

Anomaly-preserving contrastive neural embeddings for end-to-end model-independent searches at the LHC

Kyle Metzger^{1,*}, Lana Xu,² Mia Sodini,² Thea K. Årrestad^{1,†}, Katya Govorkova^{2,‡},
Gaia Grosso^{2,3,4,§} and Philip Harris^{2,3,||}

¹*ETH Zürich, Zürich, Switzerland*

²*Massachusetts Institute of Technology, Cambridge, Massachusetts, USA*

³*NSF AI Institute for Artificial Intelligence and Fundamental Interactions,
Cambridge, Massachusetts, USA*

⁴*School of Engineering and Applied Sciences, Harvard University,
Cambridge, Massachusetts, USA*



(Received 21 March 2025; accepted 5 September 2025; published 23 October 2025)

Anomaly detection—identifying deviations from Standard Model predictions—is a key challenge at the Large Hadron Collider due to the size and complexity of its datasets. This is typically addressed by transforming high-dimensional detector data into lower-dimensional, physically meaningful features. We tackle feature extraction for anomaly detection by learning powerful low-dimensional representations via contrastive neural embeddings. This approach preserves potential anomalies indicative of new physics and enables rare signal extraction using novel machine-learning-based statistical methods for signal-independent hypothesis testing. We compare supervised and self-supervised contrastive learning methods, for both Multilayer perceptron- and transformer-based neural embeddings, trained on the kinematic observables of physics objects in LHC collision events. The learned embeddings serve as input representations for signal-agnostic statistical detection methods in inclusive final states. We achieve significant improvement in discovery power for both rare new physics signals and rare Standard Model processes across diverse final states, demonstrating its applicability for efficiently searching for diverse signals simultaneously. We study the impact of architectural choices, contrastive loss formulations, supervision levels, and embedding dimensionality on anomaly detection performance. We show that the optimal representation for background classification does not always maximize sensitivity to new physics signals, revealing an inherent trade-off between background structure preservation and anomaly enhancement. We demonstrate that combining compression with domain knowledge for label encoding produces the most effective data representation for statistical discovery of anomalies.

DOI: [10.1103/5n77-ynsp](https://doi.org/10.1103/5n77-ynsp)

I. INTRODUCTION

The Large Hadron Collider (LHC) at CERN is the world's largest particle accelerator, designed to explore the fundamental properties of matter by colliding protons and heavy ions at high energies. The LHC allows physicists to probe the deep structure of the Universe, searching for new particles and phenomena that could extend our

understanding of the Standard Model of particle physics. Anomaly detection in high-energy physics is crucial for identifying events that might indicate the presence of new unexpected phenomena beyond the ones described by the Standard Model. Given the immense volume of data generated by the LHC, advanced techniques for anomaly detection are essential for sifting through the data to find rare signals that could point to ground-breaking discoveries. Such methods must be highly sensitive and capable of distinguishing true anomalies from background noise, making them a powerful tool in the ongoing search for new fundamental particles and forces.

A crucial challenge for anomaly detection on LHC data is the step of *feature learning*, constructing a low-dimensional set of informative variables from the high-dimensional set of observables that describes each collision event. Raw collision data at the LHC consist of millions of sensor measurements, which are processed using clustering algorithms to reconstruct individual particles. The data are then

*Contact author: kyle.metzger@bluewin.ch

†Contact author: thea.aarrestad@cern.ch

‡Contact author: ekaterina.govorkova@cern.ch

§Contact author: gaia.grosso@cern.ch

||Contact author: philip.coleman.harris@cern.ch

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

further reduced to physically meaningful summary statistics, such as particle energy, momentum, mass, and direction. These summary statistics serve as inputs to data analyses aimed at detecting deviations from expected backgrounds, which in turn could hint at new physics. The process of transforming high-dimensional detector data into compact, but informative representations is therefore well established in high-energy physics. Machine learning (ML) has demonstrated high effectiveness in extracting powerful representations from high-dimensional data [1]. Such neural embeddings can similarly be leveraged for data compression, mapping raw data to lower-dimensional spaces that retain essential features for downstream tasks [2]. In this work, we study whether neural embeddings, when trained appropriately, can outperform classical feature representations commonly used in LHC data analysis, particularly in the context of anomaly detection.

