Semi-visible jets, energy-based models, and self-supervision

Luigi Favaro[1], Michael Krämer[2], Tanmoy Modak[1], Tilman Plehn[1], and Jan Rüschkamp[1]

1 Institut für Theoretische Physik, Universität Heidelberg, Germany
2 Institute for Theoretical Particle Physics and Cosmology, RWTH Aachen, University, Germany

December 14, 2023

Abstract

We present DarkCLR, a novel framework for detecting semi-visible jets at the LHC. DarkCLR uses a self-supervised contrastive-learning approach to create observables that are approximately invariant under relevant transformations. We use background-enhanced data to create a sensitive representation and evaluate the representations using a normalized autoencoder as a density estimator. Our results show a remarkable sensitivity for a wide range of semi-visible jets and are more robust than a supervised classifier trained on a specific signal.

Contents

---

1 Introduction

Model agnostic searches are of paramount importance for the current and future LHC physics program. The independence from signal hypothesis allows this approach to extend the coverage of possible new physics scenarios. Machine learning can provide a unique platform for this strategy by providing access to high-dimensional correlations and low-level data modeling.

The well-established approaches for model agnostic searches through anomaly detection are based on scores from density estimates or classification in semi-supervised settings between background and signal-enriched regions, see [1] for a recent review and [2] for an up-to-date list of relevant references.

Density-based scores select anomalies by identifying low-density regions of the data. Early research used the reconstruction error of an auto-encoder as a proxy for density [3,4]. In recent years [5–20], this approach has been continuously refined with better density estimates, such as in the normalized autoencoder (NAE) [21,22], normalizing flow techniques [17,23–25], energy flow polynomials [26], background estimation with ABCD methods [27], and network interpretability [28]. Anomaly detection through density estimation and semi-supervised learning has already been applied in recent ATLAS analyses [29,30]. More details on the different methods and architectures can be found in recent white papers [31,32].

However, the definition of an anomaly based on low-density regions of the data is not invariant under coordinate transformations [17,33]. Therefore, each step in the preprocessing chain can change what are considered inliers and outliers. To remedy this problem, we propose a framework for constructing a representation space suitable for anomaly detection in jet physics. We avoid the use of hand-crafted transformations of the data by creating observables based on physical invariances and a few assumptions about the signal hypothesis.

We develop our framework within a self-supervised contrastive learning representation (CLR) method CLR [34]. Self-supervision provides a unique way to detect anomalous objects in high-dimensional data. We generate "pseudo-labels" derived from the data, allowing the optimization of neural networks without relying on ground truth labels. This approach, similar to contrastive learning, can establish connections between original and augmented events, facilitating the discovery of novel phenomena. Learning invariances to transformation with contrastive learning has already been shown to be powerful in JetCLR [35], AnomalyCLR [36], and resonant anomaly detection [37]. The latter introduces "anomalous" augmentations for anomaly detection applications on reconstructed high-level objects. These augmentations intentionally introduce variations in event kinematics that may resemble features found in anomalous events. Their definition follows general features of a new physics scenario and preserves the model agnostic aspect of an unsupervised anomaly detection tool.

In this work, we apply the concept of anomalous enhancements to the detection of semi-visible jets [38–44]. Semi-visible jets arise in models of strongly interacting dark sectors, which in turn belong to the general class of Hidden Valley models [45–47]. Distinguishing such semi-visible jets from large QCD backgrounds is difficult and represents a major challenge for jet classification. We call our framework DarkCLR, an extended representation space for studying and finding semi-visible jets within LHC jets. We show that the latent space learned by DarkCLR provides informative representations of semivisible jets for downstream tasks. We propose two scores for anomaly detection: an anomaly score defined in the representation space, and the reconstruction error of a normalized autoencoder trained on the representations.

Our paper is organized as follows. We describe the background data and signal benchmarks in Sec. 2. Then, Sec. 3 introduces DarkCLR, the network architecture, and the physical and anomalous extensions. We present the anomaly scores in Sec. 4, and finally, we look at the tagging performance in Sec. 5, examining the discriminative power of the representations, the robustness of the anomaly scores, and the dependence on the main training hyperparameters.

2 Dark jets

Jets are a prevalent signature of several new physics models, such as Hidden Valley models, which can lead to tantalizing semi-visible jet signatures at the LHC. In this work, we are interested in Hidden Valley models that consist of a strongly coupled dark sector with dark quarks coupled to the SM through a vector mediator. As a result, jets can be produced by the dark quarks from the decay of the vector mediator. The shower in this case would involve radiation into the dark sector, resulting in jets that are called semi-visible or dark jets, depending on the phenomenology of the signal.

2.1 Datasets

For our purposes, we consider a benchmark signal scenario with an underlying dark sector as introduced in [15, 17, 43]:

$$pp \to Z' \to q_d \bar{q}_d, \quad \text{with} \quad m_{Z'} = 2\,\text{TeV} \quad \text{and} \quad q_d = 500\,\text{MeV}, \tag{1}$$

where Z' is the mediator between the dark sector and the SM quarks, charged under a $U(1)'$ gauge group, and q_d is a dark quark charged under a dark $SU(3)_d$. The dark sector hadronizes to dark pions ($\pi_d = 4$ GeV) and dark rho mesons ($\rho_d = 5$ GeV). The neutral dark rho mesons mix with the Z' and can thus decay into SM quarks. The other dark mesons are stable and escape detection. In our benchmark scenario the fraction of invisible particles in a shower is given by $r_{\text{inv}} = 0.75$ [15, 43]. This dark sector model then leads to semi-visible jets and can be simulated with the Pythia Hidden Valley module [48, 49]. We will refer to this benchmark scenario as the "Aachen" dataset in the remainder of the paper.

