

Self-supervised anomaly detection for new physics

Barry M. Dillon,^{1,*} Radha Mastandrea^{2,3,†} and Benjamin Nachman^{3,4,‡}

¹*Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16,
D-69120 Heidelberg, Germany*

²*Department of Physics, University of California, Berkeley, California 94720, USA*

³*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

⁴*Berkeley Institute for Data Science, University of California, Berkeley, California 94720, USA*



(Received 2 June 2022; accepted 5 August 2022; published 8 September 2022)

We investigate a method of model-agnostic anomaly detection through studying jets, collimated sprays of particles produced in high-energy collisions. We train a transformer neural network to encode simulated QCD “event space” dijets into a low-dimensional “latent space” representation. We optimize the network using the self-supervised contrastive loss, which encourages the preservation of known physical symmetries of the dijets. We then train a binary classifier to discriminate a beyond the standard model resonant dijet signal from a QCD dijet background both in the event space and the latent space representations. We find the classifier performances on the event and latent spaces to be comparable. We finally perform an anomaly detection search using a weakly supervised bump hunt on the latent space dijets, finding again a comparable performance to a search run on the physical space dijets. This opens the door to using low-dimensional latent representations as a computationally efficient space for resonant anomaly detection in generic particle collision events.

DOI: [10.1103/PhysRevD.106.056005](https://doi.org/10.1103/PhysRevD.106.056005)

I. INTRODUCTION

A central goal of high energy physics is to find the theory that will supersede the Standard Model. A number of competing models exist, with many involving new resonant particles [1,2] such as supersymmetric partners or weakly interacting massive particles. However, the number of candidate particles is too large to justify a hunt-and-pick procedure of data analysis. Therefore it is advantageous to consider methods of anomaly detection that are agnostic to a particular underlying model of new physics.

There are a number of recent machine learning (ML) based anomaly detection proposals designed to reduce model dependence (see Refs. [3–6] for overviews). Notably, most existing methods (including the first result with data from ATLAS [7]) in the field are best-performing in low-dimensional spaces. However, a single event from a particle collision experiment can have on the order of a thousand degrees of freedom.

A resolution to this tension is to reduce the dimensionality of an entire particle collision event while preserving its essential character such that a search for anomalies can be done in the reduced-dimension space. A number of methods exist for carrying out this phase space reduction. For example, one could choose a set of observables (e.g., mass, multiplicity) and perform anomaly detection on this set. However, attempts to reduce dimensionality through selecting a choice of observables implicitly favors a certain class of models.

Dimensionality reduction can also be performed with unsupervised ML techniques, which ensure a model-agnostic approach. A common tool in the ML literature for this compression is the autoencoder (AE). An autoencoder is a pair of neural networks whereby one function encodes *event space* data into a *latent space* and a second function decodes the latent space back into the event space. No labels are needed because the AE is trained to ensure that the composition of the encoder and decoder is close to the identity to produce high reconstruction efficiency. While effective (see e.g., [8–11]), AE-based tools may not be ideal for anomaly detection. For one, there is nothing in their architectures or loss functions that ensure that anomalies, or basic physical (i.e., geometric) properties, of events are preserved by the encoder. Additionally, most AE studied so far cannot process a variable number of inputs per event, which would require the decoder to generate a variable number of outputs and the loss to compare events with a variable dimensionality.

*dillon@thphys.uni-heidelberg.de

†rmastand@berkeley.edu

‡bpnachman@lbl.gov

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

These problems can be circumvented by making use of self-supervised contrastive learning techniques, which use only the inherent symmetries of physical data to perform a dimensionality reduction. Recent studies have explored this; for example with astronomical images in Ref. [12], and with constituent level jet data in Ref. [13]. In the former, a ResNet50 architecture [14] is used to map each astronomical image to a latent representation, while in the latter, a permutation-invariant transformer-encoder architecture maps jet constituents to a latent representation. The networks are trained on the contrastive loss, which ensures that the latent space representations faithfully model the physical symmetries of the original objects. Further analysis is then done directly on the latent space representations.

In this paper, we continue the explorations of latent space representations of particle collisions originally carried out in Ref. [13]. We first demonstrate that particle collisions can be well-modeled in a latent space representation with a dimensionality that is an order of magnitude smaller than that of the original events. As part of this work, we extend the per-jet work of Ref. [13] to a per-event structure. We then conduct a low-dimension model-agnostic anomaly search in the latent space representations of the particle collision events. For this we use the classification without labels (CWoLa) technique [15–17], which uses deep neural-network classifiers to distinguish between anomaly-enriched events and anomaly-depleted events. We conduct these studies on dijet resonance events in the LHC Olympics dataset [18].

The structure of this paper is as follows. In Sec. II, we motivate the relevance of contrastive learning to modeling particle collisions and introduce a dataset of dijet events. We further outline a set of *symmetry augmentations* for the contrastive loss function that leave the essential character of dijet events invariant and explain how these are used in the contrastive learning approach. Lastly, in this section we outline the CWoLa anomaly detection method where we will use the self-supervised event representations. In Sec. III, we implement the contrastive learning method using a transformer neural network to map the dijet events from the event space into a latent space and evaluate the efficiency of the encoding. In Sec. IV, we use the CWoLa method to perform a relevant, but simplified, anomaly bump-hunt analysis using the latent space representations for the dijet events. The paper ends with conclusions and outlook in Sec. V.

II. METHODS

Our overarching goal is to optimize a mapping from the event space of particle collision events (i.e., the representation in the space of the individual particles) to a new latent space representation. The mapping between the event and latent space representations of particle collisions should maximally exploit the physical symmetries of particle collision events. In this way, the dimensionality of the events can be reduced from a few hundred degrees of freedom (corresponding to the momentum 4-vectors of the particles) to a few tens.

Such a mapping can be realized by using a *transformer-encoder* neural network architecture [19]. Event space events are fed into a transformer neural network where they are embedded into a reduced-dimension latent space. A distinguishing feature of the transformer architecture is permutation invariance: the latent space representations are invariant with respect to the order that the event constituents are fed into the network. The transformer-encoder network consists of 4 heads each with 2 layers. The output of the transformer-encoder network is fed into a “tail” network, which consists of two layers of fully connected networks of the same latent space size. These architecture parameters were not heavily optimized.

See Fig. 1 for a schematic of the full network. We find, as do the authors of Refs. [13,19], that the latent space representation of the output of the first fully connected layer in the tail network gives a better representation than that of the final output layer, or that of the transformer-encoder output.

A. Data selection and preparation

For this paper, we focus on the LHC 2020 Olympics R&D dataset [3,18]. The full dataset consists of 1,000,000 background dijet events (standard model quantum chromodynamic (QCD) dijets) and 100,000 signal dijet events. The signal comes from the process $Z' \rightarrow X(\rightarrow q\bar{q})Y(\rightarrow q\bar{q})$, with three new resonances Z' (3.5 TeV), X (500 GeV), and Y (100 GeV). The only trigger is a single large-radius jet ($R = 1$) trigger with a p_T threshold of 1.2 TeV. The events are generated with PYTHIA 8.219 [20,21] and DELPHES 3.4.1 [22]. Each event contains up to 700 particles with three degrees of freedom (DoF) p_T, η, ϕ . The average number of nonzero DoF per event is 506 ± 174 .