When solving supervised problems, the information loss due to feature reduction can be curated to minimally impact the specific objective of the training task. Conversely, for semisupervised tasks, like anomaly detection at the LHC, the absence of *a priori* knowledge about the signal increases the risk of performing a disruptive feature reduction.

Contrastive learning, when used for feature reduction in the context of anomaly detection, offers a compelling approach to learn a meaningful form of data organization in an unsupervised or semisupervised fashion. In the context of anomaly detection, the goal of contrastive learning is to create a latent space where common data points are grouped closely together. Meanwhile, anomalies, which are points that deviate from the bulk of the data, are pushed further apart. By training the model to maximize the similarity between representations of normal data and minimize it for data that are distinct or unusual, the model learns to highlight the features that distinguish anomalies from the regular pattern. This process is particularly advantageous in high-dimensional and complex datasets, such as those encountered in high-energy physics or other scientific domains, where anomalies may be subtle or rare. Once the latent space is structured in this way, unsupervised clustering techniques can be employed to further isolate and identify potential anomalies. In this context, contrastive learning helps construct well-defined clusters in the latent space, easing the tasks of outliers or statistical anomaly detection.

This paper presents a comprehensive study focused on the development and evaluation of neural embeddings for anomaly detection, particularly in the context of high-energy physics. Section II provides an overview of existing methods and approaches in anomaly detection and contrastive learning. In Sec. III, we go into the details of how different models and loss functions are employed to pretrain embeddings that can effectively represent the underlying data structure. In Sec. IV, we apply these

pretrained embeddings to the task of identifying anomalies, with a focus on evaluating how well the learned latent spaces can separate anomalous events from the normal data distribution. In Sec. VI, we analyze the results, comparing the performance of different training strategies and architectures and discussing the implications of our findings for future research in both machine learning and high-energy physics. Finally, Sec. VII summarizes the key findings from our work and outlines potential directions for further investigation, emphasizing the importance of robust anomaly detection techniques in advancing our understanding of complex LHC datasets.

II. RELATED WORK

Machine learning has long been an integral component of anomaly detection in high-energy physics. Traditionally, anomaly detection in this field has relied on the formulation of a theory model for the signal, used by physicists to define specific criteria or regions of interest within the data space (signal-dependent searches). The advent of machine learning has enabled the development of flexible methods, capable of learning optimal signal to background multidimensional discriminants directly from data and simulation. These approaches mainly include supervised learning techniques, where models are trained on labeled datasets to distinguish between normal and prespecified anomalous events. These methods, while effective in targeted scenarios, are limited by their dependency on predefined signal models.

On the other hand, looking for new physics in the absence of a signal model (signal-agnostic searches) is extremely hard due to the lack of information on how to build relevant summary statistics. Following the steps of the D0 and CDF Collaborations at Tevatron and the H1 Collaboration at HERA [3–9], both the ATLAS and CMS experiments at CERN have developed strategies to scan the experimental phase space searching for anomalies prior to the advent of deep learning [10,11]. In recent years, semisupervised, weakly supervised, and unsupervised learning methods have emerged as powerful tools to enhance these approaches. Deep learning techniques have shown great promise in identifying anomalies in complex, high-dimensional datasets and have been recently deployed in both the CMS and ATLAS experiments [12–16]. Techniques such as autoencoders, generative adversarial networks, and contrastive learning have been explored for their ability to learn meaningful representations of data and detect deviations from expected patterns, without the need of strict *a priori* assumptions on the nature of the anomalous signal (see [17] for a review on the topic). Despite the recent advancements, there remain significant challenges—particularly, the interpretability of these ML methods, as well as their practical application in the real-world context of LHC experiments, that requires the handling of vast and complex datasets. Challenges in this

field include (i) mitigating sensitivity degradation caused by the curse of dimensionality [18,19], (ii) speeding up training schemes to handle extremely large sample sizes [20,21], and (iii) uncertainty quantification [22,23].

