Tag N’ Train: A Technique to Train Improved Classifiers on Unlabeled Data

Oz Amram\textsuperscript{1} Cristina Mantilla Suarez\textsuperscript{1}

\textsuperscript{1}Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218, USA

E-mail: oamram1@jhu.edu

ABSTRACT: There has been substantial progress in applying machine learning techniques to classification problems in collider and jet physics. But as these techniques grow in sophistication, they are becoming more sensitive to subtle features of jets that may not be well modeled in simulation. Therefore, relying on simulations for training will lead to sub-optimal performance in data, but the lack of true class labels makes it difficult to train on real data. To address this challenge we introduce a new approach, called Tag N’ Train (TNT), that can be applied to unlabeled data that has two distinct sub-objects. The technique uses a weak classifier for one of the objects to tag signal-rich and background-rich samples. These samples are then used to train a stronger classifier for the other object. We demonstrate the power of this method by applying it to a dijet resonance search. By starting with autoencoders trained directly on data as the weak classifiers, we use TNT to train substantially improved classifiers. We show that Tag N’ Train can be a powerful tool in model-agnostic searches and discuss other potential applications.
1 Introduction

Despite numerous searches for physics beyond the standard model [1–5], the experiments at the Large Hadron Collider have not yet provided evidence for new particles or new fundamental forces of nature. While ATLAS, CMS and LHCb have explored many possible signatures, their searches have often been tailored with a particular signal model in mind and have left unexplored a variety of final states [6, 7]. Given that it is impractical to perform a dedicated search for every possible new physics signature, it is natural to consider designing model-agnostic searches that make minimal assumptions about the nature of the signal. Such searches have traditionally been performed at collider experiments by comparing distributions in data to simulation in many different final states [8–19]. However, this technique is insensitive to signals with very small cross sections or in final states not well modeled by simulation.

At the same time, classifiers used to tag hadronic jets of different types have greatly increased in performance thanks to the use of machine learning [20–31], but in almost all applications, classifiers have been trained using simulation. Because simulations do not perfectly model the actual radiation pattern in jets [32], their use in training will lead to sub-optimal performance on real data. Training a model directly on data seems to be the only way not to compromise performance, but this is challenging due to the lack of true labels.

These two predicaments naturally motivate further explorations of how to directly train classifiers on actual data, and how to use them to perform model-agnostic searches on LHC data.

In [33], the Classification Without Labels (CWoLa) approach was introduced as a way to train classifiers directly on data. Instead of relying on fully labeled examples, the CWoLa
approach trains by using statistical mixtures of events with different amounts of signal; allowing the use of many of the techniques from fully-supervised training. To apply this technique in practice, one must find information orthogonal to the classification task that can be used to select the mixed samples in data. One application of this technique has been in a model-agnostic dijet resonance search [34, 35]. In this approach, called CWoLa hunting, a particular invariant mass region is used to select the potentially signal-rich sample and neighboring sidebands are used to select the background-rich sample. This method has been shown to be highly sensitive to resonant new physics, but it is unclear how to extend the approach to non-resonant signals where the anomaly is not localized in a particular kinematic distribution, such as the resonance mass. The technique further relies on the information used in classification being completely uncorrelated with the resonance mass. Slight correlations may lead to the classifier preferentially selecting background events at a particular resonance mass as signal-like, distorting the distribution of background events and complicating the background estimate.

Another approach that has been explored [36–39] is to scan for anomalous events by using autoencoders trained directly on data. Autoencoders are a type of network that learns how to compress an object to a smaller latent representation and then decompress it to reconstruct the original object. An autoencoder trained on a background-rich sample can learn how to compress and decompress objects in background events, but will not learn to do the same for anomalous events. The reconstruction loss, the difference between the original and reconstructed representation of the object, can then be used as a classification score that selects anomalous signals. While the autoencoders have the advantage of making very minimal model assumptions about the signal, their output as a signal-vs-background classifier is sub-optimal. This is because their training aim is compression and decompression, not classification; they learn how to effectively represent the data’s dominant component (i.e. background events), but they do not learn anything about what the sought-after signal looks like.

There have also been recent proposals to use other machine learning techniques for model independent searches[40–50]. See [49] for an overview of the various techniques and their degree of model-independence.

