MACHINE LEARNING
Science and Technology

PAPER

Triggering dark showers with conditional dual auto-encoders

Luca Anzalone[1,3,*] ⓘ, Simranjit Singh Chhibra[1,2,5] ⓘ, Benedikt Maier[2,4] ⓘ, Nadezda Chernyavskaya[2] ⓘ and Maurizio Pierini[2] ⓘ

1 Department of Physics and Astronomy (DIFA), University of Bologna, Bologna, Italy
2 European Organization for Nuclear Research (CERN), Geneva, Switzerland
3 Istituto Nazionale di Fisica Nucleare (INFN), Bologna, Italy
4 Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
5 Queen Mary University of London (QMUL), London, United Kingdom
* Author to whom any correspondence should be addressed.

E-mail: luca.anzalone2@unibo.it

Abstract

We present a family of conditional dual auto-encoders (CoDAEs) for generic and model-independent new physics searches at colliders. New physics signals, which arise from new types of particles and interactions, are considered in our study as anomalies causing deviations in data with respect to expected background events. In this work, we perform a normal-only anomaly detection, which employs only background samples, to search for manifestations of a dark version of strong force applying (variational) auto-encoders on raw detector images, which are large and highly sparse, without leveraging any physics-based pre-processing or strong assumption on the signals. The proposed CoDAE has a dual-encoder design, which is general and can learn an auxiliary yet compact latent space through spatial conditioning, showing a neat improvement over competitive physics-based baselines and related approaches, therefore also reducing the gap with fully supervised models. It is the first time an unsupervised model is shown to exhibit excellent discrimination against multiple dark shower models, illustrating the suitability of this method as an accurate, fast, model-independent algorithm to deploy, e.g. in the real-time event triggering systems of large hadron collider experiments such as ATLAS and CMS.

1. Introduction

Model-independent searches are becoming a valid alternative to model-dependent searches at colliders, aiming to discover new physics beyond the standard model (BSM) governed by a vast parameter space[6]. Hence, an enormous number of model-dependent analyses would be required to unravel such a vast parameter space in its entirety because each analysis need to target a specific signal: in this regard, model-independent searches provide a great, more flexible, alternative. The conventional cut-based analyses involve physics experts inspecting the distributions of various physical parameters to find discriminating characteristics. Once identified, the best threshold is determined, above which the events are considered signal-like. This part can be automatized by training a machine learning (ML) [1] or deep learning (DL) classifier [2–4] separating simulated background and signal events. Subsequently, a rigorous statistical test [5] determines the significance of the classified signal events: if above a certain threshold, the signal is declared to exist; if too low, the signal can be confidently excluded to exist at all.

Both the cut- and supervised ML-based search techniques are *model-dependent*, i.e. they assume a particular scenario for new physics, thus being signal-specific. For the ML-based approach, the classifier

---

6 A search aims to discard as many background events as possible while preserving the most signal: the background represents what is already well known to exist, i.e. standard model (SM) processes or detector effects. The signal, which may or may not exist in Nature, is the object of the search. The scope of analysis is to determine how plausible the existence of a specific signal is.

inherently adapts its learned parameters to be sensitive to specific signal features. However, it does not necessarily generalize towards unknown signals. Moreover, a supervised approach requires accurate signal and background simulation and robustness against systematic uncertainties. To mitigate such limitations, we propose a data-driven model-independent search strategy powered by conditional dual (Variational) auto-encoders (CoDAEs) and normal-only anomaly detection (AD) [6], demonstrating generalization over multiple signals despite not being trained on them.

In this article, we focus on two important and highly challenging manifestations of hidden valley (HV) models [7], more specifically, of dark quantum chromodynamics (QCDs), namely soft unclustered energy patterns (SUEPs) [8] and semi-visible jets (SVJs) [9–11]. The HV models are a type of BSM physics describing the existence of a new sector of particles and forces with new gauge groups and a mediator to the SM, which the large hadron collider (LHC) [12] can produce. The HV models have been mainly developed to address the origin of dark matter [13], whose experimental signatures often feature non-isolated objects with high-multiplicity and/or low-energy final states, representing a challenging target for existing analyses at the LHC [14].

Our proposed models can detect both SUEP and SVJ signals in highly sparse raw detector image data, constructed from the trigger system information, within the time budget of the high-level trigger (HLT) step [15][7], being trained only on the simulated QCD events: the class of data considered as not anomalous. Without making strong assumptions about the signals we avoid problem-specific pre-processing, on which discrimination performance can be highly dependent [16], and further reduce the dependency on the physics model. Our novel architecture can learn a two-dimensional (2D) *auxiliary* latent space through conditioning [17], capturing intrinsic information of the input that can be visualized, interpreted, and, in principle, employed for AD. Our contributions can be summarized as follows:

- We frame the new physics search problem as a normal-only AD task, making minimal assumptions on the nature of the signals. We only assume: 1) to have access to *normal* (i.e. not anomalous) data samples, and 2) that the signals can be revealed through tracking information.
- We propose a novel architecture that combines two encoders through spatial conditioning, in order to learn additional criteria for discriminating between signal and background.
- We perform a comprehensive comparison of anomaly scores, evaluating both scores derived from reconstructed images and the latent spaces.
- We ultimately show that our novel auto-encoder can reconstruct the target images with a much higher quality than compared approaches, which can also help human experts when visually inspecting anomalies.

Compared to both weakly-supervised (e.g. [18]) and classification methods (e.g. [2, 3]), which require partial or full knowledge of the signal(s), our approach assumes only the knowledge about the background events. Therefore, potentially enabling generic physics searches for unknown signals. In the following two sections we provide some further physics background relevant to understand our work.

1.1. The new physics search scenario: HV models

The HV models can produce dark quarks in proton-proton collisions at the LHC, leading to a dark shower and the production of a large number of dark hadrons (ϕ_D), analogous to QCD jets [7, 8]. Depending on the details of the theory, the dark showers can follow large-angle emission and dark hadrons do not form narrow QCD-like jets. The decay of dark hadrons results in dark photons (Z_D), which further decay to low-energy SM particles with transverse energy (E_T) of $\mathcal{O}(10^2)$ MeV, whose final experimental signature being high-multiplicity spherically-symmetric SUEPs [8]. Through their decay to SM particles via some portal state, like a dark photon, these processes become visible and in principle detectable in 4π-detectors at the LHC such as a toroidal LHC apparatus (ATLAS) [19] and compact muon solenoid (CMS) [20]. We focus on a well-motivated scenario where SUEP is produced in exotic Higgs (H) boson decays via gluon–gluon fusion and all dark hadrons decay promptly and exclusively to pions and leptons, an experimental nightmare scenario because of an overwhelming multi-jet QCD background.

Another manifestation of HVs can be SVJs [9, 11], a phenomenon in which energetic particles are emitted in a spray of stable invisible dark matter along with unstable states that decay back to SM. These showers are partially detectable, with the visible components looking like QCD showers [9]. This partial visibility makes it challenging to identify and study these particles thoroughly, having a low acceptance with current methods.

---

[7] We considered the compact muon solenoid (CMS) experiment, a general-purpose detector at the LHC [12], as the reference experiment for this study. The CMS trigger system is a two-tiered event selection system. The electronics-based first level (L1) uses information from the calorimeters and muon detectors and reduces the event rate from 40 MHz to around 100 kHz within a time interval of 4 microsec. The second level, known as the HLT, runs a version of the full event reconstruction software optimized for fast processing on a farm of processors. The HLT reduces the event rate to about 1 kHz within $O(10^2)$ ms, and the selected events are transferred to storage.
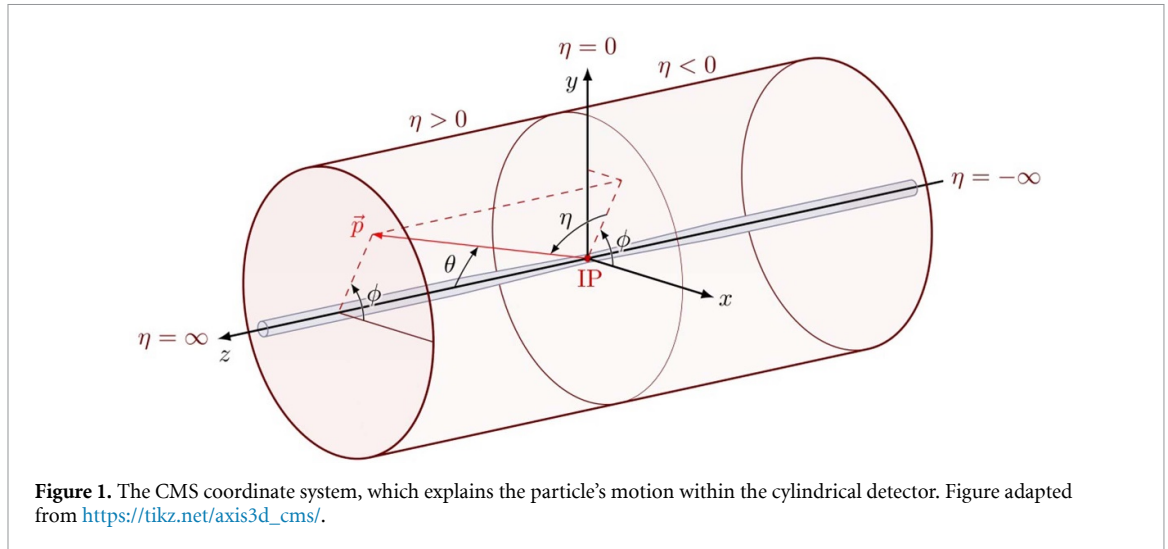
Figure 1. The CMS coordinate system, which explains the particle's motion within the cylindrical detector. Figure adapted from https://tikz.net/axis3d_cms/.

1.2. The CMS detector and simulated samples

The CMS experiment [20] is designed to explore the physics of proton-proton collisions through a system of different sub-detectors, each designed to measure different aspects of the particles produced in a collision. Given its *cylindrical* design, as we can see in figure 1, it is often convenient to adopt a polar coordinate system (θ, ϕ) where: $0 \leqslant \theta \leqslant \pi$ is the polar angle, and $0 \leqslant \phi \leqslant 2\pi$ is the azimuthal angle. From these coordinates, it is possible to explain the particle's kinematic as (p_T, y, ϕ, m): where m is the invariant mass, p_T the transverse momentum, and y the rapidity. A quantity related to the rapidity is the *pseudo-rapidity η*, which is a measure of the angle of the particle's motion relative to the beam line. The images employed in our study are represented in the η-ϕ plane, therefore considering the pseudo-rapidity and azimuth.

