in the loss function for training, and the Anomaly Metric column shows the distance metric used at inference time. The bold blue entries mark the highest AUCs and signal efficiencies overall. We indicate the p-Wasserstein distance metric as Wass(p), and the MPE with power $\alpha = 1, 2$ by MAE and MSE, respectively. The number in parenthesis in the Down Sampling column denotes the value of β which yields the lowest total loss on the validation set for the given number of down samplings.

JHEP03(2022)066

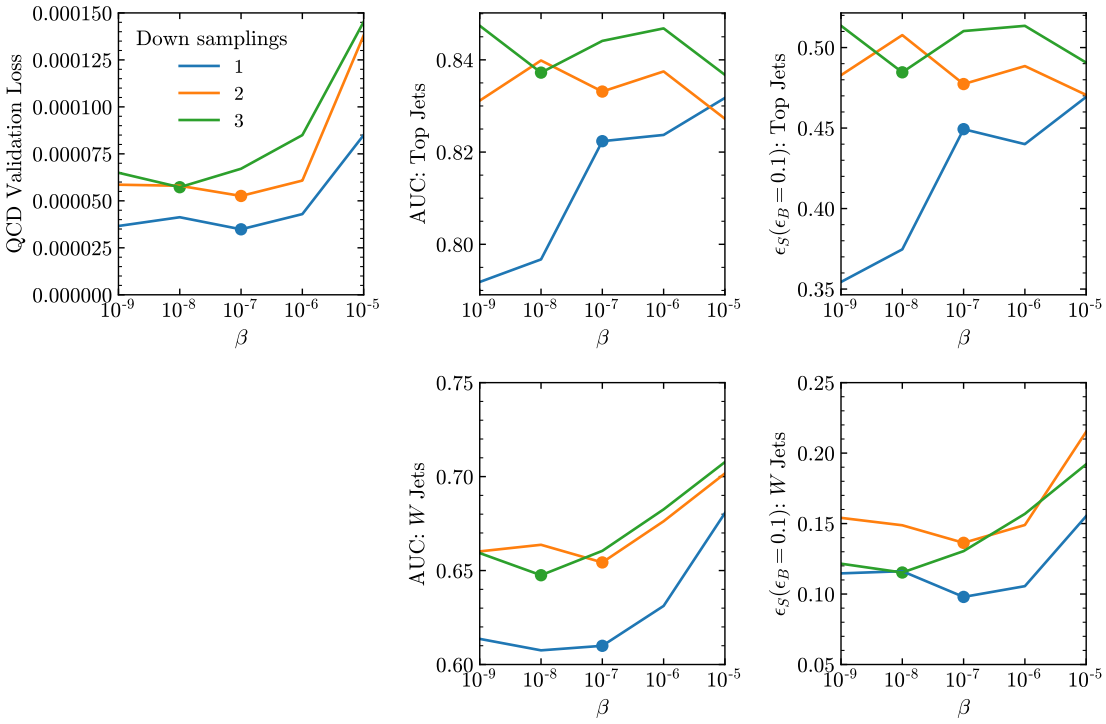

Figure 3. Results from scanning over β. The value of β which minimizes the validation loss does not yield the highest AUCs or signal acceptence for fixed background rejection for either the top or W samples. If one were to use one of the signal samples to chose the value of β, it can lead to worse results on the other signal.

hyperparameters of a universal anomaly detector. Examining only performance on the top jets, it would be tempting to pick the three down sample networks with a value of $\beta = 10^{-9}$, as this gives the best AUC and signal acceptance for the fixed background rejection for the tops. However, these particular networks have poor score for the W jets. This is the challenge of signal independent searches; without a signal model in mind, optimizing analysis strategies is hard to do in a principled manner.

The network trained with $d^{(2)}_{\text{MPE}}$ with a small KL divergence term yields the best anomaly detection performance. Therefore, we expect that it is learning a good representation of the underlying background distribution. We next explore this hypothesis by examining event-to-event distances among different metrics.

4.2 What has the VAE learned?

In order for a variational autoencoder to be able to judge the likelihood of an event given the assumption that it came from the set of training data, it must have a representation of the probability distribution of events in the training sample. Moreover, since it first maps events to a lower-dimensional latent space, the information about the relative likelihood should be encoded in the latent space in some way. It would make sense if the network places similar events *nearby* in the latent space, and dissimilar events far apart. In this

JHEP03(2022)066

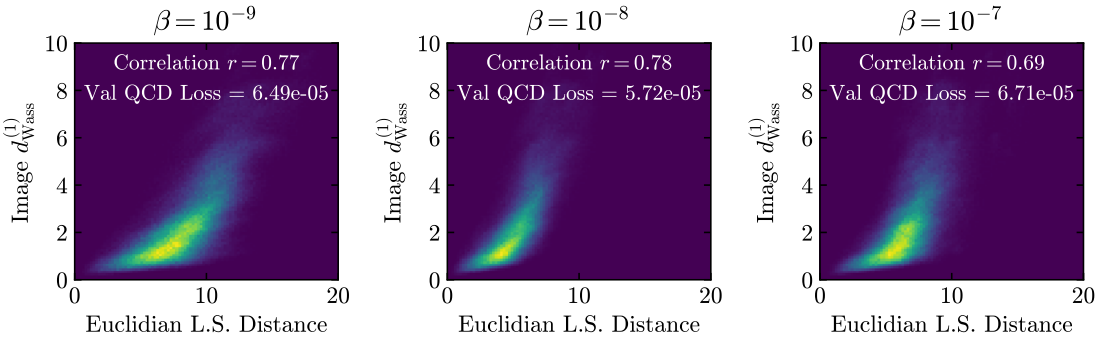

Figure 4. Each panel shows correlations between pair-wise distances of events in the QCD test set. The y-axis always denotes the $d^{(1)}_{\text{Wass}}$ distance. The x-axis denotes the Euclidean distance in the latent space. The representation learned by the network is more correlated with the $d^{(1)}_{\text{Wass}}$ distances than the MSE distances (see figure 1). The latent space distances were computed from networks trained with an MSE loss function, with two downsampling steps.

section we attempt to quantify if this is indeed true by comparing to the more physical Wasserstein distance.

Since each input is mapped to a (Gaussian) distribution in the latent space, we use the Euclidean distance between the means of these distributions, which is a simple measure of the distance in the latent space.[7] In figure 4, we show the correlations between the Euclidean latent space distance and the 1-Wasserstein event distance among all the $\sim 10^6$ pairings of 1000 events in the QCD test set for various values of the VAE parameter β. For this study, the events are passed through the encoder part of a VAE with three down sampling layers, down to a 64 dimensional latent space where the Euclidean distance is computed. As the value of β is increased, the network goes from having little regularization to being forced to approach a Gaussian. Correspondingly, the correlation initially grows as the structure is forced upon the latent representation, and then decreases as β becomes so large that the regularization dominates and the distribution becomes nearly Gaussian. We observe similar results for the networks with one and two downsampling layers that are trained with $d^{(2)}_{\text{MPE}}$ in the loss function. In this figure, the value of β which gives the minimum loss corresponds to the β with maximum correlation, but we do not find this trend to hold in general. It seems that the VAE with an intermediate value of β that balances the $d^{(2)}_{\text{MPE}}$ and KLD terms in the loss function creates a latent space where distances between events are correlated with the $d^{(1)}_{\text{Wass}}$ distance in the image space.