We propose a technique for training classifiers on data that utilizes the CWoLa paradigm to improve weak classifiers. The method, called Tag N’ Train (TNT), is based on the assumption that signal events contain two separate objects; and thus the appearance of these objects is correlated. By using the weak classifier and one of the objects, one can tag events as signal-like or background-like. This then provides samples of signal-rich and background-rich events which can be used to train a classifier for the other object. This technique has a natural application to a model-agnostic searches for new physics in di-object events. We explore a dijet resonance search based on TNT that uses autoencoders as the initial weak classifiers. We find that its sensitivity compares favorably to that of CWoLa hunting and autoencoder-based approaches. We also highlight that the TNT approach naturally allows data-driven ways to estimate QCD backgrounds in anomaly searches.

This paper is organized as follows. Section 2 outlines the Tag N’ Train technique and its key assumptions. The remainder of the paper illustrates the power of TNT through an
example based on a dijet search using jet substructure. Section 3 describes the simulation and deep learning setup. Section 4 emulates an LHC dijet anomaly search and includes signal sensitivity comparisons of the TNT technique to the CWoLa hunting and autoencoder based searches. Conclusions and possible future applications of the TNT approach are discussed in 5.

2 Tag N’ Train

The Tag and Train technique is a method for training classifiers directly on data. The technique assumes that the data has 2 distinct objects and, each of the objects can be “tagged”, i.e. each has a weak classifier can select signal-like events. It takes as input a set of unlabeled data events and the initial classifiers, and outputs two new classifiers which may be substantially improved. The original classifiers might be trained directly on data with an unsupervised approach (e.g. autoencoders), trained on simulation that is known to mis-model data, or might be single features of the data which are known a priori to be useful in distinguishing signal vs. background.

The main idea behind the approach is to exploit the paired nature of the data, where one can use one sub-component of the data to tag examples as signal-like or background-like. These signal-rich and background-rich samples can then be used to train a classifier for the other sub-component. The approach bears some similarity to the commonly used Tag and Probe technique, that uses the two body decays of resonances to measure efficiencies in data [51, 52].

TNT needs a consistent decomposition of the data into two sub-components hereafter named Object-1 and Object-2. It assumes one has initial classifiers for Object-1 and Object-2. The procedure to train new classifiers is as follows:

1. Classify events as signal-like or background-like using the Object-1’s in each event.

2. Train a classifier to distinguish between the Object-2’s in the signal-rich sample and the Object-2’s in the background-rich sample.

3. Repeat the procedure, constructing samples by classifying the Object-2’s in each event, and training a classifier for Object-1 using these samples \(^1\).

The Tag N’ Train sequence is shown graphically in Figure 1.

We stress that the signal-rich and background-rich samples, obtained from the original classifiers do not need to be very pure for the technique to work. In our example use case, detailed in Sec 4, we find that even for a few percent signal in the signal-rich sample, TNT can improve on the original classifiers. This allows one to use weak classifiers as inputs, and/or apply the technique to data samples with very small amounts of signal present.

One of the main assumptions behind this technique is that the information used for the classification tasks of Object-1’s and Object-2’s is uncorrelated in background events.

\(^1\)One could also use the new classifier trained for Object-2 to train the classifier for Object-1, avoiding the need for initial classifiers for both objects. But this risks correlating the two classifiers.
Figure 1. An illustration of the Tag N’ Train technique. Here O1 and O2 represent Object-1 and Object-2, the two components of the data one wishes to train classifiers for.

That is, the background objects in the signal rich and background rich samples must have identical distributions even though they were selected using the other object in the event. If this is not the case, then the classifier may learn the difference between the background objects in these two samples, instead of learning information about the signal. If this condition is satisfied, then the requirements of the CWoLa paradigm are fulfilled and the classifier will be asymptotically optimal. Additionally, the scores of the classifier for object 1 and object 2 on background events will remain uncorrelated, which is often desirable. In practice, one can afford to slightly violate this condition and still achieve good results, as long as the difference between the background in the two samples is smaller than the difference caused by the presence of signal.

The technique works better if the initial classifier can create a larger separation between signal and background in the mixed samples used for training. Thus, if the TNT’s output classifiers achieve better separation than the starting ones, multiple iterations of this technique can further improve classification performance until a plateau is reached.

3 Methods

3.1 Sample Generation

To test our search strategy we use the research development dataset from the LHC Olympics 2020 challenge [53]. The dataset consists of 1M QCD dijet events and 100k $W' \rightarrow X Y$ events, both produced with PYTHIA 8 [54, 55] with no pileup or multiple parton interactions included. The $W'$ has a mass of 3.5 TeV, and the X and Y have masses of 500 GeV and 100 GeV respectively and both decay promptly to pairs of quarks. Because of the large Lorentz boost, the hadronic decays of the X and Y bosons can each be captured in a single large radius jet.