For each event, we cluster the jets using FastJet [23,24] with a radius $R = 0.8$. We select the two highest-mass jets from each event, and we select the 50 hardest (highest p_T) constituents from each event, zero-padding for any jet with fewer than 50. Note that the average number of constituents per event is 81 with a standard deviation of 16.15. However, we found that including more than 50 constituents per jet did not lead to an appreciable improvement in the performance of the transformer-encoder network. The constituents are assumed to be massless, and so the relevant degrees of freedom for each constituent are (p_T, η, ϕ) .

For analysis, we select jets from the windows $p_T \in [800, 3000]$ GeV and $\eta \in [-3, 3]$. The cut on p_T was chosen such that invariant dijet mass m_{JJ} has a lower bound at approximately 2 TeV. This cut removes approximately 12% of eligible events from the LHCO dataset and has the benefit of removing a small tail of events with m_{JJ} below 2 TeV that could appear to be artificially anomalous to the transformer-encoder network.

B. Contrastive learning

The contrastive learning method is self-supervised, meaning that it is trained using “pseudolabels” rather than truth labels. Supervised approaches use truth labels which exactly

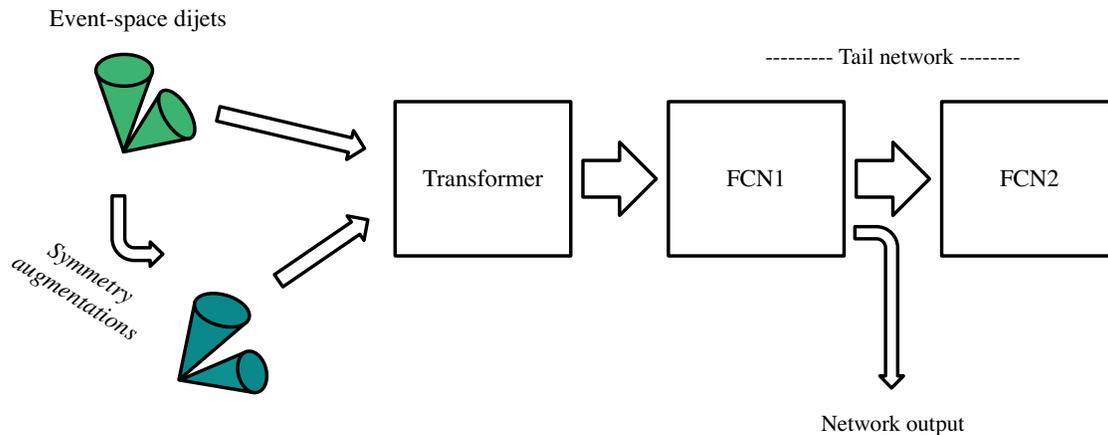


FIG. 1. A schematic of the full transformer-encoder network. Event space dijets and their symmetry-augmented versions are fed as input into the network, which creates a mapping into the latent space by training on the contrastive loss function. The output of the transformer-encoder network is then passed through a tail network, consisting of two fully connected layers (FCN1 and FCN2). In practice, the representations from the first fully connected layer perform the best in signal versus background classification tasks.

identify the truth label of the data. Pseudolabels are artificial labels created from the data alone, without access to the truth labels. This means that the contrastive learning method is also unsupervised and receives no information as to whether the training samples are signal or background. Following JetCLR [13], the pseudolabels are used to identify jets which are related to each other via some augmentation, for example a symmetry transformation. Using the pseudolabels, this technique aims to construct a latent space representation of events that exploits their physical symmetries.

As an example, consider the transformer-encoder network's encoding of a dijet event \mathbf{r}_j , and the encoding of an *augmented* version of that event \mathbf{r}'_j . The exact physical symmetries considered in this analysis are outlined in Sec. II C, but for this example, let the augmentation be a random rotation of the dijet event about the beam axis. These two events represent the same underlying physics, as we expect physical events to be symmetric about the beam axis. Hence, \mathbf{r}_j and \mathbf{r}'_j are often called *positive pairs*. Therefore we would want the transformer-encoder network to map the event and its augmented version into similar regions of the latent space. In contrast, we would expect the transformer to map the jet event \mathbf{r}_j and a different jet event \mathbf{r}_k into different points of the latent space, since we do not expect a high degree of similarity between two arbitrary events. Therefore \mathbf{r}_j and \mathbf{r}_k are often called *negative pairs*. These positive and negative pairs are exactly the pseudolabels for the contrastive learning method.

These requirements on the transformer-encoder mapping motivate the expression for the contrastive loss:

$$\mathcal{L}(\mathbf{r}_j, \mathbf{r}'_j, \mathbf{r}_k, \mathbf{r}'_k, \tau) = -\log \left(\frac{\exp(\text{sim}(\mathbf{r}_j, \mathbf{r}'_j))}{\sum_{j \neq k} [\exp(\text{sim}(\mathbf{r}_j, \mathbf{r}_k)) + \exp(\text{sim}(\mathbf{r}_j, \mathbf{r}'_k))]} \right). \quad (1)$$

We can interpret this loss function as follows: $\text{sim}(\mathbf{r}_j, \mathbf{r}'_j)$ calculates the similarity between two latent space representations, where

$$\text{sim}(\mathbf{r}_1, \mathbf{r}_2) = \frac{\mathbf{r}_1 \cdot \mathbf{r}_2}{\tau \|\mathbf{r}_1\| \|\mathbf{r}_2\|}. \quad (2)$$

The similarity is parametrized by a temperature τ , which balances the numerator and denominator in the contrastive loss. The numerator of the contrastive loss optimizes for *alignment*, which tries to map jets and their augmented

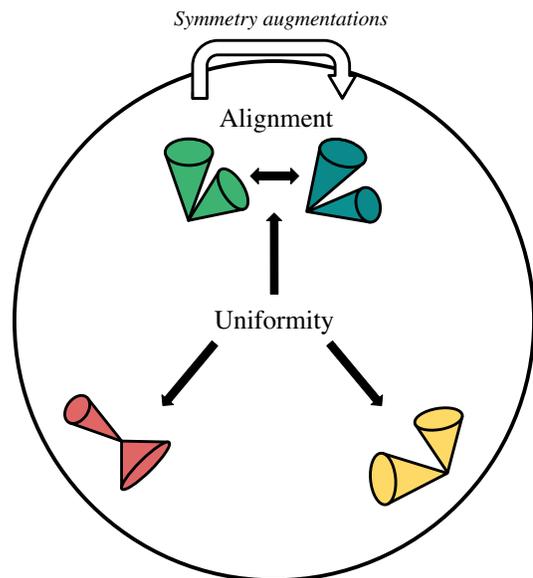


FIG. 2. An illustration of the latent space of the dijet events, built by a transformer-encoder network trained on the contrastive loss. The loss function optimizes for both alignment of dijets and their symmetry-augmented versions, and uniformity of physically distinct dijets.

versions to similar regions in the latent space. The denominator of the contrastive loss maximizes the *uniformity*, which tries to use up the entirety of the latent space when creating representations (see Fig. 2).

C. Event augmentations

We now outline the list of symmetry augmentations used to create physically equivalent latent space jets.

