

A robust anomaly finder based on autoencoder

Tuhin S. Roy^a and Aravind H. Vijay^b

^a*Department of Theoretical Physics, Tata Institute of Fundamental Research, Mumbai 400005, India*

^b*Department of High Energy Physics, Tata Institute of Fundamental Research, Mumbai 400005, India*

E-mail: tuhin@theory.tifr.res.in, aravind.vijay@tifr.res.in

ABSTRACT: We propose a robust method to identify anomalous jets by vetoing QCD jets. The robustness of this method ensures that the distribution of the discriminating variable, which allows us to veto QCD-jets, remain rather unaffected even if QCD-jets from different m/p_T bins are used as control samples. This suggest that using our method one can look for anomalous jets in high m/p_T bins, by simply training on jets from low m/p_T bins, where the data is surplus and pure in background. The robustness follows from coupling a simple fully connected autoencoder to a novel way of preprocess jets. We use momentum rescaling followed by a Lorentz boost to find the frame of reference where any given jet is characterized by predetermined mass and energy. In this frame we generate the jet image via constructing a set of orthonormal basis vector using the Gram-Schmidt method to span the plane transverse to the jet axis. Due to our preprocessing, the autoencoder loss function does not depend on the initial jet mass, momentum, or orientation while still offering remarkable performance. When combined only with the jet mass, our method performs equally well with state-of-the-art top taggers, which uses a large amount of physics information associated with top decays.

Contents

1	Introduction	1
2	Our Proposed Analysis	4
2.1	From an infrared safe jet to the jet image	4
2.1.1	Rescale the jet	5
2.1.2	Lorentz boost the jet	5
2.1.3	Forming the images	6
2.2	The Autoencoder	7
2.3	Summary of the whole algorithm	8
3	Results	8
3.1	Simulation Details	9
3.2	Robustness	10
3.2.1	Robustness at $\sqrt{s} = 13$ TeV	11
3.2.2	Robustness at $\sqrt{s} = 100$ TeV	13
3.2.3	Cross validation	15
3.3	Performance	16
3.3.1	Benchmark with top jets	17
3.3.2	Benchmark with W -jets	20
3.3.3	Benchmark with di- W jets from new physics	22
4	Conclusion	23

1 Introduction

With the recent discovery of the 125 GeV Higgs by CMS [1] and ATLAS [2] collaborations of the LHC, the last missing link in the standard model (SM) of particle physics is completed. The focus has now shifted to find physics beyond the standard model (BSM), for which there are plethora of candidates all designed to solve various short comings of SM. Unfortunately, we have failed to observe any trace of physics beyond the standard model (BSM) so far in all our searches for these “well-motivated” scenarios. At this juncture, it may well be equally justifiable to concentrate on the “Bottom-up” approach of looking for the “unexpected” in data, apart from the usual “Top-down” approach of first hypothesizing an extension to the SM and then predicting it’s phenomenology. Even though this “Bottom-up” approach is nothing other than looking for deviations from the SM background and there has been remarkable progress recently in automated precision (higher order) calculations (radiative corrections due to quantum chromodynamics (QCD) and electroweak (EW) interactions as well), it is still extremely difficult (impractical) to precisely estimate all backgrounds due

to SM *without* using data. In this context, searching for new physics (NP) as anomalies directly from data depict a qualitatively new and highly promising direction.

Note that the philosophy of hunting for NP as anomalies where data is used to estimate background is already a common practice in jet physics. A simple example is the search for boosted massive resonances using jet-mass m_J . The background is generically due to QCD-jets. The well-studied and well-understood variable $\frac{d\sigma}{dm_J}$ for QCD-jets has a continuously falling distribution at high m_J/p_T region. Any massive resonance decaying hadronically and boosted enough such that all its decay products is contained in a jet, shows up as a bump peaked around its mass. Looking for these bumps over a falling spectrum provides one such simplistic example where one does not need to rely on the exact nature of the massive resonance (as long as it decays hadronically). Knowing the falling distribution due to QCD is enough, in principle, to look for unknown new particles. Additionally, jet masses for boosted massive resonances are robust under various jet grooming procedure [3–6], owing to their substructures, as compared to jetmasses for QCD-jets. Searching for bumps in the distribution for groomed jetmasses, therefore, provide a model-independent way to look for hadronically decaying NP particles with substantial discovery potential. One can even take, massive resonances of SM like the W^\pm , Z gauge bosons, Higgs particle h , and top quark t , as test cases to hone this search procedure. A broader generalization of this philosophy, is given in Ref [7], where the authors use a large set of jet-shape observables to construct a veto for standard objects such as QCD-jets, photons, leptons, etc. Jets that pass this veto, would be termed anomalies. Examples of such anomalies include jets containing decay products (not restricted to hadrons only) of NP resonances, long-lived particles etc.

Machine learning can be of a big help in looking for such anomalies by constructing an anti-QCD tagger [8]. The traditional applications [9–28] of machine learning techniques for classification problems in collider physics are, however, centered around the idea of supervised learning where the training data is labeled into various categories and the model tries to correctly classify the training data into these categories. However, this method is inadequate when a model needs to be trained directly on data which is not labeled. This is where autoencoders, which are unsupervised learning algorithms [29–33], can be useful. Because of training, an autoencoder learns to reproduce QCD-jets efficiently, and it is expected to be not as effective at reproducing jets originating from other processes, resulting in larger “loss-functions”. Therefore, the loss function of the autoencoder can be used to find anomalies [31, 32]).

An essential requirement for a high performance machine learning based anomaly finder, is that it needs to be trained with control samples of QCD-jets. However, note that one expects anomalies to appear within samples of jets at high m/p_T bins, whereas it is difficult to obtain sufficiently pure samples of QCD-jets with high m/p_T from data itself. Apart from contamination effects, it is also difficult to gather sufficient training data in these high mass bins. This problem can be partially mitigated if one can use trainings on jets from low m/p_T bins, which are relatively more pure samples of QCD-jets and at the same time are easily available. In order to implement this strategy, however, we require an anomaly hunter that is *robust*. Technically speaking, by robustness we here refer to the

criteria that the distribution of the loss-function for QCD-jets remains unaltered even if training samples from different m/p_T bins are used.

Unfortunately, when an autoencoder is trained in a naive way, it tends to learn features which have large correlations with the jetmass. This is expected since the jet mass is a powerful observable to characterize a jet. Therefore, generically speaking, one does not expect a simple autoencoder based anomaly hunter or anti-QCD tagger to be robust. It is not a surprise that previous attempts in this direction [31, 32] find significant dependence of autoencoder loss on jetmasses. Consequently, cuts on such jetmass-depended loss functions necessarily sculpt the jetmass distribution itself, which makes techniques like bump hunts (for boosted massive resonances) and side band analyses much more challenging. Ref. [31] suggests using convolution autoencoders, since these are less affected by the above problem. However, the improvement in robustness is marginal[31]. On the other hand, Ref. [32] tries to alleviate this problem using adversary training to reduce the dependence of jet mass on autoencoder response. This technique requires an additional network to try and guess the mass from the output of the first network, which is added complexity requiring significantly larger computation to train the network and leads to decrease in transparency. Also, it is not a general purpose method in that it is not applicable for autoencoders based on Lorentz layers [32].

In this work, we present a robust anti-QCD tagger based on autoencoders. The robustness comes due to a series of novel preprocessing stages, and therefore can be easily generalized for many other cases such as an autoencoder based on the Lorentz layer or even supervised learning problems. Summarizing, the method proceeds as follows:

- Given a jet, the algorithm first rescales and then boosts to a frame of reference where the mass and energy of the jet gets fixed to predetermined values (denoted here by m_0 and E_0 respectively).
- The Histogram, representing the jet image is filled by using the dimensionless number E_i/E_0 , where E_i represents the energy of each constituent in this new frame. Also, in order to remove residual rotational symmetries present in the plane transverse to the jet-axis, the transverse plane is spanned using an orthonormal set of basis vectors, constructed using Gram-Schmidt procedure on either two hardest subjets, or two hardest constituents.

We find that this method has a remarkable effect that it makes QCD-jets in different m/p_T bins to appear similar as viewed by the network. As a result, an autoencoder trained on jet data in the lower m/p_T bins (where data is surplus and pure) can be used in searches for anomalies in a set of jets at the higher m/p_T bins (where data is scarce and may be contaminated). Computationally, it is a far simpler and cheaper solution for removing the dependence of autoencoder loss on the jet mass compared to the proposal in Ref. [32]. At the same time, this method yields significant more robust performance than Ref. [31]. These highly desirable characteristics in this proposal make it highly portable and easily usable in many important techniques like mass bump hunt and side band analyses at all energy scales.

We also take this opportunity to study the performance of a simple fully-connected autoencoder as an anomaly finder, where the jets are preprocessed using the methods proposed here. In order to benchmark, we treat topjets and W -jets as anomalies. Here topjets and W -jets refer to jets containing decay products of hadronically decaying boosted top and W respectively. We find that our anomaly finder, trained on only QCD-jets, yields performance comparable to Ref. [31]. On the other hand, when combined with the mass of jets, it tags top as efficiently as HEPTopTagger2 [34–37], even though it uses a large amount of physics information specific to top decays. Additionally, when combined with Nsubjettiness, it substantially outperform HEPTopTagger2 [34–37]. Summarizing, we propose a new anomaly finder that yields remarkable robustness without sacrificing performance. Additionally, we also benchmark the effectiveness of the anomaly finder when jets consisting of all decay products of a boosted NP particle decaying to W^+W^- , each of which in turn, decay hadronically, is treated as an anomalous jet. Being remarkably robust to giving high discovery potentials for finding anomalous jets, we find our method to generate impressive performances.

The rest of this document is organized as follows: in Sec. 2, we describe in detail our method of preprocessing, and give a brief description of the autoencoder we use for this study; in Sec. 3, we present main results of this study after specifying all simulation details; and finally we conclude in Sec. 4.

2 Our Proposed Analysis

Our method can be broadly divided into two main parts. In the first part, we take jets from the event as input and produce images corresponding to these jets as described in Subsec. 2.1. The second part of our method involves the autoencoder. In Subsec. 2.2, we describe the general features of the autoencoder and explain how it can be trained.

2.1 From an infrared safe jet to the jet image

Our method requires jets as inputs. A jet is basically a collection of four momenta of its constituents which are clustered according to any of the infrared-safe or IR-safe [38] sequential jet clustering algorithm like Anti- k_t [39], Cambridge/Aachen (C/A) [40] and k_t [41, 42] or IR-safe cone based algorithms (like SISCone) [4, 43]. In this work, we use the notation P_J^μ (\vec{P}_J) for the four-momentum (three-momentum) of the jet, p_i^μ (\vec{p}_i) for the four-momentum (three-momentum) of the i^{th} constituent in the jet. We use “E-scheme” [44], where the four-momenta of the jet is simply the vector sum of four-momenta of its constituents.

$$P_J^\mu = \sum_{i=1}^{N_J} p_i^\mu, \quad m_J^2 \equiv |P_J^\mu|^2 = (P_J^0)^2 - \left(\vec{P}_J\right)^2, \quad \text{and} \quad E_J \equiv P_J^0, \quad (2.1)$$

where N_J is the total number of constituents, m_J the jet mass, and E_J is the jet energy.