The downsampling operations are critical to the production of the latent space. As they combine information from neighboring pixels, they introduce an element of scale which MPE would not exhibit. To verify the importance of downsampling, we show in figure 5 the pair-wise event distance correlations for the same network at different depths into the encoder. In the first panel, distances on the x-axis are computed in the first downsampling

[7] One could also try to take into account the variance of the distributions, by e.g., taking the KL divergence between the two distributions.

JHEP03(2022)066

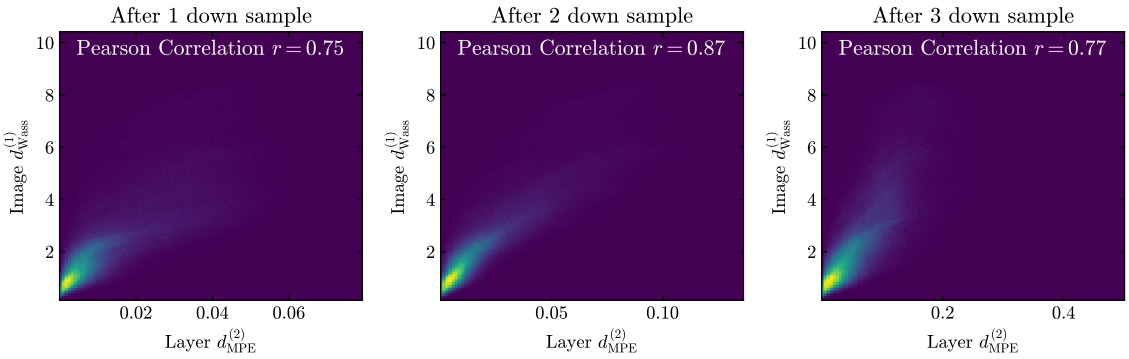

Figure 5. Correlations of the pair-wise event distance in the image space with the interior activations of the network after different numbers of downsampling blocks.

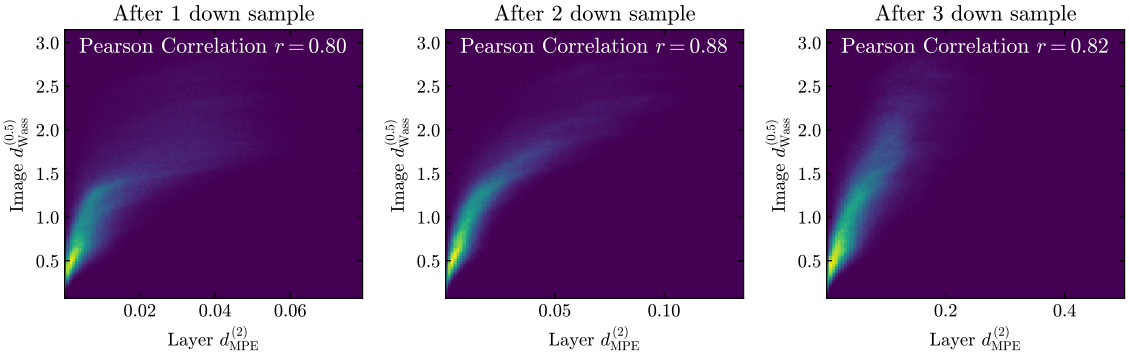

Figure 6. Correlations of the pair-wise event distance in the image space with the interior activations of the network after different numbers of down sample blocks. This is the same network as in figure 5, but now the image distances use a different power of p. The correlation is higher with $p = 0.5$ than $p = 1$.

block, where the events are represented as $20 \times 20 \times 5$ tensors and the $d^{(2)}_{\text{MPE}}$ goes across all 2000 "pixels" (see figure 2). The correlation between the distance in this first downsampling layer and the Wasserstein distance of the events is much larger than the MSE distance between the original events. The correlation further increases from the first down sample block to the second. The correlation then decreases after a third downsampling. Then, when the information is further reduced to the latent space, we get smaller correlations than seen in the early stages of the network.

Although the EMD metric is a p-Wasserstein metric with $p = 1$, there is no clear reason why $p = 1$ should be preferred to other values. So, next we consider $p = 0.5$. In figure 6, we show the correlations for the same network with three down sampling layers but now using $d^{(0.5)}_{\text{Wass}}$ distances along the y-axis. The distances between events at different layers in the network are $\sim 5\%$ more correlated with $d^{(0.5)}_{\text{Wass}}$ than $d^{(1)}_{\text{Wass}}$.

In this section, we have explored the representation of the QCD event distribution that variational autoencoders learn. Our conclusion from this study is that the Euclidean distances between QCD events in the latent space are highly correlated with the p-Wasserstein

JHEP03(2022)066

distances between the events themselves. This is particularly compelling because the VAE is trained with the MSE metric for its loss function and has no direct access to any p-Wasserstein metric. A related question is how the correlations would look if a p-Wasserstein metric were used for training. In that case, we find that the Euclidean distances between events in the latent space of the Wasserstein trained networks are even more correlated with the Wasserstein distances in the image space. Thus, it could be argued that the Wasserstein trained networks learn an even better representation of the QCD distribution than the MSE trained networks. However, the VAE with MSE training worked better for finding the top- and W-jet signals than those trained with a Wasserstein loss. The fact that the method with the "best" latent representation does not yield the best signal separation highlights the challenges of model agnostic anomaly detection.

5 Event-to-ensemble distance

In the previous section we showed that VAEs tend to work better when MSE loss is used for training than when Wasserstein metrics are used for training and that the Euclidean distance in the latent space correlates strongly with the Wasserstein metric on the data, regardless of the metric used for training. If the power of the VAE for anomaly detection in physical problems stems from it implicitly learning aspects of the p-Wasserstein metrics, we can then ask if there may be a way to use these metrics more directly for anomaly detection, sidestepping the VAE entirely. One way to do this is to use the metric to compute an event-to-ensemble distance, as we explore in this section.

We would like to use the p-Wasserstein distance, or another metric, to characterize the distance of an event to the background ensemble. There are already several options for using Wasserstein distances to characterize different types of events in the literature, such as k nearest neighbor (kNN) classifiers [56], "linearized" optimal transport [58], where all the events are compared to a single reference event and this is used to define a new distance, and event isotropy, which compares a given event to an isotropic configuration [77]. Our goal, using a method like these, is to extract from the background ensemble one or more representative images and to compute the distance of a given signal or background event to those images. This direct event-to-ensemble distance measure can then be compared to the VAE anomaly score, which is also effectively an event-to-ensemble distance.

To compute the direct event-to-ensemble distance we need an algorithm to select or construct fiducial events from the ensemble and a metric with which to compute the distance. As with the VAE architecture, there may be no choice that is optimal for all signals. In choosing the fiducial events, we must decide which sample to choose events from, how to select those events, how many events to use, how to represent the fiducial events (e.g. as images), and how to combine the distances to the different events. To make a fair comparison to the VAE approach, we would like our algorithm for generating fiducial events to depend only on the background sample, independent of what anomalous signal we might search for. Thus we choose the QCD jet event ensemble as our reference sample. To select events from the sample, the simplest possibility is to arbitrarily select some number of random images. However, despite occasionally giving a large AUC for classification, results with

JHEP03(2022)066

random images are very sensitive to fluctuations between the random images. A second possibility that may seem sensible is to take the average of all events in the sample. A third option, which we find to be the most natural, is to use k medoids as we now explain.