We define the following single-jet augmentations:

- (1) *Rotation*: each jet is randomly (and independently) rotated about its central axis in the $\eta - \phi$ plane. This is not an exact symmetry, but correlations between the radiation patterns of the two jets are negligible.
- (2) *Distortion*: each jet constituent is randomly shifted in the $\eta - \phi$ plane. The shift is drawn from a Gaussian of mean 0 and standard deviation $\sim 1/p_T$, where p_T is the transverse momentum of the constituent being shifted. This shift represents the smearing from detector effects.
- (3) *Collinear split*: a small number of the jet constituents (“mothers”) are split into two constituents (“daughters”) such that the daughters have η and ϕ equal to that of the mother, and the transverse momenta of the daughters sum to that of the mother.

We define the following event-wide augmentations:

- (1) *η -shift*: The dijet event is shifted in a random η direction.
- (2) *ϕ -shift*: The dijet event is shifted in a random ϕ direction.

Augmentations are applied to each training batch of the transformer. Each jet in the dijet event receives all three of the single-jet augmentations. The full event then receives both event-wide augmentations. See Fig. 3 for a visualization of the jet augmentations in the $\eta - \phi$ plane.

The jet augmentations are meant to not modify any of the important physical properties of the jets. As a test, we plot the jet masses of the hardest and second hardest jets from a subset of the LHC Olympics dataset in Fig. 4(a), as well as the nsubjettiness variables τ_{21} and τ_{32} in Figs. 4(b) and 4(c), both before and after receiving the jet augmentations and find no significant change in the distributions.

We also plot m_{JJ} for the dijet system in Fig. 5, again finding good agreement before and after the augmentations are applied. This confirms that our set of jet augmentations can be seen as true symmetry transformations of the dijet events.

D. Training procedure

We train the transformer-encoder network on a dataset of 50,000 background dijet events and up to 50,000 signal events, optimized on the contrastive loss in Eq. (1). The batch size is set to 400, which is the largest possible given the computing resources available. The network is trained with a learning rate of 0.0001, an early stopping parameter

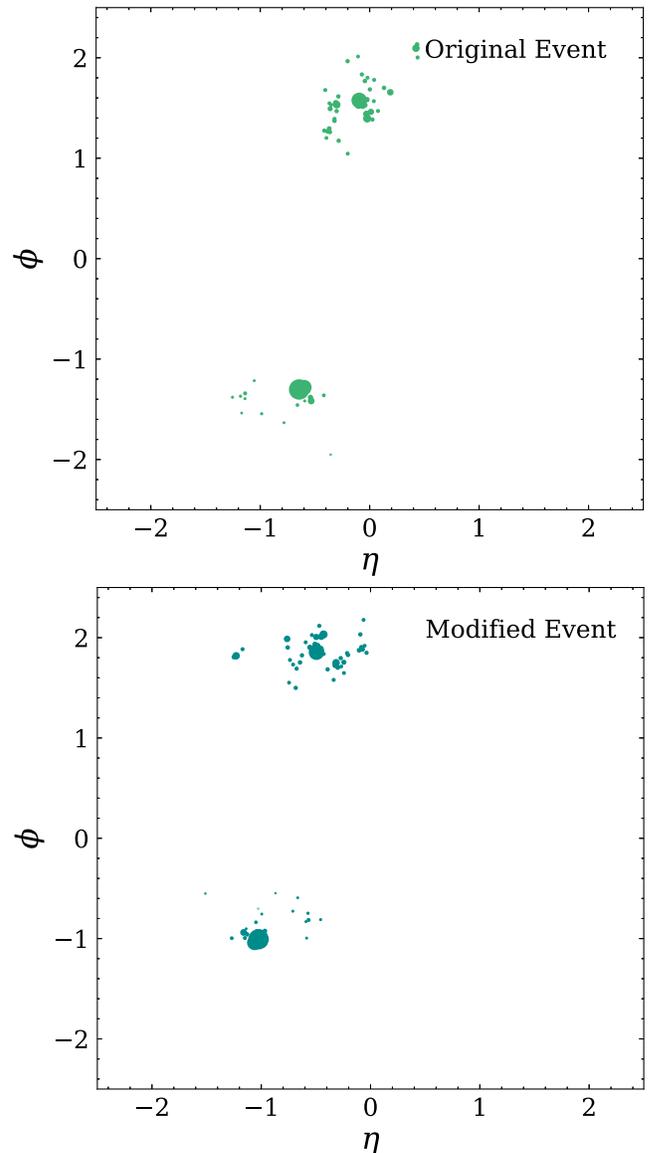


FIG. 3. A dijet event before (top) and after (bottom) receiving the set of single-jet and event-wide augmentations. Note that the upper and lower jets have visibly been rotated about their central axes, and the full event has been shifted in the upper-right direction of the $\eta - \phi$ plane.

of 20 epochs, and a temperature parameter τ of 0.1. All of these hyperparameters were empirically found to deliver the best transformer performance. The transformer-encoder network is implemented using PyTorch 1.10.0 [25] and optimized with ADAM [26]. Jet augmentations are applied batchwise, with each dijet event receiving a different randomized augmentation.

We also construct a binary classification dataset used to evaluate the latent space jet representations. This dataset consists of 85,000 signal and background dijet events each. We consider two types of binary classification tasks. Fully connected binary classifiers (FCN’s) are implemented in

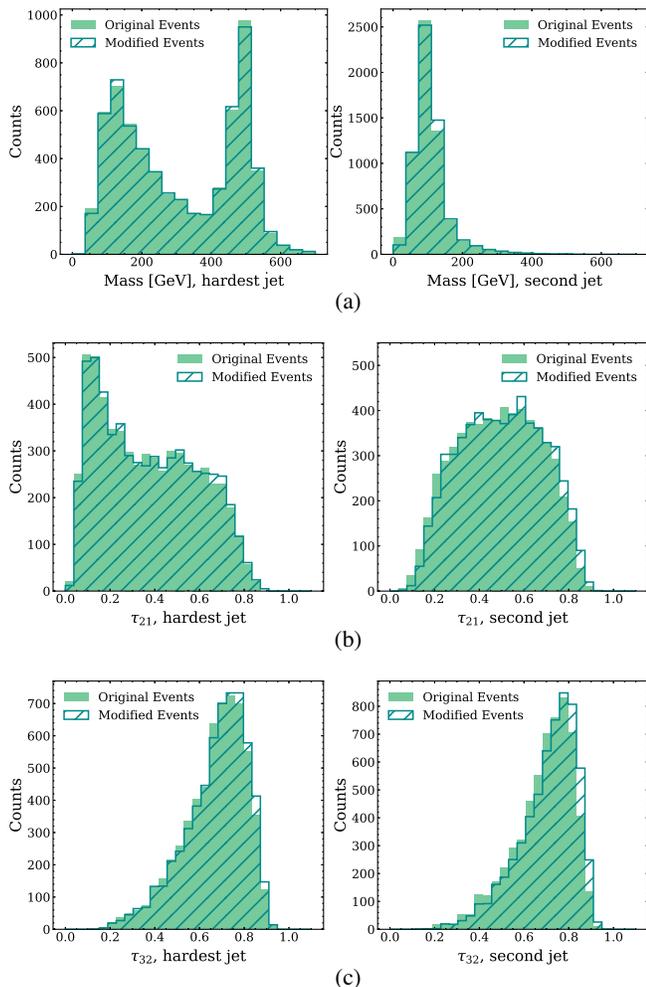


FIG. 4. Jet observable distributions for a sample of dijet events, before and after receiving the symmetry augmentations. (a) Jet mass distributions, (b) τ_{21} distributions, and (c) τ_{32} distributions.