Detector simulation is performed with Delphes 3.4.1 [56] and particle flow objects are clustered into jets using the FASTJET [57, 58] implementation of the anti-$k_t$ algorithm [59].
with a radius parameter of \( R = 1.0 \).

For every event, we construct separate jet images \([23, 27–30]\) for the two highest \( p_T \) jets, to be used in event classification. Following \([30]\), we apply pre-processing steps to our jets images before they are pixelated. We center the image based on the \( p_T \) weighted center of the jet constituents and rotate the jet so that the principle axis is in the 12 o’clock position. Then the image is flipped along both axes so that the hardest \( p_T \) constituent of the jet is in the upper left quadrant of the image. After these steps, the image is pixelated into a 40 x 40 pixel image. The image covers an \( \eta \) range and \( \phi \) range of -0.7 to 0.7 around the center of the jet. In order to reduce dependence on the \( p_T \) of the jet, each image is normalized so that the sum of all the pixel intensities sums 1. The sample of images is then normalized so that each pixel has zero mean and unit variance.

### 3.2 Architectures

We use neural networks built and trained in Keras \([60]\) with a TensorFlow \([61]\) backend for all the classifiers considered in this work. All networks are trained with the Adam optimizer with a learning rate of 0.001, first and second moments decay rates of 0.8 and 0.99 respectively and a learning rate decay of 0.0005. Unless otherwise stated, all nodes use a Rectified Linear Unit (ReLu) activation function.

As this work is a proof of concept for the Tag N’ Train technique, none of the network architectures have been optimized.

To train the autoencoders we use a convolutional network with filter sizes of 3x3 where the image’s dimensionality is reduced through max pooling layers after each convolutional layer. The output is then fed through a dense layer which outputs the latent representation. Based on \([36, 37]\) we choose the size of our latent dimension to be 16 as this was seen to be within the performance plateau in both. Then the architecture is mirrored, with 2D sampling layers in place of the max pooling layers to output an image of the same dimensions. We use a Mean Squared Error loss function during the classifier training.

To train the image based classifiers we also use a convolution network with filter sizes of 3x3 for all convolutional layers followed by dense layers. The final layer has a sigmoid activation function. A binary cross-entropy loss is used during training.

### 4 A Dijet Anomaly Search

#### 4.1 Search Strategy

One exciting application of the Tag N’ Train technique is two-body searches at the LHC. We consider specifically a dijet anomaly search where one uses an autoencoder trained directly on data as the initial classifier, then uses the TNT technique to train improved classifiers. These improved classifiers are then used to suppress QCD backgrounds and a resonance is searched for in the invariant mass of the dijet events.

We implement the search strategy as follows. For each event, we consider the highest two \( p_T \) jets to be the dijet candidate. In order to apply the Tag N’ Train technique, we treat our Object-1 as the more massive jet in each event and Object-2 as the less massive jet. For each event we then have a separate image for the heavier and lighter jet. To evaluate
how well the strategy works with varying levels of signal, we vary the amount of signal present in the dataset by filtering out signal events. We run the search in the case where 9%, 1%, 0.3% and 0.1% of the events in the dataset are signal.

Using an initial sample of 200k events we train separate autoencoders for the heavier and lighter jets in the sample. We use the autoencoder architecture described in section 3.2 for both autoencoders, use 10% of events for validation, and train for 30 epochs. We then use these autoencoders and a new sample of 200k events to train new classifiers with the Tag N’ Train technique. Specifically, we define our signal-rich samples as the 20% of events with the highest autoencoder loss, and the background-rich samples as the 40% of events with the lowest autoencoder loss. We iterate the TNT procedure for a total of 3 iterations, each time using the classifiers from the previous iteration and a new set of 200,000 events as the inputs. For the second iteration onwards we use the 10% events with the highest classifier score for our signal-rich sample, but still use the same selection for the background-rich sample. We did not extensively optimize these selections, but did check that the performance of the technique is robust to the exact values used.