With a given metric, which we call the medoid metric, we can compute the pairwise distance $d(x_i, x_j)$ between any two events in the ensemble. Then for each event x we can sum over all the distances to all other events $d(x) = \sum_j d(x, x_j)$. The *medoid* of the ensemble is the event x that minimizes $d(x)$. k *medoids* generalizes this to finding the k events for which the sum of the distances of each event to the closest of the k medoids is minimized. Thus the event fragments into a set of clusters, with each cluster closest to one of the k medoids. k-medoids clustering is similar to k-means clustering when the medoid metric is chosen to be the Euclidean metric, except that k-medoid clustering actually requires the medoid to be one of the events in the set. Medoids have previously been explored in other contexts in refs. [56, 57, 90, 91]. To use k medoids, we need to choose a value for k and a medoid metric. Then it is natural to take for the event-to-ensemble distance the distance of an event to its closest medoid. Although one could in principle use a different metric to compute the event-to-ensemble distance, it is also most natural to use the same medoid metric that determines the medoids.

Choosing the number of medoids k is challenging to do in a signal independent way. One approach is the elbow method, a common heuristic for determining how many clusters are in a dataset. In our case, to use the elbow method we scan over the number k of medoids, and for each k compute the distances of all the events in our sample to the nearest medoid and histogram the results. There will be a small number of events very close to a medoid and a small number very far from all medoids, so the histograms will have a peak at some finite value of the distance. Moreover, the distance to the peak will decrease monotonically as the number of medoids increases. In many applications the decrease is rapid for small k, but at some k this behavior abruptly slows down. Thus the peak distance as a function of k often has an elbow shape. To determine the elbow location algorithmically we perform a linear regression to an elbow function (two straight lines), and take the first integer value after the elbow as the suggested value of k. The result can be seen in figure 7. The idea behind the elbow method is that increasing k past the location of the elbow should not give much improvement. Moreover, in the case of anomaly detection, if we have too many medoids, we can get one medoid that looks "signal-like" rather than "background-like". We find that typically $k \sim 2 - 4$ medoids is selected according to this elbow method.

The main advantage of the elbow method is that it can be automated and used independent of the sample or the use case. However, there often is not a clear elbow. In figure 7, the elbow is only apparent because we have fit to a piecewise linear function. The data seems to follow more of a power law behavior. In addition, the location of the elbow can be affected by the maximum number of medoids we include in the fit. Additionally, the elbow can only be computed once we've already made the arbitrary choice of the medoid metric, and of the metric being used for the comparison between the full sample and the reference sample. Finally, there is no reason to expect that the elbow choice of k, which is determined only by the background sample, would be optimal for anomaly detection tasks. Thus, we also consider values of k not determined by the elbow method for this study.

JHEP03(2022)066

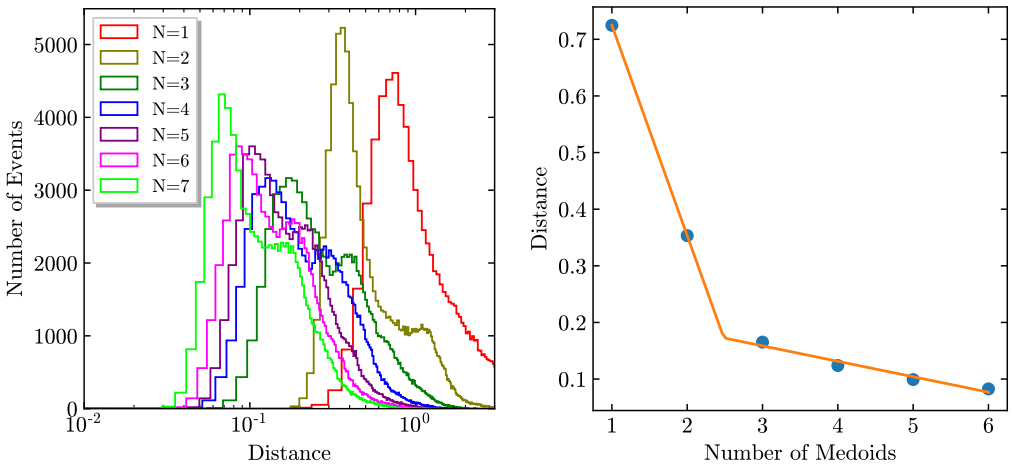

Figure 7. Example of the elbow method. Left shows histograms of the 1-Wasserstein distance to the closest medoid, with colors corresponding to the number of medoids. Right plots the peak of each of these histograms, as a function of the number of medoids.

The top of table 2 shows results for top-jet vs. QCD-jet discrimination and W-jet vs. QCD-jet discrimination when QCD jets are used for the reference sample. We show results for different values of k with medoids, using different medoid metrics. We also show the result from using the distance to a single composite average event determined by averaging each pixel intensity over all events in the reference sample. For each case, we include both the AUC and the signal efficiency at a cut which allows only 10% of the QCD events to pass. When we study the elbow for the most common 1-Wasserstein metric, we see reasonable performance for both QCD and top jets, though it is best for neither of them. This is in line with what we expect for unsupervised anomaly detection. For simplicity, we report results where the metric used to select the medoids is the same one used to compute our observable. We could have chosen two different metrics for the medoid metric and that used to compute the event-to-ensemble distance, but restricting to the case where they are the same does not qualitatively change our results. Table 2 shows that the number of medoids and the choice of metric matters substantially.

We find that the QCD medoids typically perform better than the average QCD jet. This is not surprising, since the average QCD jet is much more concentrated in the center of the image than any real QCD jet, as can be seen in figure 8. In this figure, the color shows the fraction of the total p_T in each pixel on a logarithmic scale. We also find better performance for anomaly detection with 5–6 medoids, rather than the 2–4 medoids suggested by the elbow method. When detecting top jets with QCD reference images, we get the best results when the p-Wasserstein metric with $p = 1$ is used to compare images, though we also find reasonable performance for the p-Wasserstein metric with $p = 0.5$ or $p = 2$ (not shown in the table) and when the MAE metric is used. Although MAE is not a physically motivated metric, the performance in this case is not surprising because MAE between events is highly correlated with p-Wasserstein distances between events for QCD jets (the Pearson correlation coefficient between MAE and the 1-Wasserstein distance is 0.87).