2.1.1 Rescale the jet

In the first step, all the jet constituent four-momenta are rescaled such that the new jet mass is given by a preassigned m_0 :

$$p_i^\mu \rightarrow p_i'^\mu = \frac{m_0}{m_J} \times p_i^\mu \quad \Rightarrow \quad P_J^\mu \rightarrow P_J'^\mu = \frac{m_0}{m_J} P_J^\mu. \quad (2.2)$$

Note that the energy of the “rescaled jet” $P_J'^\mu$, is given by $E_J' = E_J m_0 / m_J$.

2.1.2 Lorentz boost the jet

The next step involves performing a Lorentz boost (Λ_ν^μ) to a frame such that, in the new frame, all jets have a given energy E_0 . In other words, the rescaled jets ($P_J'^\mu$) are Lorentz transformed to boosted jets (\mathbf{P}_J^μ) such that:

$$p_i'^\mu \rightarrow \mathbf{p}_i^\mu = \Lambda_\nu^\mu p_i'^\nu \quad \Rightarrow \quad P_J'^\mu \rightarrow \mathbf{P}_J^\mu = \Lambda_\nu^\mu P_J'^\nu \quad (2.3)$$

such that: $\mathbf{E}_J \equiv \mathbf{P}_J^0 = E_0$.

Since this is a critical step in our method, we present a simple derivation. We need to determine both the direction of the boost and the γ factor required for the boost. Starting with the rescaled jet (with the given jet mass m_0 and energy E_J'), we can split the situation into 2 cases:

1. $E_J' > E_0$: the boost is along the three-momentum direction of the jet and
2. $E_J' < E_0$: the boost is opposite to the three-momentum direction of the jet

In order to determine the γ factor required for the boost consider the first case $E_J' > E_0$, where impose the condition that the boosted energy must be E_0 .

$$E_0 = \gamma E_J' - \beta \gamma |\vec{P}_J'| \quad \text{where:} \quad \beta^2 = 1 - \frac{1}{\gamma^2}. \quad (2.4)$$

This gives:

$$\left(\frac{m_0^2}{|\vec{P}_J'|^2} \right) \gamma^2 + \left(-\frac{2E_J'E_0}{|\vec{P}_J'|^2} \right) \gamma + \left(\frac{E_0^2}{|\vec{P}_J'|^2} + 1 \right) = 0, \quad (2.5)$$

where, we have used $m_0^2 = (E_J')^2 - |\vec{P}_J'|^2$. Solving this quadratic equation in Eq. (2.6) and picking the smaller solution (so that we do not boost past the rest frame of the jet), we get:

$$\gamma = \frac{1}{m_0^2} \left(E_J'E_0 - P_0 |\vec{P}_J'| \right), \quad (2.6)$$

where $P_0^2 = E_0^2 - m_0^2$ is the magnitude of three-momentum of the jet in the boosted frame.

In case $E_J' < E_0$, we get the exact same solution for γ but with the boost direction being opposite to the jet momentum. After boosting, we now have a jet with four-momentum \mathbf{P}_J^μ such that $\mathbf{E}_J = E_0$ and $|\mathbf{P}_J^\mu|^2 = m_0^2$.

2.1.3 Forming the images

A jet image is basically just a 2 dimensional histogram of the energy (or p_T) of it's constituents with respect to the plane transverse to the jet axis. We use the Gram-Schmidt procedure to determine the optimal set of basis vectors in the transverse plane while also removing the residual rotation symmetry still present in the constituents.

Denoting the set of all constituent momenta in the boosted frame $\{\mathbf{p}_1^\mu, \mathbf{p}_2^\mu, \dots, \mathbf{p}_N^\mu\}$ such that $\mathbf{p}_1^0 \geq \mathbf{p}_2^0 \geq \dots \geq \mathbf{p}_N^0$, we construct the new set of basis vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ obtained by the Gram-Schmidt method:

$$\hat{e}_1 \equiv \frac{\vec{\mathbf{P}}_J}{|\vec{\mathbf{P}}_J|}, \quad (2.7)$$

$$\hat{e}_2 \equiv \frac{\vec{\mathbf{p}}_1 - (\vec{\mathbf{p}}_1 \cdot \hat{e}_1) \hat{e}_1}{|\vec{\mathbf{p}}_1 - (\vec{\mathbf{p}}_1 \cdot \hat{e}_1) \hat{e}_1|}, \quad (2.8)$$

$$\hat{e}_3 \equiv \frac{\vec{\mathbf{p}}_2 - \hat{e}_1 (\vec{\mathbf{p}}_2 \cdot \hat{e}_1) - \hat{e}_2 (\vec{\mathbf{p}}_2 \cdot \hat{e}_2)}{|\vec{\mathbf{p}}_2 - \hat{e}_1 (\vec{\mathbf{p}}_2 \cdot \hat{e}_1) - \hat{e}_2 (\vec{\mathbf{p}}_2 \cdot \hat{e}_2)|}. \quad (2.9)$$

Noted here that we approximate the constituent four-momenta to be massless. The first step of the Gram-Schmidt ensures that the jet axis is along the origin, the second and third step ensure the major and minor axis of the jet is along the x and y axis. We are aware that the above method might not be IR safe for some situations. Therefore, we also use three-moments of 2 hardest subjects (after the original jet constituents are reclustered to find 3 exclusive k_T -subjects) to instead of \mathbf{p}_1 and \mathbf{p}_2 – we get similar results.

We convert the full set of constituents to a jet image using the new basis. Note that the jet image is basically a 2D histogram $\{I, J, F_{IJ}\}$, where I and J are integers, and represent the pixel coordinates. Also, for a image of $N_H \times N_H$ pixels, $1 \leq I, J \leq N_H$. On the other hand, F_{IJ} is a real number representing intensity of the (I, J) -th pixel. In this work we derive F_{IJ} from all of the constituents according to,

$$F_{IJ} = \sum_{i=1}^{N_J} \frac{\mathbf{p}_i^0}{E_0} \delta_{IX_i} \delta_{JY_i}, \quad \text{where} \quad (2.10)$$

$$X_i = \left\lfloor \frac{1}{2} (N_H + 1) + \frac{1}{2} (N_H - 1) \left(\frac{\hat{e}_2 \cdot \vec{\mathbf{p}}_i}{\mathbf{p}_i^0} \right) \right\rfloor$$

$$Y_i = \left\lfloor \frac{1}{2} (N_H + 1) + \frac{1}{2} (N_H - 1) \left(\frac{\hat{e}_3 \cdot \vec{\mathbf{p}}_i}{\mathbf{p}_i^0} \right) \right\rfloor.$$

In the above $\lfloor f \rfloor$ is the usual floor function representing the greatest integer less than or equal to f . Further, it is easy to verify that by construction

$$F_{IJ} \geq 0 \quad \forall I, J \quad \text{and} \quad \sum_{I=1}^{N_H} \sum_{J=1}^{N_H} F_{IJ} = 1. \quad (2.11)$$

and hence can be interpreted as the discrete probability distribution function (PDF). The 2D histogram (image) is flattened to a 1D vector and used as input to a fully connected autoencoder.

2.2 The Autoencoder

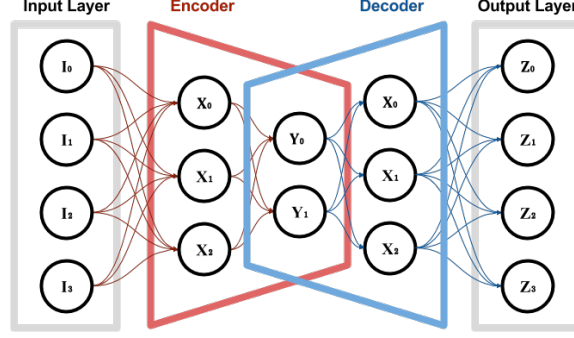


Figure 1. A schematic representation of an autoencoder (image was taken from Ref. [45]).

An autoencoder consists of two segments of neural network. The first segment is called the encoder, which leads to a bottleneck with the number of nodes being less than the input. The second segment is a decoder which takes the output from the encoder bottleneck and tries to reconstruct the original input at the final output. A schematic diagram representing an autoencoder is shown in Figure. 1. Note that the size of the bottleneck is an important characteristic of the autoencoder.

In the simplest case, both encoder and decoder are fully connected (dense) neural networks. In this study we consider only fully connected autoencoders for simplicity. It consists of multiple layers of neural networks with the output of one layer serving as input to the next layer:

$$O_i^L = \sigma \left(\sum_j W_{ij}^L I_j^L + B_i^L \right), \quad (2.12)$$

where, O_i^L and I_j^L are the output and input vectors respectively, W_{ij}^L is the weight matrix, and B_i^L is the bias vector for the neural network layer L . Note that, for internal layers I_j^L is the output from the previous layer, or $I_j^L = O_j^{L-1}$. Finally, σ is the activation function. We use ReLU [46] as the activation functions for all the intermediate layers and SoftMax for the final layer defined as

$$\text{ReLU}(x_i) = \begin{cases} 0 & \text{if } x_i < 0 \\ x_i & \text{otherwise} \end{cases}, \quad \text{and} \quad \text{SoftMax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^N e^{x_j}}. \quad (2.13)$$

Further note that the weight matrix W_{ij}^L is of dimension $M^L \times N^L$, where M^L is the dimension of the output vector O_i^L and N^L is the dimension of the input vector I_j^L . For an example, in the first layer, $N^1 = N_H^2$. Given a network with K layers, we additionally impose $M^K = N^1$, which ensures that the final output vector has the same dimensionality as the input and hence can be readily compared. Further, since we use SoftMax for the final layer, we ensure that the final output (namely, O_i^K), is also a discrete probability distribution function. This allows us to compare the final output to the initial image by a

simple L_2 norm function, which we use as the autoencoder loss function per input vector:

$$\epsilon \equiv \sqrt{\sum_i (O_i^K - I_i^1)^2} \quad (2.14)$$

Note that even though we focus on the simplest form of autoencoders, there has also been various improvements to the autoencoder brought in by using convolution layers [47, 48], or variational decoding [49]. For further information see references and articles cited in Refs. [31, 32, 47–51]

2.3 Summary of the whole algorithm

Before concluding this section, we summarize the full algorithm in brief:

- Start with jets clustered using any of the standard IR-safe jet algorithms. We represent the jet momentum by P_J^μ and its constituent momenta by p_i^μ (see Eq. (2.1)).
- Rescale p_i^μ (and as a result P_J^μ) such that the the new jet momentum $P_J'^\mu$ is characterized by a given mass m_0 . We also denote the rescaled component momenta by $p_i'^\mu$ (see Eq. (2.2)).
- Lorentz boost either parallel or anti-parallel to the jet momenta as prescribed in Eq. (2.3) such that the energy of the jet in the new frame is given as a predetermined E_0 . The boost factor required is given in Eq. (2.6). We denote the jet momentum and the constituent momenta in the new frame by \mathbf{P}_J^μ and \mathbf{p}_i^μ respectively.
- Find the optimal set of basis vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ as described in Eqs. (2.7 - 2.9). The vectors \hat{e}_2 and \hat{e}_3 span the plane transverse to the jet axis, which is given by \hat{e}_1 .
- Use the constituent momenta \mathbf{p}_i^μ , and the basis vectors \hat{e}_2 and \hat{e}_3 , to construct the final jet image (or rather a 2D histogram) of a given size $N_H \times N_H$ by using Eq. (2.10).
- After flattening the 2D histogram for the image, the resultant 1D vector becomes the input to a fully connected autoencoder.