In this work we tackle the first of the aforementioned challenges by exploring neural embeddings for anomaly-preserving data compression using contrastive learning techniques. Several recent results in particle physics have highlighted the success of contrastive learning and neural embedding strategies to train models capable of performing various tasks [24–26] and foundation models providing data representations suitable for many tasks at one [27–31]. Focusing on anomaly detection applications, Refs. [32–34] showed that pretrained foundation models can enhance the performance of weakly supervised classifiers for anomaly detection. In Ref. [35], physics-informed self-supervised strategies are used to define a contrastive space for anomaly detection. Whereas, in Ref. [36], self-supervised strategies are explored in the context of real-time anomaly detection in jets as a way to perform feature extraction for an autoencoder, in this work, we take a further step by exploring the application of contrastive learning methods to a different data representation, namely an event-level representation of proton-proton collisions at the LHC, where inclusive final states are characterized by the 3-momenta of particles reconstructed from the collision products. Furthermore, we compare two different contrastive approaches, supervised and self-supervised, and we demonstrate that supervised contrastive learning yields significantly greater performance gains compared to self-supervised approaches, as it leverages available domain knowledge to guide the organization of the representation space. To the best of our knowledge, this study presents the first application of supervised contrastive learning to high-energy physics data and provides the first evidence that the availability of labeled background data can effectively steer discovery within unlabeled datasets. Moreover, whereas weakly supervised methods rely on defining a signal sideband and thus incorporate some signal assumptions, we propose an end-to-end, fully signal-agnostic technique applicable to generic final states.

III. CONTRASTIVE-BASED NEURAL EMBEDDINGS

Meaningful data representations play an essential part in how performant the ML algorithm is [37]. Therefore, a lot of emphasis has been put on processing data before giving it to an ML model. Representation learning describes the extraction of features using deep learning methods. A neural network ϕ_θ parametrized by a set of trainable parameters θ maps the events from an original high-dimensional space $\mathcal{X}$ to a lower-dimensional latent space $\mathcal{Y}$. This mapping is formalized as

$$\phi_\theta: (\mathcal{X}, d_\mathcal{X}) \xrightarrow{\phi} (\mathcal{Y}, d_\mathcal{Y}), \quad (1)$$

where $d_\mathcal{X}$ and $d_\mathcal{Y}$ are metrics in the original space and the latent space, respectively. The learned representation can then be used for downstream tasks, such as classification. While representation learning is usually accompanied by dimensionality reduction, the defining aspect is the extraction of features that best describe the input data with regards to the downstream task. This step is often referred to as *feature learning*. When the downstream task is signal-agnostic anomaly detection, the optimal representation for the signal cannot be known *a priori* and, ultimately, it would depend on the composition of the real data, namely which signal, if any, is present.

The approaches to learn neural embeddings considered in this work are based on contrastive learning, originally designed to solve problems of computer vision. The methods are constructed according to the Siamese principle, where the input to a weight-sharing model consists of two (or more) views of the same data point or, in this case, physics events [38]. The outputs are subsequently compared by a measure of similarity, such that similar instances are pulled together in the latent space, whereas dissimilar ones are pushed apart.

The real LHC collision dataset is several orders of magnitude larger than its simulated counterpart. As a result, it may be valuable to pretrain neural embeddings in a self-supervised manner using unlabeled data. In a self-supervised setting the views of an event are constructed using *augmentations*, namely applying to the event representation a set of transformations that leave the physics nature of the event invariant. A collision event is augmented twice and the resulting views are grouped together in the latent space by the loss objective. The rest of the events in the batch are seen as “negative” examples, and therefore they are pushed apart from the pair of augmented views. As a consequence, even when they represent the same underlying physics, different collision events are pushed apart. This is potentially a problem, as we would like to build a neural embedding where data are organized according to “concepts,” namely the underlying physics underlying the events. One way to circumvent this problem is adopting the supervised extension of the self-supervised approach outlined above, where similar instances (e.g., “positive” examples) are defined using their physics process labels. This variant is referred in the literature as *supervised contrastive learning* [39]. It should be noticed that this approach can only be applied if the physics labels are available, namely with simulated datasets, thus preventing the use of large-size real LHC data in the training stage. Given the agnostic nature of the downstream task considered in this work, our trainings in this setup will rely only on simulated data of the known background physics processes, endowed with labels.