JHEP03(2022)066

				Top jet		W jet	
Reference Sample	Metric	Number of medoids	Method	AUC	$\epsilon_S(\epsilon_B = 0.1)$	AUC	$\epsilon_S(\epsilon_B = 0.1)$
Supervised	-	-	-	**0.94**	**0.81**	**0.96**	**0.91**
QCD Reference	Wass(1)	-	Avg	0.81	0.33	0.62	0.02
		1	Medoid	0.83	0.28	0.63	0.02
		3 (elbow)	Medoids (min)	0.85	0.43	0.67	0.04
		5	Medoids (min)	**0.87**	**0.54**	0.60	0.05
		7	Medoids (min)	0.87	0.54	0.61	0.05
	Wass(5)	-	Avg	0.53	0.10	0.60	0.03
		1	Medoid	0.62	0.21	0.36	0.03
		3	Medoids (min)	0.66	0.19	0.41	0.05
		4 (elbow)	Medoids (min)	0.67	0.22	0.41	0.04
		5	Medoids (min)	0.71	0.24	0.43	0.04
	MAE	-	Avg	0.83	0.47	0.71	0.08
		1	Medoid	0.82	0.40	**0.71**	0.07
		3 (elbow)	Medoids (min)	0.82	0.49	0.61	**0.08**
		5	Medoids (min)	0.83	0.48	0.67	0.08
		7	Medoids (min)	0.83	0.48	0.65	0.08
Top Reference	Wass(1)	-	Avg	0.69	0.17	0.69	0.04
		1	Medoid	0.58	0.20	0.79	0.31
		3 (elbow)	Medoids (min)	0.32	0.07	0.79	0.53
		5	Medoids (min)	0.45	0.12	**0.84**	**0.62**
		7	Medoids (min)	0.49	0.13	0.83	0.60
	Wass(5)	-	Avg	0.72	0.18	0.40	0.01
		1	Medoid	0.53	0.12	0.52	0.05
		2 (elbow)	Medoids (min)	0.72	**0.32**	0.70	0.06
		3	Medoids (min)	0.66	0.20	0.61	0.04
		5	Medoids (min)	0.61	0.16	0.54	0.03
	Wass(5)	3 (elbow)	Medoids (sum)	0.66	0.27	0.66	0.06
		5	Medoids (sum)	**0.73**	0.30	0.58	0.02
		7	Medoids (sum)	0.75	0.31	0.60	0.02
	MAE	-	Avg	0.48	0.05	0.57	0.05
		1	Medoid	0.29	0.04	0.64	0.23
		3 (elbow)	Medoids (min)	0.25	0.03	0.36	0.03
		5	Medoids (min)	0.31	0.10	0.58	0.31

Table 2. Results for QCD vs. signal classification, with signal labeled in the top row. Top rows use a QCD reference sample, and bottom rows use a top reference sample (assuming W events are more "top-like" than QCD events). When there are multiple medoids, distances are combined either by taking the minimum or the sum of the distances to the k different medoids, as denoted in the table. Medoids are selected with the same metric as the one used to compare images. For each metric, we note which number of medoids corresponds to the elbow. The best AUC and $\epsilon_S(\epsilon_B = 0.1)$ values for each reference sample are denoted in blue. We indicate the p-Wasserstein distance metric as Wass(p), and the MPE with power $\alpha = 1, 2$ by MAE and MSE, respectively.

JHEP03(2022)066

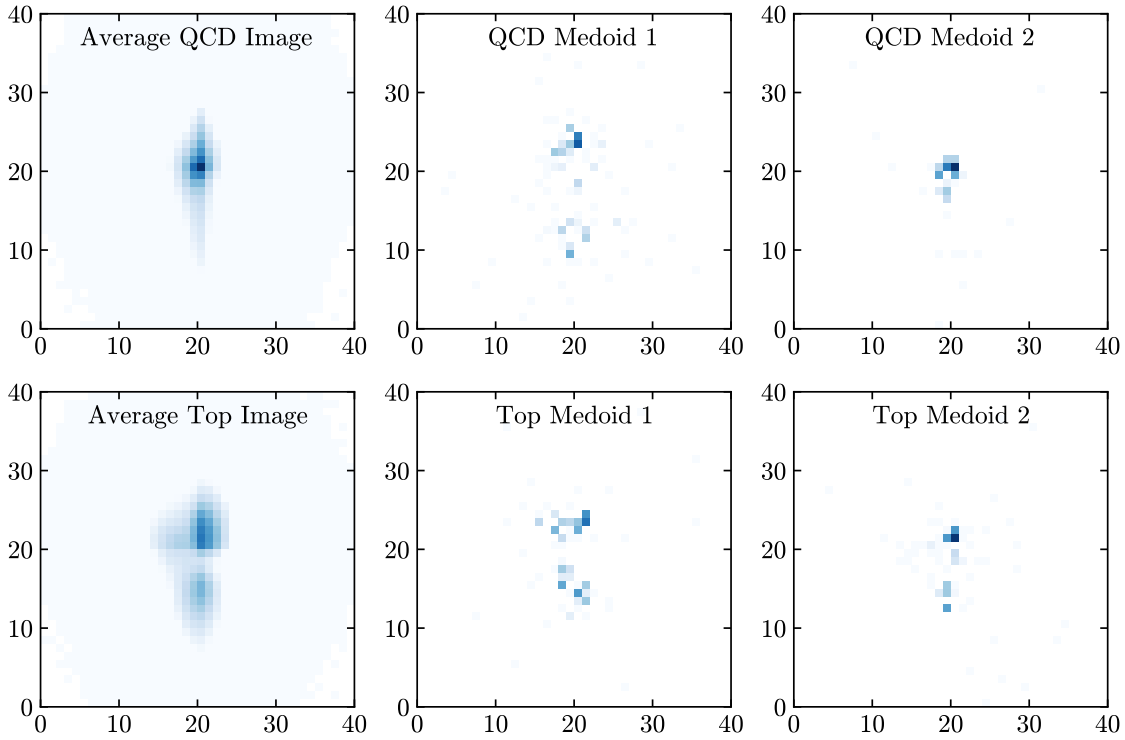

Figure 8. Images of example QCD and top events. The top row shows QCD events; the bottom row shows top events. The left column shows the average image in each case, and the other two columns show two medoids computed with the 1-Wasserstein metric. Note medoids are more sparse and varied than average images, and that one of the top medoids appears "QCD-like" when we include multiple medoids.

That our results depend on the exponent p is suggestive. For the p-Wasserstein metric with larger p we get substantially decreased performance when comparing to QCD reference jets. This suggests that the ideal value of p is related to the relevant scales in the problem: a smaller value of p places comparatively larger emphasis on pixels with smaller differences. This is consistent with results such as ref. [46], which finds better AE performance when pixel intensities are rescaled to emphasize dim pixels. When we choose a better (smaller) value of p, the results are also slightly less sensitive to exactly which QCD reference images are chosen than when a larger value is chosen.

While autoencoders are often trained on a QCD background, several studies have explored trying to train an AE on alternative samples. One example is ref. [46], which showed that AEs perform poorly when tagging QCD jets if trained on a top jet sample. This can be attributed to top jets being more complex than QCD jets, so that an AE trained on top jets can also reconstruct QCD jets despite them being out of distribution samples. While modifications can be made so that an AE trained on top jets can tag QCD jets [46, 48], requiring sample dependent optimization defeats the point of unsupervised anomaly detection.