The dataset is generated using Madgraph5 [50] for the hard process. The generated events are then interfaced with Pythia 8.2 [51] for showering and hadronization and finally fed to Delphes 3 for fast detector simulation [52]. The jets are reconstructed using the anti-k_T algorithm [53] with radius parameter $R = 0.8$ in FastJet and satisfy:

$$p_T^j = 150...300 \text{ GeV} \quad \text{and} \quad |\eta^j| < 2. \tag{2}$$

The most important phenomenological parameters for Hidden Valley models are the invisible fraction of the constituents, r_{inv}, and the mass of the dark mesons, $m_{\pi/\rho}$. To test the model dependence of our approach, we generate several data sets with the following parameter choices: starting from our benchmark signal, we first vary only the mass of the dark mesons and the confinement scale Λ as $m_{\pi_d} = m_{\rho_d} = \Lambda = 10\,\text{GeV}, 20\,\text{GeV}$. In addition, for our default choice of dark meson masses, we change the invisible fraction r_{inv} by allowing all dark mesons to decay back to SM quarks with a given probability. To explore the region where the number of visible jet constituents is closer to the QCD background, we reduce the invisible fraction to $r_{\text{inv}} = 0.5, 0.2$. The light QCD background is generated from leading order di-jet events.

The selection of the jets at detector level is done by calculating the ΔR between the reconstructed fat jets and the dark quarks at parton level and ensuring that $\Delta R < 0.8$. On the selected fat jets we apply the kinematic selection in p_T and η.

3 DarkCLR

3.1 Contrastive Learning Representation

Contrastive Learning of Representations (CLR) is a method for learning representations of the training data in high-dimensional spaces. These representations can then be used for any

downstream task, from classification to unsupervised learning. CLR falls into the category of self-supervised learning, i.e. it does not require "truth" labels of the training data.

In CLR, a function $f(\cdot)$ maps from the data space $\mathcal{D}$ to a representation space $\mathcal{R}$, where the function is optimized to solve an auxiliary task for which we define pseudo-labels. In this work, we focus on performing anomaly detection on the representations. Therefore, the function that performs the mapping from $\mathcal{D}$ to $\mathcal{R}$ is trained only on background data. Since collider events or objects such as jets typically consist of unordered sets of particles, we opt for a permutation invariant architecture. Specifically, we use a transformer encoder network to learn the mapping.

To overcome the lack of signal in our training data and to keep the approach model agnostic, we use only augmentations of the background data. These augmentations are used to define two types of pseudo-labels:

- **Positive-pair:** x_i, x_i'. This pair is constructed from a data point and an augmented version of itself via a positive augmentation;
- **Anomaly-pair:** x_i, x_i^*. This pair is constructed from a data point and an augmented version of itself via an anomalous augmentations.

Once we have defined the pseudo-labels, we minimize the following loss function [36]:

$$\mathcal{L}^{+}_{\text{AnomCLR}} = -\log\, e^{\left(s(z_i,z_i')-s(z_i,z_i*)\right)} = s(z_i,z_i^*) - s(z_i,z_i'), \tag{3}$$

where $z_i = f(x_i)$, $z_i' = f(x_i')$, $z_i^* = f(x_i^*)$ and $s(\cdot,\cdot)$ is the cosine similarity, a measure of proximity between points in a compact $\mathbb{S}^{d-1}$ representation space. The function $f(\cdot)$ then maps the raw data into the representation space such that positive pairs are close in $\mathcal{R}$ while anomaluous pairs are pushed apart. The first objective is commonly known as alignment and the latter one ensures separation between objects in the anomalous pair. While the term $s(z_i, z_i')$ ensures the alignment, i.e. different objects are mapped onto the same point in the compact latent space, the term $s(z_i, z_i^*)$ maximizes the distance between anomalous pairs while keeping the representation space informative about the anomalous augmentations. The chosen transformations are intended to be alterations of the original data that preserve the fundamental physics, such as the symmetries of the system. More details about the applied augmentations are given in the next section.

Note that $\mathcal{L}^{+}_{\text{AnomCLR}}$ is a modified version of the original CLR loss function [36] and has two special features that we can exploit. First, it contains only the invariances we want to impose and the anomalous features we want to distinguish from the background. Therefore, the representation space will be approximately invariant to the symmetries of the data we require during training, and it will be exposed to potential new physics signals through the anomalous augmentations. Second, as shown in Eq. (3), the loss function scales as N_{batch}, as opposed to the N^2_{batch} scaling of the original CLR loss function [36], and is therefore less computationally expensive. Although the partial removal of the uniformity requirement could potentially lead to a collapse of the representation space to a single point, this is not observed in our numerical analysis. We suggest that the large variety in the training data combined with the use of multiple augmentations prevents mode collapse and information loss.

3.2 Augmentations

Here we discuss the augmentations we use during training. We start with the positive (or, synonymously, physical) augmentations. These are easy to implement approximate symmetries of a jet:

- **Rotations:** We rotate each jet in $\eta - \phi$ by an angle which is chosen randomly between $[0, 2\pi]$. Note that the angle is chosen randomly for each jet, i.e., each constituent inside a jet is rotated by the same angle.