The CMS detector consists of several layers that are used to measure various properties of particles produced in high-energy collisions. The ones [21] relevant to our work are the:

- **Inner tracking system**, which measures the momentum of particles by their curvature radius through the magnetic field. The tracker can monitor the paths of charged particles. This sub-detector covers a pseudo-rapidity region of up to $|\eta| < 2.5$, being made of 66M silicon pixel detectors ($100 \times 150 \, \mu\text{m}^2$ in size) for accurate measurement of the particle's trajectory.
- **Calorimeters**, consisting of an electromagnetic calorimeter (ECAL), and a hadron calorimeter (HCAL). The calorimeters can measure the direction and energy of both charged and neutral particles. The two sub-detectors have different granularity: for the ECAL, the granularity of $0.0174 \times 0.0174 \, \text{rad}^2$ results in $286\eta \times 360\phi$ bins for the size of the images, whereas the HCAL is 25 times less granular, i.e. $0.087 \times 0.087 \, \text{rad}^2$. Therefore, each HCAL image is up-sampled by a factor of 25 in the preprocessing step, giving $1/25$th of the energy to each pixel.

We employ the Delphes v.3.4.3pre1 fast detector simulation [22] with the CMS Run-2 detector model to obtain the tracker, ECAL, and HCAL images. Samples of SM multijet events as well as for SVJ and SUEP signal processes have been generated with the Pythia v8.244 event generator [23].

2. Related work

In this section we review the relevant literature about AD in high-energy physics (HEP). AD [24] is the task of determining which samples violate some notion of normal behavior: once identified, these samples will be referred to as *outliers* or *anomalies*. We assume a normal-only setting, in which the training is performed only on background data representing the already known (i.e. not anomalous) behavior, being a good approximation to what occurs in practice, i.e. having the background contaminated with a little fraction of unknown signals. (Variational) auto-encoders (V/AEs) [25–27] are a popular mean to perform AD: the model is trained to minimize the reconstruction error of the normal samples, which is then used to score the novel data. Anomalies are found by thresholding such error. A general challenge is about designing anomaly scores that best separate the normal data from the anomalies [28]; to this end, V/AEs allow conceiving two main classes of anomaly scores, as described in the next two sections.

2.1. Reconstruction-based AD

Reconstruction-based anomaly scores are obtained by comparing the reconstructions, $\hat{x}$, with the inputs, x, of the V/AE. Different scores can be determined according to the distance or similarity function used to compare images. Heimel *et al* [6] introduce the benchmark dataset of QCD vs top jets. Their approach heavily relies on a specific particle-based processing of the raw collisions, which greatly simplifies the problem. Their LoLa AE, which is based on jet-level kinematics features, is able to beat an image-based AE by a large margin even with a smaller latent space, although at the cost of introducing an even larger dependency on the jet mass. Finke *et al* [29] discuss the limitations of using AEs on the same kind of data. The authors propose the kernel-MSE loss function, which is less sensitive than MSE, encouraging the network to learn dim pixels even in presence of sparsity. Recently, Dillon *et al* [30] propose to use a normalized auto-encoder (NAE) [31] to identify anomalous jets symmetrically. The NAE maximize the likelihood of the data through the minimization of an energy function. Under this probabilistic formulation, the NAE is forced to inhibit the reconstruction of an outlier, since it has to maximize the likelihood of the normal data, guaranteeing a low reconstruction error only for them. Although, NAEs are well-suited for AD, avoiding their training instabilities is still a practical challenge.

2.2. Latent-based AD

Latent-based anomaly scores are defined from the latent space captured by the encoder network: directly using the learned latent representation to flag anomalies can be difficult due its high-dimensionality, therefore combining the information carried by each latent component may require explicit supervision [32, 33]. Dillon *et al* [18] proposed to use a Dirichlet VAE [34] to learn a bi-modal, one-dimensional latent space that naturally encodes the two classes: signal and background. The authors show that the Dirichlet prior on the latent space naturally leads to mode separation, something that was not observed for both the regular VAE [25] and the Gaussian-mixture VAE [35], without enforcing any additional loss term. The proposed Dirichlet VAE reaches high class separation performance although weak-supervision is still required. Bortolato *et al* [36] propose to use the Kullback–Leibler divergence (KLD) between the learned and prior Gaussian distributions as an anomaly score to detect anomalous jets. Dillon *et al* [37] compared the effectiveness of using low-dimensional latent space representations instead of the event space features to perform model-agnostic AD. They trained a transformer encoder [38] to optimize the JetCLR's contrastive objective [39], where symmetry augmentations were employed to define positive and negative pairs for the contrastive learning. Through a binary classification test, the authors discovered that a sufficiently large latent space (e.g. of size 512) is required to encode the physical symmetries of jets. Finally, the CWoLa [32] method was used to perform model-agnostic AD, showing that still a significant fraction of signal events is required to achieve meaningful class separation. Govorkova *et al* [40] demonstrate a real-world deployment of a VAE on FPGA hardware for real-time AD at the LHC [12]. The authors compared the performance of both reconstruction- and KL-based anomaly scores, for both AE and VAE models. They concluded that with a minor loss in performance, the scores based on the KL divergence allowed them to only deploy the VAE's encoder on the FPGA, thus saving both hardware resources and latency costs. Recently, Cheng *et al* [41] enhanced a VAE with a technique known as outlier exposure (OE) [33], which makes use of an auxiliary set of out-of-distribution (OOD) data to improve the sensitivity to anomalies. An auxiliary loss term is computed from OOD predictions, which ensured a good compromise between high separation of anomalies and jet mass decorrelation. Although the promising results, it is not yet clear if data from the same physics domain is enough to be considered as OOD.

2.3. Related HEP analyses

The analyses conducted in [42] and [43] are related to ours, since it is assumed a similar signal setting. In particular, Barron *et al* [42] target the same SUEPs scenario in which the signal decays to exotic Higgs, and all the dark hadrons to SM hadrons. The authors identify three observables: charged particle multiplicity, event isotropy, and inter-particle distance. These are used to build the input features for their unsupervised fully-connected auto-encoder. Canelli *et al* [43], instead, study the SVJs signature by training a fully-connected auto-encoder on jet-level and jet substructure variables, minimizing the mean absolute error. Compared to these two studies, we neither rely on high-level nor engineered particle-based features but, instead, learn from raw detector images. Moreover, our models are evaluated against both signals, demonstrating anomaly scores that can identify both.

3. Simulated dataset of particle collisions

The dataset employed for our study contains simulated images of size $360 \times 286 \times 3$, for a total of about 615k samples, divided in: 442k QCD, 67k SUEPs, and 106k SVJs. The image channels represent 2D E_T (energy)

Algorithm 1. Image Pre-processing.

Input: a batch of images $I \in \mathbb{R}^{B \times H \times W \times C}$, kernel size K

Output: pre-processed images $I_M^{trk} \in \mathbb{R}^{B \times \lceil H/K \rceil \times \lceil W K^{-1} \rceil \times 1}$

   /* Depth-wise convolution to down-sample each channel by a factor of K */

1 $\text{kernel} = \texttt{tf.ones}((K, K, C, 1))$

2 $I' = \texttt{tf.nn.depthwise\_conv2d}(I, \text{filter} = \text{kernel}, \text{strides} = (1, K, K, 1), \text{padding} = \text{`SAME'})$

   /* Consider only the tracker channel, discarding the other two */

3 $I^{trk} = I'[\ldots, 0, \texttt{tf.newaxis}]$

   /* Compute the *mask* image */

4 $I_M^{trk} = \texttt{tf.cast}(I^{trk} > 0, \text{dtype} = \text{float})$

5 return I_M^{trk}

---

deposits in the η-ϕ plane, which are measured by the Inner tracker (Trk), ECAL, and HCAL sub-detectors [21] of CMS [20], respectively. Moreover, each image is annotated with a:

- **Class label.** There are three of them in total: the label 0 indicates the QCD background, the label 1 is associated to SUEP signal samples, and the label 2 refers to the second SVJ signal.
- **Mass label.** Signal samples only are identified by the mediator masses m_H [8] and $m_{Z'}$ [43] at which these were generated. In particular, SUEPs were generated at $m_H = \{125, 200, 300, 400, 700, 1000\}$ GeV and SVJs at $m_{Z'} = \{2.1, 3.1, 4.1\}$ TeV. For the rest of the paper, we refer to a particular signal sample by its mediator mass, such as SUEP(m_H GeV) and SVJ($m_{Z'}$ TeV).
- **Number of tracks.** This is a model-independent quantity that best approximates the number of decay products, obtained by the particle-flow reconstruction algorithm [21]. We refer to this variable as nTracks. It should be noticed that computing this quantity is expensive, being not feasible for real-time inference at the HLT.

Since the images are very sparse, having about 99.4% of zero pixels, and also moderately large, we employ a simple pre-processing (as described by algorithm 1) that down-scales the images, thus reducing sparsity while also preserving their total energy. The down-scaling is performed by convolving a 5×5 kernel with all ones on the input images along the channel dimension (i.e. in a depth-wise fashion), in a non-overlapping manner with a stride equal to the kernel size, yielding a $25 \times$ reduction in spatial resolution while preserving the sum of the energy deposits: the sparsity is also reduced to 96%; zero-padding is also applied to let the output size be divisible by the kernel size. The last step of the pre-processing is to discard the HCAL and ECAL channels, considering only the tracker one, resulting in images of size $72 \times 58 \times 1$: smaller images are faster to process by the network and require less storage, allowing to save up parameters, memory, computation and time.
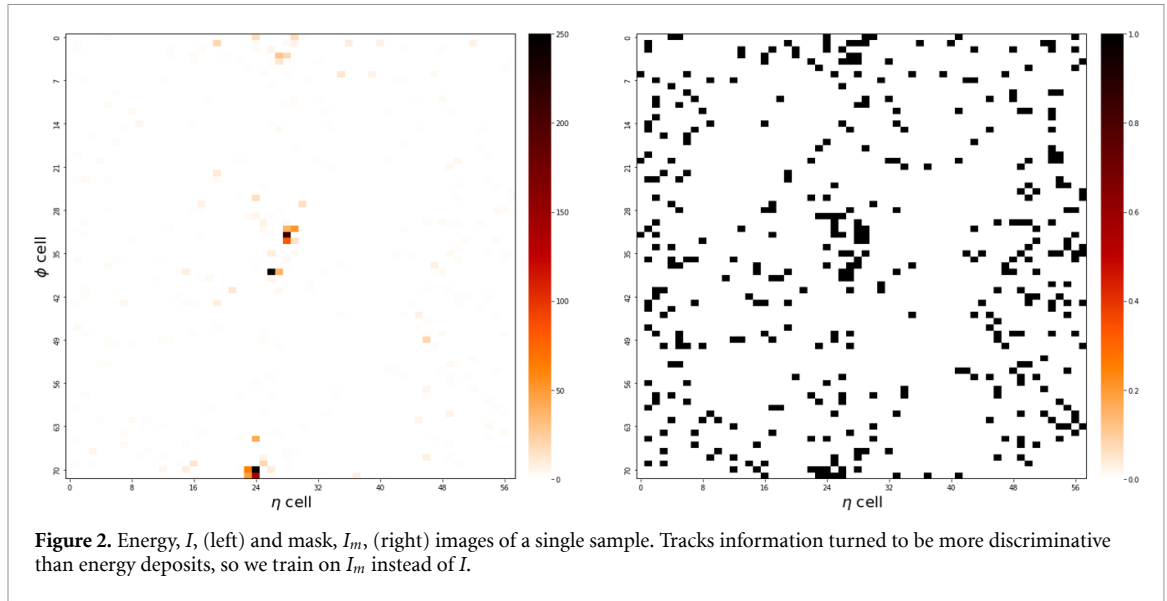
4. Method

In this section we detail our CoDAE architecture and training procedure, the physics-inspired image augmentations applied to the pre-processed input images, and also define a variety of reconstruction- and latent-based anomaly scores.