JHEP03(2022)066

Unlike in the case of an AE, event-to-ensemble distances can be directly applied to other reference samples, assuming an applicable metric is selected. We include in the bottom half of table 2 results using a top-jet reference sample for concreteness and brevity, but the method can be easily applied to other reference samples. We find that for top reference samples, the best metric is not the same as for QCD reference samples. In contrast to the case of the QCD reference sample, we find the p-Wasserstein metric with higher p does better for QCD vs. top classification than that with $p = 1$. We suspect that top reference samples usually prefer a large value of the exponent p because a larger power prioritizes larger differences in pT over smaller ones, and effective classification using top reference samples should deprioritize small pT differences, since the top reference samples are spread out with substantial pixel-to-pixel variation. In contrast, the QCD reference sample case has fewer populated pixels with less variation, so a smaller value of the exponent p is needed. However, this result is signal dependent, in addition to being dependent on the background sample: when trying to use the top reference sample to distinguish QCD vs. W-jet events, we find using the p-Wasserstein metric with $p = 1$ is better than higher p. We also find that whether the average event or minimum distance to the medoids does better depends on the signal sample — for QCD versus top classification with a top reference sample the average event does better than medoids (unlike the QCD reference sample case), but the opposite is true for QCD vs. W classification. Furthermore, we find the somewhat counter intuitive result that the sum of the distance to the medoids does better than the minimum for top vs. QCD classification with top reference jets, which is surprising because only the distance to the closest event is actually used when determining the medoids. For QCD vs. top classification, using a QCD reference sample still outperforms the top reference sample, but the opposite is true when doing QCD vs. W classification.

Our best results using the event-to-ensemble distance approach are comparable to and even slightly exceed the performance of the VAE in the previous section, which can be seen from comparing table 2 to table 1. This suggests that if we choose a smart, physically motivated metric like the p-Wasserstein distance then we can use the medoid method, which is much faster and simpler than the VAE, and avoids optimizing all the hyperparameters in the VAE network architecture. The trade-off is that we need to put effort into optimizing the metric and choice of k for the medoid approach. This is not surprising, since as we saw in the previous section, distances in the VAE latent space between two separate encoded latent representations are fairly correlated with the p-Wasserstein distance between the original images. This equivalence is further supported by our study of the number of downsampling layers in the previous section. Like in the case of the p-Wasserstein metrics, locality is incorporated in the convolutional VAE on scales other than the arbitrary pixel size due to the convolutions and down samplings.

The ease and speed of using the event-to-ensemble approach is a distinct advantage when compared to AEs, where the architecture, normalization, and parameters all need to be optimized. However, since the ideal metric and fiducial sample selection depend on both the background and signal samples, the signal dependence of the event-to-ensemble approach further suggests that there may be advantages of weakly/semi-supervised learning as compared to unsupervised learning, and that weakly/semi-supervised methods should

JHEP03(2022)066

be explored further. The potential advantages of semi-supervised learning can be further seen from the fact that the best QCD vs. W AUC values come from top reference samples, rather than QCD reference samples.

6 Conclusions

Using an autoencoder for anomaly detection is particularly challenging, since it must be trained well enough to reconstruct the background, but not so well that it also reconstructs the signal. There are many details about the network configuration that need to be optimized, such as the network size, metric used to compare input and output images, the definition of the anomaly score, and hyperparameters. Many of the results currently in the literature do not sufficiently emphasize these difficulties, so we have attempted in this paper to characterize and resolve them. To be concrete, we considered detecting boosted hadronic top and W jets over a QCD-jet background. We considered using different metrics for training the VAE; we found that using the more physically-motivated optimal transport-based metrics did not outperform the simpler mean-squared-error metrics, and actually performed slightly worse. We found that the optimal values of various hyperparameters depend on the signal that we are trying to detect and that the optimal hyperparameters for describing the QCD sample are not necessarily those that detect anomalies the best.

In order to understand what the autoencoder has learned, we also studied the autoencoder latent space. The latent space provides a representation of any particular event which can be used to study the background distribution. In order to characterize this latent space, we computed the distance between distinct events. One way to do this is by using the Euclidean distance between quantities in the latent space. Alternatively, if we rely on a more physical, optimal transport based metric, we can compute the distances between images directly. When we compared the two, we found that the event-to-event optimal transport based distances between the background events are highly correlated with the Euclidean distances between events in the latent space of the autoencoder. This suggests that the autoencoder is learning some aspects of optimal transport, despite being trained with only a mean-squared-error based loss function.

This motivated us to develop methods that use optimal transport more directly. By choosing a representation of the QCD background distribution, such as the average QCD image or several medoids of the set of QCD jets, we can directly compute the optimal transport distance to this fiducial sample and use it as an anomaly score. We found that this method is at least as effective as the autoencoder, with the added benefits of being easier and faster to optimize, and generalizing more easily than the autoencoder to more complicated background distributions. We also found that the best choice of optimal transport metric depends on both the new physics signal and the qualities of the expected background distribution. Before using this medoid method in practice, we suggest additional studies of this method, including exploring background sculpting, testing the effects of signal contamination when selecting a background reference sample, and studying the correlation of this anomaly score with known observables.

JHEP03(2022)066

Although we have shown that the performance of variational autoencoders can be reproduced, and improved upon, by the relatively simpler medoid method, neither approach is very close to optimal for signal detection. To be quantitative, when trained on a QCD sample, the best autoencoder performance we found gave an AUC of 0.65 for W detection (see table 1). The best performance using medoids with a QCD background gave a slightly better AUC of 0.71 (see table 2). These are both worse than the performance of a fully supervised network which gave a nearly perfect AUC of 0.96. Somewhat surprisingly, we found that when the medoids method was used on a top-jet background sample, it found W jets over QCD better (AUC of 0.84) than when trained on a QCD background. This is comparable to what a supervised network trained to find tops over QCD gives when tested on W vs. QCD (AUC of 0.86). These observations suggest that a path forward might be to use a semi-supervised approach [43, 92–94], where a network is trained with an example signal in mind, and then used for anomaly detection more broadly.

Acknowledgments

We thank Jack Collins, Philip Harris, and Sang Eon Park for useful discussions and comments on a previous version of this manuscript. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/) The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. BO was partially supported by the Department of Energy (DOE) Grant No. DE-SC0020223. KF is supported in part by NASA Grant 80NSSC20K0506 and in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745303. SH is supported in part by the DOE Grant DE-SC0013607, and in part by the Alfred P. Sloan Foundation Grant No. G-2019-12504. RKM is supported by the National Science Foundation under Grant No. NSF PHY-1748958 and NSF PHY-1915071.

A Variational inference for autoencoders

The idea behind variational inference for anomaly detection is to estimate the true probability distribution of the background, $p(\mathbf{x})$. Assuming we have an underlying latent space of elements z, we can write $p(\mathbf{x})$ as

$$p\left(\mathbf{x}\right) = \mathbb{E}_{p(z)}\left[p\left(\mathbf{x}|z\right)\right] \equiv \int p\left(\mathbf{x}|z\right)\, p(z)\, dz, \tag{A.1}$$

where $\mathbb{E}$ denotes the expectation value, $p\left(\mathbf{x}|z\right)$ is the probability of $\mathbf{x}$ given z, and $p(z)$ is the prior likelihood of the latent data. We can take the latent space prior to be a set of independent Gaussians with zero mean and unit standard deviation, $z_i \sim \mathcal{N}(0,1)$, where i runs over the dimension of the latent space. At this point $p(\mathbf{x}|z)$ is an unknown and intractable distribution.