Because we are searching for a resonance we have additional information about the nature of our signal: it is likely to be localized to particular region in the dijet mass spectrum. In such cases, where one has a priori assumptions about the anomalous events, one can add them as additional selection requirements for events to be signal-like. Specifically, we require the events in the signal-rich sample to to fall within a dijet invariant mass window in addition to the cut on the classifier score. In a real search in data, one would scan the dijet mass range, training a separate set of networks for each dijet mass window (as is suggested in [34, 35]), but we simplify things here by just requiring the dijet mass to be near the resonance mass of 3.5 TeV, i.e. within 3.3 and 3.7 TeV. We do not apply any dijet mass selection to the background-rich sample, so that events from this sample populate this mass window as well. These additional requirements improve the fraction of signal events in the signal-rich sample. That is, for the dataset with 0.3% signal events, there is 1% signal in the signal-rich samples without a dijet mass cut, 1.5% signal in the mass window with no additional requirement, and 5% when both are used together.

Because the TNT technique creates separate classifiers for each jet, one must combine them in a sensible way in order to select anomalous events. Given the unsupervised nature of the search, one will not know a priori what the optimal event selection will be for each object, and likely multiple criteria will be tried. We select events by choosing a certain percentile $X$ and requiring that the respective classifier score for both jets in that event to be above that percentile in their respective distributions. For example, if we pick a percentile selection of $X = 20\%$ for our search, we select events where jet-1 must be in the top 20% of jet-1 scores and jet-2 must be in the top 20% of jet-2 scores.

---

2For future work, it might be interesting to explore the use of a “soft” labeling scheme where the label for each event is not required to be strictly 0 or 1, but allowed to be a value in between based on the score of the initial classifier.

3This approach for combining the jet scores makes sense when both scores have roughly equal discrimination power and a “tight” selection is desired. In other use cases one should pursue a different strategy.
For the example search, we try different percentile cuts and search for a resonance in the resulting dijet mass spectrum. After selecting the signal-like events in the validation sample using the combination of the classifier scores, we bin this dataset in $M_{jj}$ and perform the signal extraction. We model the shape of the falling QCD background component with a third degree polynomial and propagate the uncertainties on the polynomial parameters as systematic uncertainties of the fit. We assume the signal is relatively narrow, and fit the resonant signal with a Gaussian peak. We do not attempt to perform a more complicated parametric fit to the background nor explore a different functional form. We rather assume that because the mass distribution is smoothly falling after the TNT selection, this will allow for a sufficient number of sideband events, that will reliably constrain the QCD multijet background.

Although we have not done so here, in a real implementation of this technique one would want to use a cross-validation scheme so all the data could be used for the search and none is “wasted” by the training (as in [34, 35]). This would involve splitting up the data into multiple samples and cycling through which samples are used for training and which were used to search for a signal. Then one would simultaneously fit all of the signal regions to achieve full sensitivity.

4.2 Results

We use an independent sample of 100k events to evaluate the performance of the pure TNT classifiers, and the classifiers trained using TNT and a dijet mass cut (hereafter called TNT + $M_{jj}$ classifiers). In Fig. 2, we compare their performance against fully supervised classifiers, those trained using the CWoLa hunting method [36, 37] and autoencoders. We also compare the performance of supervised classifiers trained using the images of both jets at the same time and classifiers trained on each jet individually and combined in the same way as the TNT classifiers. One can see that there is a noticeable drop in performance when separating the jets, but that good classification performance is still possible. Although the 9% signal test is rather optimistic from an anomaly search perspective, it shows that the both the TNT and TNT + $M_{jj}$ converge to the performance of a fully supervised classifier given sufficient signal. For the 1% signal test, the TNT classifier is somewhat worse than the TNT + $M_{jj}$ classifier, but still has significantly improved performance with respect to the autoencoder. Finally, for the 0.3% and 0.1% signal tests, we can see that there is too little signal for the TNT classifier to learn from, and TNT performs significantly worse than the autoencoder. The TNT + $M_{jj}$ classifier performs similarly to CWoLa hunting for the 3 tests with larger signal, but for the 0.1% test the TNT + $M_{jj}$ is able to maintain better performance better than the CWoLa hunting method, but without improving with respect to the autoencoders approach.