The dynamics of contrastive learning rely on the interplay of attractive forces and repulsive forces, subject to the constraints set by a form of regularization. We consider two

loss functions serving this purpose: the SimCLR loss [40] and the VICReg loss [41]. The two loss functions mainly differ in how they approach the issue of mode collapse, namely when the inputs are mapped to a constant. SimCLR uses a contrastive loss based on the cosine similarity metric $\text{sim}(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\| \|x_j\|}$, resulting in

$$l_{\text{SimCLR}}(x_i, x_j) = -\log \frac{e^{\text{sim}(\phi_\theta(x_i), \phi_\theta(x_j))/\tau}}{\sum_{k=1}^N e^{\text{sim}(\phi_\theta(x_i), \phi_\theta(x_k))/\tau}}. \quad (2)$$

The loss temperature τ is a tunable hyperparameter for regularization. In Eq. (2), the sum at the denominator includes all possible pairs of the x_i example with the N examples in the batch. The presence of negative pairs act against mode collapse, eventually preventing it if the size of the batch is large enough. On the other hand, the similarity metric in VICReg has three terms:

$$\begin{aligned} l_{\text{VICReg}}(x_i, x_j) = & \lambda s(\phi_\theta(x_i), \phi_\theta(x_j)) \\ & + \mu [v(\phi_\theta(x_i)) + v(\phi_\theta(x_j))] \\ & + \nu [c(\phi_\theta(x_i)) + c(\phi_\theta(x_j))]. \end{aligned} \quad (3)$$

The first term, s , is a mean squared error measure that encourages invariance between positive pairs, whereas the variance term v and covariance term c act as regularizers against mode collapse. The variance term holds the standard deviation along the batch size for every feature above a certain threshold; the covariance term decorrelates the embedding variables between pairs along the batch size. The hyperparameters λ , μ and ν determine the relative contribution of each term to the overall loss function and are optimized via grid search. In both cases the similarity metric is averaged across all positive pairs $P(i)$ of the i th example in the batch and summed over all of the N training examples in the batch:

$$L_{\text{CL}}[\phi_\theta] = \sum_{i=1}^N \frac{1}{|P(i)|} \sum_{x \in P(i)} l_{\text{CL}}(x_i, x) \quad (4)$$

with $\text{CL} \in [\text{SimCLR}, \text{VICReg}]$. In the self-supervised case, each example x_i in the minibatch has exactly one positive pair, denoted $x_{j(i)}$, which corresponds to the other augmented view generated from the same underlying event. Accordingly, the set of positive pairs is defined as

$$P(i) := \{x_{j(i)}\}.$$

In the supervised extension, all instances x within the minibatch with the same class labels c as x_i are considered positive pairs to x_i , such that

$$P(i) := \{x | c_x = c_{x_i}\}.$$

IV. ANOMALY DETECTION ON NEURAL EMBEDDINGS

In this work, we consider the detection of anomalous signal events as the downstream task of the feature learning step. A sample of experimental data is passed through the neural embedding map and its latent representation is analyzed to detect and statistically quantify data departures from their expected behavior. We assume that the expected behavior is faithfully represented by means of a mixture of simulated background processes (B) depicting the Standard Model of particle physics. In realistic scenarios, the model representing the background processes is not known exactly but rather affected by systematic uncertainties that can impact on the robustness of the anomaly detection task. While a way to incorporate systematic uncertainties for the method considered in this work has been proposed [22], in this work we do not address this problem and focus on the impact of feature learning on the anomaly detection task. We leave the study of systematic uncertainties to future work. We assume that the signal (S), if present, manifests itself as a mild overdensity in the data density distribution. The exact location and shape of the signal contribution is *a priori* unknown, giving rise to the challenging problem of engineering the right statistical test for novelty discovery. In a d -dimensional latent space, the density of the embedded data y under their true (T) generative model can be modeled as the sum of the signal and background components:

$$n(y|T) = n(y|B) + n(y|S). \quad (5)$$

The aim of the anomaly detection task boils down to recognizing the presence of signal events and estimating their significance, in other words, determining $n(y|S)$ to maximize the likelihood of the observed data.

To address this problem, we consider the signal-agnostic Neyman-Pearson goodness-of-fit test performed by the new physics learning machine (NPLM) algorithm [20,42].