To make progress, we introduce a new *tractable* distribution $q_\phi(z|\mathbf{x})$, where ϕ are some parameters to be optimized over. In an autoencoder architecture, this is the encoder. We

JHEP03(2022)066

can then write eq. (A.1) in a more useful form:

$$p(\mathbf{x}) = \int q_\phi(z|\mathbf{x}) \frac{p(\mathbf{x}|z)}{q_\phi(z|\mathbf{x})} p(z) dz = \mathbb{E}_{q_\phi(z|\mathbf{x})} \left[\frac{p(\mathbf{x}|z)p(z)}{q_\phi(z|\mathbf{x})} \right] . \tag{A.2}$$

The log likelihood, $\log p(\mathbf{x})$, is then given as

$$\log p(\mathbf{x}) = \log \mathbb{E}_{q_\phi(z|\mathbf{x})} \left[\frac{p(\mathbf{x}|z)p(z)}{q_\phi(z|\mathbf{x})} \right] \tag{A.3}$$

$$\geq \mathbb{E}_{q_\phi(z|\mathbf{x})} \left[\log \left(\frac{p(\mathbf{x}|z)p(z)}{q_\phi(z|\mathbf{x})} \right) \right] \tag{A.4}$$

$$= \mathbb{E}_{q_\phi(z|\mathbf{x})} \left[\log p(\mathbf{x}|z) - \log \left(\frac{q_\phi(z|\mathbf{x})}{p(z)} \right) \right] , \tag{A.5}$$

where we have used Jensen's inequality in the second line above. Let's first consider the first term in the last line in eq. (A.5). It is the expectation value of $\mathbf{x}$ given z when z is sampled from $q_\phi(z|\mathbf{x})$ (which is a distribution in z given $\mathbf{x}$). This term can be interpreted as a (negative) reconstruction error term. If we approximate $p(\mathbf{x}|z)$ by a decoder part of the architecture $p_\theta(x|z)$ (where θ is to be optimized over), $\mathbb{E}_{q_\phi(z|\mathbf{x})}(p_\theta(\mathbf{x}|z))$ is the usual (negative) reconstruction error term in the loss function for an autoencoder with decoder $p_\theta(\mathbf{x}|z)$ and encoder $q_\phi(z|\mathbf{x})$.

The second term is by definition the Kullback-Leibler divergence (KLD) between the distributions $q_\phi(z|\mathbf{x})$ and $p(z)$. Recall that $p(z) \sim \mathcal{N}(0, 1)$. We take $q_\phi(z|\mathbf{x})$ to also be a Gaussian distribution, but with a unknown mean and standard deviation (to be fixed by the optimization), i.e. $q_\phi(z|x) = \mathcal{N}(\mu(x), \sigma^2(x))$. The KLD between these two distributions is then given exactly by eq. (3.4). Using the reparameterization trick [60, 85], we can write $q_\phi(z|x)$ in terms of a standard normal:

$$z \sim q_\phi(z|x) \, , z = \mu(x) + \sigma(x)\epsilon \, , \epsilon \sim \mathcal{N}(0, 1) \, . \tag{A.6}$$

Using the reparameterization trick allows for more efficient training of the network, as the back propagation of the gradients extends to the parameters of the distribution (μ and θ) even though a random draw from the distribution is passed to the decoder.

It's now clear that the last line in eq. (A.5) is the negative loss for a VAE architecture. By training the VAE, we are minimizing the loss. By the inequality in eq. (A.5), the last line is also a lower limit for the log likelihood. The optimized VAE therefore gives a *maximized* lower bound to the log likelihood, the so called Evidence LOwer Bound (ELBO). Notice that in this discussion it is imperative to use the full VAE loss in order for it to have the variational inference interpretation.

B Supervised results

It is well known that anomaly detection is sub-optimal for looking for any particular model; if the signal is known before-hand, supervised classification will yield the best results. We use a similar setup for our supervised classification as we did for the VAEs. The network

JHEP03(2022)066

consists of 1, 2, or 3 convolution blocks. Each block is made of two successive convolutional layers with 5 filters with a kernel size of 3 pixels, followed by an ELU activation function. After the convolutions, the data is down sampled with a 2×2 average pool operation. Following the convolution blocks, the data is flattened to a vector and a fully connected layer reduces the output to a single number with a sigmoid activation.

The networks are trained using 50000 events from the QCD sample and 50000 events from either the top or W samples. Similarly, 5000 events from each dataset are used for validation and to stop the network training when the validation loss has stopped improving. The training minimizes the binary cross entropy. After training, the network is applied to the test data of 5000 events in each class. We find that the network with three down sample layers achieves the best AUCs, with a score of 0.94 for top tagging and 0.96 for W tagging.

References

[1] G. Kasieczka et al., *The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics*, *Rept. Prog. Phys.* **84** (2021) 124201 [arXiv:2101.08320] [INSPIRE].

[2] T. Aarrestad et al., *The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider*, *SciPost Phys.* **12** (2022) 043 [arXiv:2105.14027] [INSPIRE].

[3] J.H. Collins, K. Howe and B. Nachman, *Anomaly Detection for Resonant New Physics with Machine Learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

[4] R.T. D'Agnolo and A. Wulzer, *Learning New Physics from a Machine*, *Phys. Rev. D* **99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].

[5] A. De Simone and T. Jacques, *Guiding New Physics Searches with Unsupervised Learning*, *Eur. Phys. J. C* **79** (2019) 289 [arXiv:1807.06038] [INSPIRE].

[6] A. Casa and G. Menardi, *Nonparametric semisupervised classification for signal detection in high energy physics*, arXiv:1809.02977 [INSPIRE].

[7] J.H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, *Phys. Rev. D* **99** (2019) 014038 [arXiv:1902.02634] [INSPIRE].

[8] B.M. Dillon, D.A. Faroughy and J.F. Kamenik, *Uncovering latent jet substructure*, *Phys. Rev. D* **100** (2019) 056002 [arXiv:1904.04200] [INSPIRE].

[9] A. Mullin, S. Nicholls, H. Pacey, M. Parker, M. White and S. Williams, *Does SUSY have friends? A new approach for LHC event analysis*, *JHEP* **02** (2021) 160 [arXiv:1912.10625] [INSPIRE].

[10] R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer and M. Zanetti, *Learning multivariate new physics*, *Eur. Phys. J. C* **81** (2021) 89 [arXiv:1912.12155] [INSPIRE].

[11] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*, *Phys. Rev. D* **101** (2020) 075042 [arXiv:2001.04990] [INSPIRE].

JHEP03(2022)066

[12] A. Andreassen, B. Nachman and D. Shih, *Simulation Assisted Likelihood-free Anomaly Detection*, *Phys. Rev. D* **101** (2020) 095004 [arXiv:2001.05001] [INSPIRE].

[13] ATLAS collaboration, *Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector*, *Phys. Rev. Lett.* **125** (2020) 131801 [arXiv:2005.02983] [INSPIRE].

[14] B.M. Dillon, D.A. Faroughy, J.F. Kamenik and M. Szewc, *Learning the latent structure of collider events*, *JHEP* **10** (2020) 206 [arXiv:2005.12319] [INSPIRE].

[15] K. Benkendorfer, L.L. Pottier and B. Nachman, *Simulation-assisted decorrelation for resonant anomaly detection*, *Phys. Rev. D* **104** (2021) 035003 [arXiv:2009.02205] [INSPIRE].

[16] V. Mikuni and F. Canelli, *Unsupervised clustering for collider physics*, *Phys. Rev. D* **103** (2021) 092007 [arXiv:2010.07106] [INSPIRE].