In addition to achieving good classification performance, we also highlight that neither the TNT or TNT + $M_{jj}$ classifiers significantly sculpt the QCD dijet mass distribution. In Figure 3, we show the QCD dijet mass distribution after applying various cuts using the TNT + $M_{jj}$ classifier. We can see that the shape of the distribution is not altered by any of these cuts. This is crucial because it allows the use of data driven estimates of the QCD background which rely on the smoothness of the dijet mass distribution. The lack
Figure 2. Classification performance for distinguishing signal and background events. The numbers next to each label show the Area Under Curve (AUC) of each classifier. We show results for when the fraction of signal events is 9%, 1%, 0.3% and 0.1% in the top left, top right, bottom left and bottom right respectively. We show the performance of the Tag N’ Train technique with and without the dijet mass cut, and the CWoLa hunting and autoencoders. Two fully supervised classifiers trained with ground-truth labels, are shown for reference. The one labeled ‘separate’ consists of 2 classifiers, trained on each jet separately and combined in the same way as the TNT classifiers, and the one labeled ‘both jets’ is trained on both jets at the same time.

of sculpting is due to our choice of classifier inputs, we normalize each jet image so that the sum of all pixel intensities is 1. This means that each image does not carry very much information about jet $p_T$ which can be used to sculpt the dijet mass distribution. But it is also important to point out algorithmic differences that can mitigate the risk of sculpting, between the TNT approach with a dijet mass cut and the CWoLa hunting. The first is that TNT selects the background-like and signal-like events using more information than just the dijet mass cut. This means that there are background events that populate the signal window and that are used in the training with a dijet mass in the signal-window, whereas in the CWoLa hunting approach all background events are in dijet mass sidebands. The other advantage is that by training a classifier for each jet separately, one can try to explicitly decorrelate the classifier’s dependence on jet $p_T$ through one of the techniques that have
Figure 3. The QCD dijet mass distribution after applying different selections on the TNT + $M_{jj}$ classifiers, trained with the 1% signal in the dataset. The TNT selection is based on an $X\%$ percentile cut on the separate jet scores as described in the text. The QCD dijet mass distribution remains smooth, allowing the use of data-driven background estimates.

been used in supervised jet classification [62–66]. We explored reweighting events in the background-rich sample to have the same $p_T$ distribution as the signal-rich region, but as there was not much mass sculpting to begin with, there were no significant differences in the mass sculpting or classification performance.

Another key feature to point out is that the TNT procedure maintains the independence of jet 1 and jet 2 scores on background events. In Fig. 4 we show the correlation between the jet 1 and jet 2 classification scores for classifiers trained with TNT. We also show a similar figure for ‘pure CWoLa’ classifiers trained using randomly selected mixed samples with signal-rich and background-rich samples with similar S/B’s to the TNT training. We can see that the TNT classifier produces roughly similar distribution of jet 1 and jet 2 scores as the pure CWoLa classifier. We also compute the Pearson correlation coefficient between jet 1 and jet 2 scores for background events. We can see that the TNT classifier does not develop a significant correlation between jet 1 and jet 2 scores. This is desirable from an event-level classification standpoint because it means when a background event passes the selection for one of the classifiers it is not biased to be more likely to pass the other. Additionally this independence allows the use of background estimation techniques that rely on creating control regions by inverting the selection on one of the jets.

We test the fitting procedure detailed in the last section and fit for the presence of a resonant signal at 3.5 TeV. We see that when no signal is present, no significant bump is created by our procedure. When a signal is present there is a significant dijet mass bump

\[^4\text{In principle it is possible to do this in the CWoLa hunting approach as well if one trains a classifier for each jet separately, and attempts to decorrelate classification score from jet } p_T.\]
Figure 4. Correlation between classifier scores for the TNT classifier trained on a dataset with 1% signal (left) and a classifier trained using CWoLa on mixed samples with similar signal to background ratio as the samples used in the TNT training (right). Blue dots are QCD background and red dots are signal. In the top right we report the Pearson correlation coefficient of the jet 1 and jet 2 scores for background events.

Figure 5. $m_{jj}$ distribution of dijet events (including injected signal, indicated by the green gaussian shape) before and after applying tight percentile cuts using the TNT classifier outputs. We look for a $m_{res} = 3.5$ TeV signal mass hypothesis. The blue lines indicate the fit to the background component of the data, with the yellow bands representing the parametric fit uncertainties. The TNT classifiers and the search strategy were performed in a test dataset with only signal fractions of 1% (left) and 0.3% (right).

that forms at the signal hypothesis, after a tight percentile selection on our data. We use test datasets and train TNT classifiers with events that only contain 1% or 0.03% of the signal included in the original dataset. In the dataset with 1% of signal, we use a tight 1% percentile selection and see a large signal significance that reaches $7.7\sigma$, see Fig. 5. In the dataset with 0.3% of signal, we use a 10% percentile selection and observe a signal bump at $6.3\sigma$ level.