3 Results

Before proceeding, note that our tasks are divided into two main categories:

1. The main task is to show that our procedure is robust. The robustness feature reflects the fact that the autoencoder loss function as defined in Eq. (2.14) does not vary as jets from different m/p_T bins are used to train the autoencoder.
2. Also we benchmark the performance of our method as an anomaly finder. We carry this out in two parts:
 - A) In the first part, we treat topjets and W -jets as anomalies. This also allows us to compare the performance of our method with tools based on physics information, as well as with previous proposals.

- B) In the second part, we benchmark the performance of our method in after introducing a NP particle decaying to di- W -jets giving rise to jets consisting of decay products of two W -bosons.

In the rest of this section we begin with the simulation details in Subsec. 3.1, while results for task 1 and task 2 are presented in Subsec. 3.2 and Subsec. 3.3 respectively.

3.1 Simulation Details

In this study we generate all events using PYTHIA8-8.2.26 [52, 53] and use Delphes-3.4.1 [54] to simulate detector environment. We compare results before and after adding various levels of complexity, such as turning on multi parton interaction (MPI) or detector effects.

We begin by clustering final state detectable particles from the output of PYTHIA8 into jets Anti- k_t [39] of $R = 1.0$ using fastjet-3 [44]. We only consider the hardest jet from an event in the central region $|\eta| < 2.0$. On the other hand, we use eflow elements such as `tracks`, `neutrals` and `photons` from the output of Delphes to find jets, when we simulate detector effects.

The jet mass and transverse momentum we have considered in our study vary over a wide range as discussed later. Using these jets as input, jet images are formed using the method detailed in Subsec. 2.1 with the parameters $m_0 = 1/2$ GeV, $E_0 = 1$ GeV and $N_H = 40$. The full structure of the network can be read from the source code linked at the end of the document. Basically the input consists of a 40×40 pixel flattened image, each layer decreases both the row and column by 7 until the narrowest bottleneck of 5×5 is reached. The network expands in the same way again by 7 for both column and row, each layer has the ReLU activation, only the final layer has a softmax activation. We try to give a short hand notation for the structure of our network in Figure. 2.

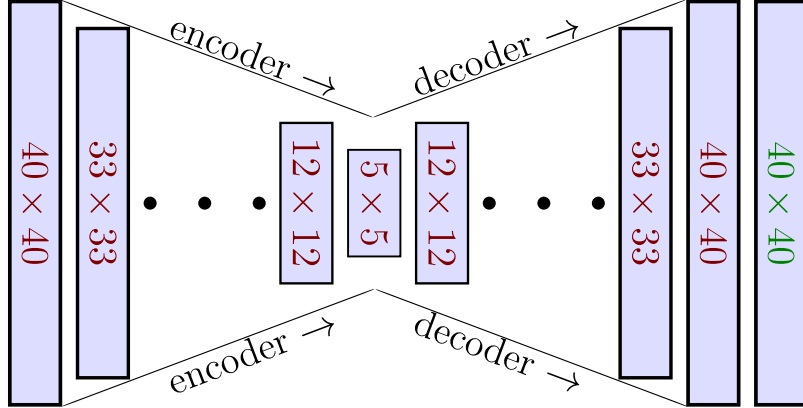


Figure 2. A schematic of the autoencoder structure used in our work. Layers with text in red have ReLU activation while the last layer with text in green has SoftMax activation (see Eq. (2.13)).

Note that the jet image used as input to the autoencoder is normalized to 1 and each bin (pixel) is non-negative, *i.e.*, it represents a discrete PDF. These properties might be lost when the image is propagated through the neural network. The softmax activation is

used in the last layer to restore the normalization of energy clusters and also to make them non-negative. This forces the network to learn the main features of the image instead of just reproducing the magnitude and improves the overall training speed.

We arrive at this structure by checking through a few configurations mainly determined by the size of the bottleneck (1 through 8×8) and settled on 5×5 , as increasing it further did not yield any significant improvement on reconstruction (loss function) of QCD in the first few training epochs. Also, for training this network, we use the `Adam` [55] gradient descent with a batch size of around 128 to 256 and learning rate (step size of the gradient descent) of around 0.001 to 0.0001.

We have written and trained the network using [the gluon API of Apache MXNet](#) [56] in python. This was chosen over `TensorFlow` [57] because of superior training performance on the CPU and easier imperative API, although we verified that the error convergence and prediction results were similar between this and `TensorFlow` using the `TFLearn` [58] wrapper API in the first few epochs of training.

3.2 Robustness

The main goal of this section (as stated before) is to establish that the autoencoder loss function obtained using our method is rather insensitive to the jet masses of the control sample. We demonstrate this is by training the autoencoder network on QCD jets from a particular region of the phase space and by testing the trained network on QCD jets from a different region of the phase space, we divide our result into three parts:

1. We train the autoencoder network on QCD-jets with $800 \text{ GeV} < p_{TJ} < 900 \text{ GeV}$ produced at $\sqrt{s} = 13 \text{ TeV}$ collision energy (with no cuts on the jetmass). The data is then divided into bins depending on jetmasses. The loss function (from the autoencoder trained in the previous step) is evaluated on events from each bin and compared with each other as well as with the distribution of the whole sample (Subsec. 3.2.1).
2. We use QCD-jets produced at $\sqrt{s} = 100 \text{ TeV}$ with $p_{TJ} > 10 \text{ TeV}$, and then divide the full sample into bins depending on jet masses. We train an autoencoder network using QCD-jets from one of these mass bins and test with the others. (Subsec. 3.2.2).
3. As extreme comparisons, we train on jets from $\sqrt{s} = 13 \text{ TeV}$ collider, and test on data from $\sqrt{s} = 100 \text{ TeV}$ collider, and vice versa (Subsec. 3.2.3).

3.2.1 Robustness at $\sqrt{s} = 13$ TeV

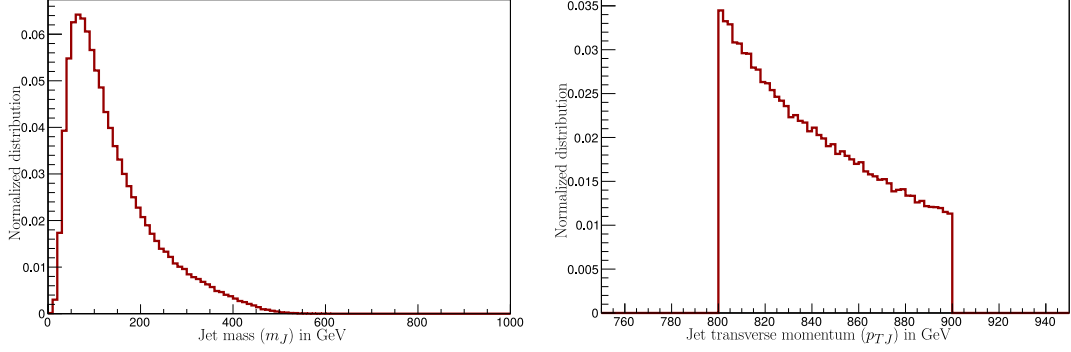


Figure 3. The mass (left) and transverse momentum (right) distribution of jets we use for our study.

In this part, we study QCD jets at $\sqrt{s} = 13$ TeV. To be particular, we consider QCD di-jet events and study the jet hardest in p_T from each event. We begin by studying jets with $800 \text{ GeV} < p_{T,J} < 900 \text{ GeV}$ generated without MPI. The jet mass and transverse momentum distribution of the sample is shown in Figure. 3. The transverse momentum distribution is, as expected, lying in the window of 800 GeV to 900 GeV due to the imposed phase space cuts and the jet mass gives the expected falling distribution.

The jet images (as explained in Subsec. 2.1.3) are basically a map from a set of two dimensional integers (representing the pixels or the bins (I, J)) to the set of real numbers between 0 and 1 (representing the intensities, F_{IJ}) given by Eq. (2.10).

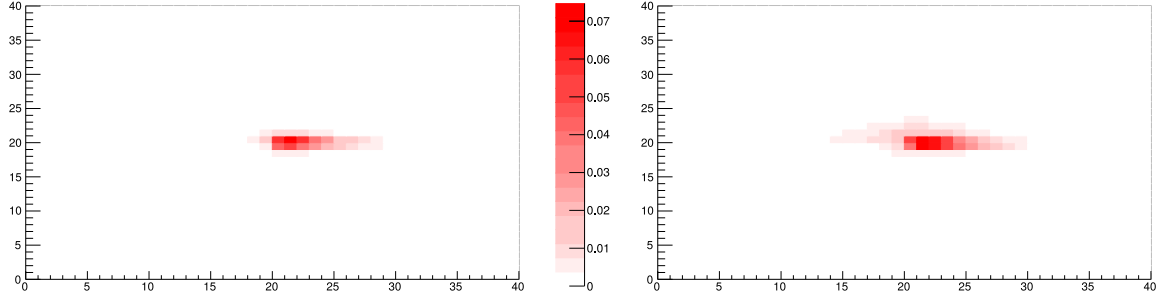


Figure 4. The Image of a QCD-jet (averaged over ≈ 500000 jets) obtained after our pre-processing method without(**left**) and with(**right**) detector effects (using **Delphes**). The image is a 3D plot with the color density (the z axis) representing intensity. To be specific, the axis represent: $(x, y, z) \rightarrow (I, J, \langle F_{IJ} \rangle)$.

In Figure. 4, we present the averaged (over 500000 jets) jet image of QCD-jets using our method. The color intensity of a bin (pixel) with coordinate (I, J) , represents the average $\langle F_{IJ} \rangle$ (the intensity of each pixel or bin is averaged across many jets). The aver-

aging procedure captures the broad features of the event type while smoothing over sharp fluctuations from individual events.

The image faithfully reproduces the simple dipole emission structure of QCD jets, where the hardest emission from the parton initiating the jet decides the major (elongated) axis of the jet (which is brought along the x axis of the image Figure. 4 due to the Gram-Schmidt procedure in Eq. (2.8)) while most subsequent emissions tend to lie along this axis (owing to the color connected [59] structure of QCD interactions). We also observe that the main impact of detector simulation seems to smear the energy distribution of the jet constituents.

At this point, we divide our task into two parts:

- (A) Train the autoencoder network on the full set without any cut on the jet mass and then test the variation of the autoencoder loss function on various mass bins.
- (B) Train the autoencoder network on jets with $m_J < 200$ GeV and test the trained network on jets with $300 \text{ GeV} < m_J < 500$ GeV.