[17] G. Stein, U. Seljak and B. Dai, *Unsupervised in-distribution anomaly detection of new physics through conditional density estimation*, in *34th Conference on Neural Information Processing Systems*, Online Conference Canada (2020) [arXiv:2012.11638] [INSPIRE].

[18] J. Batson, C.G. Haaf, Y. Kahn and D.A. Roberts, *Topological Obstructions to Autoencoding*, *JHEP* **04** (2021) 280 [arXiv:2102.08380] [INSPIRE].

[19] A. Blance and M. Spannowsky, *Unsupervised event classification with graphs on classical and photonic quantum computers*, *JHEP* **21** (2020) 170 [arXiv:2103.03897] [INSPIRE].

[20] B. Bortolato, B.M. Dillon, J.F. Kamenik and A. Smolkovič, *Bump Hunting in Latent Space*, arXiv:2103.06595 [INSPIRE].

[21] J.H. Collins, P. Martín-Ramiro, B. Nachman and D. Shih, *Comparing weak- and unsupervised methods for resonant anomaly detection*, *Eur. Phys. J. C* **81** (2021) 617 [arXiv:2104.02092] [INSPIRE].

[22] T. Dorigo, M. Fumanelli, C. Maccani, M. Mojsovska, G.C. Strong and B. Scarpa, *RanBox: Anomaly Detection in the Copula Space*, arXiv:2106.05747 [INSPIRE].

[23] S. Volkovich, F. De Vito Halevy and S. Bressler, *A Data-Directed Paradigm for BSM searches: the bump-hunting example*, arXiv:2107.11573 [INSPIRE].

[24] A. Hallin et al., *Classifying Anomalies THrough Outer Density Estimation (CATHODE)*, arXiv:2109.00546 [INSPIRE].

[25] J.A. Aguilar-Saavedra, J.H. Collins and R.K. Mishra, *A generic anti-QCD jet tagger*, *JHEP* **11** (2017) 163 [arXiv:1709.01087] [INSPIRE].

[26] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty Detection Meets Collider Physics*, *Phys. Rev. D* **101** (2020) 076015 [arXiv:1807.10261] [INSPIRE].

[27] T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson, *QCD or What?*, *SciPost Phys.* **6** (2019) 030 [arXiv:1808.08979] [INSPIRE].

[28] M. Farina, Y. Nakai and D. Shih, *Searching for New Physics with Deep Autoencoders*, *Phys. Rev. D* **101** (2020) 075021 [arXiv:1808.08992] [INSPIRE].

[29] O. Cerri, T.Q. Nguyen, M. Pierini, M. Spiropulu and J.-R. Vlimant, *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, *JHEP* **05** (2019) 036 [arXiv:1811.10276] [INSPIRE].

[30] T.S. Roy and A.H. Vijay, *A robust anomaly finder based on autoencoders*, arXiv:1903.02032 [INSPIRE].

JHEP03(2022)066

[31] A. Blance, M. Spannowsky and P. Waite, *Adversarially-trained autoencoders for robust unsupervised new physics searches*, *JHEP* **10** (2019) 047 [arXiv:1905.10384] [INSPIRE].

[32] M. Romão Crispim, N.F. Castro, R. Pedro and T. Vale, *Transferability of Deep Learning Models in Searches for New Physics at Colliders*, *Phys. Rev. D* **101** (2020) 035042 [arXiv:1912.04220] [INSPIRE].

[33] O. Amram and C.M. Suarez, *Tag N' Train: a technique to train improved classifiers on unlabeled data*, *JHEP* **01** (2021) 153 [arXiv:2002.12376] [INSPIRE].

[34] M. Crispim Romão, N.F. Castro, J.G. Milhano, R. Pedro and T. Vale, *Use of a generalized energy Mover's distance in the search for rare phenomena at colliders*, *Eur. Phys. J. C* **81** (2021) 192 [arXiv:2004.09360] [INSPIRE].

[35] O. Knapp, O. Cerri, G. Dissertori, T.Q. Nguyen, M. Pierini and J.-R. Vlimant, *Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark*, *Eur. Phys. J. Plus* **136** (2021) 236 [arXiv:2005.01598] [INSPIRE].

[36] M. Crispim Romão, N.F. Castro and R. Pedro, *Finding New Physics without learning about it: Anomaly Detection as a tool for Searches at Colliders*, *Eur. Phys. J. C* **81** (2021) 27 [*Erratum ibid.* **81** (2021) 1020] [arXiv:2006.05432] [INSPIRE].

[37] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette and T. Golling, *Variational Autoencoders for Anomalous Jet Tagging*, arXiv:2007.01850 [INSPIRE].

[38] C.K. Khosa and V. Sanz, *Anomaly Awareness*, arXiv:2007.14462 [INSPIRE].

[39] P. Thaprasop, K. Zhou, J. Steinheimer and C. Herold, *Unsupervised Outlier Detection in Heavy-Ion Collisions*, *Phys. Scripta* **96** (2021) 064003 [arXiv:2007.15830] [INSPIRE].

[40] J.A. Aguilar-Saavedra, F.R. Joaquim and J.F. Seabra, *Mass Unspecific Supervised Tagging (MUST) for boosted jets*, *JHEP* **03** (2021) 012 [*Erratum ibid.* **04** (2021) 133] [arXiv:2008.12792] [INSPIRE].

[41] A.A. Pol, V. Berger, G. Cerminara, C. Germain and M. Pierini, *Anomaly Detection With Conditional Variational Autoencoders*, in *Eighteenth International Conference on Machine Learning and Applications*, 10, 2020 [arXiv:2010.05531] [INSPIRE].

[42] M. van Beekveld et al., *Combining outlier analysis algorithms to identify new physics at the LHC*, *JHEP* **09** (2021) 024 [arXiv:2010.07940] [INSPIRE].

[43] S.E. Park, D. Rankin, S.-M. Udrescu, M. Yunus and P. Harris, *Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge*, *JHEP* **21** (2020) 030 [arXiv:2011.03550] [INSPIRE].

[44] P. Chakravarti, M. Kuusela, J. Lei and L. Wasserman, *Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests*, arXiv:2102.07679 [INSPIRE].

[45] D.A. Faroughy, *Uncovering hidden new physics patterns in collider events using Bayesian probabilistic models*, *PoS* **ICHEP2020** (2021) 238 [arXiv:2012.08579] [INSPIRE].

[46] T. Finke, M. Krämer, A. Morandini, A. Mück and I. Oleksiyuk, *Autoencoders for unsupervised anomaly detection in high energy physics*, *JHEP* **06** (2021) 161 [arXiv:2104.09051] [INSPIRE].

JHEP03(2022)066

[47] O. Atkinson, A. Bhardwaj, C. Englert, V.S. Ngairangbam and M. Spannowsky, *Anomaly detection with convolutional Graph Neural Networks*, *JHEP* **08** (2021) 080 [arXiv:2105.07988] [INSPIRE].

[48] B.M. Dillon, T. Plehn, C. Sauer and P. Sorrenson, *Better Latent Spaces for Better Autoencoders*, *SciPost Phys.* **11** (2021) 061 [arXiv:2104.08291] [INSPIRE].