It is necessary to remark that the p-value computed in our results is only local. The translation to a global p-value would depend on the procedure used to scan over the full dijet mass range and how many different selections are tried in each mass window\textsuperscript{5}.

\textsuperscript{5}Because the p-values observed with different selections in the same mass window will be correlated,
Figure 6. Correlation between $\tau_{21}$ and jet mass for the dataset with 1% signal. The top row corresponds to the heavier jet and the bottom row the lighter jet. On the left (in red) are the 1% of events the TNT classifier found to be most signal-like and on the right (in blue) the true signal events.

If a local p-value is below some threshold, it would crucial to characterize the nature of the signal. While there are some known strategies that can be used to understand what a deep CNN has learned [67–69], a simple approach would be to just examine the events that the classifier has found to be most signal like. In Fig. 6 we compare the characteristics of the events the TNT classifier found to be most signal-like to the characteristics of the true signal events. We show 2D scatter plots of jet mass and the N-subjetiness ratio $\tau_{21}$ [70]. Despite not using the jet mass or N-subjetiness as direct inputs to the network, one can see that the TNT classifier has learned the correct masses of the X and Y jets and that they are both two pronged. Characterizing the signal in this way would also give an analyzer confidence they had truly found evidence of new physics rather than a unknown feature of the detector.

5 Conclusions

We have introduced a new method of training classifiers directly on data called Tag N’ Train. It relies on decomposing the data into two distinct sub-objects which can be classified separately. One can then use one of the sub-objects to tag events as signal-like or computation of a global p-value is non-trivial. We leave the exploration of this for future work.
background-like, and those samples can be used to train a classifier for the other object. Here we have explored the possibility of using the Tag N’ Train technique to perform a dijet anomaly search by using autoencoders trained directly on data as the initial classifiers. We demonstrate that given sufficient signal in the data, the TNT technique is able to produce classifiers that perform significantly better than the autoencoders. When a cut on the dijet mass is used in addition to the autoencoders to select signal-like events, the TNT classifiers achieve similar performance to those trained using the CWoLa hunting technique.

As this work was meant to be a proof of concept for the Tag N’ Train method, we believe there is substantial room to improve the performance of the TNT dijet anomaly search, both by optimizing the initial network used to detect anomalous jets, and the architecture of the classifier trained with TNT. An obvious direction to explore would be other variants of the autoencoder architecture, such as variational autoencoders [42, 71] or normalizing-flow based autoencoders [72], but in principle any anomaly detection method that is able to isolate signal events using only one jet at a time could work as an initial classifier. Also, while using low-level inputs like jet images to the TNT classifiers offers robustness to many types of signals using higher-level features may offer advantages as well. If there was only a small amount of signal present, it would likely be easier for the network to learn if higher level features were used. However, by restricting the information given to a hand-selected subset of variables, one may lose sensitivity to anomalies exhibiting novel features. We leave the exploration of these ideas for future work.

Although here we have applied the TNT technique to a dijet resonance search, the performance of the TNT technique without using dijet mass window, shows it could be applied to a non-resonant anomaly search as well. Because the classifier scores for the two jets are independent, data-driven background estimations are possible. For example, in an anomalous pair production search one could measure the rate of QCD jets passing the heavier jet selection in a sample of data events where the light jet is required to be background-like, and vice versa. After measuring both rates, one could estimate the amount of QCD jets in the signal region. A key point to explore would be how the Tag N’ Train technique performs in the presence of sub-dominant SM backgrounds with interesting sub-structure, such as top quark pairs or $W+\text{Jets}$ production. Preventing the TNT technique from learning these events as signal-like may require additional control regions to be added to the background-rich sample in training, or explicitly veto-ing events that look like from known SM backgrounds from the signal-rich sample during training.

The Tag N’ Train framework could also be applied to model-specific search as well. Running the Tag N’ Train technique with classifiers for Standard Model jets as inputs, while scanning for a resonance, could target models covered by existing searches. The possible advantage of the Tag N’ Train framework would be that by training new classifiers directly on data, one would mitigate the effects of imperfections in the simulation used in training. It would be interesting to compare how the performance of this sort of Tag N’ Train search compares to a traditional supervised search if there were significant mis-modeling of the signal or background in simulation.
Code and data availability

Code to reproduce all of our results can be found on Github and the dataset used is available on Zenodo [53].