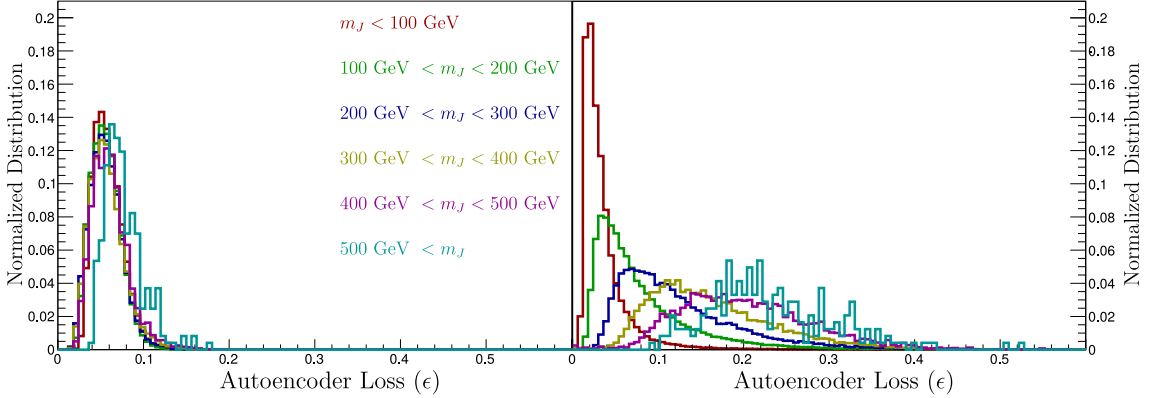


Figure 5. The effect of mass cuts on the autoencoder response (for QCD jets) using our method (**LEFT**) and using dense autoencoder from Ref. [31] (**RIGHT**).

For task (A), we plot the autoencoder loss functions using our method in Figure. 5 (**LEFT**), which clearly shows that the autoencoder loss function is mostly independent of jet masses. For comparing and benchmarking, we also plot responses where we use dense autoencoder from Ref. [31] to calculate loss (**RIGHT**), here the dependence of jet mass on autoencoder loss is apparent. The change in the peak position of the autoencoder loss distribution for different jet masses using our method is within 10% – 15%. On the other hand, when we use the method from Ref. [31], the peak position shifts by factors of 4 to 5. Similar trends are also seen in the widths of the distribution which hardly changes for different mass bins when using our method while changing significantly (by several factors) for the method proposed in Ref. [31]. Although, we have compared only the dense autoencoder from Ref. [31] here, the robustness of convolution autoencoders are found to be only slightly better (order 30% for $m_J > 300$ GeV)[31]. However, as one can

see from Figure. 6 (right), 20% to 30% improvement is not sufficient. Our simple method is significantly more robust

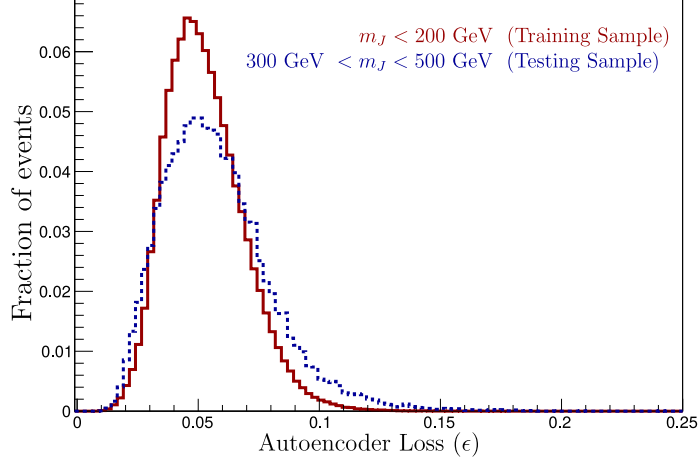


Figure 6. The performance variation of the autoencoder when it is trained on jets with $m_J < 200$ GeV and tested on jets with $300 \text{ GeV} < m_J < 500$ GeV.

We present the results for task (B) in Figure. 6. Again as expected (claimed), we confirm the independence of autoencoder loss with respect to the jet mass. Notice a tiny loss in performance due to the difference in evaluation sample as compared to the training sample is manifest as a small broadening of the loss distribution from the testing sample. The width of the loss distribution in testing case is $\sim 5\%$ larger than the width for the training case, even though the peak position essentially does not change.

3.2.2 Robustness at $\sqrt{s} = 100$ TeV

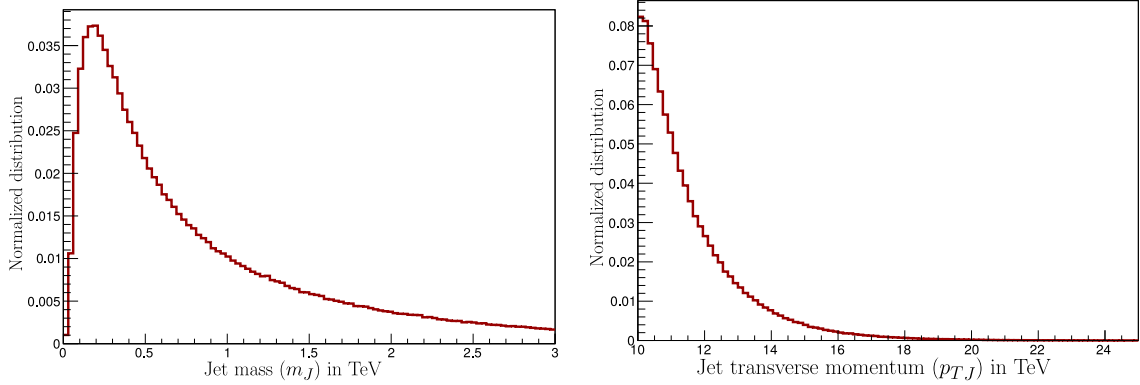


Figure 7. The mass (left) and transverse momentum (right) distribution of jets we have used for our study at $\sqrt{s} = 100$ TeV.

In this subsection we study the robustness with QCD-jets of high transverse momentum and jet mass. For this purpose, we consider jets with $p_{T,J} > 10$ TeV from QCD-dijet events at a $\sqrt{s} = 100$ TeV collider, generated without MPI. For these event we do not simulate any detector effects. The m_J and $p_{T,J}$ distributions of these jets are presented in Figure. 7,

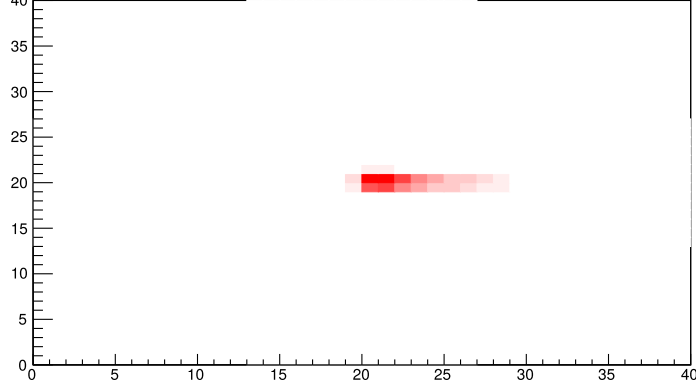


Figure 8. The image of QCD jets at $\sqrt{s} = 100$ TeV obtained using our method without MPI or detector effects.

Also, the image of these jets obtained using our method are shown in Figure. 8. Clearly, the image is almost identical to the image in Figure. 4 obtained for QCD jets from $\sqrt{s} = 13$ TeV collisions further fortifying our claims. Note that the most energetic part of the image (as measured using the scales on the x and y axis) is identical between Figure. 4 and Figure. 8. This demonstrates that our method always resolves the jet at the same length scales irrespective of the jet mass or momentum (boost).

Before proceeding further, we divide the full set of jets into subcategories according to:

- | | |
|--|--|
| 1. No cut on m_J (all jets are considered) | 4. $400 \text{ GeV} < m_J < 600 \text{ GeV}$ |
| 2. $m_J < 200 \text{ GeV}$ | 5. $600 \text{ GeV} < m_J < 800 \text{ GeV}$ |
| 3. $200 \text{ GeV} < m_J < 400 \text{ GeV}$ | 6. $800 \text{ GeV} < m_J$ |

In all of these case, we form images using our method and analyze further using the autoencoder in various permutations. In order to check for robustness, we follow a strategy analogous to Subsec. 3.2.1:

- A) Train an autoencoder on jets from each of the mass bins mentioned above.
- B) Evaluate the performance of these networks on jets from each of the mass bins listed above.

We show the results from this study in Figure. 9, which again strongly support the claim that our method mostly eliminates the dependence of jet mass on the autoencoder loss function. As can be observed from the distributions Figure. 9, most of the peak position lie within $\approx 20\%$ of each other and the distributions largely overlap.