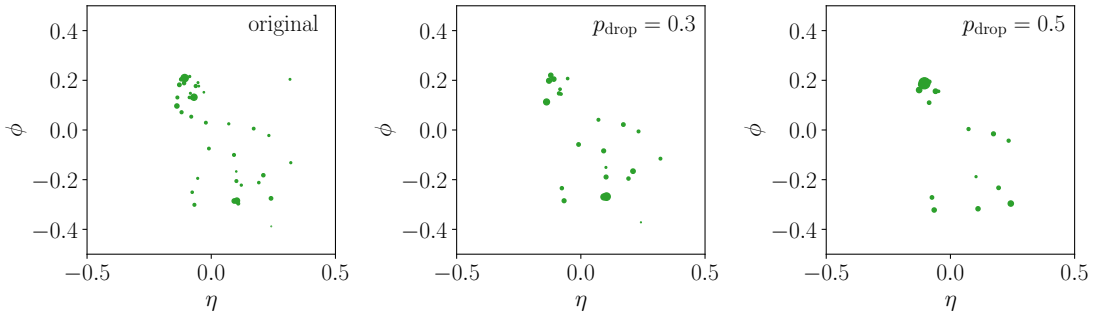

Figure 1: Example of an anomalous transformation on a QCD jet. The left panel shows the original background jet while the middle and right panels show the same jet after applying the augmentations with $p_{\text{drop}} = 0.3$ and $p_{\text{drop}} = 0.5$ respectively.

- **Translations:** We shift each constituent in the $\eta - \phi$ plane by randomly choosing a shift in a window with size given by the distance between the two furthest constituents.

After applying these two transformations to the original jet x_i, we obtain the augmented version x_i' and the positive pair $\{x_i, x_i'\}$.

Semi-visible jets, as discussed earlier, have fewer constituents than QCD jets. Therefore, we consider the dropping of constituents as a **anomaly augmentation**. The transformation is implemented as follows: We drop each component of the jet with a fixed probability p_{drop}, and the p_T of the augmented jet is rescaled to match the original p_T.

Fig. 1 shows an example transformation of a QCD jet used during training with $p_{\text{drop}} = 0.3$ and $p_{\text{drop}} = 0.5$.

3.3 Network architecture

As the first step of the CLR network, an embedding layer maps each constituent, consisting of $[p_T, \eta, \phi]$, to a larger vector with 128 dimensions. The input vector has a fixed size of 50 constituents, where we zero-pad jets with less number of constituents. We ensure that the zero padding does not affect the transformer by masking the zero p_T entries. This stops the propagation of information from zero value constituents, more details on the implementations are provided in [35]. The embedded constituents are then passed through the transformer encoder with a feed-forward network between each transformer layer. The output of the encoding has a dimension of $[\alpha, \text{model dimension}]$, where α is the number of constituents per jet. As a crucial next step, this output is now summed over α to induce permutation symmetry between the constituents. Finally, this summed output is passed to a final fully connected head network. The output of the head network then serves as the representation and input to the contrastive loss function of Eq. 3. Unless otherwise noted, the set of parameters used to train the transformer network is summarized in Tab. 1.

4 Anomaly scores

CLR anomaly score We study the effect of the CLR transformation by analyzing the CLR embedding space. We find that although the representation before the head network is more informative, the output of the head network encodes useful information for out-of-distribution detection. We first note that one way to reduce the loss is to simply increase the length of the vector so that jets with different properties are separated in the non-normalized space and

Hyper-parameter	Value
Model (embedding) dimension	128
Feed-forward hidden dimension	512
Output dimension	512
# self-attention heads	4
# transformer layers (N)	4
# head architecture layers	2
Dropout rate	0.1
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	5×10^{-5}
Batch size	256
# constituents (α)	50
# jets	100k
# epochs	150

Table 1: Default configuration of the transformer encoder and the training process.

close to each other after projection. Therefore, we expect the norm of the representation vector to be a discriminative variable and propose it as a CLR-based anomaly score that can show the effect of DarkCLR. Namely:

$$s_{\text{CLR}} = ||z||_{L_2}, \qquad z \in \mathbb{R}^D, \tag{4}$$

where D is the embedding dimension.

Before using this anomaly score, a small modification is needed. Since our loss Eq. 3 is norm-free, the ordering between background and signal norms is not a priori fixed. This ambiguity, which can spoil applications in anomaly detection, is resolved by introducing a regularization term which penalizes background representations with large norms. This ensures that anomaly detection associates high norm with outlier data. The implementation is done by adding to the loss function the L_2 norm of the representations of the background batch. We find empirically that this new term does not affect the similarity, and therefore the loss, of the training.

NAE The second anomaly score we consider is the reconstruction error of an autoencoder. In an autoencoder we define an unsupervised learning task by constructing an encoder and a decoder network trained only on the background data. The compression that takes place in the encoder forces the network to learn the manifold of the dataset in a latent space from which the decoder has to reconstruct the original input. This is achieved by minimizing the reconstruction error of the input, where we follow the standard practice of using the mean squared error as a measure of the reconstruction quality. After training, we can use the same quantity as an anomaly score, since off-manifold events are not reconstructed by the decoder, and thus give a large reconstruction error.

A normalized autoencoder (NAE) [21, 22] promotes classical auto-encoding training to an energy-based model by fixing the energy function to be the reconstruction error of the network. A NAE shares the same structure of a standard AE with the added robustness of Maximum Likelihood Estimate (MLE) training. The underlying probability distribution is a Boltzmann distribution p_θ with energy E_θ:

$$p_\theta(x) = \frac{e^{-E_\theta(x)}}{\Omega}, \qquad E_\theta(x) = ||x - x'||_2, \tag{5}$$

where θ are the trainable parameters of the network.

Performing MLE on the probability distribution translates to minimizing the sum of the reconstruction error and the normalization factor Ω. However, computing Ω becomes easily intractable for high-dimensional spaces, so we do not explicitly minimize this quantity. Instead, we rewrite the gradient of the loss function in a computationally feasible manner as:

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{x\sim p_d}[\nabla_\theta E_\theta(x)] - \mathbb{E}_{x\sim p_\theta}[\nabla_\theta E_\theta(x)]. \tag{6}$$

This allows us to reformulate the optimization as a min-max problem, where samples from the model distribution substitute the expensive integral. We obtain samples from p_θ using Langevin Markov Chains (LMC). An LMC process follows the equation:

$$x_{t+1} = x_t - \lambda \nabla_x \log p_\theta(x) + \sigma\epsilon \qquad \epsilon \sim \mathrm{N}(0,1), \tag{7}$$

and does not require an estimate of the integral due to the independence of the latter from the input x.