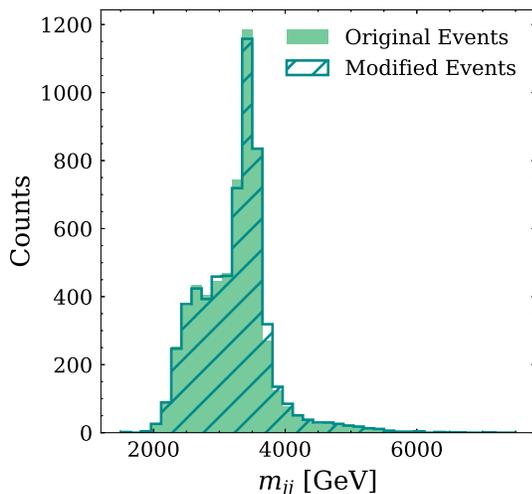


FIG. 5. Dijet mass (m_{JJ}) distributions for a sample of dijet events, before and after receiving the symmetry augmentations.

PyTorch and optimized with ADAM. These networks consist of three linear layers of sizes (64, 32, 1) with ReLU activation, a dropout of 0.1 between each layer, and a final sigmoid layer. The FCN is trained with a batch size of 400, a learning rate of 0.001, and an early stopping parameter of 5 epochs. Linear classifier tests (LCT's) are implemented in `scikit-learn` [27]. Both binary classification tasks discriminate signal from background in the latent space and have access to signal and background labels (i.e., are fully supervised).

We further define a “standard test” dataset consisting of 10,000 signal and background events each. There is no overlap of events in the standard test dataset with those in the transformer training or binary classifier training sets.

E. Anomaly detection

The usefulness of the latent space dijet representations is evaluated in a realistic model agnostic anomaly detection search setup. CWoLa (classification without labels) [15] is a weakly supervised training method that allows for signal versus background discrimination in cases where training samples of pure signal and background cannot be provided. Such a scenario might occur in resonant bump-hunting, where it is common to define signal and sideband regions, both of which will have a non-negligible fraction of background events [16,17]. The authors of Ref. [15] show that a classifier that is trained on two mixed samples (each with a different signal fraction) is in fact maximally discriminatory for classifying signal from background.

In Sec. IV, we run a CWoLa training procedure on the latent space representations. Our mixtures consist of one background-only sample and one sample with a suppressed signal fraction representing a mixture of background and a rare unknown anomaly. This is the ideal anomaly detection setup in which a sample of pure background can be generated (and is also the starting point of Refs. [28–30]). In practice, this may not be possible, and other methods must be used to obtain this dataset directly from data using sideband information [31–36].

III. EVALUATING THE LATENT SPACE REPRESENTATIONS

We evaluate the ability of our transformer-encoder network to faithfully translate dijet events into a latent space in the following way: we first train the network to embed event space particle collisions into a reduced-dimension latent space. We then perform a binary classification task on the latent space. We test the sensitivity of this setup to the amount of signal present in the training of the transformer as well as in the training of the classifier. The latter test demonstrates the anomaly detection capability of the approach.

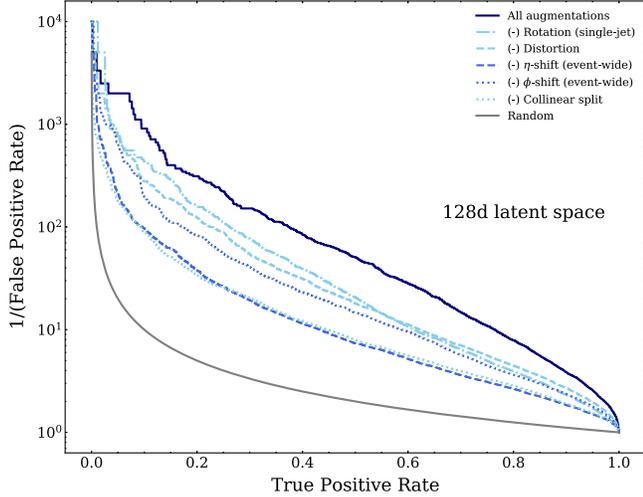


FIG. 6. Classifier efficiency curves for a fully connected binary classifier (FCN) run on latent space dijet representations trained with all except the indicated augmentation. All of the five augmentations appear to contribute significantly to the transformer-encoder network performance.

A. Quantifying the effect of each augmentation

As a first study, we explore the importance of each of the five symmetry augmentations outlined in Sec. II C. In Fig. 6, we plot the rejection ($1/\text{false positive rate}$) versus the true positive rate for a FCN trained on the latent space dijet representations. Each curve in the figure represents a transformer-encoder network trained with all of the symmetry augmentations *except* the one indicated. The transformer-encoder network is trained on 50,000 signal and 50,000 background dijet events, and the dimension of the transformer latent space is held at 128 dimensions.

In general, the performance of the transformer-encoder (as quantified by the receiver operating characteristic (ROC) area under the curve (AUC) for the FCN) drops sizably if any of the symmetry augmentations is not used during the transformer-encoder network training. The worst-performing transformers are those that do not receive the event-wide η -shift or collinear split augmentations.

TABLE I. ROC AUC and $\max(\text{SIC})$ scores for a binary classifier trained to discriminate signal from background on the LHC Olympics dataset. The performance scores decrease sizably if any of the five symmetry augmentations are not used when training the transformer-encoder network.

Removed augmentation	AUC	$\max(\text{SIC})$
(None)	0.918	3.805
Distortion	0.872	2.376
Rotation (single-jet)	0.860	2.677
ϕ -shift (event-wide)	0.851	1.949
Collinear split	0.800	1.430
η -shift (event-wide)	0.791	1.387

However, the decrease in AUC is significant for every removed augmentation. It is likely that the addition of symmetry augmentations would lead to further improvement in the transformer performance.

The AUC scores for the FCN's trained on latent space representations are given in Table I. We additionally include as a performance metric the maximum of the significance improvement characteristic ($\max(\text{SIC})$), defined as $\max(\frac{\text{true positive rate}}{\sqrt{\text{false positive rate}}})$. The $\max(\text{SIC})$ can be seen as the multiplicative factor by which signal significance improves after performing a well-motivated cut on the dataset.

B. Exploring the dimensionality of the latent space

We next gauge how the size of the latent space affects the usefulness of the representations. The latent space jet representations would ideally be lower in dimension than the physical space versions so as to save computational processing time by removing nonessential degrees of freedom from a given dataset. However, a latent space embedding with too few dimensions might not contain enough parameters to encode the essential physical dynamics of the jets.

In Fig. 7, we plot the rejection versus the true positive rate for a FCN trained on the latent space dijet representations. (See Fig. 10 in Appendix A for curves from a linear classifier.) We scan the latent space size in powers of two from 512 down to 8 dimensions. For all latent space dimensions, the transformer-encoder network is trained on 50,000 signal and 50,000 background dijet events.