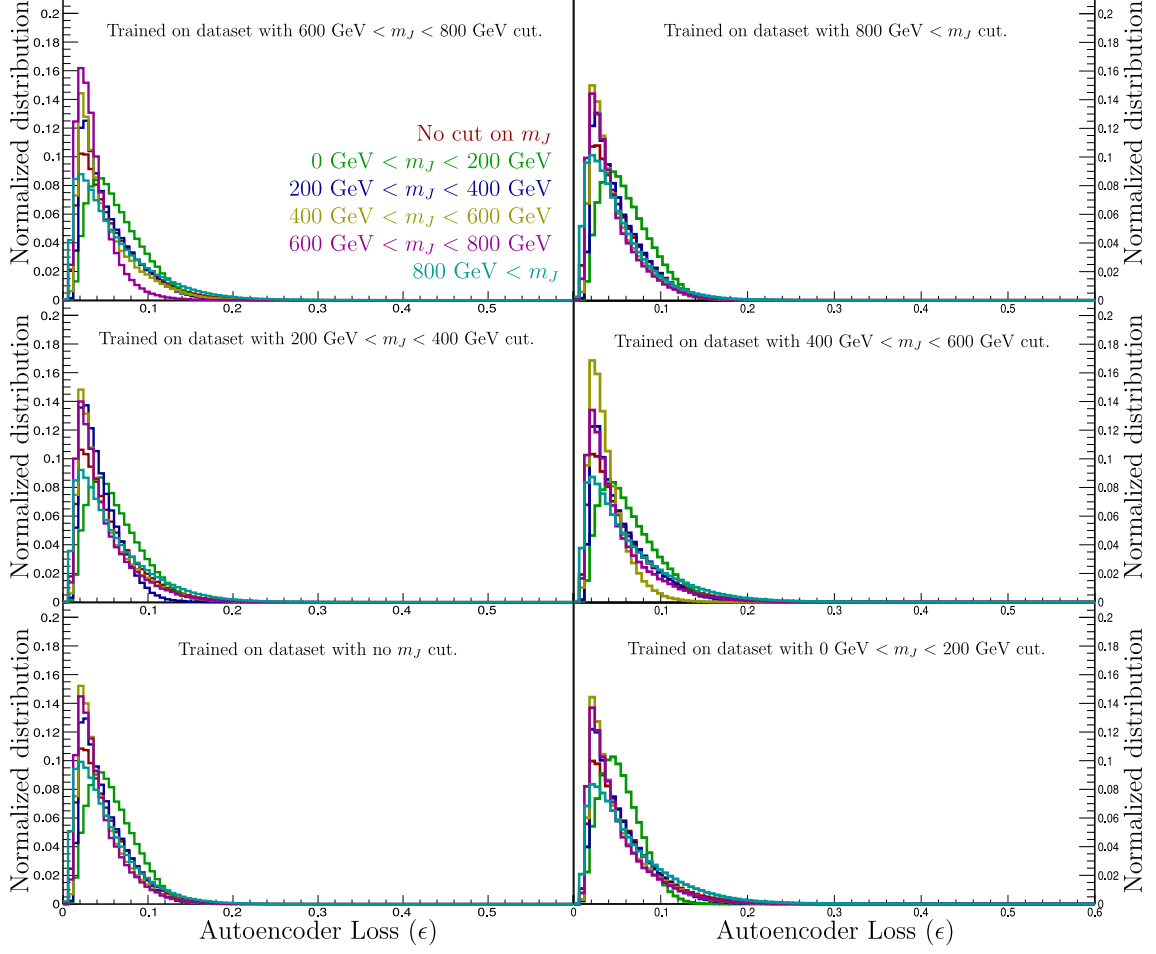


Figure 9. The (normalized) distribution of autoencoder loss function for QCD jets ($p_{TJ} > 10$ TeV) in various mass bins, the autoencoder with jets in a particular mass bin (indicated in each of the figures) and tested on all the other mass bins. The colors of the histogram represent the mass bins as labeled.

3.2.3 Cross validation

In this subsection, we present results from two studies:

1. Train the autoencoder on $\sqrt{s} = 13$ TeV QCD jet data and test it on QCD jets at $\sqrt{s} = 100$ TeV.
2. Train the autoencoder on QCD jets from $\sqrt{s} = 100$ TeV collisions and test it on the QCD jets produced at $\sqrt{s} = 13$ TeV.

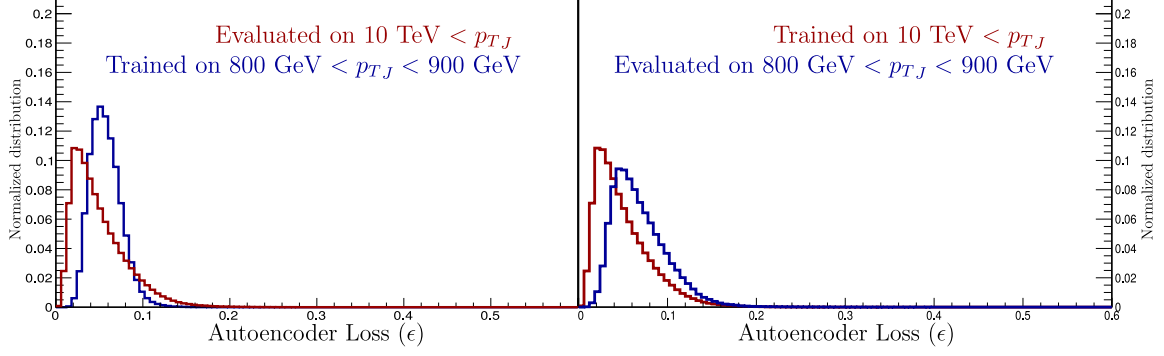


Figure 10. Dependence of autoencoder loss function on jet mass when the network is trained on low energy jets and evaluated on high energy jets (**LEFT**) and when network is trained on high energy jets and evaluated on low energy jets (**RIGHT**).

Even with these extreme comparisons of jets with drastic differences in kinematics, our method yields largely insensitive results as can be seen in Figure. 10. Note that the shift in peak positions in both cases are less than 50% of the width. Most of the actual distributions overlap and, consequently, widths and averages in all cases are within 20% of each other.

3.3 Performance

The task in this subsection is to examine and benchmark the performance of our method as an anomaly finder for various types of jets. As stated before, the autoencoder loss function can be used as a measure to determine how different a jet “looks like” as compared to the jets which the autoencoder was trained on. We consider topjets and W -jets as anomalies for benchmarking purposes and finally examine the performance of our method on a jet due to a new physics process (di- W jet). The loss function for these three different event types using the autoencoder which was trained on QCD jets with $800 \text{ GeV} < p_{T,J} < 900 \text{ GeV}$ and without MPI or detector effects (from Subsec. 3.2.1) is presented in Figure. 11. It is clearly seen that the autoencoder loss increases as more complex structure is added to the jets. QCD-jets (on which the autoencoder was trained on and also have the simplest 1-prong dipole emission substructure) have, on average, the smallest loss function. W -jets, on the other hand, have a 2-prong structure which is slightly more “complex” than the 1-prong QCD-jet structure, and therefore have larger loss functions on average. Top jets come next in the hierarchy as they mostly have a 3-prong structure (with an onshell W) making it more “complex” than a W -jet. Finally, we show jets originating due to boosted $\phi \rightarrow W^+W^-$. This gives rise to jets with 4-prong structure with 2 onshell W resonances. Not surprisingly, we find these jets with the highest loss functions.

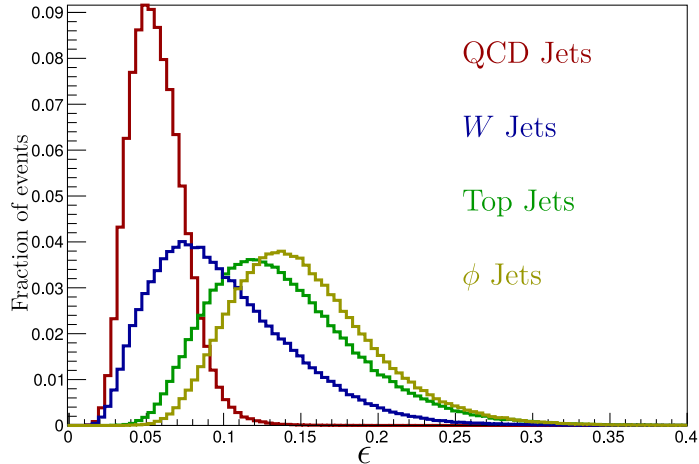


Figure 11. The autoencoder loss function (ϵ , Eq. (2.14)) obtained from the autoencoder using our method of pre-processing for the case with no detector effects or multi parton interaction.

In order to quantify the comparison between the performance of various methods, we compare the receiver operating characteristic (ROC) curves. In our study, these curves are presented with the convention that the x axis represent the signal selection efficiency and the y axis represent the reciprocal of the background fake rate (in log scale). Therefore, the algorithm producing curves closer towards the top right region of the graph would mean a better performing algorithm. It is also possible to compare two algorithms based on the area under the ROC curves of the two algorithms, larger area would generally mean better performance.

Next, we examine each of the cases in greater detail, benchmarking our method against some standard techniques which are well established and also against some recent techniques found in literature.

3.3.1 Benchmark with top jets

The problem of discriminating jets originating from boosted tops against jets from QCD is extremely well studied theoretically [60–63], numerically [34–37, 64, 65] and experimentally [66, 67]. We use this opportunity to benchmark our method by comparing it against the established algorithm `HEPTopTagger2` [34–37]. Also, like in the previous sections, we compare our method with the one proposed in Ref. [31] using dense autoencoders and also show the performance curves for both dense and convolutional autoencoders (as claimed in Ref. [31]).

Top jets are obtained from $t\bar{t}$ events as detailed in Subsec. 3.1. We present the results for three main cases:

- (A) Top jets with $800 \text{ GeV} < p_{TJ} < 900 \text{ GeV}$ simulated without MPI or detector effects.
- (B) Top jets with $800 \text{ GeV} < p_{TJ} < 900 \text{ GeV}$ simulated without MPI but with detector effects.

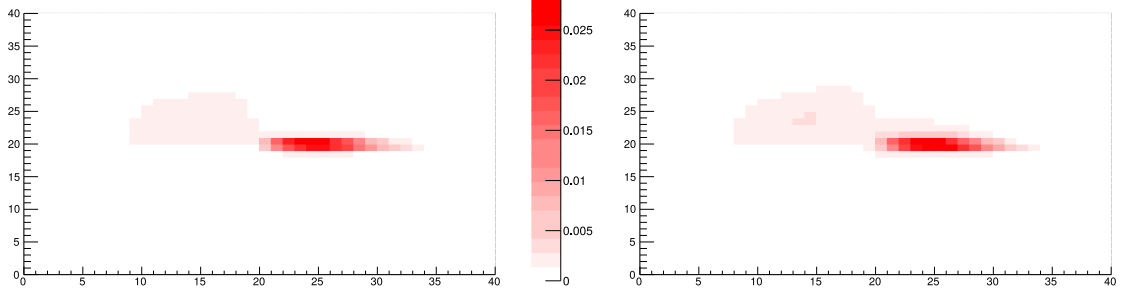


Figure 12. The images of top-jets (averaged over ≈ 500000 events) obtained after our preprocessing method without (left) and with (right) detector effects (using `Delphes`).

(C) Top jets with $p_{T,J} > 400$ GeV simulated with MPI and detector effects.

While generating these events in cases. (A) and (B), we impose an additional merge requirement between the jet and the partonic top along with its decay products (b, q, \bar{q}') by demanding that they lie within the jet at `PYTHIA8` level.

$$\begin{aligned} \Delta R(p_J^\mu, p_{\text{top}}^\mu) &< 0.6 & \Delta R(p_J^\mu, p_q^\mu) &< 0.6 \\ \Delta R(p_J^\mu, p_b^\mu) &< 0.6 & \Delta R(p_J^\mu, p_{\bar{q}'}^\mu) &< 0.6, \end{aligned} \quad (3.1)$$

Which ensures a consistent comparison with Ref. [31]. The merge requirement is not imposed for case. (C).

The jet images obtained using our method with and without detector effects is shown in Figure. 12, clearly the structure is much more “complex” than the case of QCD jets. Even though the three prong structure is not pronounced (due to the three body decay) in the image, it is still visible in a subtle way. The extended horizontal lump and lobe approximately constitute two prongs while the third is from the vertical extension of the lobe towards the left. Again, the effect of detector simulations seem just to smear the momentum distribution of the jet constituents.

We present the results (ROC curves) of our comparison for case. (A) and case. (B) in Figure. 13. Note that the results from `HEPTopTagger2` contribute a single point to this graph as shown. It is observed that our method performs reasonably well for discriminating top jets from QCD. By itself the performance of our autoencoder loss ϵ is comparable to the autoencoder loss of the dense network without our preprocessing (we refer to it as ϵ_{FNS}). When combined with jetmass (m_J), our autoencoder loss is as good a top-tagger as the simplest (default) form in `HEPTopTagger2` [34–37]. Note that the default version functions without Qjets [68, 69] or any other variables (like N-subjettiness [65] or Energy Correlation [70]), but uses the jet mass information in the mass drop step and also many other features from the physics of top quark decay kinetics. As a further demonstration we additionally show ROC with ϵ combined with m_J and N-subjettiness $\{\tau_1, \dots, \tau_5\}$. Not surprisingly, we find significantly better performance – it produces an improvement by a factor of 2.5 over the `HEPTopTagger2` in signal selection efficiency for a given background

rejection. It shows that our technique is extremely general and its performance will be competitive with specialized taggers designed for each event topology.