In particular, we utilize the Contrastive Divergence [54] MCMC scheme. Given a transition kernel T_θ for the data distribution p_D, the following loss function has a zero only for $p_\theta(x) = p_D(x)$ [55]:

$$KL(p_D||p_\theta) - KL(T_\theta^t(p_D)||p_\theta). \tag{8}$$

Therefore, we can run short Langevin Markov Chains with steps t, which define the transition kernel T_θ^t, and estimate the gradients of Eq. 6 as:

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{x\sim p_d}[\nabla_\theta E_\theta(x)] - \mathbb{E}_{x\sim T_\theta^t p_D}[\nabla_\theta E_\theta(x)]. \tag{9}$$

Note that Eq. 9 ignores an additional term as pointed out in [54]. We find that this approximation does not affect the convergence of our model and therefore we use the base CD loss.

The procedure defined above stabilizes the training and corrects for the mismodeling of the density estimate introduced by the mere minimization of the reconstruction error. The epoch with the energy difference closest to zero defines the best loss, and we select the corresponding model for evaluation. Before turning on the regularization term, we pre-train the autoencoder for 200 epochs then continue training according to Eq. 6 for another 100 epochs. The architecture of the encoder network is a simple feed-forward network with five layers with neurons from 8 to 128 in powers of two and a three-dimensional bottleneck. The decoder mimics the encoder network, this time up-sampling from 8 to 128 dimensions in powers of two.

5 Results

In this section, we show results using DarkCLR on the benchmark signal. First, we compare our results with previous methods tested on the same dataset. We then perform studies to test the robustness of our results with respect to variation of the semi-visible jet model parameters. Finally, we discuss the dependence of the performance on the main network parameters.

5.1 Improved performance

First, we discuss the base pipeline of our procedure and compare the results with other methods. We train the transformer encoder network with the hyper-parameters as specified in Tab. 1. The chosen embedding space uses 512 dimensions, and the augmentations follow the implementation described in Sec. 3, where $p_{\text{drop}} = 0.5$. Note that the size of the embedding space must be large enough to contain the information passed from the head to the output

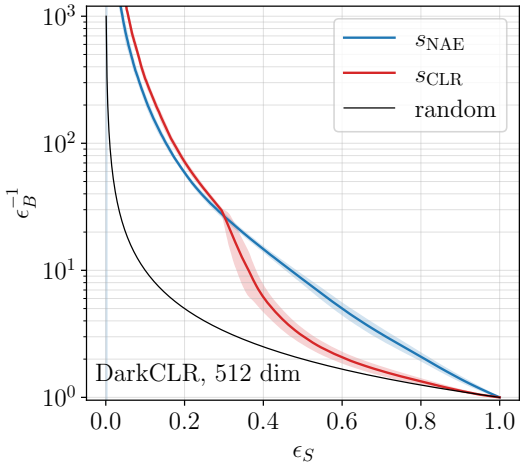

Figure 2: ROC curves of background suppression ϵ_B^{-1} versus signal efficiency ϵ_S, computed from the L_2 norm of the representations, s_{CLR} (red), and from the MSE of an NAE trained on the representations from DarkCLR, s_{NAE} (blue).

	DVAE [28]	INN [17]	NAE Jet images [21]	DarkCLR
AUC	0.71	0.73	**0.76(1)**	**0.76(1)**
$\epsilon_B^{-1}(\epsilon_S = 0.2)$	36	39	41(1)	**59(1)**

Table 2: Summary of AUCs and background rejections at low signal efficiencies for DarkCLR compared to other methods.

layer. As we show in App. A, our results are not sensitive to the specific choice of the embedding dimension, as long as it is sufficiently large. We show Receiver Operator Characteristic (ROC) curves for the CLR latent score s_{CLR} and the NAE score s_{NAE}. In addition, we report the low signal efficiency background rejection as a measure of the purity of a signal sample in the low background region. The error bands on s_{CLR} are taken from 5 runs of CLR training with different initializations. From each of these representations, we train 3 autoencoders for a total of 15 s_{NAE} scores, which are used to compute the mean and standard deviation. Note that no transformations are applied to the representations before training the autoencoder, thus limiting the preprocessing to the mere p_T rescaling and the physically guided CLR transformation.

Fig. 2 shows the ROC curves obtained with our method. The new embedding space greatly improves the background rejection ϵ_B^{-1}, in particular in the region of low signal efficiency as estimated by $\epsilon_B^{-1}(\epsilon_S = 0.2)$. We find that the transformer network does indeed encode information in the norm to discriminate between jets. In particular, it improves purity in the low background region, as shown by the background rejection of s_{CLR} at low signal efficiency. However, due to the high dimensionality of the representations, many jets will share the same norm in the bulk of the distribution, causing the s_{CLR} ROC curve to drop off at $\epsilon_S = 0.3$. We also observed similar problems when training a standard autoencoder. This is solved by a more precise density estimator like the NAE. The resulting s_{NAE} ROC curve is much more stable with an average AUC of 0.76 and a $\epsilon_B^{-1}(\epsilon_S = 0.2) = 59$.