Acknowledgments

We thank Petar Maksimovic for useful input and his continuous support. We also would like to thank the organizers of the ML4Jets conference, and in particular Gregor Kasieczka, Ben Nachman, and David Shih for organizing the LHC Olympics 2020 challenge that was the impetus for this work.

References

[1] ATLAS collaboration, “Exotic physics searches.”
https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults (2020).

[2] ATLAS collaboration, “Higgs and diboson searches.”
https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults (2020).

[3] CMS collaboration, “Cms exotica public physics results.”
https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO (2020).

[4] CMS collaboration, “Cms beyond-two-generations (b2g) public physics results.”
https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G (2020).

[5] LHCb collaboration, “Publications of the qcd, electroweak and exotica working group.”
http://lhcbproject.web.cern.ch/lhcbproject/Publications/LHCbProjectPublic/
Summary_QEE.html (2020).

[6] J. H. Kim, K. Kong, B. Nachman and D. Whiteson, The motivation and status of two-body resonance decays after the LHC Run 2 and beyond, 1907.06659.

[7] N. Craig, P. Draper, K. Kong, Y. Ng and D. Whiteson, The unexplored landscape of two-body resonances, Acta Phys. Polon. B50 (2019) 837 [1610.09392].

[8] B. Abbott, M. Abolins, V. Abramov, B. S. Acharya, D. L. Adams, M. Adams et al., Search for new physics in eµx data at dØ using sleuth: A quasi-model-independent search strategy for new physics, Physical Review D 62 (2000) .

[9] B. Abbott, A. Abdesselam, M. Abolins, V. Abramov, B. S. Acharya, D. L. Adams et al., Quasi-model-independent search for new high pt physics at d0, Physical Review Letters 86 (2001) 3712–3717.

[10] V. M. Abazov, B. Abbott, A. Abdesselam, M. Abolins, V. Abramov, B. S. Acharya et al., Quasi-model-independent search for new physics at large transverse momentum, Physical Review D 64 (2001) .

[11] H1 collaboration, A General Search for New Phenomena at HERA, Phys. Lett. B674 (2009) 257 [0901.0507].

[12] A. Aktas, V. Andreev, T. Anthonis, A. Asmone, A. Babaev, S. Backovic et al., A general search for new phenomena in ep scattering at hera, Physics Letters B 602 (2004) 14–30.
[13] T. Aaltonen, A. Abulencia, J. Adelman, T. Akimoto, M. G. Albrow, B. A. González et al., Model-independent and quasi-model-independent search for new physics at cdf, Physical Review D 78 (2008).

[14] T. Aaltonen, J. Adelman, T. Akimoto, M. G. Albrow, B. A. González, S. Amerio et al., Global search for new physics with 2.0 fb$^{-1}$ cdf, Physical Review D 79 (2009).

[15] CMS collaboration, MUSiC, a Model Unspecific Search for New Physics, in pp Collisions at $\sqrt{s} = 8$ TeV.

[16] CMS collaboration, Model Unspecific Search for New Physics in pp Collisions at $\sqrt{s} = 7$ TeV.

[17] ATLAS collaboration, A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment, Eur. Phys. J. C79 (2019) 120 [1807.07447].

[18] ATLAS collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV.

[19] ATLAS collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV.

[20] A. J. Larkoski, I. Moult and B. Nachman, Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning, 1709.04464.

[21] L. G. Almeida, M. Backović, M. Cliche, S. J. Lee and M. Perelstein, Playing Tag with ANN: Boosted Top Identification with Pattern Recognition, JHEP 07 (2015) 086 [1501.05968].

[22] 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 (2016) 094034 [1603.09349].

[23] G. Kasieczka, T. Plehn, M. Russell and T. Schell, Deep-learning Top Taggers or The End of QCD?, JHEP 05 (2017) 006 [1701.08784].

[24] G. Louppe, K. Cho, C. Becot and K. Cranmer, QCD-Aware Recursive Neural Networks for Jet Physics, JHEP 01 (2019) 057 [1702.00748].

[25] A. Butter et al., The Machine Learning Landscape of Top Taggers, SciPost Phys. 7 (2019) 014 [1902.09914].

[26] A. Butter, G. Kasieczka, T. Plehn and M. Russell, Deep-learned Top Tagging with a Lorentz Layer, SciPost Phys. 5 (2018) 028 [1707.08966].