NPLM is a semisupervised approach to agnostic signal detection that aims to quantify the statistical significance of an observed deviation in the data [43]. Inspired by the Neyman-Pearson approach to hypothesis testing based on the log-likelihood ratio [44], NPLM compares the background hypothesis B and an unknown alternative H_W , which is learned from the data $\mathcal{D}$ in a machine-learning fashion. A model $f_w(y)$ with trainable parameters w is defined on the space of the data y to parametrize the density distribution of the data in the family of alternatives

$$n(y|H_W) = n(y|B) e^{f_w(y)}. \quad (6)$$

Since the total number of events in the experimental dataset from the LHC is a Poissonian random variable whose expectation depends on the theory model, the Neyman-Pearson test statistic is computed as the extended log-likelihood ratio, which in terms of f_w is written as

$$t(\mathcal{D}) = 2 \max_{\mathbf{w}} \left[N(\mathbf{B}) - N(\mathbf{H}_{\mathbf{w}}) + \sum_{x \in \mathcal{D}} f_{\mathbf{w}}(y) \right], \quad (7)$$

with $N(\mathbf{B})$ and $N(\mathbf{H}_{\mathbf{w}})$ the expectation values of the total number of events in $\mathcal{D}$ under the reference and the alternative hypotheses, respectively. The maximization problem described by Eq. (7) can be solved by a machine-learning task that learns the optimal values of the trainable parameters of f , $\hat{\mathbf{w}}$, to approximate the log-density ratio of the two samples:

$$f_{\hat{\mathbf{w}}}(y) = \log \frac{n(y|\mathbf{H}_{\hat{\mathbf{w}}})}{n(y|\mathbf{R})} \approx \log \frac{n(y|\mathbf{T})}{n(y|\mathbf{R})}. \quad (8)$$

In this work we adopt the fast version of the NPLM algorithm originally introduced in [20] which relies on the FALKON package [45]. The latter is based on kernel methods whose coefficients $\mathbf{w}$ are trained with an L2 regularized weighted binary cross-entropy loss function

$$L_{\text{FLK}}[f_{\mathbf{w}}] = \sum_{y \in \mathcal{R}} w_{\mathcal{R}} \log(1 + e^{f_{\mathbf{w}}(y)}) - \sum_{y \in \mathcal{D}} \log(1 + e^{-f_{\mathbf{w}}(y)}) + \lambda \sum \|\mathbf{w}\|^2, \quad (9)$$

where the weight $w_{\mathcal{R}}$ is introduced to match the size of sample of reference $\mathcal{R}$ with the expected experimental size in absence of signal, $N(\mathbf{B})$. The test statistic is then computed as in Eq. (7), using the learned model $f_{\hat{\mathbf{w}}}$.

As for any test statistic, the value of the test obtained for $\mathcal{D}$ has to be compared to the distribution of the test under the background hypothesis, $p(t|\mathbf{R})$, and the p value is finally used as a metric of significance. In the results presented in Sec. V, $p(t|\mathbf{R})$ is estimated both empirically, using 1000 pseudoexperiments, and analytically relying on the asymptotic χ^2 behavior of the test. To help the interpretation, p values resulting from our experiments will be reported in their standardized version of Z scores, defined as the quantile Φ^{-1} of the normal distribution at the p complement to 1:

$$Z_p = \Phi^{-1}(1 - p). \quad (10)$$

V. NEURAL EMBEDDINGS ON LHC DATA

In this section we apply the representation learning approaches outlined in Sec. III to a synthetic dataset representing realistic LHC collision events. We train various neural embeddings with different learning schemes in a signal-agnostic way, namely only using background data. We evaluate the quality of their representation as a backbone for both the classification of background physics processes and for the detection of anomalies. For the first task, we use a simple one-layer linear evaluation as a downstream task. For the second problem, we consider

the anomaly detection approach described in Sec. IV. For both tasks we compare the performances among different learning strategies and rates of compression.