The performance of the transformer-encoder improves as the dimension of the latent space increases. We find that a

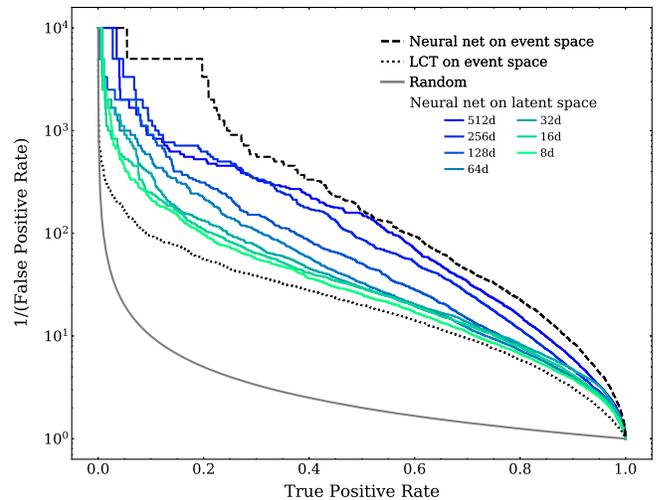


FIG. 7. Classifier efficiency curves for a FCN run on the latent space dijet representations. The performance of the binary classifier increases with the dimension of the latent space. For comparison, we also provide efficiency curves for a FCN and a linear classifier test (LCT) run on the event space dijets.

TABLE II. ROC AUC and max(SIC) scores for a binary classifier trained to discriminate signal from background on the LHC Olympics dataset. FCN = fully connected (dense) neural network; LCT = linear classifier test; Trans + BC = transformer architecture trained on the binary cross entropy loss. The particle space training dimension is (average number of nonzero entries) \pm (standard deviation of nonzero entries) per LHC Olympics event.

Training set	Training dimension	Classifier	AUC	max(SIC)
Particle space	506 \pm 174	FCN	0.958	15.401
		LCT	0.883	2.277
Latent space	8	FCN	0.904	2.542
		LCT	0.841	1.882
		Trans + BC	0.955	7.608
	64	FCN	0.915	3.163
		LCT	0.816	1.799
		Trans + BC	0.960	7.238
512	FCN	0.945	6.396	
	LCT	0.926	4.624	
	Trans + BC	0.968	13.862	

FCN trained on latent space jet representations cannot outperform a FCN trained on event space jet representations, but it can outperform a LCT trained on event space jet representations. Perhaps more striking is that the linear classifier trained on the compressed latent representations outperforms the linear classifier trained on the full event space data. This indicates that the self-supervised representations are highly expressive despite being compressed, and it agrees with the top-tagging results obtained in Ref. [13].

A selection of AUC scores for the FCN’s trained on latent space representations are given in Table II. The table also contains scores for a LCT trained on the latent space representations, as well as a binary classifier constructed from the transformer architecture with an additional sigmoid function as the final layer (Trans + BC), trained using the binary cross entropy loss. The Trans + BC network has access to all of the input particles and is not trained post-hoc on the self-supervised latent space, so we expect it to perform the best of all configurations. The hope is that the FCN performance is as close as possible to the performance of the Trans + BC (on the largest latent space).

Table II does show a performance gap between a FCN trained on the latent space and the Trans + BC network. In Appendix B, we evaluate the performance of both such networks when trained on increasing amounts of data. In both cases, we find that the networks are “data-hungry”; in other words, the classifier performances increase with the amount of training data, and the performances do *not* saturate when trained on the 85,000 signal and 85,000 background dijet events sampled from the LHC Olympics dataset. Therefore the performance of the FCN could likely reach

that of the Trans + BC network with a larger training dataset than what was used in this study.

C. Varying the amount of training signal

In practice, we want to use the transformer-encoder network for model-agnostic anomaly detection. In this case, we would not be able to train the transformer on a known signal fraction, as the training data would contain an unknown (and extremely tiny, if any) percentage of beyond the standard model (BSM) signal. It is therefore useful to see if the transformer-encoder network is effective at translating rare events into a latent space. We might expect this to be true if the transformer is learning only generic features about collider events that hold for both signal and

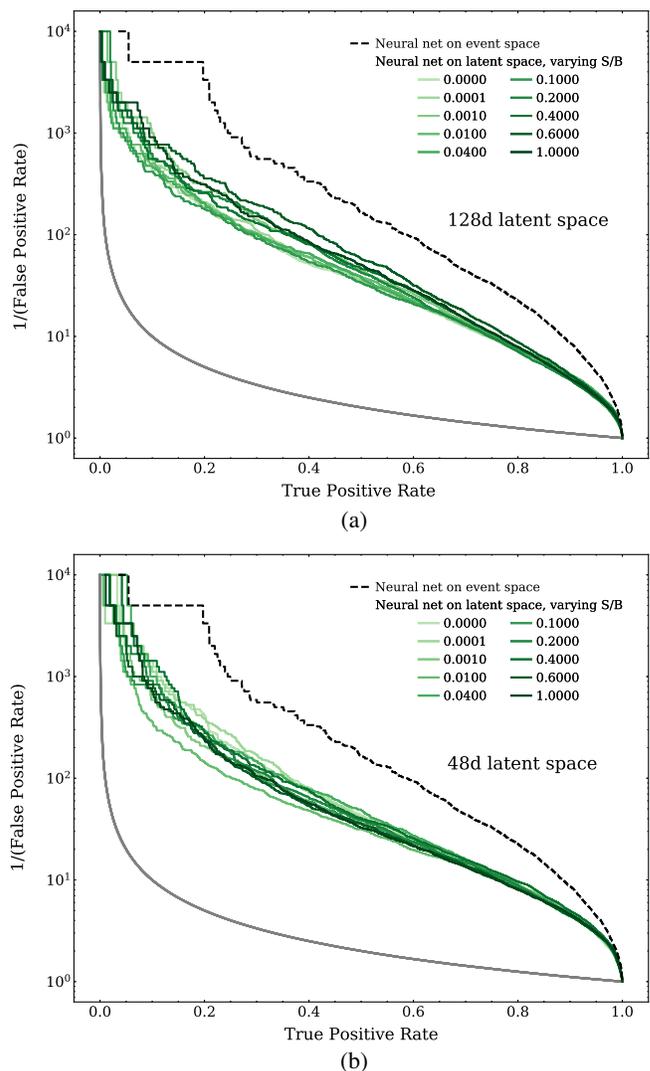


FIG. 8. Classifier efficiency curves for a latent space FCN classifier trained on varying amounts of signal fraction. The classifier performance is robust with respect to the signal fraction of the training data. This implies that the transformer-encoder network can be trained on background alone. (a) 128d latent space and (b) 48d latent space.