Tab. 2 summarizes the AUC and the background rejection $\epsilon_B^{-1}(\epsilon_S = 0.2)$ for DarkCLR and compares them to previous methods: an NAE trained on jet images [21], a Dirichlet variational autoencoder [28], and an invertible neural network [17]. While the best AUC is similar for all

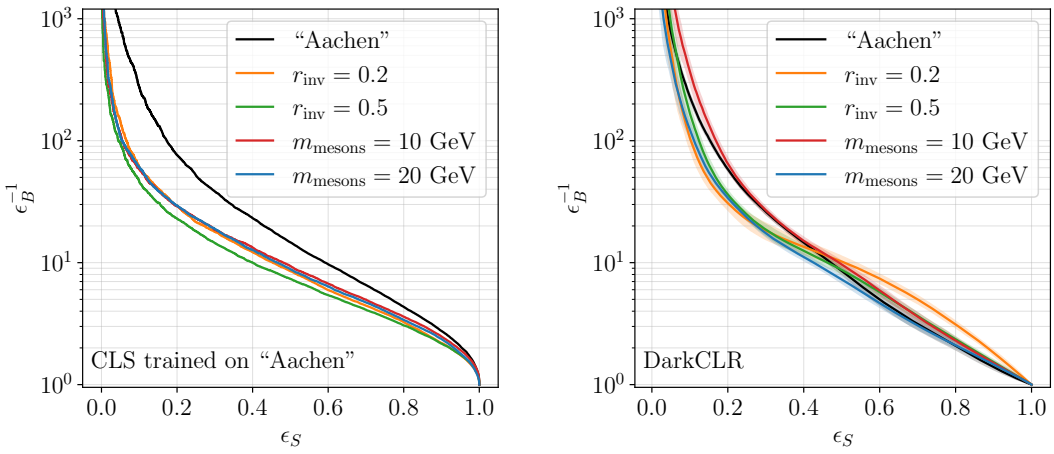

Figure 3: Left panel: ROC curves of a supervised classifier trained on the "Aachen" benchmark signal and tested on datasets with different dark shower model parameters. Right panel: ROC curves obtained from DarkCLR after training on the QCD background only and tested on additional datasets.

	$\epsilon_B^{-1}(\epsilon_S = 0.1)$				
	"Aachen"	$r_{\text{inv}} = 0.2$	$r_{\text{inv}} = 0.5$	$m_{\text{mesons}} = 10$ GeV	$m_{\text{mesons}} = 20$ GeV
CLS ("Aachen")	**258**	68	47	61	63
DarkCLR	230(13)	**110(20)**	**173(25)**	**390(70)**	**130(18)**

Table 3: Summary of the results presented in Fig. 3 for the background rejection ϵ_B^{-1} at a signal efficiency of $\epsilon_S = 0.1$.

methods, with DarkCLR we find much stronger background rejection at low signal efficiency, and we do not rely on image-based representations or any specific preprocessing steps.

5.2 Robustness of DarkCLR

Dependence on the dark shower signal As a next step, we study the robustness of our method with respect to the main phenomenological parameters of the semi-visible jet as described in Sec. 2. We set up a benchmark by training a transformer classifier with 100k jets equally divided between the QCD background and the "Aachen" dataset. We then use the classifier score to detect the signals with different invisible fraction r_{inv} and dark meson mass scale m_{mesons}. The classifier uses the same backbone transformer architecture of Sec. 3 where the head network is replaced by a two-layer MLP with ReLU nonlinearities and a single output. We train the network for 300 epochs, minimizing the binay cross-entropy loss, and refer to the validation loss to select the best model.

Fig. 3 shows the results of the supervised classifier (left panel) compared to DarkCLR trained only on the QCD background and tested on all signals (right panel). The supervised classifier shows a large drop in performance when applied to datasets with different model parameters, see also [15]. Instead, our DarkCLR method performs well on different semi-visible jet signals, as expected from the unsupervised training approach.

The small differences between the DarkCLR ROC curves for the various signals can be understood by analyzing the phenomenological aspects of the different semi-visible jet models.

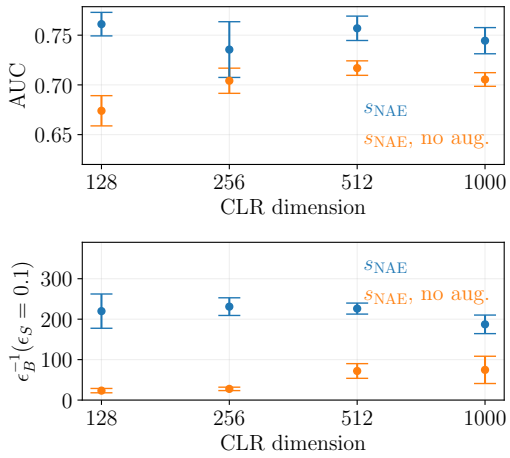

Figure 4: CLR and NAE AUC (upper panel) and background rejection at low signal efficiency (lower panel) for different embedding dimensions.

As we reduce the invisible fraction r_{inv}, the signal becomes more similar to a QCD jet, increasing the overlap between the two distributions and thus reducing the detection efficiency. Similarly, increasing the confinement scale and thus the mass of the dark hadrons leads to an earlier hadronization of the dark quarks. Therefore, the visible SM decays continue to shower down to the QCD confinement scale, again more closely resembling a QCD background jet initiated by light quarks. We observe this effect when we increase the energy scale from the default choice of the Aachen benchmark dataset to $m_{\pi_d} = m_{\rho_d} = \Lambda = 10\,\text{GeV}$ and $20\,\text{GeV}$.

For a summary of the background suppression at low signal efficiency, see Tab. 3. The generalization capabilities of DarkCLR outperform the supervised classifier for all signal models, especially in the more interesting low signal efficiency region.

Impact of Anomaly augmentation To validate the use of anomalous augmentations, we compare DarkCLR with the standard JetCLR training. The latter is trained only on QCD jets using the set of physical augmentations. We refer to previous work for the implementation and training of JetCLR [35]. After creating the new representations, we train an NAE using the same procedure. Fig. 4 shows the performance of JetCLR compared to DarkCLR in terms of AUC and background rejection for the benchmark dataset. Without anomalous pairs, the results vary between different embeddings and underperform in both figures of merit. Notably, DarkCLR improves detection at low signal efficiency even for small embedding dimensions, while without augmentation we observe a small increase in sensitivity only for large embedding spaces.