A. Dataset

We evaluate our models on the *ADC2021* dataset, specifically designed for anomaly detection in high-energy particle physics [46]. This dataset emulates realistic proton-proton collisions at the LHC with a center-of-mass energy of 13 TeV, applying an event selection criterion that mimics an online filter, requiring at least one electron (e) or muon (μ) with transverse momentum $p_T > 23$ GeV. Standard Model background processes are generated using PYTHIA [47], with their detector responses simulated via DELPHES [48]. For each collision event at most 19 objects are reconstructed: the missing transverse energy, the four most energetic electrons and muons and the ten most energetic jets. For each object the kinematics variables (p_T, η, ϕ) are provided. Data are provided in a 19×3 tabular format where zero padding is used for events with less than 19 reconstructed objects. The dataset is partitioned into training and validation sets using fivefold cross-validation. During neural embedding training, each fold serves as the validation set once, while the remaining four folds are combined to form the training set. Additionally, 20% of the Standard Model background processes are withheld from the embedding training and reserved as the test dataset for anomaly detection. The Standard Model events are labeled according to the underlying physics process generating them. The dataset contains four Standard Model (background) classes in different proportions representing the expected relative abundance: W boson production (59.2%), Z boson production (6.7%), $t\bar{t}$ production (0.3%) and QCD multijet production (33.8%). Alongside the Standard Model dataset, four new physics scenarios are also simulated and used for evaluating the anomaly detection performance. These include a neutral scalar boson (A) with a mass of 50 GeV decaying to two off-shell Z bosons that further decay to leptons ($A \rightarrow 4l$), a leptoquark (LQ) with a mass of 80 GeV decaying into a b quark and a τ lepton, a scalar boson h^0 with a mass of 60 GeV decaying to two τ leptons ($h^0 \rightarrow \tau\tau$), and a charged scalar boson $h^\pm$ with a mass of 60 GeV decaying to a τ lepton and a neutrino ($h^\pm \rightarrow \tau\nu$). At test time, these are injected into the Standard Model test dataset such that they make up a fraction ranging from 0.1% to 1% of the total dataset size.

B. Embedding models

We trained various contrastive-based neural embeddings, varying one of the following aspects at a time: the model architecture [Multilayer perceptron (MLP)- or transformer-based], the loss function (SimCLR or VICReg), the training scheme (supervised or self-supervised), and the input data

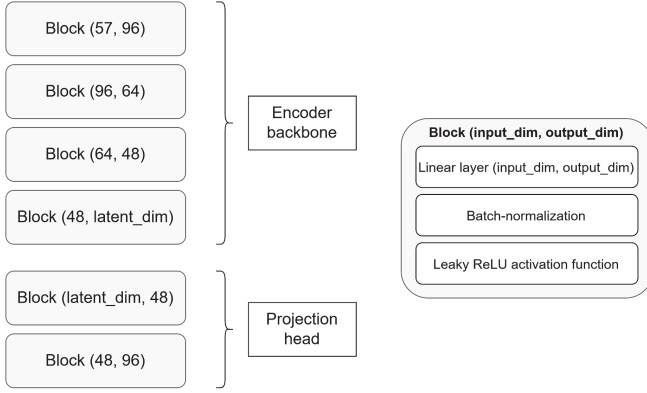


FIG. 1. MLP-based neural network architecture used for training the supervised and self-supervised embeddings with SimCLR. Illustration made with Ref. [49].

composition (balanced background classes or a realistic background “cocktail”).

With SimCLR loss we trained both MLP- and transformer-based architectures and a realistic composition of input background data. The network architecture used for both the supervised and self-supervised MLP-based models is depicted in Fig. 1. On the right side of the figure, we show the main building block of the architecture, namely a batch-normalized linear layer activated by leaky ReLU. On the left side we represent the backbone encoder and the projection head used for training the neural embeddings. The former is composed of four blocks, whereas the latter is composed of two blocks. As per [40], adding a projection head during training leads to improved embeddings. For inference (i.e., background classification and anomaly detection), the projection head is dropped.

In contrast, the transformer-based supervised SimCLR implementation employs a standard transformer encoder [50,51] (see dimensions in Table I), and its projection head is based on the distillation with no labels (DINO) head [52] with parameters detailed in Table II. The 19 particles in the source representation are input as 19 different tokens with the kinematic variables (p_T, η, ϕ) as features. In this architecture, an additional classification token is used to aggregate information from all the tokens before feeding it to the projection head. The training hyperparameters are summarized in Table III. The learning rate follows a linear warmup for the first ten epochs and a subsequent cosine decay to 1/1000 of the base learning rate. The MLP-based training uses the Layer-wise Adaptive Rate Scaling (LARS) optimizer [53], while the transformer-based training utilizes the Adam optimizer [54] for applying the gradient updates.