4.1. Image feature-engineering

Both the energy deposits and the nTracks variable can be seen as physics-motivated discriminators. Moreover, the number of tracks is much more sensitive to the searched signals than the energy, as stated in [8] and confirmed by our prior experiments, representing a better input for our models. Therefore, we devised a simple way to approximate the nTracks information by 'feature-engineering' the energy images, I, where each pixel depicts an E_T deposit, without running track reconstruction algorithms. The resulting *mask image*, I_m, is obtained by determining whether a pixel depicts a non-zero energy value: $I_m = \mathbf{1}[I > 0]$, where $\mathbf{1}[\cdot]$ is an indicator function applied to each pixel of I. Each pixel in I_m represents whether or not a single track has occurred, so its value can be at most one: a comparison of both kinds of images is shown in figure 2.

A mask image, if summed, denotes the number of non-zero deposits associated with sensors in the detector that measured some energy. This quantity is similar to the nTracks variable, but not equivalent since, depending on the granularity used to yield the images, two or more tracks can fall in the same bin (pixel) thus being not distinguished when counting non-zero pixels. In particular, we consider the mask image computed from the energy deposits of the tracker channel only, as the calorimeter information turned to be not enough informative: this fact was validated by our prior experiments, in which one possible explanation provided by [8] is that at the calorimeter level the SUEP resembles the pile-up since lacking hard

Figure 2. Energy, I, (left) and mask, I_m, (right) images of a single sample. Tracks information turned to be more discriminative than energy deposits, so we train on I_m instead of I.

and isolated objects, therefore, being more noisy than informative. Moreover, pixels in I_m have the additional benefit of being either zero or one, avoiding the need of normalizing E_T deposits which are often large in range and skewed towards small values.

4.2. Image augmentations

Since our data have physical properties like total energy, and the E_T deposits are arranged according to the design of the detector (other than being the result of a physics phenomenon), we cannot simply apply the usual off-the-shelf image augmentations like random crop, cutout, rotation, and jittering [44] that would reorganize the image' pixels without following the underlying physics and also without preserving both the individual and overall value of energy deposits. For this reason, we design novel data augmentations that preserve the physical meaning of the images by working on the η-ϕ plane, hence, also respecting the geometry of the detector.
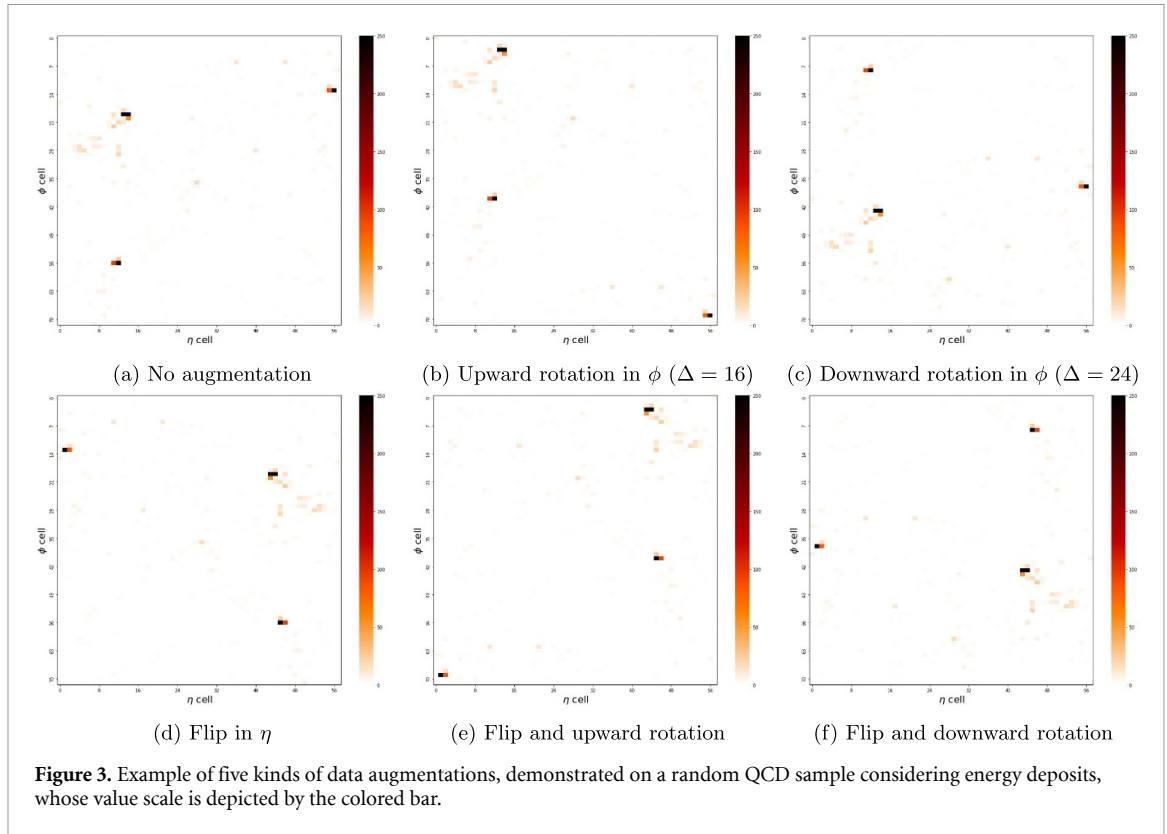
In particular, one kind of image augmentation involves a flipping in η, while the other is a rotation in ϕ: considering the figure 1 as a visual reference, the former can be interpreted as considering particles moving in the opposite direction with respect to the beam line (i.e. along the detector's z-axis), whereas the latter as rotating clockwise or anticlockwise the whole collision event on the detector's x-axis. The η-flip augmentation is simply implemented by mirroring the x-axis from left to right, where the ϕ-rotation is a little more complex. Rotation in ϕ (i.e. along the image's y-axis[8]) can occur both upward (i.e. anticlockwise) and downward (clockwise), in which a portion of the image moves up (or down) and the part in excess (the one that would fall off vertically from the image boundaries) is then attached to the bottom (or top). From a practical perspective, the ϕ-rotation is done in chunks of Δ rows, in which the chunk size is uniformly sampled for each image that should be rotated from the set, $\Delta \in \{8, 16, \ldots, 56\}$, whose values are only multiples of eight: a hyper-parameter value found to work well experimentally.

By combining these two kinds of image augmentations, flipping and rotation, it is possible to yield a total of five combinations of augmentations: 1) upward rotation, 2) downward rotation, 3) flipping, 4) flipping and upward rotation, and lastly 5) flipping with downward rotation. The image augmentations presented in figure 3 can be applied to raw images of particle collisions, regardless the specific kind of signal and background processes. Moreover, these are also designed to encourage the model to be invariant with respect the detector geometry: learning the properties of the detector's coordinate space is useful not only for AD but also for classification and regression problems.

4.3. The dual encoders

Capturing a latent space that is both discriminative for AD, and high-capacity for accurate pixel-level reconstructions can be challenging due to a trade-off between the size of the latent space and the reconstruction quality. Large latent spaces yield accurate reconstructions but cannot be used directly as AD scores unless summarized in some way. Conversely, small latent spaces can encode discriminative features but at the cost of poor reconstructions due to the low-dimensionality that does not retain pixel-level details:

---

[8] We refer to Cartesian x and y axes in the context of images, rather than x and y (i.e. rapidity) as detector coordinates.

Figure 3. Example of five kinds of data augmentations, demonstrated on a random QCD sample considering energy deposits, whose value scale is depicted by the colored bar.

this fact can stop the training prematurely, resulting in sub-optimal anomaly scores that may not even discriminate the anomalies. Since the model is trained to maximize the reconstruction fidelity, ensuring a good convergence of the loss can indirectly improve the anomaly scores too since these are, even if not optimized directly, defined from the reconstructions: reaching a better optimum entails a lower reconstruction loss on the normal examples, which, in the context of AEs, means obtaining a better (i.e. more structured, rich) latent space and a more accurate decoder.

As we want to capture both the detail and discriminative features, we define two encoders, f_R and f_m, trying to disentangle these two notions without any additional supervision. The encoder f_R is a residual network [45] that embeds the input images in a large convolutional latent space, $Z = f_R(I_m)$, of size $|Z| = 5 \times 4 \times 64$: given its large capacity, the latent components are expected to retain enough information to let the decoder yield high-quality reconstructions. The mask encoder f_m, instead, is a shallower convolutional network aimed at learning a compact and discriminative auxiliary latent space, $Z_m = f_m(I_m)$ where $|Z_m| = 2$, such that its components can be directly used as anomaly scores. Both encoders receive the same mask image, I_m, as input. Furthermore, the two networks have different architectures to induce a bias during training, established with prior experiments: f_R is high capacity and its skip connections can propagate information deeply in the layers' hierarchy helping to retain pixel-level details, whereas the max pooling layers in f_m are meant to consider only the most important activations to enhance the discrimination power of Z_m. For clarity, we also refer to Z as the *convolutional latents*, and Z_m as the *auxiliary or compact latents*.

4.4. The conditional decoder

The conditional decoder D is a residual network [45] whose main input is the convolutional latent space, Z (i.e. the output of the residual encoder, f_R), from which it tries to reconstruct the input mask images, I_m. The latent space, Z, is sufficiently large to provide enough information to the decoder to enable high-quality reconstructions; but the question is about how to enable the mask encoder, f_m, to learn a compact latent space, Z_m. The answer is provided by *conditioning* [17, 46], which establishes a dependency between the decoder, D, and the auxiliary latent space, Z_m, allowing the gradients of the loss to flow through f_m without any direct supervision. We call the whole auto-encoder architecture a *Conditional Dual Auto-Encoder* (or CoDAE): described in figure 4.