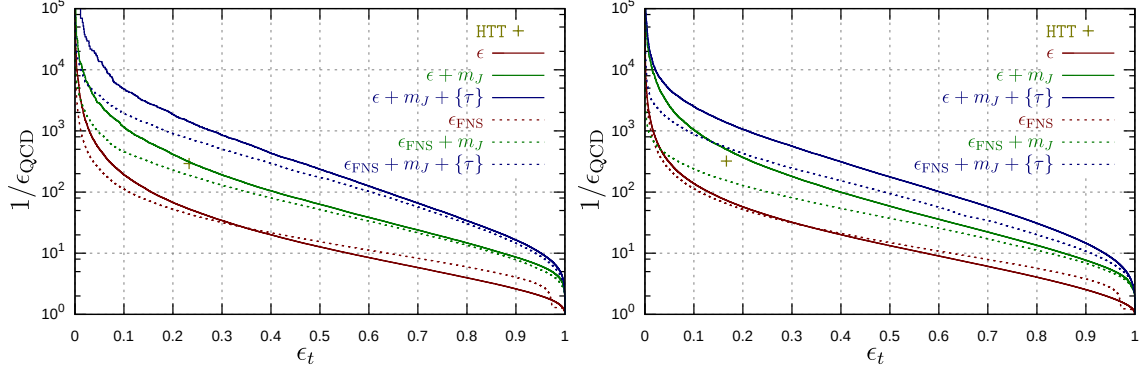


Figure 13. ROC for discriminating QCD vs. top using our auto encoder loss function ϵ , as well as combining ϵ with jet mass and N-subjettiness variables $\{\tau_1, \dots, \tau_5\}$. For comparison, we also produce the same but using ϵ_{FNS} , loss functions trained using Ref. [31] These results are presented without **(left)** and with **(right)** detector effects.

Since we remove mass information (using the preprocessing procedure described in Subsec. 2.1) we expect the loss ϵ to be significantly less correlated with jetmass or N-subjettiness variables. In fact, we expect ϵ to be less correlated than ϵ_{FNS} . To demonstrate this fact we plot the correlation matrix for QCD-jets in Ref. 14. As we expect, ϵ is completely uncorrelated with m_J (less than 10%). It is more correlated to τ_N than to m_J . However, the correlation coefficients never reaches even 30%. On the other hand, correlation coefficients for ϵ_{FNS} remains in the range $\sim 45\%$ to 65% for all these variables.

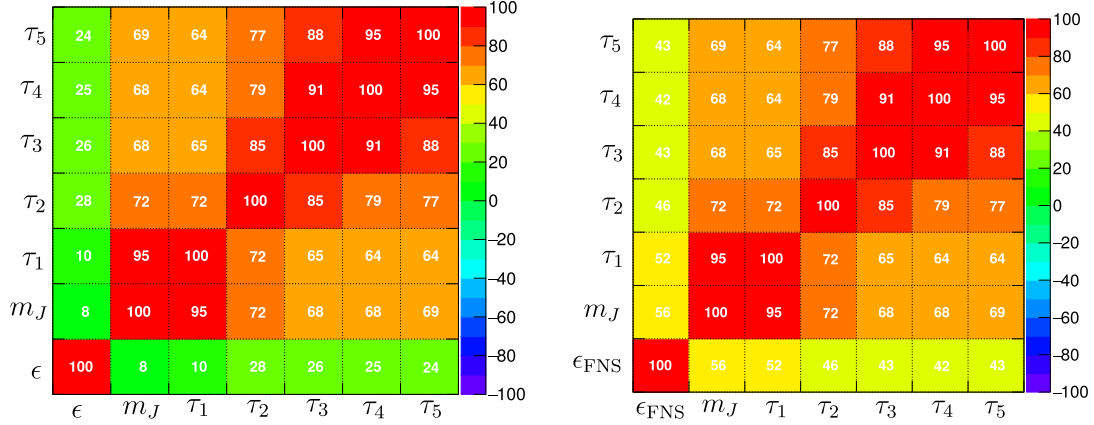


Figure 14. Linear correlation coefficients (in %) for QCD/top discrimination using ϵ and ϵ_{FNS} , for QCD-jets.

An additional point to note is that our technique yields comparable performances compared to the one proposed in Ref. [31]. This suggests that while preserving the mass

invariance as demonstrated in Subsec. 3.2 we do not need to sacrifice its performance. The detector simulation seems to affect the performance of all three methods by 10% – 20%. Using our method, for a signal selection efficiency of $\approx 80\%$, background fake rates as low as 5% to 10% can be achieved.

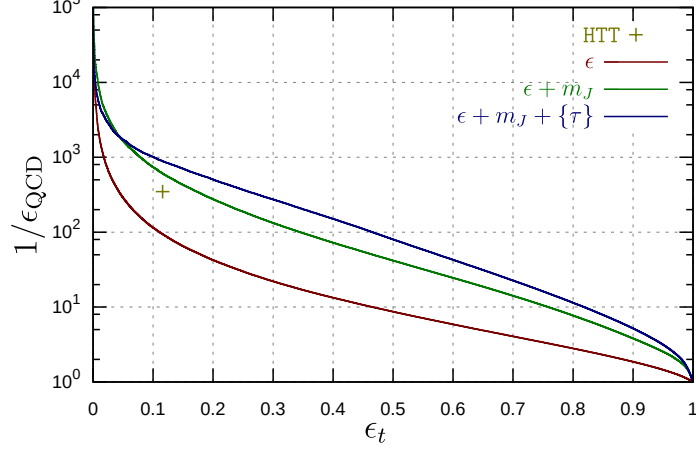


Figure 15. ROC for QCD vs top discrimination using our method with MPI and detector effects.

Finally, in order to demonstrate the fact that our method preserves its performance in a more realistic situation, we benchmark its performance for QCD-jets vs topjet discrimination on events simulated with MPI and detector effects. We also discard all parton level merge requirements, imposed earlier, as explained in case. (C). Additionally, we use a much larger phase space region with $p_{TJ} > 400$ GeV. We show the results from this study in Figure. 15. In the same plot we also show the performance of HEPTopTagger2 for comparing. We see that much of the performance seen in the previous case is still retained even after MPI and detector effects are turned on (see, Figure. 13 for comparison).

3.3.2 Benchmark with W -jets

In this section, we consider the problem of discriminating W jets against QCD jets. This is again a well studied problem and variables like N -subjettiness [71, 72] have been observed to work reasonably well. The CMS collaboration for LHC, for example, used τ_2/τ_1 W -tagging [73]. We use W -jets as anomalies and compare the performance of our methods with that of τ_2/τ_1 .

In this work we consider a sample of boosted W jets with $800 \text{ GeV} < p_{TJ} < 900 \text{ GeV}$, generated without MPI or detector effects. We also impose a merge requirement (as was the case for top) that the parent W and its decay products lie within the jet $[\Delta R(p_J^\mu, p_a^\mu) < 0.6]$, where a stands for decaying W , as well as quarks from the decay.

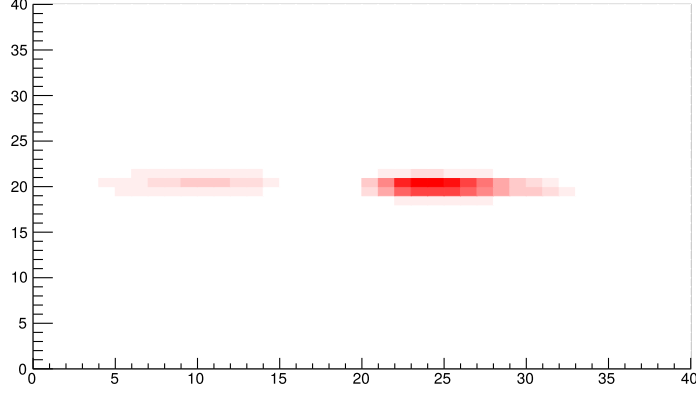


Figure 16. The average jet image of W -jet obtained after our pre-processing method.

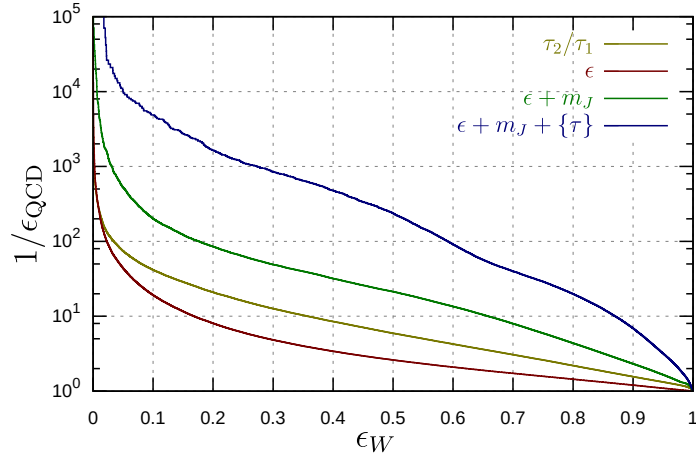


Figure 17. ROCs for discriminating QCD-jets vs W -jets using our anomaly finder and a standard N -subjettiness variable.

The jet images obtained using our method is presented in Figure. 16. Notice that the two prong structure of the jet is clearly visible. The asymmetry in the intensities of the two prongs is due to the Gram-Schmidt procedure Eq. (2.8), which always brings the harder prong along the right side of the horizontal axis by construction.

As before, we use an ROCs to quantify the performance of our method and compare it against an existing standard. The ROC in Figure. 17 shows the achievable background rejection efficiency as a function of signal efficiency, when we use autoencoder loss function alone as well as when we combine it with m_J and N -Subjettiness variables. Figure. 17 also shows the ROC when we use τ_2/τ_1 alone as a discriminating variable. We find that our method performs reasonably well as an anomaly finder only when it is combined with m_J . At the level of 1% QCD-jet reduction it manages to yield a signal selection efficiency of $\sim 20\%$. On the other hand, using ϵ along with m_J and N -Subjettiness as well, yields signal selection efficiency as high as $\approx 60\%$ for the same QCD-jet rejection.

3.3.3 Benchmark with di- W jets from new physics

In this section, we study the efficiency of finding jets consisting of decay products a NP particle (namely ϕ , a di- W resonance) as anomalies. To be specific, we use a scalar with mass 180 GeV. We use the effective two Higgs doublet model, where ϕ is the heavy scalar Higgs, for generating events. The actual process we consider is given below

$$p + p \rightarrow Z(\nu_e \bar{\nu}_e) + \phi(W^+W^-) . \quad (3.2)$$

The events are generated at $\sqrt{s} = 13$ TeV without MPI or detector effects. We consider jets with $800 \text{ GeV} < p_{TJ} < 900 \text{ GeV}$ that satisfy the usual merge requirement that the heavy parent ϕ , W^+ , W^- and the four final partons lie inside the jet, namely $[\Delta R(p_J^\mu, p_a^\mu) < 0.6]$, where a stands for decaying ϕ , as well as all the partons from W -decays.