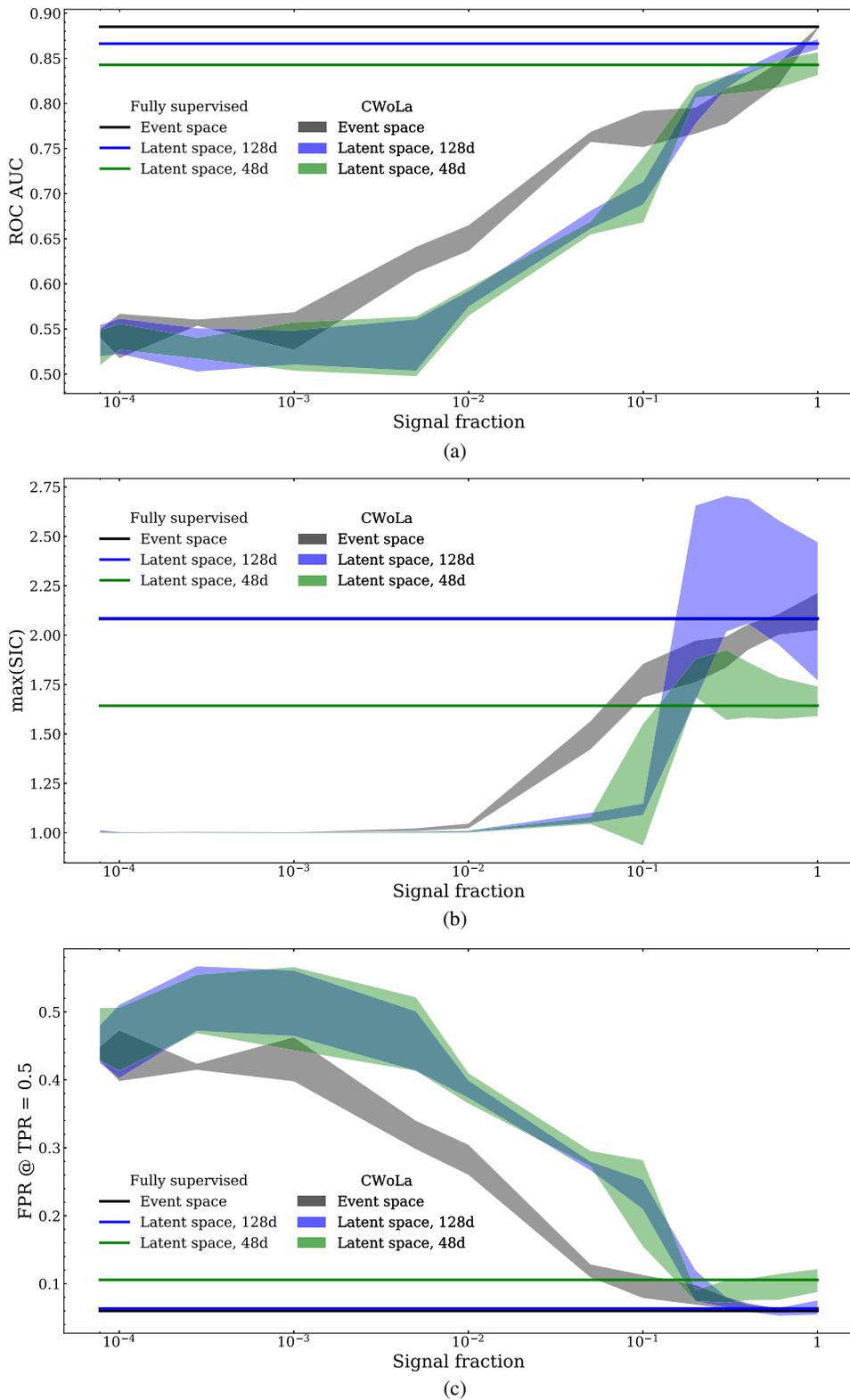


FIG. 9. Various metrics for evaluating a CWoLa weakly-supervised classifier trained to discriminate a background-only sample from a signal + background sample with a variable signal fraction. The classifier run on the event space representations slightly outperforms one run on the latent space representations, especially at low signal fractions. (a) ROC AUC training metric, (b) max(SIC) training metric, and (c) False positive rate at a true positive rate of 50%.

background events. This is encouraged by the universality of the symmetry augmentations in the contrastive loss.

In Fig. 8(a), we hold the dimension of the transformer latent space fixed at 128, then scan the signal to background ratio S/B down from 1.0 to 0.0. In Fig. 8(b), we repeat the previous steps for a transformer latent space of dimension 48. (See Fig. 11 in Appendix A for curves from a linear classifier.) Note that for this study, the transformer-encoder network is always trained on 50,000 background dijet events, but the number of signal dijet events changes with the signal S/B ratio. We find that the classifier performance is robust with respect to the signal to background ratio, as was found in Ref. [13]. This demonstrates that the transformer-encoder network can be trained on background alone and still faithfully model rare signal events.

IV. ANOMALY DETECTION

We now test the usefulness of the latent space jet representations in a more practical setting by performing a CWoLa-style anomaly search. To create the latent space representations, we use a transformer-encoder network trained on a background-only sample of 50,000 dijet events. As before, we use the “standard test” dataset of 10,000 signal and 10,000 background events for all binary classifier tests.

We first create a baseline against which to compare the CWoLa analyses by carrying out a self-supervised binary classification task in the event space representation. For this study, we use 42,500 signal and 42,500 background dijet events.

For the anomaly-detection analysis, we set one CWoLa “mixed sample” to be a set of 42,500 background-only dijets (the same as in the self-supervised task). The other mixed sample is a mixture of 42,500 signal and background dijets, with the signal fraction scanned from 0% to 100%. We run the analysis three times, once for the event space dijets and once each for latent space dijets at 128 and 48 dimensions. The evaluation of the performance is always computed with pure signal and background labels. Comparisons of the CWoLa classifier performances are shown in Fig. 9. In Fig. 9(a), we use the ROC AUC as a metric for evaluating the CWoLa classifier; in Fig. 9(b), we use instead use the max(SIC) as the metric; in Fig. 9(c), we provide the false positive rate at a fixed true positive rate of 50%.

We find that the CWoLa weakly supervised classifier performance of a small dimensional latent space is comparable to (but cannot match) that of the full particle event space, with an improvement in performance for a larger dimension latent space as evaluated by the max(SIC) metric. The most notable difference between the classifier performance on latent space versus event space is that in the former case, the classifier performance diminishes to no better than random at a higher signal fraction for the training data. (This is indicated by the ROC AUC dropping to 0.5, the max(SIC) dropping to 1, and the

FPR @ TPR = 0.5 dropping to 0.5.) More specifically, the small-dimension latent space classifier hits random performance at a signal fraction of just below 10^{-2} , while the event space classifier does better than random at all nonzero signal fractions.

Overall, the classifier performances at anomaly-level signal fractions as shown in Fig. 9 are lower than what has been seen in other recent anomaly-detection methods on the LHCO data. In fact, the SIC curves shown in Refs. [31–36] are typically an order of magnitude greater than those in Fig. 9. However, such curves were constructed by training on standard jet observables (e.g. m_J , τ_{12}), and thus represent training methods that are inherently model-dependent. Evidently, this performance decrease from model-dependent anomaly detection methods is the price to pay on this particular signal for using a more widely-applicable, model-agnostic method.

There exist a number of avenues for future work to improve on this contrastive-learning trained classifier. For one: we have considered a small set of symmetry augmentations specific to dijet events. However, additional augmentations for dijet events could be added to the contrastive loss. Alternatively, a different selection of augmentations that leads to an even more general event representation could be chosen. As another avenue: we mentioned earlier (and illustrate in Appendix B) that the transformer-encoder network is data-hungry. It would therefore be reasonable to expect an improvement in classifier performance if the training dataset were made larger.

V. CONCLUSIONS

In this paper, we have used transformer-encoder neural networks to embed entire collider events into low and fixed-dimensional latent spaces. This embedding is constructed using self-supervised learning based on symmetry transformations. Events that are related by symmetry transformations are grouped together in the latent space while other pairs of events are spread out in the latent space.