During training, the conditioning mechanism propagates the reconstruction error also to the mask encoder, without any additional loss term or extra supervision, providing feedback to learn Z_m such as to maximize the reconstruction quality. Turns out that Z_m alone is not enough for pixel-accurate
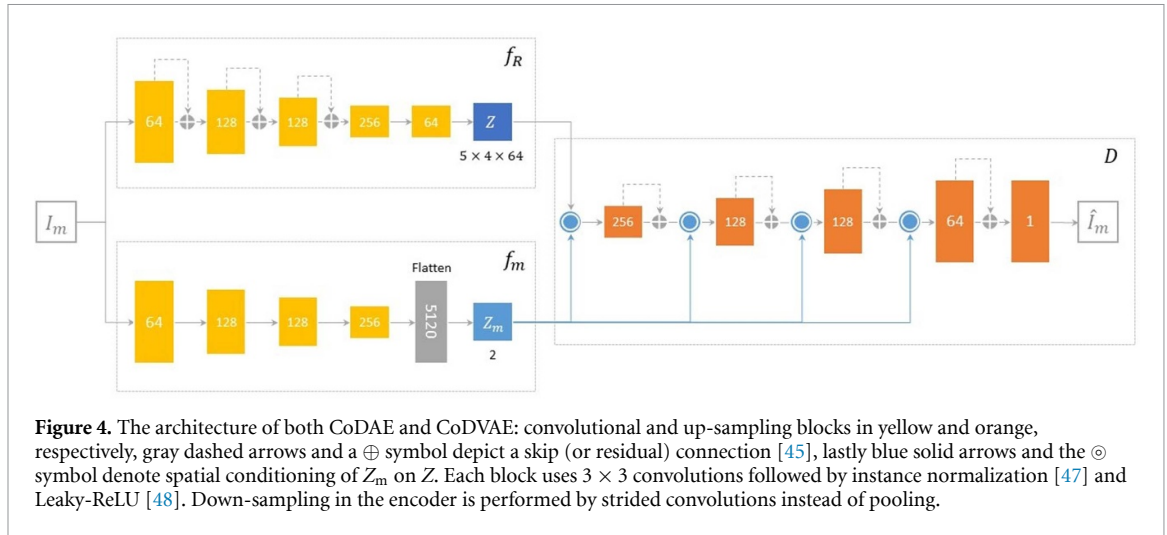
Figure 4. The architecture of both CoDAE and CoDVAE: convolutional and up-sampling blocks in yellow and orange, respectively, gray dashed arrows and a $\oplus$ symbol depict a skip (or residual) connection [45], lastly blue solid arrows and the $\odot$ symbol denote spatial conditioning of Z_m on Z. Each block uses 3×3 convolutions followed by instance normalization [47] and Leaky-ReLU [48]. Down-sampling in the encoder is performed by strided convolutions instead of pooling.

Algorithm 2. Spatial multiplicative conditioning.

Input: latents $Z_m \in \mathbb{R}^2$, tensor $h_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, kernel size K
Output: spatially conditioned representation $r_i \in \mathbb{R}^{H_i \times W_i \times C_i}$
/∗ See algorithm 1 in [49] ∗/
1 $z = \texttt{SpatialBroadcast}(Z_m, W_i, H_i)$
/∗ Expand channels of z to match h_i, through a linear convolution ∗/
2 $z = \texttt{Conv2D}(\text{filters} = C_i, \text{kernel\_size} = K, \text{padding} = \text{'same'})(z)$
/∗ Multiplicative conditioning: Hadamard product $z \odot h_i$ ∗/
3 $r_i = \texttt{tf.multiply}(z, h_i)$
4 **return** r_i

---

reconstructions[9], and so we also need to learn the high-capacity latent space, Z. Intuitively, we write $\hat{I}_m = D(Z \mid Z_m)$ to highlight that the compact latents, Z_m, must influence the reconstructions, $\hat{I}_m$, in order to represent meaningful and not just random encoded features. For such reason, the conditioning should be strong enough to prevent D from completely relying only on the convolutional latents: in its base form, Z is modulated by conditional scaling [17] (i.e. a conditioning mechanism that establish a dependency through multiplications), which occurs at multiple levels of the decoder hierarchy.

In particular, our form of conditioning combines *spatial broadcast* [49] with a feature-wise transformation [17]: element-wise multiplication or scaling. We will refer to this operation as *spatial conditioning*: described by equation (1) and algorithm 2. The spatial broadcast (SB) operation provides an inductive bias to the convolutional encoder, f_m, for learning *disentangled* latent factors in Z_m (which should encourage to capture independent features over the latent components) while, at the same time, modulating Z and the subsequent hidden feature maps. Spatial conditioning is performed at multiple spatial resolutions of the decoder's hierarchy of layers. Initially, at stage $i = 0$, the conditioning is performed on the convolutional latents (i.e. $h_0 = Z$) which are then fed to the first layers of the decoder. Subsequently ($i > 0$) the hidden feature maps (output of the previous residual block in stage $i - 1$), h_i, are conditioned on the same Z_m. In this way, the auxiliary latent space effectively modulates the decoder at different spatial resolutions. The operation performed on a generic tensor h_i, of size $H_i \times W_i \times C_i$, can be written as:

$$r_i = \text{Conv}(\text{SB}(Z_m)) \odot h_i, \tag{1}$$

where SB (see algorithm 1 in [49]) replicates and expands Z_m to match the shape of $H_i \times W_i$, the subsequent convolution (Conv) linearly expands the channels of the intermediate result to C_i, to finally perform the Hadamard product (denoted by $\odot$) with h_i, yielding the conditioned representation r_i at stage i.

In general, our spatial conditioning operation is not limited to only multiplicative interactions. Other simple conditioning mechanisms applied on feature maps are possible, like addition (also called biasing), concatenation, or even an affine-like operation that combines both multiplication (also known as scaling) and biasing. In principle, it is also possible to exploit the domain knowledge of the problem and data to

---

[9] Even with a high-capacity residual encoder, reconstructing from only the two components of Z_m results in reconstructions that look like just *average images*, thus without pixel-level details.

devise an application-specific conditioning mechanisms that utilizes such knowledge: for example, we may think of conditioning on a relevant physics observable.

4.5. Categorical CoDVAE

The described CoDAE architecture can be easily extended to variational auto-encoders [25] (VAE). In particular, we explored the Categorical VAE [50, 51] in which the convolutional latents, Z, are sampled from a Gumbel-Softmax (or Concrete) distribution: $Z \sim Cat(\alpha_R, \tau_R)$, where $\alpha_R, \tau_R = f_R(I_m)$ are the learned logits and temperature. The motivation is that the Categorical VAE can learn discrete features, like counts, whereas the regular (i.e. Gaussian) VAE captures continuous quantities which are not completely related to the nTracks variable. We call this model the *Categorical Conditional Dual Variational Auto-Encoder*, or CoDVAE in short. Likewise, in previous approaches [52, 53], we let the residual encoder also output a temperature τ_R, which can be difficult to tune properly, defining the degree of relaxation of the Categorical distribution (approximated by the Gumbel-Softmax) parameterized by α_R. Basically, we add a second point-wise convolution with a softplus activation, to ensure positive values, on the last residual block of figure 4(diagram section about f_R). Then, a base temperature, τ_0, which is a hyper-parameter set to 1, is added to τ_R to avoid gradient instabilities due to small numbers as $\tau_R \to 0$ recovers the true Categorical distribution: conversely, if $\tau_R \to \infty$ the distribution approaches a Uniform one. Since we learn the temperature, we define a more complex prior $p(Z)$ as a uniform mixture of N Gumbel-Softmax with different temperatures sampled uniformly in $[0.1, 1]$, to provide more flexibility to the latent space:

$$p(Z) = \frac{1}{N} \sum_i^N Cat(\log 1/C, \tau_i), \quad \tau_i \sim U(0.1, 1). \tag{2}$$

We then approximate the KL divergence (D_{KL}) between the prior and the learned distribution, $q = Cat(\alpha_R, \tau_R)$, with a Monte–Carlo estimate ($M = 10$), as follows:

$$D_{KL}(q \| p) \approx \frac{1}{M} \sum_{z \sim q(x)}^M \log q(z \mid x) - \log p(z). \tag{3}$$

To enable sharp reconstructions of mask images, we opt for learning a probabilistic decoder in which each output pixel is independently governed by a Bernoulli distribution, which is able to represent pixel values that are either zero or one, as in I_m. A single Bernoulli distribution can be denoted as $\text{Bern}(p_\theta)$ where p_θ is learned and specific for a single pixel. For sharp predictions we take the *mode* of such distribution since it is 1 if $p_\theta > \frac{1}{2}$ and 0 otherwise: its mean, instead, is suitable to represent smooth values in $[0, 1]$ like normalized energy deposits.

4.6. Anomaly scores

Auto-encoders provide the opportunity to define a variety of scores to perform AD. Our CoDAE architecture learns an auxiliary low-dimensional latent space, allowing each component to be an anomaly score: in our case $|Z_m| = 2$, so we have two discriminators[10]. In addition to this, VAEs provide a natural way to discriminate on the latent space through the KLD [36, 40] between the learned posterior and the prior distribution, i.e. $D_{KL}(q(z \mid x) \| p(z))$. Yet another option is provided by the fact that the KLD is not *symmetrical*: $D_{KL}(q \| p)$ and $D_{KL}(p \| q)$ have two different meanings. The former promotes mode-seeking behavior (called the *reverse* KL) and the latter encourages coverage of probability mass (known as *forward* KL.) We exploit these two additional scores (denoted as KL-R and KL-F) with our Categorical CoDVAE model. Let be x an input image and $\hat{x}$ its reconstruction. To ease the notation, we denote the set of pixel indexes $P = \{i, j, k \mid i = 0, \ldots, H-1, j = 0, \ldots, W-1, k = 0, \ldots, C-1\}$ for images of size $H \times W \times C$, and define our reconstruction-based scores on that as follows:

- $\text{BCE}(x, \hat{x}) = -\sum_{p \in P} x_p \log \hat{x}_p + (1 - x_p) \log 1 - \hat{x}_p$. It denotes the sum of the *binary-cross entropy* between the true and predicted pixels. Since the pixel values in each mask image are either zero or one, this measure is well-defined, without any normalization.
- $\text{SSE}(x, \hat{x}) = \sum_{p \in P} (x_p - \hat{x}_p)^2$. It represents the *sum of squared errors* between the true and reconstructed pixels.

---

[10] In principle, the latent space Z_m could have an arbitrary number of dimensions but there is a trade-off: learning either too numerous or too few (e.g. one) latent components may result in individual scores with a weak discriminatory power.