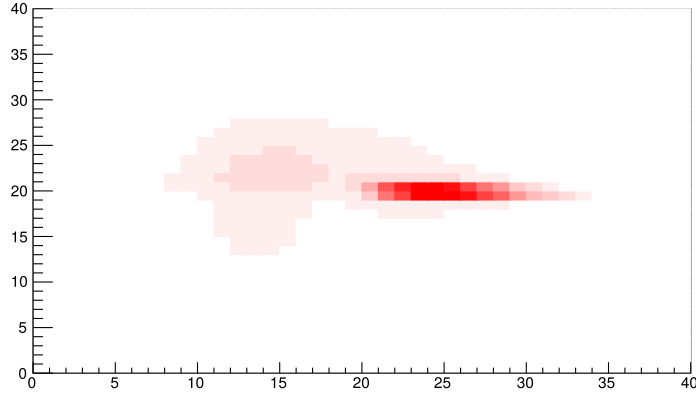


Figure 18. The average jet image for a di- W resonance with mass 180 GeV after our pre-processing method.

The jet image formed using our method is presented in Figure. 18. Even though the four prong structure is not pronounced, it is still visible in a subtle way. As in the case of Figure. 12, the horizontal lump and the extended lobe towards the left constitute two prongs while the third and fourth prongs overlap with the first two leading to the halo around the lump towards the right and a bright spot on the left lobe.

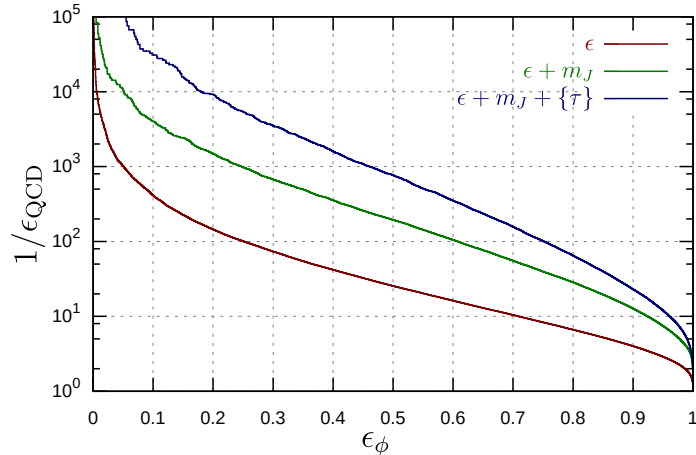


Figure 19. ROC for discriminating QCD vs di- W jets using our anomaly finder.

Finally, we show the ROCs for discriminating these ϕ -jets as anomalies in Figure. 18. The performance of our method is highly promising. For a signal selection efficiency of $\approx 20\%$, it suffers a fake rate of only $\approx 0.5\%$ due to QCD-jets. When combined with mass and N-subjettiness information the acceptance rate for the di- W jets can be increased to order 70% keeping the same fake rate.

4 Conclusion

In this paper, we propose a new anomaly finder or rather an anti-QCD tagger based on autoencoder. While the proposal uses the simplest version of an autoencoder using the fully connected network structure, its novelty lies in the preprocessing stage, where we exploit the full Lorentz symmetry apart from symmetries of shift, rotation and rescaling. Our method gives comparable performances as compared to some existing techniques in the literature, while, at the same time, removes the dependence of autoencoder loss function on the jet mass. Consequently, it gives a straight forward way to find pure control samples to train. Our anomaly finder can be trained using jets in low m/p_T bins, which is rich in QCD-jets, and can be used to analyze jets from high m/p_T bins to find anomalies (or, signature of NP). The lack of jet mass dependency of the autoencoder loss brings in additional advantages – it can be readily combined with existing discriminating variables to improve performances of even supervised learning to a much greater effect.

Apart from this robustness feature, we also find that detector effects or multi parton interactions have little effects in the performance of the anomaly finder. We have not explicitly tested our method against pileup in this study. However, we see no reason why this method can not be made pileup robust by simply using the standard pileup mitigation techniques [74–76] before forming the jet image, or even after the jet image is formed as shown in Ref. [77].

Problems in machine learning require large amount of computation which is generally achieved using GPUs for improved speed. This allows one to optimize on the structure of

the neural network. However, even with our naive optimization methods, we manage to get remarkable results as compared to existing techniques in literature. Our method provides an extremely simple yet a powerful technique with great potential for generalization as it can be readily incorporated into any existing neural network based algorithms such as image based top tagging [20–25], Lorentz layers [14], or recurrent networks [16] with minimal effort. Because of its simplicity, the preprocessing stage can even be extended to existing boosted jet substructure classification algorithms that uses only physics information. With more improvements and optimizations, it might even be possible to use the autoencoder loss obtained with our method (combined with the jet mass) as a spectroscopy like observable to find anomalous events.

Source code for most of the content discussed here can be found at:

[https://github.com/aravindhv10/CPP_Wrappers/tree/master/AntiQCD4]

Acknowledgments

We are grateful to Tanmoy Bhattacharya and Michael Graesser who originally suggested using Auto-encoders for anomaly hunting at colliders. This work also benefitted from stimulating conversations with Harrison Prosper and Deepak Kar. TSR was supported in part by the Early Career Research Award by Science and Engineering Research Board, Dept. of Science and Technology, Govt. of India (grant no. ECR/2015/000196). We also acknowledge the workshop “Beyond the Standard Model: where do we go from here?” hosted at the Galileo Galilei Institute for Theoretical Physics (GGI), as well as the workshop “International Meeting on High Energy Physics” hosted at Institute of Physics (IOP) where parts of this work was completed.

References

- [1] **CMS Collaboration** Collaboration, S. Chatrchyan *et al.*, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,” *Phys.Lett.* **B716** (2012) 3061, [arXiv:1207.7235 \[hep-ex\]](#). <https://arxiv.org/abs/1207.7235>.
- [2] **ATLAS Collaboration** Collaboration, G. Aad *et al.*, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,” *Phys.Lett.* **B716** (2012) 129, [arXiv:1207.7214 \[hep-ex\]](#). <https://arxiv.org/abs/1207.7214>.
- [3] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, “Jet substructure as a new Higgs search channel at the LHC,” *Phys. Rev. Lett.* **100** (2008) 242001, [arXiv:0802.2470 \[hep-ph\]](#).
- [4] G. P. Salam, “Towards Jetography,” *Eur. Phys. J.* **C67** (2010) 637–686, [arXiv:0906.1833 \[hep-ph\]](#).
- [5] D. Krohn, J. Thaler, and L.-T. Wang, “Jet Trimming,” *JHEP* **02** (2010) 084, [arXiv:0912.1342 \[hep-ph\]](#).

- [6] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, “Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches,” *Phys. Rev.* **D81** (2010) 094023, [arXiv:0912.0033 \[hep-ph\]](#).
- [7] A. Chakraborty, A. M. Iyer, and T. S. Roy, “A Framework for Finding Anomalous Objects at the LHC,” *Nucl. Phys.* **B932** (2018) 439–470, [arXiv:1707.07084 \[hep-ph\]](#).
- [8] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, “A generic anti-QCD jet tagger,” *JHEP* **11** (2017) 163, [arXiv:1709.01087 \[hep-ph\]](#).
- [9] T. Cohen, M. Freytsis, and B. Ostdiek, “(Machine) Learning to Do More with Less,” *JHEP* **02** (2018) 034, [arXiv:1706.09451 \[hep-ph\]](#).
- [10] H. Qu and L. Gouskos, “ParticleNet: Jet Tagging via Particle Clouds,” [arXiv:1902.08570 \[hep-ph\]](#).
- [11] K. Datta and A. Larkoski, “How Much Information is in a Jet?,” *JHEP* **06** (2017) 073, [arXiv:1704.08249 \[hep-ph\]](#).
- [12] L. Moore, K. Nordström, S. Varma, and M. Fairbairn, “Reports of My Demise Are Greatly Exaggerated: N -subjettiness Taggers Take On Jet Images,” [arXiv:1807.04769 \[hep-ph\]](#).
- [13] P. T. Komiske, E. M. Metodiev, and J. Thaler, “Energy Flow Networks: Deep Sets for Particle Jets,” *JHEP* **01** (2019) 121, [arXiv:1810.05165 \[hep-ph\]](#).
- [14] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, “Deep-learned Top Tagging with a Lorentz Layer,” *SciPost Phys.* **5** no. 3, (2018) 028, [arXiv:1707.08966 \[hep-ph\]](#).
- [15] P. T. Komiske, E. M. Metodiev, and J. Thaler, “Energy flow polynomials: A complete linear basis for jet substructure,” *JHEP* **04** (2018) 013, [arXiv:1712.07124 \[hep-ph\]](#).
- [16] G. Louppe, K. Cho, C. Becot, and K. Cranmer, “QCD-Aware Recursive Neural Networks for Jet Physics,” *JHEP* **01** (2019) 057, [arXiv:1702.00748 \[hep-ph\]](#).
- [17] M. Erdmann, E. Geiser, Y. Rath, and M. Rieger, “Lorentz Boost Networks: Autonomous Physics-Inspired Feature Engineering,” [arXiv:1812.09722 \[hep-ex\]](#).
- [18] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, “TMVA: Toolkit for Multivariate Data Analysis,” *PoS ACAT* (2007) 040, [arXiv:physics/0703039](#).
- [19] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, “Jet Constituents for Deep Neural Network Based Top Quark Tagging,” [arXiv:1704.02124 \[hep-ex\]](#).
- [20] L. G. Almeida, M. Backovi, M. Cliche, S. J. Lee, and M. Perelstein, “Playing Tag with ANN: Boosted Top Identification with Pattern Recognition,” *JHEP* **07** (2015) 086, [arXiv:1501.05968 \[hep-ph\]](#).
- [21] S. Macaluso and D. Shih, “Pulling Out All the Tops with Computer Vision and Deep Learning,” *JHEP* **10** (2018) 121, [arXiv:1803.00107 \[hep-ph\]](#).
- [22] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, “Deep-learning Top Taggers or The End of QCD?,” *JHEP* **05** (2017) 006, [arXiv:1701.08784 \[hep-ph\]](#).
- [23] 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 \[hep-ph\]](#).
- [24] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, “Jet Substructure Classification in High-Energy Physics with Deep Neural Networks,” *Phys. Rev.* **D93** no. 9, (2016) 094034, [arXiv:1603.09349 \[hep-ex\]](#).