We have shown that the latent space preserves the essential properties of the events for distinguishing certain BSM events from the SM background. This latent space can then be used for a variety of tasks, including anomaly detection. We have shown that anomalies can still be identified in the reduced representation as long as there is enough signal in the dataset. For the particular signal model studied, the required amount of signal is much higher than reported by other studies using high-level features. This illustrates the tradeoff between signal sensitivity and model specificity. Our reduced latent space knows nothing of particular BSM models and is thus broadly useful but not particularly sensitive. Future work that explores the continuum of approaches by adding more augmentations to the contrastive learning may result in superior performance for particular models in the future.

VI. CODE AVAILABILITY

The code can be found at [37].

ACKNOWLEDGMENTS

We thank Jernej Kamenik and David Shih for useful feedback on the manuscript. B. M. D was supported by a Postdoctoral Research Fellowship from the Alexander von Humboldt Foundation. B. N. and R. M. were supported by the Department of Energy, Office of Science under Contract No. DE-AC02-05CH11231. This material is based upon work supported by the National Science

Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

APPENDIX A: EVALUATING THE LATENT SPACE WITH A LINEAR CLASSIFIER TEST

In this section, we provide the analogs to Figs. 7 and 8 with the transformer-encoder efficiency curves calculated for a binary linear classifier test run on the latent space representations (rather than a binary FCN). This aligns with the field-standard way to evaluate representations of jets, through a LCT. However, these plots (Figs. 10 and 11) are not shown in the main text of this report as a realistic anomaly detection analysis would be carried out using fully connected networks.

APPENDIX B: HOW DATA-HUNGRY ARE THE NEURAL NETWORKS?

In this section, we provide plots illustrating the data-hungry nature of the transformer-encoder network. The performance of the binary classifier trained on the latent space dijet representations, as shown in Fig. 9, is admittedly low. However, it is likely that the performance could improve if the classifiers were trained on a larger amount of data.

In Fig. 12(a), we train a FCN on a varying fraction of the available dijet dataset (a 100% training fraction makes use of all 85,000 signal and 85,000 background events). In Fig. 12(b), we repeat this procedure for a Trans + BC network. In both cases, the ROC AUCs of the trained binary classifiers do not appear to be saturated when trained on the full dataset.

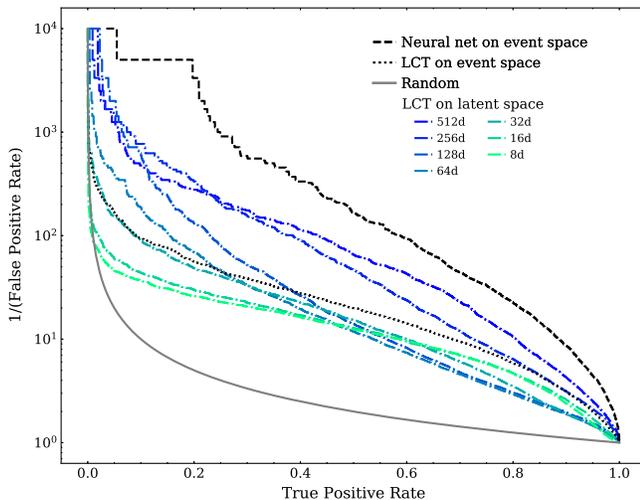
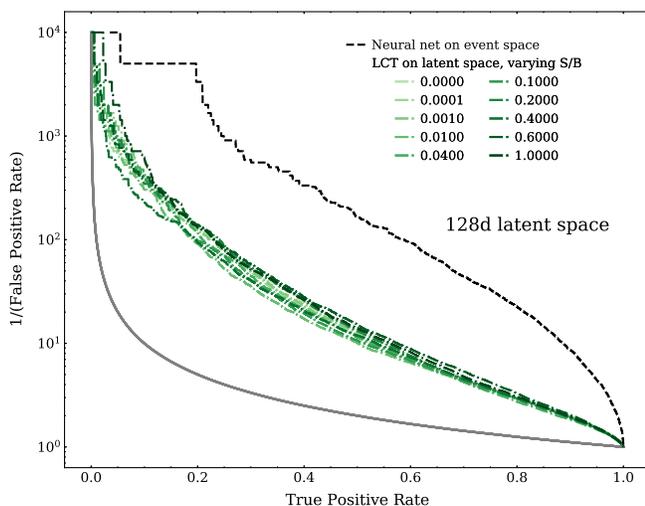
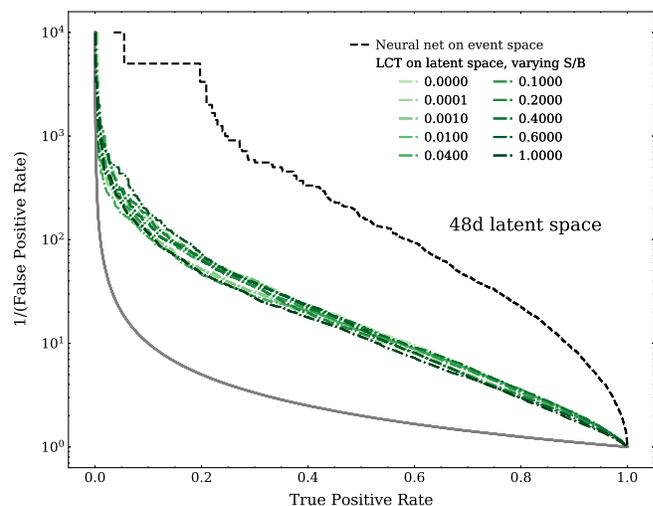


FIG. 10. Classifier efficiency curves for a linear classifier test (LCT) run on the latent space dijet representations. For comparison, we also provide efficiency curves for a FCN and a LCT run on the event space dijets.



(a)



(b)

FIG. 11. Classifier efficiency curves for a latent space linear classifier test trained on varying amounts of signal fraction. The classifier performance is again robust with respect to the signal fraction of the training data. (a) 128d latent space and (b) 48d latent space.

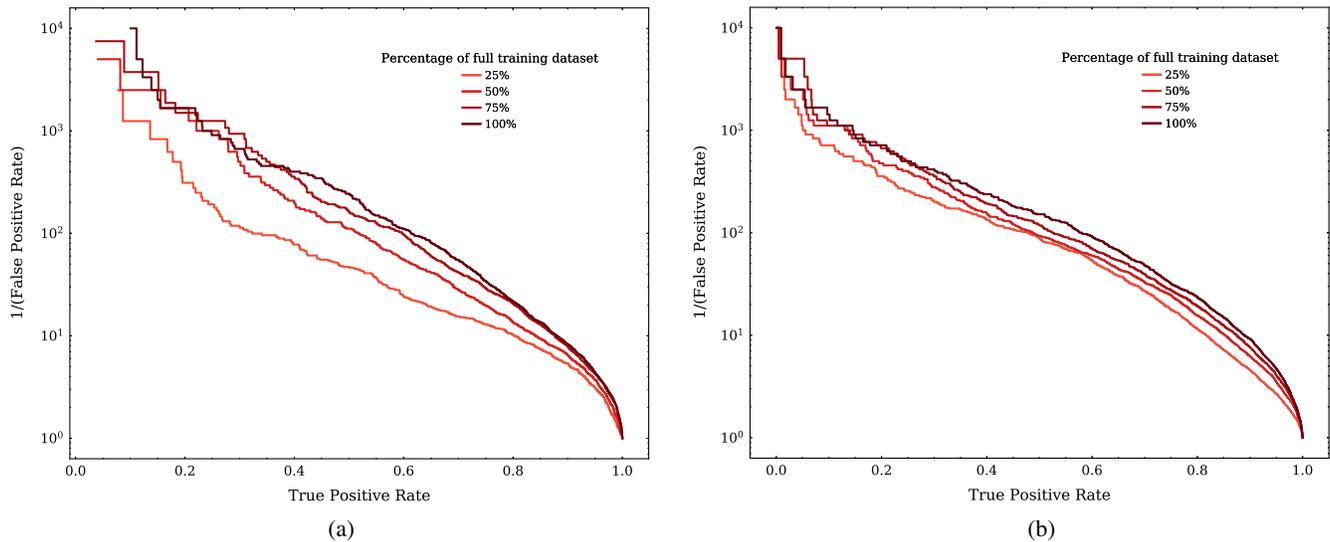


FIG. 12. Classifier efficiency curves for a binary classifier trained on an increasing percentage of the full dijet dataset. (a) FCN network architecture. The network performance shows only a mild degree of saturation and (b) Trans + BC network architecture. The network performance steadily increases with amount of training data, showing no signs of saturation.