- $\text{SAE}(x,\hat{x}) = \sum_{p \in P} |x_p - \hat{x}_p|$. It depicts the *sum of absolute errors*, which is the absolute difference between true and predicted pixels. It is worth noticing that if the Categorical CoDVAE model is evaluated on mask images, this metric is equal to the SSE since both squared and absolute differences can be either zero or one (according to whether the pixel is correctly predicted or not): this is a direct consequence of taking the mode of the learned Bernoulli decoder, whose output pixels are binary values. Thus, in such cases, we omit the results of the SAE scores. Moreover, we can define SAE on mask images too: $\text{SAE-mask}(x,\hat{x}) = \sum_{p} \left| 1[x_p > 0] - 1[\hat{x}_p > 0] \right|$.

- $\text{Dice}(x,\hat{x}) = \frac{\sum_{p \in P} x_p^2 + \sum_{p \in P} \hat{x}_p^2}{2 \sum_{p \in P} x_p \cdot \hat{x}_p}$. This score is defined as the inverse of the Dice coefficient [54], to provide a measure of dissimilarity between two sets of pixels: x and $\hat{x}$.

- $\text{PixelSum}(x,\hat{x}) = \sum_{p \in P} \hat{x}_p$. It is just defined as the sum of each predicted pixel value. This score can be interpreted as *predicted total energy* if the model is trained to reconstruct I (energy image), or as an approximation of the `nTracks` if reconstructing I_m (mask image), instead. We also investigated the total number of predicted non-zero pixels, i.e. $\sum_p 1[x_p > 0]$, but we did not report results about it because its discriminatory capability is quite weak.

Both categories of anomaly scores have their pros and cons. Reconstruction scores are in general easier to define, for example from either common loss functions or metrics, but are slower to compute them since it is required to forward the full model (encoder and decoder) to reconstruct the samples. Instead, latent-based scores involve only the encoder predictions which are more suited for inference, although possibly more difficult to define (e.g. analytical or empirical KL) and visualize (e.g. dimensionality reduction on high-dimensional latent space.)

4.7. Training procedure and model acceleration

During training, we make use of the set of augmentation functions, $\mathcal{T}$, defined in section 4.2. At each mini-batch of mask images I_m a random augmentation is sampled $t \sim \mathcal{T}$ and applied to it. The augmented images, $\tilde{I} = t(I_m)$, are then fed to both residual and mask encoders; the decoder is then trained to reconstruct $\tilde{I}$. The data augmentations enable the CoDAE models to learn invariances related to the coordinates η and ϕ of the detector, which can be also seen as an implicit way to impart some physics notions to the model. In addition, we perform model selection according to the value of the *structural similarity* (SSIM) [55] metric between the true and reconstructed images, which provides a more human-aligned measure of image quality, computed on a validation set of only background samples.

The whole CoDAE models are learned end-to-end using the `AdamW` optimizer [56], whose learning rule decouples the weight decay regularization term from the main objective [57] making it easier to tune, minimizing the binary cross-entropy loss[11]. The optimizer is left with default parameters and learning rate, except for the weight decay coefficient set to 10^{-4}. Furthermore, to improve training stability, we limit the l_2-norm of each gradient to be at most one. Lastly, all the weights are initialized by following the `he_uniform` [58] scheme with zeros biases except for the decoder, whose biases are initialized to -1 to provide a better starting point for the initial reconstructions.

We use *TensorFlow lite* [59] for both model compression and acceleration of our CoDAE models. In particular, we employ a very light `float16` weight and activation quantization that resulted in a $5.7 \times$ reduction in model size and lower inference time, without any reduction in both AD and reconstruction performance. In this way we are able to achieve inference latency on a single consumer CPU hardware of about: 3ms (i.e. $6-8\times$ faster) for the mask encoder, and 40ms (i.e. $3.2-4.1\times$ faster) for a full forward pass on the whole model; therefore well under the 100ms time-limit of the HLT. This also demonstrates that our model architecture is very easy to optimize for deployment.

5. Results

In this section we present our evaluation protocol, and show the obtained results of our experiments comparing physics-motivated baselines, prior approaches, and our models.

5.1. Evaluation metrics

For the evaluation of both baselines (defined in the following section) and our proposed models, we treat the anomalies (the two signals) as the positive class and employ two popular metrics in OOD detection [33]: the area under the receiving operating characteristic curve (AUROC), and the false positive rate at $N\%$ of the true positive rate (FPRN). The AUROC summarizes the performance of the discriminator across multiple

---

[11] The loss is summed over spatial dimensions (height, width, and channels), and averaged over the batch size.

thresholds while the FPRN evaluates the performance at one specific threshold value: such threshold is often specific for the application and domain requirement. In our case, we choose $N\% = 40\%$ which targets a signal efficiency of 40%. Consequently, the metric is named FPR40.

The AUROC can be considered as the probability that a signal sample is assigned a higher AD score than a background example. Thus, higher AUROC values are better, depicting a higher retention of the signal at a lower background efficiency (and so at a higher rejection rate of the background.) The FPRN metric, instead, is more suited to compare strong models: interpreted as the probability that a normal (background) sample is flagged as an anomaly (so as a signal) when the 40% of signals are correctly detected. Since we want to decrease such false alarm probability, lower FPR40 values indicate a better model.

Moreover, since the performance of deep neural networks are dependent on many stochastic factors (e.g. sampling of the data, random weight initialization, dropout, etc) and specific (architecture and optimization) hyper-parameters, making difficult to conclude which model is actually better than another according to few (average) scores, especially when these show counter-intuitive results: e.g. one score is the best on a signal, and the worst on the other. Hence, we employ the *Almost Stochastic Order* (ASO) test [60, 61] as implemented by [62] which provides a statistically significant result from which it can be decided the best performing algorithm. The ASO test is specifically designed for neural networks, building on the concept of *stochastic order* in which one distribution of scores (e.g. AUROC) is said to stochastically dominate another one if the cumulative distribution of the former is lower than the latter for every point. Since the stochastic dominance is too strict to be practical, the *almost stochastic dominance* is used instead, which quantifies the extent to which stochastic order is violated. Therefore, the ASO test returns an upper bound, called $\epsilon_{\min}$, expressing the amount of violation: if $\epsilon_{\min} < \tau$ (where τ is the *rejection threshold* usually set to 0.5 or less, like 0.2 for a more confident result), then the former algorithm is superior than the latter. The value $\epsilon_{\min}$ can be interpreted as a confidence score: the lower it is, the more sure we can be about the dominance of one model over another. Experimentally, we follow the best practice suggested in [62] by fixing one set of hyper-parameters and comparing multiple runs of the same model where possible. Moreover, for each benchmark mass we build an empirical distribution of the AUROC by computing this metric on a thousand of random, class-balanced subsets (with $2k$ samples each) of the data.

5.2. Baseline discriminators

For a fair comparison and assessment of our method, we determined various baseline discriminators: two physics-motivated ones, a fully supervised classifier, a convolutional auto-encoder (CAE), and two AD models. The two physics discriminators are respectively based on the *total energy* (i.e. the sum of energy deposits, E_T, in each image channel), and the `nTracks` variable. Specifically, the `nTracks` is a model-independent classical variable corresponding to the total number of tracks per event [19, 20]. Such quantity is related to the number of decay products providing the best approximation of such to discern the signal particles, being also independent of the binning used to discretize the detector resolution. In particular, we define the total energy baseline as follows:

$$s^{(k)}(c) = \sum_{i}^{H} \sum_{j}^{W} x_{i,j,c}^{(k)}, \quad c \in \{0,1,2\} \tag{4}$$

where $s^{(k)}(c)$ is the score value for channel c (which denotes, respectively, the Trk, ECAL, and HCAL) of the kth image $x^{(k)}$ with height H and width W. Discrimination will be then performed according to the obtained scores, s, per channel.

Furthermore, we also consider the performance of a *supervised* classifier, therefore assuming an ideal setting in which we would have perfect knowledge of the data. Such a supervised baseline will provide a good approximation regarding upper-bound discrimination performance that our unsupervised model may achieve at its best. In particular, we consider a robust model, a *Compact Convolutional Transformer* (CCT) [63] that already outperformed a simpler convolutional network in our prior experiments.

The next baseline is a CAE derived from our CoDAE: it has the same architecture and hyper-parameters, lacking only the second (smaller) encoder network and the spatial conditioning mechanism in the decoder since there is no additional input (i.e. CoDAE's the auxiliary latent space) to perform conditioning on. This CAE model is trained in exactly the same way our models are.

The last baselines are two popular AD models: an unsupervised AE inspired[12] by [6], and the Dirichlet VAE from [18]. In particular, the AE model has a total of 600k parameters, a latent space of size 32, and was

---

[12] To the best of our knowledge, the authors provide only a figure outlining their model architecture and not a comprehensive description; we did our best to mimic their approach.

Table 1. Comparison of anomaly detection baselines and models on test-set, best scores only. AUROC metric, higher is better: mAUC denotes the AUROC averaged over all the mediator masses for a given signal. Entries associated to an (*) were averaged over three random seeds: 42, 51, and 73. Top three best results in boldface.

Model	SUEP (GeV)							SVJ (TeV)			
	125	200	300	400	700	1000	mAUC	2.1	3.1	4.1	mAUC
Total energy (`Trk`)	45.18	48.4	54.13	58.51	67.39	70.89	57.42	55.86	72.13	81.46	69.82
`nTracks`	78.68	92.39	98.31	99.6	99.94	99.93	94.81	82.05	89.32	92.92	88.1
Supervised CCT [63]	89.72	96.58	98.88	99.42	99.87	99.93	**97.4**	96.05	98.17	98.76	**97.66**
*CoDAE (Z_2)	79.77	93	98.28	99.45	99.54	99.23	94.88	83.2	88.3	90.83	87.44
*CoDAE (`BCE`)	88.74	97.54	99.61	99.93	99.99	99.99	**97.63**	85.98	90.39	92.62	**89.66**
Cat. CoDVAE (Z_1)	77.54	91.5	97.71	99.21	99.42	99.19	94.1	81.36	86.4	88.9	85.55
Cat. CoDVAE (`KL-F`)	69.32	83.92	93.18	96.38	98.04	98.18	89.84	79.69	84.08	86.11	83.3
Cat. CoDVAE (`SSE`)	86.93	97.01	99.51	99.9	99.98	99.98	97.22	85.29	89.78	92.04	89.04
*CAE (`BCE`)	89.42	97.84	99.7	99.95	99.99	99.99	**97.81**	85.96	90.45	92.72	**89.71**
AE [6]-like (`PixelSum`)	83.89	94.6	98.69	99.68	99.95	99.94	96.13	83.26	88.72	91.44	87.81
Dirichlet VAE [18] (Z_2)	51.93	54.99	59.58	63.17	71.16	74.51	62.55	63.52	67.37	69.26	66.72