6 Summary and outlook

In this article we present DarkCLR[§], a new framework for detecting semivisible jets at the LHC, as predicted in models with a strongly interacting dark sector. DarkCLR is a self-supervised method based on contrastive learning representations. The CLR paradigm provides a new representation that is approximately invariant under physically motivated transformations of the data. In this study, a permutation invariant network learns a jet representation that is invariant to rotations and translations in the angular coordinates.

[§]The code will be made available at https://github.com/luigifvr/dark-clr

In general, preprocessing can improve the discrimination between QCD background and dark shower signals. However, the preprocessing is often hand-crafted and model-specific, and the performance of the classifier depends on the chosen transformations. We propose to introduce an augmented anomalous feature in the CLR training to learn such preprocessing based on general physical features of the signal. For semivisible jets, this is done by introducing an anomalous augmentation that drops components from the original jet. This ensures that the training uses only background events, reducing the dependence on the details of the dark sector model.

We show that the transformer network provides a discriminative representation of the data, which we use for unsupervised anomaly detection with a normalized autoencoder. Our method does not rely on hand-crafted preprocessing or an image representation of jets, and exhibits stronger background rejection at low signal efficiency compared to previous state-of-the-art methods. The probability distribution of the representations is not modified before training the autoencoder, thus limiting the effect of coordinate transformation on physically motivated CLR training.

In addition, we test the dependence of our model on the main phenomenological parameters entering the dark shower model, the invisible fraction of particles and the mass of dark mesons. We find that a supervised classifier is highly sensitive to the specific choice of signal parameters used during training, especially at low signal efficiencies. In contrast, our method, based on a density estimation of the background, is more robust to a variation of the parameters of the dark shower model, thus validating the application of unsupervised methods for model agnostic search.

We provide a proof-of-concept application of self-supervision for the detection of semivisible jets. Further studies will include the inclusion of additional augmentations for a wider coverage of signal classes where jet multiplicity is not the leading discriminative feature. We will also investigate the effect of choosing the dimensionality of the representation space and the interpretability of the latent space. More generally, although we have based our studies on simulations, we foresee the application of DarkCLR directly on data to overcome the effects of particular simulation choices, e.g. a specific hadronization model for the dark sector.

Acknowledgements

We would like to thank Barry Dillon for many useful discussions. LF would like to thank Alexander Mück and Elias Bernreuther for help during the generation of the dark showers. We would like to thank the Baden-Württemberg-Stiftung for financing through the program *Internationale Spitzenforschung*, project *Uncertainties – Teaching AI its Limits* (BWST_IF2020-010). This research is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 396021762 – TRR 257: *Particle Physics Phenomenology after the Higgs Discovery*.

A Linear Classifier Test on the dimensionality

As a final study of the separability between QCD and semi-visible jets, we train a Linear Classifier Test between background and signal. Even though we move to a supervised scenario, the network never accesses the signal data during training. This evaluation will test the separation power and the information content in the representations starting only from QCD jets and their augmentations. We disentangle the effects of the embedding dimension and the head network by selecting 128 as the embedding dimension of the transformer and scanning over the output dimension of the head network. This choice closely matches the original dimensionality of the input data. Fig. 5 (left) shows that the LCT of the head representation is informative regardless of the output dimension. The head network is affected by the projection on the hypersphere and requires a larger dimension to saturate to the same separation power. In both cases, we observe that the representation space is simpler than the original constituent-level space. The implemented LCT is a single linear layer network without non-linearities.

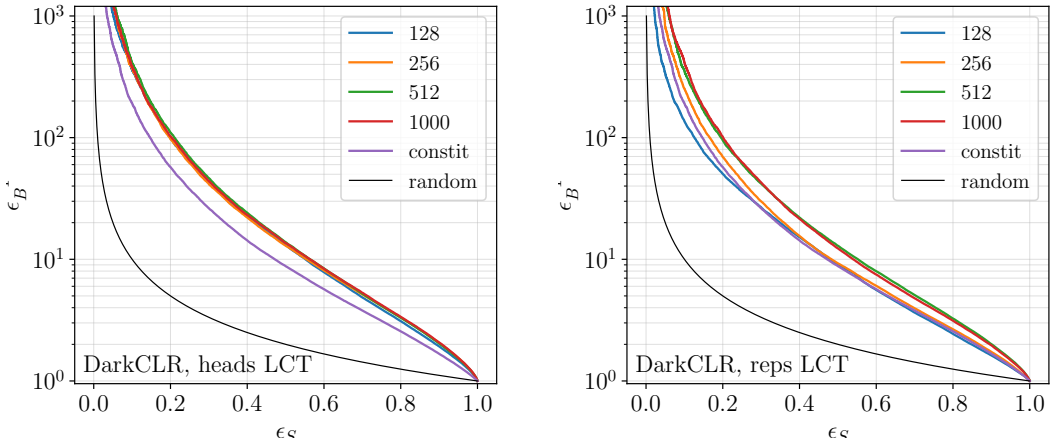

Figure 5: Linear Classifier test between the Aachen benchmark dataset and QCD jets. Head representations (left) and output representations (right) with different embedding dimensions from 128 up to 1000. The LCT on raw constituents is shown in purple.

References

[1] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman and D. Shih, *Machine learning in the search for new fundamental physics*, Nature Rev. Phys. **4**(6), 399 (2022), doi:10.1038/s42254-022-00455-1.

[2] HEP ML Community, *A Living Review of Machine Learning for Particle Physics* https://iml-wg.github.io/HEPML-LivingReview/.

[3] T. Heimel, G. Kasieczka, T. Plehn and J. M. Thompson, *QCD or What?*, SciPost Phys. **6**(3), 030 (2019), doi:10.21468/SciPostPhys.6.3.030, arXiv:1808.08979.