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

[28] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, Jet-images — deep learning edition, JHEP 07 (2016) 069 [1511.05190].

[29] P. T. Komiske, E. M. Metodiev and M. D. Schwartz, Deep learning in color: towards automated quark/gluon jet discrimination, JHEP 01 (2017) 110 [1612.01551].

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

[31] P. T. Komiske, E. M. Metodiev and J. Thaler, Energy Flow Networks: Deep Sets for Particle Jets, JHEP 01 (2019) 121 [1810.05165].
[32] J. Barnard, E. N. Dawe, M. J. Dolan and N. Rajcic, *Parton Shower Uncertainties in Jet Substructure Analyses with Deep Neural Networks*, Phys. Rev. D95 (2017) 014018 [1609.00607].

[33] E. M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, JHEP 10 (2017) 174 [1708.02949].

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

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

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

[37] T. S. Roy and A. H. Vijay, *A robust anomaly finder based on autoencoder*, 1903.02632.

[38] R. T. D’Agnolo and A. Wulzer, *Learning New Physics from a Machine*, Phys. Rev. D99 (2019) 015014 [1806.02350].

[39] R. T. D’Agnolo, G. Grosso, M. Pierini, A. Wulzer and M. Zanetti, *Learning Multivariate New Physics*, 1912.12155.

[40] 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 [1811.10276].

[41] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty Detection Meets Collider Physics*, 1807.10261.

[42] A. De Simone and T. Jacques, *Guiding New Physics Searches with Unsupervised Learning*, Eur. Phys. J. C79 (2019) 289 [1807.06038].

[43] A. Mullin, H. Pacey, M. Parker, M. White and S. Williams, *Does SUSY have friends? A new approach for LHC event analysis*, 1912.10625.

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

[45] B. M. Dillon, D. A. Faroughy and J. F. Kamenik, *Uncovering latent jet substructure*, Phys. Rev. D100 (2019) 056002 [1904.04200].

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

[47] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*, 2001.04990.

[48] A. Andreassen, B. Nachman and D. Shih, *Simulation Assisted Likelihood-free Anomaly Detection*, 2001.05001.

[49] CMS collaboration, *Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV*, JINST 13 (2018) P06015 [1804.04528].

[50] ATLAS collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} =$13 TeV*, Eur. Phys. J. C76 (2016) 292 [1603.05598].
[53] G. Kasieczka, B. Nachman and D. Shih, *R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge*, Apr., 2019. 10.5281/zenodo.2629073.

[54] T. Sjostrand, S. Mrenna and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, *JHEP* **05** (2006) 026 [hep-ph/0603175].

[55] T. Sjostrand, S. Mrenna and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852 [0710.3820].

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

[57] M. Cacciari and G. P. Salam, *Dispelling the N^{3} myth for the k_t jet-finder*, *Phys. Lett.* **B641** (2006) 57 [hep-ph/0512210].

[58] M. Cacciari, G. P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J.* **C72** (2012) 1896 [1111.6097].

[59] M. Cacciari, G. P. Salam and G. Soyez, *The anti-k_{t} jet clustering algorithm*, *JHEP* **04** (2008) 063 [0802.1189].

[60] F. Chollet et al., “Keras.” https://keras.io, 2015.

[61] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.

[62] J. Dolen, P. Harris, S. Marzani, S. Rappoccio and N. Tran, *Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure*, *JHEP* **05** (2016) 156 [1603.00027].

[63] G. Louppe, M. Kagan and K. Cranmer, *Learning to Pivot with Adversarial Networks*, 1611.01046.

[64] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul et al., *Decorrelated Jet Substructure Tagging using Adversarial Neural Networks*, *Phys. Rev.* **D96** (2017) 074034 [1703.03507].

[65] L. Bradshaw, R. K. Mishra, A. Mitridate and B. Ostdiek, *Mass Agnostic Jet Taggers*, 1908.08959.

[66] G. Kasieczka and D. Shih, *DisCo Fever: Robust Networks Through Distance Correlation*, 2001.05310.

[67] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas et al., *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)*, 2017.

[68] K. Simonyan, A. Vedali and A. Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps*, 2013.

[69] C. Olah, A. Mordvintsev and L. Schubert, *Feature visualization*, *Distill* (2017).

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

[71] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2013.

[72] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever and M. Welling, *Improving variational inference with inverse autoregressive flow*, 2016.