Table 2. Comparison of anomaly detection baselines and models on test-set, best scores only. FPR40 metric, lower is better: mFPR denotes the FPR40 averaged over all the mediator masses for a given signal. Entries associated with an (∗) were averaged over three random seeds. Top three best results in boldface.
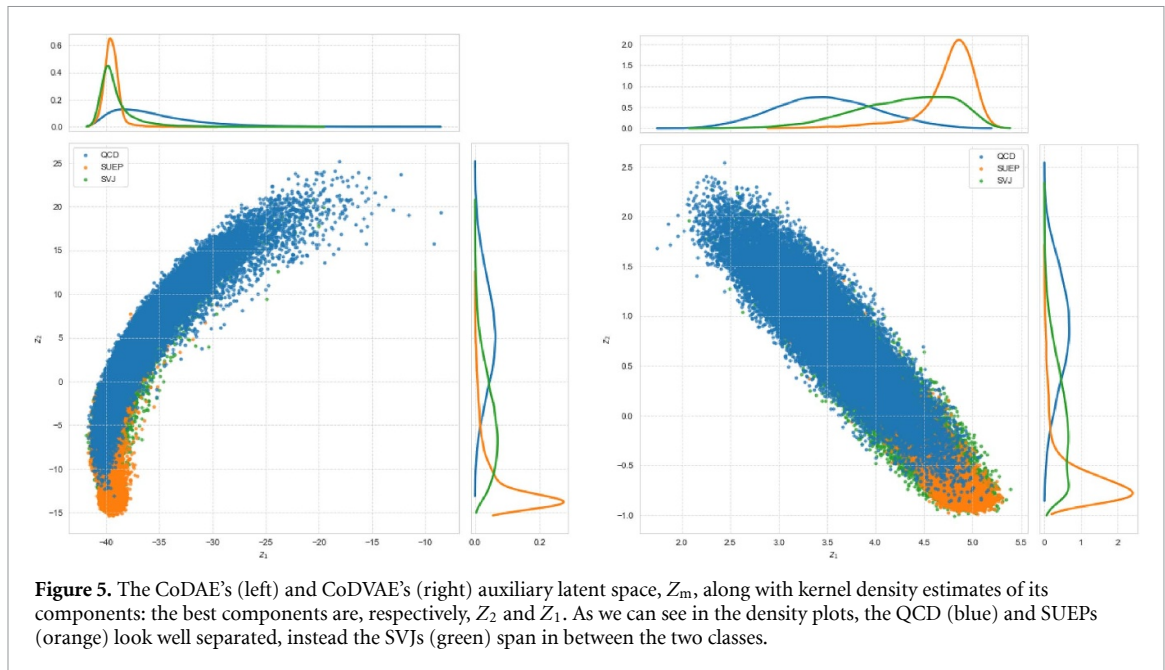
Model	SUEP (GeV)							SVJ (TeV)			
	125	200	300	400	700	1000	mFPR	2.1	3.1	4.1	mFPR
Total energy (`Trk`)	50.63	48.11	41.81	37.71	29.3	25.11	38.78	33.84	17.46	9.155	20.15
`nTracks`	11.84	2.849	0.277	0.028	∼0	∼0	2.5	5.45	1.599	0.539	2.53
Supervised CCT [63]	0.242	0.016	0.002	∼0	∼0	∼0	**0.04**	0.121	0.021	0.005	**0.05**
∗CoDAE (Z_2)	11.44	2.68	0.35	0.107	0.237	0.434	2.54	4.65	2.006	1.17	2.61
∗CoDAE (`BCE`)	5.832	0.806	0.057	0.004	∼0	∼0	**1.12**	2.829	1.027	0.523	**1.46**
Cat. CoDVAE (Z_1)	12.91	3.329	0.575	0.219	0.227	0.341	2.93	5.9	2.866	1.85	3.54
Cat. CoDVAE (`KL-F`)	19.08	7.498	2.635	1.452	0.739	0.677	5.35	7.425	4.913	4.085	5.47
Cat. CoDVAE (`SSE`)	6.652	0.944	0.071	0.007	∼0	∼0	1.28	3.065	1.176	0.588	1.61
∗CAE (`BCE`)	5.175	0.654	0.04	0.002	∼0	∼0	**0.98**	2.694	0.943	0.464	**1.37**
AE [6]-like (`PixelSum`)	8.22	1.904	0.211	0.035	∼0	∼0	1.73	4.346	1.74	0.863	2.32
Dirichlet VAE [18] (Z_2)	26.06	17.38	10.15	6.66	2.572	1.42	10.71	21.42	14.89	10.92	15.74

trained to minimize a mean squared loss: compared to our CoDAE, it lacks the second encoder and skip connections, and embeds its inputs to dense vectors while the images are reconstructed by transposed convolutions. Instead, the Dirichlet VAE is a weakly-supervised approach that, therefore, also requires a fraction of the signals for training: we assume a realistic setting in which the background is contaminated with 0.01% of the signals. This model has a three dimensional latent space, whose prior distribution is the Dirichlet [34] instead of the Gaussian, resulting in 2M parameters: during training we normalize the images to sum to one, and use the same hyper-parameters as in [18]. For both AD models, we apply the data augmentations defined in section 4.2. We trained both AE and Dirichlet VAE for 50 epochs (the latter converged earlier in training), the CoDAE for 30 epochs, and the Categorical CoDVAE for 100 of them since we observed slower convergence compared to the CoDAE. Lastly, the batch size is 128 for all models, and the weights are optimized by `AdamW` [56, 57].

5.3. AD

In this section we compared our two models (the CoDAE and Categorical CoDVAE) against the baselines defined in the previous section. For all the models and baselines we computed the anomaly scores defined in section 4.6, where possible (e.g. it is not possible to compute the KL for the AE), and evaluated their respective AUROC and FPR40. In particular, in tables 1 and 2 we provide the results for the best scores on average, i.e. the anomaly scores that achieve the best performance on both signals by considering the average over the mass points, while the full evaluation is available in the supplementary material.

Discussing about the physics baselines, for the SUEPs we can see that the sum of E_T deposits of the signal is actually lower than the one of the QCD background, resulting in shifted distributions of scores that yield

Figure 5. The CoDAE's (left) and CoDVAE's (right) auxiliary latent space, Z_m, along with kernel density estimates of its components: the best components are, respectively, Z_2 and Z_1. As we can see in the density plots, the QCD (blue) and SUEPs (orange) look well separated, instead the SVJs (green) span in between the two classes.

an AUROC below 50%. This total energy baseline indicates that one signal (the SUEPs) is less complex than the background. Instead, such issue does not occur for the `nTracks` baseline, which already provides a pretty good separation performance for both signals on both evaluation metrics, especially for the SUEPs at high mass; as stated in [8]: counting the number of tracks is particularly sensitive to high multiplicity soft particles like SUEPs, making them easier to identify. Compared to the AUROC, the FPR40 metric helps us better understand the shape of the ROC curve around our target signal efficiency of 40%, showing a very high (almost ideal) background rejection rate for the supervised classifier in all the benchmark scenarios.

Compared to the `nTracks` baseline, which requires counting the track multiplicity possible only if fully reconstructing the event as in an off-line analysis, our models are already able to get competitive performance in the latent space (showed in figure 5) and even improve by a neat margin when using reconstruction-based scores such as the `BCE` or `SSE`. We can notice that for some benchmark points, like SUEP(400 GeV), SUEP(700 GeV) and SUEP(1000 GeV), the AUROC easily saturates (table 1) attaining an almost perfect background rejection (table 2), therefore the improvement brought by a data-driven approach is negligible. Indeed, these are easier to detect since they are expected to deviate significantly from the QCD background. Instead, in the most challenging scenarios our best model achieves an AUC improvement of at most +10.1% for SUEPs and +3.9% for SVJs, as well as a reduction of FPR of at most 6% for SUEPs and 2.6% for SVJs, at the predefined signal efficiency compared to the `nTracks` baseline. Our models are able to reduce the gap with the supervised classifier despite being trained on the background class only. Next, the AE underperform our models, and the Dirichlet VAE is only competitive against the total energy baseline attaining a too low background rejection rate. The CAE, then, performs slightly better than both the CoDAE and CoDVAE: this is expected since optimizing the model is easier, in fact, it can be seen as a simplified CoDAE as it lacks the spatial conditioning mechanisms.

Lastly, in table 3 we summarize the results about the ASO test between our models against the baselines, considering the AUROC metric as score distribution for the statistical test which is performed between two distributions at a time. As we can see, our models (in particular, when using the `BCE` and `SSE` scores) can beat most of the baselines and for the SUEP signal also the supervised classifier, except for the SUEP(125 GeV) benchmark point. Furthermore, this test confirms the superiority of the CAE model over the CoDVAE but not against the CoDAE: the violation ratio fluctuates around 0.5, making difficult to decide which model is the best. This also implies that the CAE is a valid alternative to our dual-encoder models, especially in settings in which the auxiliary latent space has a low discriminatory power.

5.4. Reconstruction quality

To assess the reconstruction quality of the compared models, we evaluate two metrics: the mean squared error (MSE), and the SSIM index [55]. In particular, the popular MSE metric is useful for determining the texture quality of the reconstructions, since it penalizes pixel-level differences. Instead, the SSIM is a perceptual quality metric designed to better match the perceived visual quality of humans. These two metrics