[49] A. Kahn, J. Gonski, I. Ochoa, D. Williams and G. Brooijmans, *Anomalous jet identification via sequence modeling*, 2021 *JINST* **16** P08012 [arXiv:2105.09274] [INSPIRE].

[50] S. Caron, L. Hendriks and R. Verheyen, *Rare and Different: Anomaly Scores from a combination of likelihood and out-of-distribution models to detect new physics at the LHC*, arXiv:2106.10164 [INSPIRE].

[51] E. Govorkova, E. Puljak, T. Aarrestad, M. Pierini, K.A. Woźniak and J. Ngadiuba, *LHC physics dataset for unsupervised New Physics detection at 40 MHz*, arXiv:2107.02157 [INSPIRE].

[52] J. Gonski, J. Lai, B. Nachman and I. Ochoa, *High-dimensional Anomaly Detection with Radiative Return in e^+e^- Collisions*, arXiv:2108.13451 [INSPIRE].

[53] E. Govorkova et al., *Autoencoders on FPGAs for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider*, arXiv:2108.03986 [INSPIRE].

[54] B. Ostdiek, *Deep Set Auto Encoders for Anomaly Detection in Particle Physics*, *SciPost Phys.* **12** (2022) 045 [arXiv:2109.01695] [INSPIRE].

[55] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **03** (2011) 015 [arXiv:1011.2268] [INSPIRE].

[56] P.T. Komiske, E.M. Metodiev and J. Thaler, *Metric Space of Collider Events*, *Phys. Rev. Lett.* **123** (2019) 041801 [arXiv:1902.02346] [INSPIRE].

[57] P.T. Komiske, E.M. Metodiev and J. Thaler, *The Hidden Geometry of Particle Collisions*, *JHEP* **07** (2020) 006 [arXiv:2004.04159] [INSPIRE].

[58] T. Cai, J. Cheng, N. Craig and K. Craig, *Linearized optimal transport for collider events*, *Phys. Rev. D* **102** (2020) 116019 [arXiv:2008.08604] [INSPIRE].

[59] M.A. Kramer, *Nonlinear principal component analysis using autoassociative neural networks*, *AIChE J.* **37** (1991) 233.

[60] D.P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, arXiv:1312.6114 [INSPIRE].

[61] J. An and S. Cho, *Variational autoencoder based anomaly detection using reconstruction probability*, Special Lecture on IE (2015).

[62] S. Carrazza and F.A. Dreyer, *Lund jet images from generative and cycle-consistent adversarial networks*, *Eur. Phys. J. C* **79** (2019) 979 [arXiv:1909.01359] [INSPIRE].

[63] K. Dohi, *Variational Autoencoders for Jet Simulation*, arXiv:2009.04842 [INSPIRE].

[64] J.H. Collins, *An Exploration of Learnt Representations of W Jets*, arXiv:2109.10919 [INSPIRE].

[65] T. Cheng, *Test sets for jet anomaly detection at the lhc*, zenodo (2021).

[66] J. Leissner-Martin, T. Cheng and J.-F. Arguin, *Qcd jet samples with particle flow constituents*, zenodo (2020).

JHEP03(2022)066

[67] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[68] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[69] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].

[70] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J. C* **72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[71] M. Cacciari and G.P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, *Phys. Lett. B* **641** (2006) 57 [hep-ph/0512210] [INSPIRE].

[72] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[73] S. Macaluso and D. Shih, *Pulling Out All the Tops with Computer Vision and Deep Learning*, *JHEP* **10** (2018) 121 [arXiv:1803.00107] [INSPIRE].

[74] https://energyflow.network/.

[75] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, *JHEP* **02** (2015) 118 [arXiv:1407.5675] [INSPIRE].

[76] C. Villani, *Optimal Transport, Old and New*, Springer, Heidelberg Germany (2009).

[77] C. Cesarotti and J. Thaler, *A Robust Measure of Event Isotropy at Colliders*, *JHEP* **08** (2020) 084 [arXiv:2004.06125] [INSPIRE].

[78] C. Cesarotti, M. Reece and M.J. Strassler, *Spheres To Jets: Tuning Event Shapes with 5d Simplified Models*, *JHEP* **05** (2021) 096 [arXiv:2009.08981] [INSPIRE].

[79] C. Cesarotti, M. Reece and M.J. Strassler, *The efficacy of event isotropy as an event shape observable*, *JHEP* **07** (2021) 215 [arXiv:2011.06599] [INSPIRE].

[80] A. Paszke et al., *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in Neural Information Processing Systems 32*, H. Wallach et al. eds., Curran Associates, Inc., Red Hook U.S.A. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[81] D.-A. Clevert, T. Unterthiner and S. Hochreiter, *Fast and accurate deep network learning by exponential linear units (elus)*, arXiv:1511.07289.

[82] A. Butter et al., *The Machine Learning landscape of top taggers*, *SciPost Phys.* **7** (2019) 014 [arXiv:1902.09914] [INSPIRE].

[83] D.P. Kingma and M. Welling, *An Introduction to Variational Autoencoders*, *Found. Trends Mach. Learn.* **12** (2019) 307 [arXiv:1906.02691] [INSPIRE].

[84] D.M. Blei, A. Kucukelbir and J.D. McAuliffe, *Variational inference: A review for statisticians*, *J. Am. Stat. Assos.* **112** (2017) 859.

[85] D. Jimenez Rezende and S. Mohamed, *Variational Inference with Normalizing Flows*," arXiv:1505.05770.

[86] D.P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever and M. Welling, *Improving Variational Inference with Inverse Autoregressive Flow*, arXiv:1606.04934.

JHEP03(2022)066

[87] R. van den Berg, L. Hasenclever, J.M. Tomczak and M. Welling, *Sylvester Normalizing Flows for Variational Inference*, arXiv:1803.05649.

[88] D.P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 12, 2014 [arXiv:1412.6980] [INSPIRE].

[89] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve and G. Peyré, *Interpolating between optimal transport and MMD using Sinkhorn divergences*," in *22nd International Conference on Artificial Intelligence and Statistics*, Naha Japan (2019), pg. 2681.

[90] P.T. Komiske, R. Mastandrea, E.M. Metodiev, P. Naik and J. Thaler, *Exploring the Space of Jets with CMS Open Data*, *Phys. Rev. D* **101** (2020) 034009 [arXiv:1908.08542] [INSPIRE].

[91] M. Crispim Romão, N.F. Castro, J.G. Milhano, R. Pedro and T. Vale, *Use of a generalized energy Mover's distance in the search for rare phenomena at colliders*, *Eur. Phys. J. C* **81** (2021) 192 [arXiv:2004.09360] [INSPIRE].

[92] L.M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly Supervised Classification in High Energy Physics*, *JHEP* **05** (2017) 145 [arXiv:1702.00414] [INSPIRE].

[93] T. Cohen, M. Freytsis and B. Ostdiek, *(Machine) Learning to Do More with Less*, *JHEP* **02** (2018) 034 [arXiv:1706.09451] [INSPIRE].

[94] P.T. Komiske, E.M. Metodiev, B. Nachman and M.D. Schwartz, *Learning to classify from impure samples with high-dimensional data*, *Phys. Rev. D* **98** (2018) 011502 [arXiv:1801.10158] [INSPIRE].

JHEP03(2022)066