- [1] Nathaniel Craig, Patrick Draper, Kyoungchul Kong, Yvonne Ng, and Daniel Whiteson, The unexplored landscape of two-body resonances, *Acta Phys. Pol. B* **50**, 837 (2019).
- [2] Jeong Han Kim, Kyoungchul Kong, Benjamin Nachman, and Daniel Whiteson, The motivation and status of two-body resonance decays after the LHC Run 2 and beyond, *J. High Energy Phys.* **04** (2020) 030.
- [3] Gregor Kasieczka *et al.*, The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics, *Rep. Prog. Phys.* **84**, 124201 (2021).
- [4] T. Aarrestad *et al.*, The dark machines anomaly score challenge: Benchmark data and model independent event classification for the large hadron collider, *SciPost Phys.* **12**, 043 (2022).
- [5] Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih, Machine learning in the search for new fundamental physics, [arXiv:2112.03769](https://arxiv.org/abs/2112.03769).
- [6] Matthew Feickert and Benjamin Nachman, A living review of machine learning for particle physics, [arXiv:2102.02770](https://arxiv.org/abs/2102.02770).
- [7] ATLAS Collaboration, Dijet Resonance Search with Weak Supervision Using 13 TeV pp Collisions in the ATLAS Detector, *Phys. Rev. Lett.* **125**, 131801 (2020).
- [8] Jan Hajer, Ying-Ying Li, Tao Liu, and He Wang, Novelty detection meets collider physics, *Phys. Rev. D* **101**, 076015 (2020).
- [9] Marco Farina, Yuichiro Nakai, and David Shih, Searching for new physics with deep autoencoders, *Phys. Rev. D* **101**, 075021 (2020).
- [10] Theo Heimel, Gregor Kasieczka, Tilman Plehn, and Jennifer M. Thompson, QCD or what?, *SciPost Phys.* **6**, 030 (2019).
- [11] Blaž Bortolato, Barry M. Dillon, Jernej F. Kamenik, and Aleks Smolkovič, Bump hunting in latent space, *Phys. Rev. D* **105**, 115009 (2022).
- [12] Md Abul Hayat, George Stein, Peter Harrington, Zarija Luki, and Mustafa Mustafa, Self-supervised representation learning for astronomical images, *Astrophys. J. Lett.* **911**, L33 (2021).
- [13] Barry M. Dillon, Gregor Kasieczka, Hans Olschlagler, Tilman Plehn, Peter Sorrenson, and Lorenz Vogel, Symmetries, safety, and self-supervision, *SciPost Phys.* **12**, 188 (2022).
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Deep residual learning for image recognition (2015), [arXiv:1512.03385](https://arxiv.org/abs/1512.03385).
- [15] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler, Classification without labels: Learning from mixed samples in high energy physics, *J. High Energy Phys.* **10** (2017) 174.
- [16] Jack H. Collins, Kiel Howe, and Benjamin Nachman, Anomaly Detection for Resonant New Physics with Machine Learning, *Phys. Rev. Lett.* **121**, 241803 (2018).
- [17] Jack H. Collins, Kiel Howe, and Benjamin Nachman, Extending the search for new resonances with machine learning, *Phys. Rev. D* **99**, 014038 (2019).
- [18] Gregor Kasieczka, Benjamin Nachman, and David Shih, Official datasets for LHC olympics 2020 anomaly detection challenge (Version v6), [10.5281/zenodo.4536624](https://zenodo.org/record/4536624) (2019).

- [19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, A simple framework for contrastive learning of visual representations, [arXiv:2002.05709](https://arxiv.org/abs/2002.05709).
- [20] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands, PYTHIA 6.4 physics and manual, *J. High Energy Phys.* **05** (2006) 026.
- [21] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands, An introduction to PYTHIA 8.2, *Comput. Phys. Commun.* **191**, 159 (2015).
- [22] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [23] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez, FastJet user manual, *Eur. Phys. J. C* **72**, 1896 (2012).
- [24] Matteo Cacciari and Gavin P. Salam, Dispelling the N^3 myth for the k_t jet-finder, *Phys. Lett. B* **641**, 57 (2006).
- [25] Adam Paszke *et al.*, Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., Vancouver, Canada, 2019), pp. 8024–8035.
- [26] Diederik P. Kingma and Jimmy Ba, Adam: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [27] Fabian Pedregosa *et al.*, Scikit-learn: Machine learning in python, [arXiv:1201.0490](https://arxiv.org/abs/1201.0490).
- [28] Raffaele Tito D’Agnolo and Andrea Wulzer, Learning new physics from a machine, *Phys. Rev. D* **99**, 015014 (2019).
- [29] Raffaele Tito D’Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti, Learning multivariate new physics, *Eur. Phys. J. C* **81**, 89 (2021).
- [30] Raffaele Tito d’Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti, Learning new physics from an imperfect machine, *Eur. Phys. J. C* **82**, 275 (2022).
- [31] Benjamin Nachman and David Shih, Anomaly detection with density estimation, *Phys. Rev. D* **101**, 075042 (2020).
- [32] Anders Andreassen, Benjamin Nachman, and David Shih, Simulation assisted likelihood-free anomaly detection, *Phys. Rev. D* **101**, 095004 (2020).
- [33] George Stein, Uros Seljak, and Biwei Dai, Unsupervised in-distribution anomaly detection of new physics through conditional density estimation, in *Proceeding of the 34th Conference on Neural Information Processing Systems* (2020), [arXiv:2012.11638](https://arxiv.org/abs/2012.11638).
- [34] Kees Benkendorfer, Luc Le Pottier, and Benjamin Nachman, Simulation-assisted decorrelation for resonant anomaly detection, *Phys. Rev. D* **104**, 035003 (2021).
- [35] Anna Hallin, Joshua Isaacson, Gregor Kasieczka, Claudius Krause, Benjamin Nachman, Tobias Quadfasel, Matthias Schlaffer, David Shih, and Manuel Sommerhalder, Classifying anomalies through outer density estimation, *Phys. Rev. D* **106**, 055006 (2022).
- [36] John Andrew Raine, Samuel Klein, Debajyoti Sengupta, and Tobias Golling, CURTAINS for your sliding window: Constructing unobserved regions by transforming adjacent intervals, [arXiv:2203.09470](https://arxiv.org/abs/2203.09470).
- [37] https://github.com/rmastand/JetCLR_AD.