Table 3. Pairwise comparisons of models and baselines, performing the ASO test based on the AUROC with a 95% confidence level: the headings are our models, which are compared to the baselines (the entries). Models and baselines denoted with a (∗) were run on three random seeds. The smaller the value, the best the model compares against the respective baseline. Note: the listed values are the result of an approximate calculation, therefore there may be either false positives or negatives.
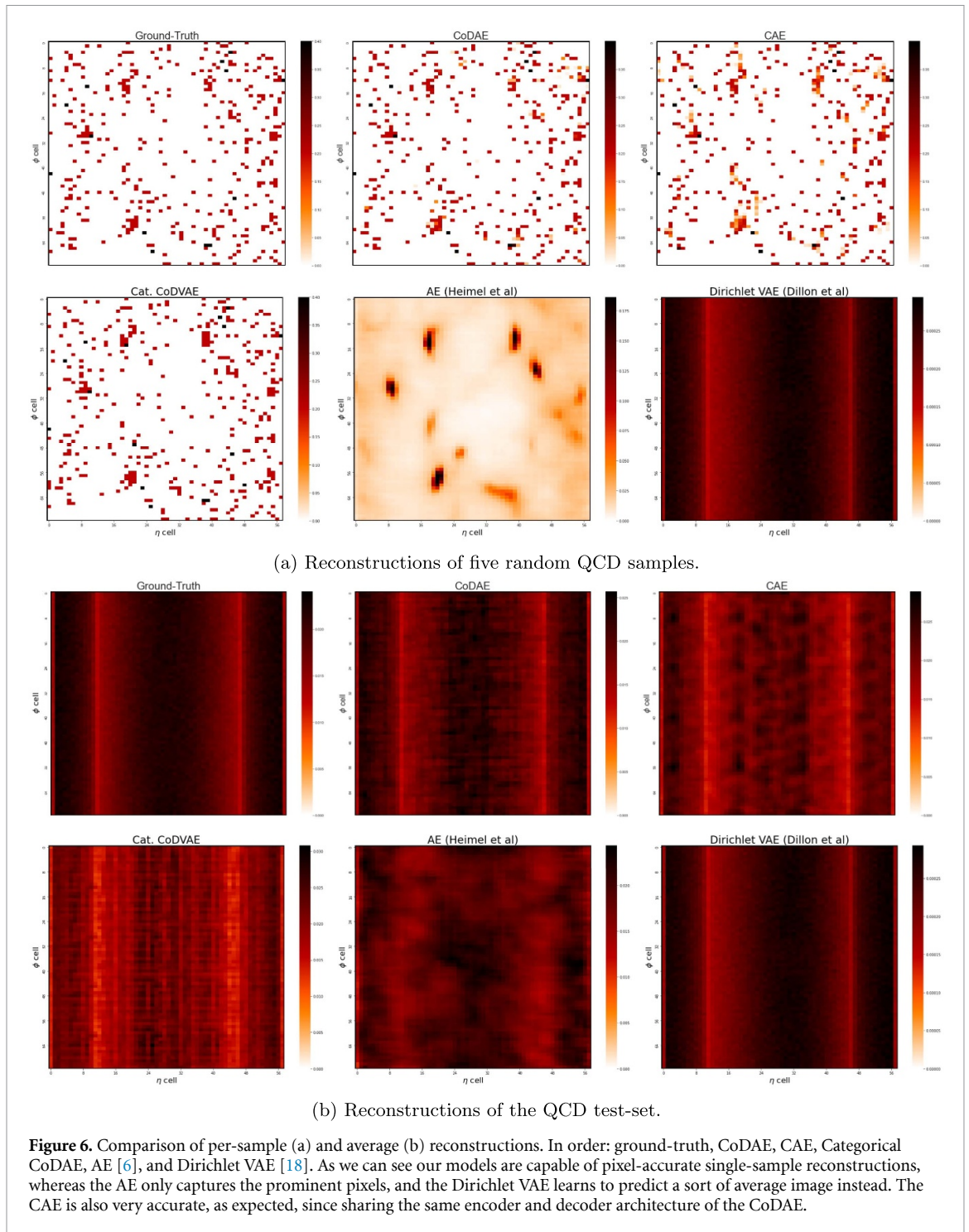
	∗CoDAE (Z_2)								
	SUEP (GeV)						SVJ (TeV)		
Baseline	125	200	300	400	700	1000	2.1	3.1	4.1
∗CAE (BCE)	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99
AE (PixelSum)	0.99	0.99	0.99	0.34	0.99	0.57	0.7	0.99	0.99
nTracks	0.66	0.65	0.99	0.99	0.99	0.99	0.33	0.99	0.99
Supervised CCT	0.99	0.99	0.99	0.45	0.69	1	0.99	0.99	0.99
	∗CoDAE (BCE)								
∗CAE (BCE)	0.55	0.55	0.55	0.55	0.53	0.65	0.43	0.54	0.39
AE (PixelSum)	0	0	0	0	0	0	0	0	∼0
nTracks	0	0	0	0	0	0	0	∼0	0.99
Supervised CCT	0	0	0	0	0	0	0.99	0.99	0.99
	CoDVAE (Z_1)								
∗CAE (BCE)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
AE (PixelSum)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
nTracks	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Supervised CCT	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	CoDVAE (KL-F)								
∗CAE (BCE)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
AE (PixelSum)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
nTracks	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Supervised CCT	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	CoDVAE (SSE)								
∗CAE (BCE)	0.99	0.99	0.99	1	0.99	0.99	1	1	1
AE (PixelSum)	0	0	0	0	0	0	0	0	0
nTracks	0	0	0	0	0	0	0	0	0.99
Supervised CCT	0.99	0	0	0	0	0	0.99	0.99	0.99

Table 4. Evaluation of reconstruction quality: QCD test images. Lower values of MSE as well higher SSIM ones are better. Each entry denotes the average metric value as well as its standard deviation (in parenthesis.) By design the Dirichlet VAE outputs images that sum to one, so we undo the normalization before computing the metrics in order to have pixel values on the same scale. Best results are shown in boldface.

Metric	CoDAE	CAE	CoDVAE	AE [6]	Dirichlet VAE [18]
MSE	**5.5 (±5.6)**	9.4 (±7.4)	17.7 (±13)	70.9 (±24.7)	79.3 (±26)
SSIM	**0.99 (±0.01)**	0.98 (±0.02)	0.95 (±0.04)	0.36 (±0.12)	0.17 (±0.11)

are complementary since the MSE looks at fine details while the SSIM at the global appearance of the images, providing a more comprehensive assessment of the quality of the reconstructed samples.

Reconstruction performance is summarized in table 4 as well visually in figure 6. From both we can deduce that our two models achieve the lowest MSE and highest structural similarity, attaining accurate single-sample reconstructions resulting in good predictions on average. As we can notice from table 4, all models show some variance in the reconstructed background, indicating that some QCD samples deviate from the ideal background event. Moreover, AE-based models show softer and smoother pixel predictions whereas the categorical CoDVAE, thanks to its Bernoulli decoder, is capable of sharp reconstructions by taking the mode of each learned distribution, one per output pixel. Furthermore, in figure 6 we can observe how the Dirichlet VAE, which is designed to capture a multi-modal latent space distribution, only captures the 'QCD mode' since each sample, regardless of being background or signal, is predicted as a sort of average of QCD images: this is confirmed by the low reconstruction metrics. We noticed a similar behavior in our

(a) Reconstructions of five random QCD samples.

(b) Reconstructions of the QCD test-set.

Figure 6. Comparison of per-sample (a) and average (b) reconstructions. In order: ground-truth, CoDAE, CAE, Categorical CoDAE, AE [6], and Dirichlet VAE [18]. As we can see our models are capable of pixel-accurate single-sample reconstructions, whereas the AE only captures the prominent pixels, and the Dirichlet VAE learns to predict a sort of average image instead. The CAE is also very accurate, as expected, since sharing the same encoder and decoder architecture of the CoDAE.

prior experiments when training regular AEs with a small latent space (e.g. 2), even with a high-capacity residual encoder.

We want to highlight the importance of accurate (or at least coherent) reconstructions. Since AD scores can be defined from such predictions, it is necessary to avoid the model learning to predict some spurious pattern or artifact instead of the inputs, otherwise, it would be difficult to understand for a human expert why a new sample deviates from the training data. Moreover, since an auto-encoder is trained to maximize the reconstruction quality which, in turn, mostly affects the reconstruction-based anomaly scores, it is equally important to avoid premature convergence of the model since this can lead to sub-optimal AD performance: employing a large latent space paired with powerful encoder and decoder networks can mitigate such issue,

as seen for our CoDAE and CoDVAE. This is essential to obtain coherent AD predictions, high discrimination performance, to have a trustworthy model, and even to debug the model itself once deployed.

5.5. Discussion

Throughout our study we showed how auto-encoders can be employed as an effective means to implement AD in HEP analyses. In particular, our models achieve higher background rejection rates meaning that the classified signal is less contaminated with false positives, which, in turn, boosts the statistical significance of actually finding a newly theorized physics signal when analyzing the real data collected during the LHC's runs: for the SUEPs, we can even equal the supervised classifier even if this model still achieves largely superior rejection rates at the target signal efficiency. In addition, our dual encoder architecture, which also learns a smaller network f_m, is particularly suitable for fast AD: the f_m model can predict in just 3ms, enabling more than real-time applications at the HLT, however at the cost of sacrificing accuracy; in principle, such a model can be further optimized to match even tighter latency requirements, for example running on a FPGA hardware like in [40, 64]. Moreover, our models just learn from raw detector images of particle collisions, requiring less or no effort to compute high-level variables (like counting the number of tracks) and objects (e.g. by a particle-based pre-processing), potentially simplifying the whole analysis setup. Lastly, our CAE model can be employed in scenarios in which the auxiliary latent space is not helpful and the latency requirements allow for a forward pass of the full model.

6. Conclusions

We demonstrate the first successful application of (variational) auto-encoders to deploy in the real-time event-triggering stages of experiments like ATLAS [19] and CMS [20] to search for two dark showers models: SUEPs and SVJs. Their discovery can potentially shed new light on the existence of dark matter and novel hidden sectors, which are currently uncovered and undercover at the LHC.

Unlike the common trend in many related works [6, 18, 29, 30, 41], we do not employ a specific particle-based pre-processing of our data, nor low- or high-level features, but instead learn directly from raw images of particle signatures obtained by discretization of the detector response, thus reducing the dependency on the physics model by only assuming tracking information to be relevant for the considered signals: although we discard the calorimeter information, we believe the ECAL and HCAL to be still useful in general, e.g. for searching long-lived particles.