[4] M. Farina, Y. Nakai and D. Shih, *Searching for New Physics with Deep Autoencoders*, Phys. Rev. D **101**, 075021 (2020), doi:10.1103/PhysRevD.101.075021, arXiv:1808.08992.

[5] A. Blance, M. Spannowsky and P. Waite, *Adversarially-trained autoencoders for robust unsupervised new physics searches*, JHEP **10**, 047 (2019), doi:10.1007/JHEP10(2019)047, arXiv:1905.10384.

[6] T. S. Roy and A. H. Vijay, *A robust anomaly finder based on autoencoder* (2019), arXiv:1903.02032.

[7] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette and T. Golling, *Variational Autoencoders for Anomalous Jet Tagging* (2020), arXiv:2007.01850.

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

[9] O. Atkinson, A. Bhardwaj, C. Englert, V. S. Ngairangbam and M. Spannowsky, *Anomaly detection with convolutional Graph Neural Networks*, JHEP **08**, 080 (2021), doi:10.1007/JHEP08(2021)080, arXiv:2105.07988.

[10] S. Tsan, R. Kansal, A. Aportela, D. Diaz, J. Duarte, S. Krishna, F. Mokhtar, J.-R. Vlimant and M. Pierini, *Particle Graph Autoencoders and Differentiable, Learned Energy Mover's Distance*, In *35th Conference on Neural Information Processing Systems* (2021), arXiv:2111.12849.

[11] V. S. Ngairangbam, M. Spannowsky and M. Takeuchi, *Anomaly detection in high-energy physics using a quantum autoencoder*, Phys. Rev. D **105**(9), 095004 (2022), doi:10.1103/PhysRevD.105.095004, arXiv:2112.04958.

[12] B. Ostdiek, *Deep Set Auto Encoders for Anomaly Detection in Particle Physics* (2021), arXiv:2109.01695.

[13] J. Barron, D. Curtin, G. Kasieczka, T. Plehn and A. Spourdalakis, *Unsupervised hadronic SUEP at the LHC*, JHEP **12**, 129 (2021), doi:10.1007/JHEP12(2021)129, arXiv:2107.12379.

[14] A. Kahn, J. Gonski, I. Ochoa, D. Williams and G. Brooijmans, *Anomalous jet identification via sequence modeling*, JINST **16**(08), P08012 (2021), doi:10.1088/1748-0221/16/08/P08012, arXiv:2105.09274.

[15] T. Finke, M. Krämer, A. Morandini, A. Mück and I. Oleksiyuk, *Autoencoders for unsupervised anomaly detection in high energy physics*, JHEP **06**, 161 (2021), doi:10.1007/JHEP06(2021)161, arXiv:2104.09051.

[16] F. Canelli, A. de Cosa, L. L. Pottier, J. Niedziela, K. Pedro and M. Pierini, *Autoencoders for semivisible jet detection*, JHEP **02**, 074 (2022), doi:10.1007/JHEP02(2022)074, arXiv:2112.02864.

[17] T. Buss, B. M. Dillon, T. Finke, M. Krämer, A. Morandini, A. Mück, I. Oleksiyuk and T. Plehn, *What's Anomalous in LHC Jets?* (2022), arXiv:2202.00686.

[18] Z. Hao, R. Kansal, J. Duarte and N. Chernyavskaya, *Lorentz Group Equivariant Autoencoders* (2022), arXiv:2212.07347.

[19] O. Atkinson, A. Bhardwaj, C. Englert, P. Konar, V. S. Ngairangbam and M. Spannowsky, *IRC-Safe Graph Autoencoder for Unsupervised Anomaly Detection*, Front. Artif. Intell. **5**, 943135 (2022), doi:10.3389/frai.2022.943135, arXiv:2204.12231.

[20] L. Bradshaw, S. Chang and B. Ostdiek, *Creating simple, interpretable anomaly detectors for new physics in jet substructure*, Phys. Rev. D **106**(3), 035014 (2022), doi:10.1103/PhysRevD.106.035014, arXiv:2203.01343.

[21] B. M. Dillon, L. Favaro, T. Plehn, P. Sorrenson and M. Krämer, *A Normalized Autoencoder for LHC Triggers* (2022), arXiv:2206.14225.

[22] S. Yoon, Y.-K. Noh and F. C. Park, *Autoencoding under normalization constraints* (2023), arXiv:2105.05735.

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

[24] P. Jawahar, T. Aarrestad, N. Chernyavskaya, M. Pierini, K. A. Wozniak, J. Ngadiuba, J. Duarte and S. Tsan, *Improving Variational Autoencoders for New Physics Detection at the LHC With Normalizing Flows*, Front. Big Data **5**, 803685 (2022), doi:10.3389/fdata.2022.803685, arXiv:2110.08508.

[25] S. Caron, L. Hendriks and R. Verheyen, *Rare and Different: Anomaly Scores from a combination of likelihood and out-of-distribution models to detect new physics at the LHC*, SciPost Phys. **12**(2), 077 (2022), doi:10.21468/SciPostPhys.12.2.077, arXiv:2106.10164.

[26] P. T. Komiske, E. M. Metodiev and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, JHEP **04**, 013 (2018), doi:10.1007/JHEP04(2018)013, arXiv:1712.07124.

[27] V. Mikuni, B. Nachman and D. Shih, *Online-compatible unsupervised nonresonant anomaly detection*, Phys. Rev. D **105**(5), 055006 (2022), doi:10.1103/PhysRevD.105.055006, arXiv:2111.06417.

[28] B. M. Dillon, T. Plehn, C. Sauer and P. Sorrenson, *Better Latent Spaces for Better Autoencoders*, SciPost Phys. **11**, 061 (2021), doi:10.21468/SciPostPhys.11.3.061, arXiv:2104.08291.