- [25] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, “Jet-images deep learning edition,” *JHEP* **07** (2016) 069, [arXiv:1511.05190 \[hep-ph\]](#).
- [26] J. H. Collins, K. Howe, and B. Nachman, “Extending the search for new resonances with machine learning,” *Phys. Rev. D* **99** no. 1, (2019) 014038, [arXiv:1902.02634 \[hep-ph\]](#).
- [27] A. Butter *et al.*, “The Machine Learning Landscape of Top Taggers,” [arXiv:1902.09914 \[hep-ph\]](#).
- [28] J. H. Collins, K. Howe, and B. Nachman, “Anomaly Detection for Resonant New Physics with Machine Learning,” *Phys. Rev. Lett.* **121** no. 24, (2018) 241803, [arXiv:1805.02664 \[hep-ph\]](#).
- [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,” [arXiv:1811.10276 \[hep-ex\]](#).
- [30] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, “JUNIPR: a Framework for Unsupervised Machine Learning in Particle Physics,” *Eur. Phys. J. C* **79** no. 2, (2019) 102, [arXiv:1804.09720 \[hep-ph\]](#).
- [31] M. Farina, Y. Nakai, and D. Shih, “Searching for New Physics with Deep Autoencoders,” [arXiv:1808.08992 \[hep-ph\]](#).
- [32] T. Heimgel, G. Kasieczka, T. Plehn, and J. M. Thompson, “QCD or What?,” [arXiv:1808.08979 \[hep-ph\]](#).
- [33] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, “Novelty Detection Meets Collider Physics,” [arXiv:1807.10261 \[hep-ph\]](#).
- [34] T. Plehn and M. Spannowsky, “Top Tagging,” *J. Phys. G* **39** (2012) 083001, [arXiv:1112.4441 \[hep-ph\]](#).
- [35] G. Kasieczka, T. Plehn, T. Schell, T. Strebler, and G. P. Salam, “Resonance Searches with an Updated Top Tagger,” *JHEP* **06** (2015) 203, [arXiv:1503.05921 \[hep-ph\]](#).
- [36] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, “Stop Reconstruction with Tagged Tops,” *JHEP* **10** (2010) 078, [arXiv:1006.2833 \[hep-ph\]](#).
- [37] T. Plehn, G. P. Salam, and M. Spannowsky, “Fat Jets for a Light Higgs,” *Phys. Rev. Lett.* **104** (2010) 111801, [arXiv:0910.5472 \[hep-ph\]](#).
- [38] G. F. Sterman, “Some Basic Concepts of Perturbative QCD,” *Acta Phys. Polon.* **B39** (2008) 2151–2172, [arXiv:0807.5118 \[hep-ph\]](#).
- [39] M. Cacciari, G. P. Salam, and G. Soyez, “The anti- k_t jet clustering algorithm,” *JHEP* **04** (2008) 063, [arXiv:0802.1189 \[hep-ph\]](#).
- [40] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, “Better jet clustering algorithms,” *JHEP* **08** (1997) 001, [arXiv:hep-ph/9707323 \[hep-ph\]](#).
- [41] S. D. Ellis and D. E. Soper, “Successive combination jet algorithm for hadron collisions,” *Phys. Rev. D* **48** (1993) 3160–3166, [arXiv:hep-ph/9305266 \[hep-ph\]](#).
- [42] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, “Longitudinally invariant K_t clustering algorithms for hadron hadron collisions,” *Nucl. Phys. B* **406** (1993) 187–224.
- [43] G. C. Blazey *et al.*, “Run II jet physics,” in *QCD and weak boson physics in Run II. Proceedings, Batavia, USA, March 4-6, June 3-4, November 4-6, 1999*, pp. 47–77. 2000.

- [arXiv:hep-ex/0005012](https://arxiv.org/abs/hep-ex/0005012) [hep-ex].
http://lss.fnal.gov/cgi-bin/find_paper.pl?conf-00-092.
- [44] M. Cacciari, G. P. Salam, and G. Soyez, “Fastjet user manual,” *The European Physical Journal C* **72** no. 3, (Mar, 2012) 1896.
<https://doi.org/10.1140/epjc/s10052-012-1896-2>.
 - [45] A. Zucconi, “AN INTRODUCTION TO AUTOENCODERS.”
<https://www.alanzucconi.com/2018/03/14/an-introduction-to-autoencoders/>.
 - [46] Hahnloser Richard H. R., Sarpeshkar Rahul, Mahowald Misha A., Douglas Rodney J., and Seung H. Sebastian, “Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit,” *Nature* **405** (Jun, 2000) 947.
<https://www.nature.com/articles/35016072#supplementary-information>.
 - [47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE* **86** no. 11, (Nov, 1998) 2278–2324.
 - [48] J. Masci, U. Meier, D. C. Ciresan, and J. Schmidhuber, “Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction,” in *21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I*, T. Honkela, W. Duch, M. Girolami, and S. Kaski, eds., International Conference on Artificial Neural Networks (ICANN). 2011.
 - [49] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *arXiv e-prints* (Dec., 2013) arXiv:1312.6114, [arXiv:1312.6114](https://arxiv.org/abs/1312.6114) [stat.ML].
 - [50] R. Prescott Adams and R. S. Zemel, “Ranking via Sinkhorn Propagation,” *arXiv e-prints* (June, 2011) arXiv:1106.1925, [arXiv:1106.1925](https://arxiv.org/abs/1106.1925) [stat.ML].
 - [51] J. W. Monk, “Deep Learning as a Parton Shower,” [arXiv:1807.03685](https://arxiv.org/abs/1807.03685) [hep-ph].
 - [52] T. Sjostrand, S. Mrenna, and P. Z. Skands, “PYTHIA 6.4 Physics and Manual,” *JHEP* **05** (2006) 026, [arXiv:hep-ph/0603175](https://arxiv.org/abs/hep-ph/0603175) [hep-ph].
 - [53] T. Sjostrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, “An Introduction to PYTHIA 8.2” *Comput. Phys. Commun.* **191** (2015) 159–177, [arXiv:1410.3012](https://arxiv.org/abs/1410.3012) [hep-ph].
 - [54] DELPHES 3 Collaboration, J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lematre, A. Mertens, and M. Selvaggi, “DELPHES 3, A modular framework for fast simulation of a generic collider experiment,” *JHEP* **02** (2014) 057, [arXiv:1307.6346](https://arxiv.org/abs/1307.6346) [hep-ex].
 - [55] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR* **abs/1412.6980** (2014), [arXiv:1412.6980](https://arxiv.org/abs/1412.6980). <http://arxiv.org/abs/1412.6980>.
 - [56] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, “MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems,” *arXiv e-prints* (Dec., 2015) arXiv:1512.01274, [arXiv:1512.01274](https://arxiv.org/abs/1512.01274) [cs.DC].
 - [57] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “Tensorflow: Large-scale machine learning on heterogeneous

- distributed systems,” *CoRR* **abs/1603.04467** (2016) , [arXiv:1603.04467](#).
<http://arxiv.org/abs/1603.04467>.
- [58] Y. Tang, “Tf.learn: Tensorflow’s high-level module for distributed machine learning,” *CoRR* **abs/1612.04251** (2016) , [arXiv:1612.04251](#). <http://arxiv.org/abs/1612.04251>.
 - [59] J. Gallicchio and M. D. Schwartz, “Seeing in Color: Jet Superstructure,” *Phys. Rev. Lett.* **105** (2010) 022001, [arXiv:1001.5027 \[hep-ph\]](#).
 - [60] M. Dasgupta, M. Guzzi, J. Rawling, and G. Soyez, “Top tagging : an analytical perspective,” *JHEP* **09** (2018) 170, [arXiv:1807.04767 \[hep-ph\]](#).
 - [61] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, “Towards an understanding of jet substructure,” *JHEP* **09** (2013) 029, [arXiv:1307.0007 \[hep-ph\]](#).
 - [62] M. Dasgupta, A. Fregoso, S. Marzani, and A. Powling, “Jet substructure with analytical methods,” *Eur. Phys. J. C* **73** no. 11, (2013) 2623, [arXiv:1307.0013 \[hep-ph\]](#).
 - [63] S. Marzani, G. Soyez, and M. Spannowsky, “Looking inside jets: an introduction to jet substructure and boosted-object phenomenology,” [arXiv:1901.10342 \[hep-ph\]](#).
 - [64] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, “Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks,” *Phys. Rev. Lett.* **101** (2008) 142001, [arXiv:0806.0848 \[hep-ph\]](#).
 - [65] J. Thaler and K. Van Tilburg, “Maximizing Boosted Top Identification by Minimizing N-subjettiness,” *JHEP* **02** (2012) 093, [arXiv:1108.2701 \[hep-ph\]](#).
 - [66] **CMS** Collaboration, C. Collaboration, “Boosted Top Jet Tagging at CMS,”.
 - [67] **ATLAS** Collaboration, M. Aaboud *et al.*, “Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC,” [arXiv:1808.07858 \[hep-ex\]](#).
 - [68] S. D. Ellis, A. Hornig, D. Krohn, and T. S. Roy, “On Statistical Aspects of Qjets,” *JHEP* **01** (2015) 022, [arXiv:1409.6785 \[hep-ph\]](#).
 - [69] S. D. Ellis, A. Hornig, T. S. Roy, D. Krohn, and M. D. Schwartz, “Qjets: A Non-Deterministic Approach to Tree-Based Jet Substructure,” *Phys. Rev. Lett.* **108** (2012) 182003, [arXiv:1201.1914 \[hep-ph\]](#).
 - [70] A. J. Larkoski, G. P. Salam, and J. Thaler, “Energy Correlation Functions for Jet Substructure,” *JHEP* **06** (2013) 108, [arXiv:1305.0007 \[hep-ph\]](#).
 - [71] I. W. Stewart, F. J. Tackmann, and W. J. Waalewijn, “N-Jettiness: An Inclusive Event Shape to Veto Jets,” *Phys. Rev. Lett.* **105** (2010) 092002, [arXiv:1004.2489 \[hep-ph\]](#).
 - [72] J. Thaler and K. Van Tilburg, “Identifying Boosted Objects with N-subjettiness,” *JHEP* **03** (2011) 015, [arXiv:1011.2268 \[hep-ph\]](#).
 - [73] **CMS** Collaboration, V. Khachatryan *et al.*, “Identification techniques for highly boosted W bosons that decay into hadrons,” *JHEP* **12** (2014) 017, [arXiv:1410.4227 \[hep-ex\]](#).
 - [74] D. Krohn, M. D. Schwartz, M. Low, and L.-T. Wang, “Jet Cleansing: Pileup Removal at High Luminosity,” *Phys. Rev. D* **90** no. 6, (2014) 065020, [arXiv:1309.4777 \[hep-ph\]](#).
 - [75] M. Cacciari, G. P. Salam, and G. Soyez, “SoftKiller, a particle-level pileup removal method,” *Eur. Phys. J. C* **75** no. 2, (2015) 59, [arXiv:1407.0408 \[hep-ph\]](#).
 - [76] D. Bertolini, P. Harris, M. Low, and N. Tran, “Pileup Per Particle Identification,” *JHEP* **10** (2014) 059, [arXiv:1407.6013 \[hep-ph\]](#).

- [77] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, “Pileup Mitigation with Machine Learning (PUMML),” *JHEP* **12** (2017) 051, [arXiv:1707.08600 \[hep-ph\]](#).