Our models are evaluated against both signals, demonstrating anomaly scores that can identify both. Our CoDAE models aim to adapt to the background samples, potentially allowing us to generalize on whatever novel signal that is diverse from what the model learned about the QCD background. Ideally, it would be possible to train a single model to reject one or more background processes, filtering only the events that resemble a potential new physics signal. Our approach can potentially enable generic physics searches for unknown, new signals from raw images only and with little-to-no assumptions about the physics model. Further research and benchmark datasets would be required to fully accomplish such an important goal.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Luca Anzalone ⓘ https://orcid.org/0000-0002-0399-8836
Simranjit Singh Chhibra ⓘ https://orcid.org/0000-0002-1643-1388
Benedikt Maier ⓘ https://orcid.org/0000-0001-5270-7540
Nadezda Chernyavskaya ⓘ https://orcid.org/0000-0002-2264-2229
Maurizio Pierini ⓘ https://orcid.org/0000-0003-1939-4268

References

[1] Chatrchyan S *et al* 2012 Observation of a New Boson at a mass of 125 GeV with the CMS experiment at the LHC *Phys. Lett.* B **716** 30–61
[2] Baldi P, Sadowski P and Whiteson D 2014 Searching for exotic particles in high-energy physics with deep learning *Nat. Commun.* **5** 4308
[3] Anzalone L, Diotalevi T and Bonacorsi D 2022 Improving parametric neural networks for high-energy physics (and beyond) *Mach. Learn.: Sci. Technol.* **3** 035017

[4] Kasieczka G *et al* 2019 The machine learning landscape of top taggers *SciPost Phys.* **7** 014

[5] Cowan G, Cranmer K, Gross E and Vitells O 2011 Asymptotic formulae for likelihood-based tests of new physics *Eur. Phys. J.* C **71** 1–19

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

[7] Strassler M J and Zurek K M 2007 Echoes of a hidden valley at hadron colliders *Phys. Lett.* B **651** 374–9

[8] Knapen S, Griso S, Papucci M and Robinson D J 2017 Triggering soft bombs at the LHC *J. High Energy Phys.* JHE08(2017)076

[9] Cohen T, Lisanti M and Lou H K 2015 Semivisible jets: dark matter undercover at the LHC *Phys. Rev. Lett.* **115** 171804

[10] Cohen T, Lisanti M, Lou H K and Mishra-Sharma S 2017 LHC searches for dark sector showers *J. High Energy Phys.* JHE11(2017)196

[11] Kar D and Sinha S 2021 Exploring jet substructure in semi-visible jets *SciPost Phys.* **10** 084

[12] Evans L and Bryant P 2008 LHC Machine *J. Instrum.* **3** S08001

[13] Denk T R A 2011 Quirky composite dark matter *Bachelor Thesis* Tech. U., Munich (https://doi.org/10.1103/PhysRevD.81.095001)

[14] Born S, Karur R, Knapen S and Shelton J 2023 Scouting for dark showers at cms and LHCB *Phys. Rev.* D **108** 035034

[15] Khachatryan V *et al* 2017 The CMS trigger system *JINST* **12** 01020

[16] Kasieczka G, Mastandrea R, Mikuni V, Nachman B, Pettee M and Shih D 2023 Anomaly detection under coordinate transformations *Phys. Rev.* D **107** 015009

[17] Dumoulin V, Perez E, Schucher N, Strub F, d. Vries H, Courville A and Bengio Y 2018 Feature-wise transformations *Distill* **3** 7

[18] Dillon B M, Plehn T, Sauer C and Sorrenson P 2021 Better latent spaces for better autoencoders *SciPost Phys.* **11** 061

[19] Collaboration T A and Aad G *et al* 2008 The ATLAS experiment at the CERN large hadron collider *J. Instrum.* **3** S08003

[20] Chatrchyan S *et al* 2008 The CMS experiment at the CERN LHC *J. Instrum.* **3** S08004

[21] Sirunyan A M *et al* 2017 Particle-flow reconstruction and global event description with the cms detector *J. Instrum.* **12** 10003

[22] de Favereau J, Delaere C, Demin P, Giammanco A, Lemaitre V, Mertens A and Selvaggi M 2014 Delphes 3: a modular framework for fast simulation of a generic collider experiment *J. High Energy Phys.* JHE02(2014)057

[23] Sjöstrand T, Ask S, Christiansen J R, Corke R, Desai N, Ilten P, Mrenna S, Prestel S, Rasmussen C O and Skands P Z 2015 An introduction to PYTHIA 8.2 *Comput. Phys. Commun.* **191** 159–77

[24] Chalapathy R and Chawla S 2019 Deep learning for anomaly detection: a survey *CoRR* (arXiv:1901.03407)

[25] Kingma D P and Welling M 2014 Auto-encoding variational bayes *2nd Int. Conf. on Learning Representations, ICLR* pp 14–16 (arXiv:1312.6114)

[26] Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks *Science* **313** 504–7

[27] Masci J, Meier U, Ciresan D C and Schmidhuber J 2011 Stacked convolutional auto-encoders for hierarchical feature extraction *Artificial Neural Networks and Machine Learning - ICANN* (Springer) pp 52–59

[28] Fraser K, Homiller S, Mishra R K, Ostdiek B and Schwartz M D 2022 Challenges for unsupervised anomaly detection in particle physics *J. High. Energy Phys.* JHE03(2022)066

[29] Finke T, Krämer M, Morandini A, Mück A and Oleksiyuk I 2021 Autoencoders for unsupervised anomaly detection in high energy physics *J. High. Energy Phys.* JHE06(2021)161

[30] Dillon B M, Favaro L, Plehn T, Sorrenson P and Krämer M 2023 A normalized autoencoder for lhc triggers *SciPost Phys. Core* **6** 074

[31] Yoon S, Noh Y and Park F C 2021 Autoencoding under normalization constraints *Int. Conf. on Machine Learning, ICML* vol 139 (PMLR) pp 12087–97 (available at: https://proceedings.mlr.press/v139/yoon21c.html)

[32] Metodiev E M, Nachman B and Thaler J 2017 Classification without labels: learning from mixed samples in high energy physics *J. High Energy Phys.* JHE10(2017)174

[33] Hendrycks D, Mazeika M and Dietterich T G 2019 Deep anomaly detection with outlier exposure *7th Int. Conf. on Learning Representations, ICLR* (https://doi.org/10.48550/arXiv.1812.04606) (OpenReview.net)

[34] Joo W, Lee W, Park S and Moon I 2020 Dirichlet variational autoencoder *Pattern Recognit.* **107** 107514

[35] Dilokthanakul N, Mediano P A M, Garnelo M, Lee M C H, Salimbeni H, Arulkumaran K and Shanahan M 2016 Deep unsupervised clustering with Gaussian mixture variational autoencoders *CoRR* (arXiv:1611.02648)

[36] Bortolato B, Smolkovič A, Dillon B M and Kamenik J F 2022 Bump hunting in latent space *Phys. Rev.* D **105** 115009

[37] Dillon B M, Mastandrea R and Nachman B 2022 Self-supervised anomaly detection for new physics *Phys. Rev.* D **106** 056005

[38] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* pp 5998–6008 (available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

[39] Dillon B M, Kasieczka G, Olischlager H, Plehn T, Sorrenson P and Vogel L 2022 Symmetries, safety and self-supervision *SciPost Phys.* **12** 188

[40] Govorkova E *et al* 2022 Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider *Nat. Mach. Intell.* **4** 154–61

[41] Cheng T, Arguin J-F, Leissner-Martin J, Pilette J and Golling T 2023 Variational autoencoders for anomalous jet tagging *Phys. Rev.* D **107** 016002

[42] Barron J, Curtin D, Kasieczka G, Plehn T and Spourdalakis A 2021 Unsupervised hadronic suep at the lhc *J. High Energy Phys.* JHEP12(2021) 129

[43] Canelli F, de Cosa A, Le Pottier L, Niedziela J, Pedro K and Pierini M 2022 Autoencoders for semivisible jet detection *J. High Energy Phys.* JHEP02(2022) 74

[44] Shorten C and Khoshgoftaar T M 2019 A survey on image data augmentation for deep learning *J. Big Data* **6** 60

[45] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *2016 IEEE Conf. on Computer Vision and Pattern Recognition, CVPR* (IEEE Computer Society) pp 770–8

[46] Mirza M and Osindero S 2014 Conditional generative adversarial nets *CoRR* (arXiv:1411.1784)

[47] Ulyanov D, Vedaldi A and Lempitsky V S 2016 Instance normalization: the missing ingredient for fast stylization *CoRR* (arXiv:1607.08022)

[48] Maas A L, Hannun A Y and Ng A Y 2013 Rectifier nonlinearities improve neural network acoustic models *Proc. ICML (Atlanta, Georgia, USA)* vol 30 p 3 (available at: http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf)

[49] Watters N, Matthey L, Burgess C P and Lerchner A 2019 Spatial broadcast decoder: a simple architecture for learning disentangled representations in vaes *CoRR* (arXiv:1901.07017)

[50] Maddison C J, Mnih A and Teh Y W 2017 The concrete distribution: a continuous relaxation of discrete random variables *5th Int. Conf. on Learning Representations, ICLR* (https://doi.org/10.48550/arXiv.1611.00712) (OpenReview.net)

[51] Jang E, Gu S and Poole B 2017 Categorical reparameterization with gumbel-softmax *5th Int. Conf. on Learning Representations, ICLR* (https://doi.org/10.48550/arXiv.1611.01144) (OpenReview.net)

[52] Havrylov S and Titov I 2017 Emergence of language with multi-agent games: Learning to communicate with sequences of symbols *Advances in Neural Information Processing Systems, NIPS* pp 2149–59 (available at: https://proceedings.neurips.cc/paper/2017/file/70222949cc0db89ab32c9969754d4758-Paper.pdf)

[53] Yan S, Smith J S, Lu W and Zhang B 2018 Hierarchical multi-scale attention networks for action recognition *Signal Process. Image Commun.* **61** 73–84

[54] Deng R, Shen C, Liu S, Wang H and Liu X 2018 Learning to predict crisp boundaries *Proc. European Conf. on Computer Vision (ECCV)* pp 562–78

[55] Wang Z, Bovik A C, Sheikh H R and Simoncelli E P 2004 Image quality assessment: from error visibility to structural similarity *IEEE Trans. Image Process.* **13** 600–12

[56] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *3rd Int. Conf. on Learning Representations, ICLR* (https://doi.org/10.48550/arXiv.1412.6980)

[57] Loshchilov I and Hutter F 2019 Decoupled weight decay regularization *7th Int. Conf. on Learning Representations, ICLR* (https://doi.org/10.48550/arXiv.1711.05101) (OpenReview.net)

[58] He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on imagenet classification *IEEE Int. Conf. on Computer Vision, ICCV* (IEEE Computer Society) pp 1026–34

[59] Abadi M *et al* 2016 Tensorflow: a system for large-scale machine learning *12th USENIX Symp. on Operating Systems Design and Implementation, OSDI* (USENIX Association) pp 265–83 (available at: https://dl.acm.org/doi/10.5555/3026877.3026899)

[60] Del Barrio E, Cuesta-Albertos J A and Matrán C 2018 An optimal transportation approach for assessing almost stochastic order *The Mathematics of the Uncertain* (Springer) pp 33–44

[61] Dror R, Shlomov S and Reichart R 2019 Deep dominance—how to properly compare deep neural models *Proc. of the 57th Conf. of the Association for Computational Linguistics, ACL* 1, ed A Korhonen, D R Traum and L Màrquez (Association for Computational Linguistics) pp 2773–85

[62] Ulmer D, Hardmeier C and Frellsen J 2022 Deep-significance-easy and meaningful statistical significance testing in the age of neural networks (arXiv:2204.06815)

[63] Hassani A, Walton S, Shah N, Abuduweili A, Li J and Shi H 2021 Escaping the big data paradigm with compact transformers *CoRR* (arXiv:2104.05704)

[64] Valente L, Anzalone L, Lorusso M and Bonacorsi D 2023 Joint variational auto-encoder for anomaly detection in high energy physics *Int. Symp. on Grids and Clouds (ISGC)* vol 19 (https://doi.org/10.22323/1.434.0014)