[29] ATLAS Collaboration, *Dijet resonance search with weak supervision using 13 TeV pp collisions in the ATLAS detector* (2020), arXiv:2005.02983.

[30] G. Aad *et al.*, *Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at* $\sqrt{s} = 13$ *TeV with the ATLAS detector* (2023), arXiv:2307.01612.

[31] G. Kasieczka *et al.*, *The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics* (2021), arXiv:2101.08320.

[32] 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**(1), 043 (2022), doi:10.21468/SciPostPhys.12.1.043, arXiv:2105.14027.

[33] G. Kasieczka, R. Mastandrea, V. Mikuni, B. Nachman, M. Pettee and D. Shih, *Anomaly detection under coordinate transformations*, Phys. Rev. D **107**(1), 015009 (2023), doi:10.1103/PhysRevD.107.015009, arXiv:2209.06225.

[34] T. Chen, S. Kornblith, M. Norouzi and G. E. Hinton, *A Simple Framework for Contrastive Learning of Visual Representations*, Proceedings of the 37th International Conference on Machine Learning, PMLR **119**, 1597 (2020), https://proceedings.mlr.press/v119/chen20j.html, arXiv:2002.05709.

[35] B. M. Dillon, G. Kasieczka, H. Olischlager, T. Plehn, P. Sorrenson and L. Vogel, *Symmetries, safety, and self-supervision*, SciPost Phys. **12**(6), 188 (2022), doi:10.21468/SciPostPhys.12.6.188, arXiv:2108.04253.

[36] B. M. Dillon, L. Favaro, F. Feiden, T. Modak and T. Plehn, *Anomalies, Representations, and Self-Supervision* (2023), arXiv:2301.04660.

[37] B. M. Dillon, R. Mastandrea and B. Nachman, *Self-supervised anomaly detection for new physics*, Phys. Rev. D **106**(5), 056005 (2022), doi:10.1103/PhysRevD.106.056005, arXiv:2205.10380.

[38] T. Cohen, M. Lisanti and H. K. Lou, *Semivisible Jets: Dark Matter Undercover at the LHC*, Phys. Rev. Lett. **115**(17), 171804 (2015), doi:10.1103/PhysRevLett.115.171804, arXiv:1503.00009.

[39] T. Cohen, M. Lisanti, H. K. Lou and S. Mishra-Sharma, *LHC Searches for Dark Sector Showers*, JHEP **11**, 196 (2017), doi:10.1007/JHEP11(2017)196, arXiv:1707.05326.

[40] A. Pierce, B. Shakya, Y. Tsai and Y. Zhao, *Searching for confining hidden valleys at LHCb, ATLAS, and CMS*, Phys. Rev. D **97**(9), 095033 (2018), doi:10.1103/PhysRevD.97.095033, arXiv:1708.05389.

[41] H. Beauchesne, E. Bertuzzo, G. Grilli Di Cortona and Z. Tabrizi, *Collider phenomenology of Hidden Valley mediators of spin 0 or 1/2 with semivisible jets*, JHEP **08**, 030 (2018), doi:10.1007/JHEP08(2018)030, arXiv:1712.07160.

[42] E. Bernreuther, F. Kahlhoefer, M. Krämer and P. Tunney, *Strongly interacting dark sectors in the early Universe and at the LHC through a simplified portal*, JHEP **01**, 162 (2020), doi:10.1007/JHEP01(2020)162, arXiv:1907.04346.

[43] E. Bernreuther, T. Finke, F. Kahlhoefer, M. Krämer and A. Mück, *Casting a graph net to catch dark showers*, SciPost Phys. **10**(2), 046 (2021), doi:10.21468/SciPostPhys.10.2.046, arXiv:2006.08639.

[44] A. Batz, T. Cohen, D. Curtin, C. Gemmell and G. D. Kribs, *Dark Sector Glueballs at the LHC* (2023), arXiv:2310.13731.

[45] M. J. Strassler and K. M. Zurek, *Echoes of a hidden valley at hadron colliders*, Phys. Lett. B **651**, 374 (2007), doi:10.1016/j.physletb.2007.06.055, arXiv:hep-ph/0604261.

[46] D. E. Morrissey, T. Plehn and T. M. P. Tait, *Physics searches at the LHC*, Phys. Rept. **515**, 1 (2012), doi:10.1016/j.physrep.2012.02.007, arXiv:0912.3259.

[47] S. Knapen, J. Shelton and D. Xu, *Perturbative benchmark models for a dark shower search program*, Phys. Rev. D **103**(11), 115013 (2021), doi:10.1103/PhysRevD.103.115013, arXiv:2103.01238.

[48] L. Carloni and T. Sjostrand, *Visible Effects of Invisible Hidden Valley Radiation*, JHEP **09**, 105 (2010), doi:10.1007/JHEP09(2010)105, arXiv:1006.2911.

[49] L. Carloni, J. Rathsman and T. Sjostrand, *Discerning Secluded Sector gauge structures*, JHEP **04**, 091 (2011), doi:10.1007/JHEP04(2011)091, arXiv:1102.3795.

[50] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli and M. Zaro, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07**, 079 (2014), doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[51] T. Sjöstrand, 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**, 159 (2015), doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012.

[52] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens and M. Selvaggi, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, JHEP **02**, 057 (2014), doi:10.1007/JHEP02(2014)057, arXiv:1307.6346.

[53] M. Cacciari, G. P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04**, 063 (2008), doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[54] G. E. Hinton, *Training products of experts by minimizing contrastive divergence*, Neural Computation **14**(8), 1771 (2002), doi:10.1162/089976602760128018.

[55] S. Lyu, *Unifying non-maximum likelihood learning objectives with minimum kl contraction*, In *Neural Information Processing Systems* (2011).