The SimCLR loss objective allows tuning for using the temperature parameter τ in Eq. (2). While in the supervised implementation a temperature parameter of $\tau = 0.1$ works well for both architectures, in the self-supervised implementation the optimal temperature depends of the latent space dimension D . The temperatures used in this case are reported in Table IV. It should be noted that an optimal choice of hyperparameters is not generally possible for anomaly detection tasks, since the target is not known *a priori*. In this work we optimize the neural embedding modeling using background classification accuracy as a figure of merit (more details in Sec. V C). While this is not guaranteed to be optimal in the broader sense of anomaly detection we consider the background classification task to be closely related to the anomaly detection task and therefore provide a set of near to optimal hyperparameters.

TABLE I. Model parameters for the vanilla transformer encoder [50,51] used as the backbone in the supervised SimCLR training.

Model dimension	No. attention heads	Feedforward dimension	No. layers	Positional encoding	Masking	Aggregate strategy	Dropout
64	8	256	4	True	False	CLS-token	0.025

TABLE II. Model parameters for the DINO head [52] used as the projection head.

Output dimension	Hidden dimension	Bottleneck dimension	No. layers	Use batchnorm	Normalize last layer
64	256	64	3	False	False

TABLE III. Training hyperparameters for the contrastive representation.

Method	Batch size	Epochs	Learning rate	Weight decay
Self-supervised SimCLR MLP	1024	50	0.8	10^{-6}
Supervised SimCLR MLP	1024	50	1.2	10^{-6}
Supervised SimCLR transformer	1024	50	0.001	10^{-6}
Supervised VICReg transformer	64	50	0.001	10^{-2}

TABLE IV. Loss temperatures (τ) and latent space dimension ($\mathbf{D}$) for different models trained with self-supervised SimCLR.

$\mathbf{D}$	4	6	8	16	32	64
τ	0.9	0.8	0.6	0.5	0.4	0.3

TABLE V. Number of heads (h) and latent space dimension ($\mathbf{D}$) for different models trained with supervised VICReg.

$\mathbf{D}$	2	4	8	16	24	32
h	2	4	4	8	8	16

With VICReg loss, we focused on transformer-based architectures [50]. Each transformer block is made of MLPs with two layers and 16 expansions and attention layers with a number of heads (h) that varies with the latent space dimension $\mathbf{D}$ (see Table V). The model consists of four transformer blocks stacked together. The batch size is 64, and the learning rate follows a CosineAnnealing schedule with initial value of 0.001 and weight decay of 0.01 (values reported in Table III). The loss function used for the training is a combination of the VICReg loss in Eq. (3) and a cross entropy (CE) loss, and the relative contribution of the two terms is regulated by a scale parameter α :

$$L = \alpha L_{\text{VICReg}} + (1 - \alpha) L_{\text{CE}}. \quad (11)$$

The scale parameter α , as well as the batch size, number of layers, expansions, learning rate and weight decay are selected after a grid search is performed to optimize the accuracy of the background classification obtained by the model.

1. Augmentations for self-supervised contrastive learning

In order to achieve meaningful feature compression with self-supervised learning a suitable augmentation pipeline is crucial. We explored domain-specific augmentations for particle physics, where the underlying physics should remain unchanged under each transformation: naive masking, where random features are zero masked with probability p ; Gaussian resampling perturbed features with a normal distribution; a variant of Gaussian resampling involving only the transverse momentum (p_T); rotation around the beam line randomly rotated events by $\delta\phi \in [0, 2\pi)$, enforcing rotational invariance; detector cropping retained only constituents within a radial cutoff ($R^* = 3$ was found to be the value performing the best on the background classification task used for evaluation). Despite these well-motivated variations, naive masking with probability $p = 0.5$ consistently yielded the best representations and was adopted as the standard augmentation. This augmentation strategy is inspired by the masking techniques used in language models training and can be viewed as a dropout technique used to increase the model's robustness to corrupted inputs. We define

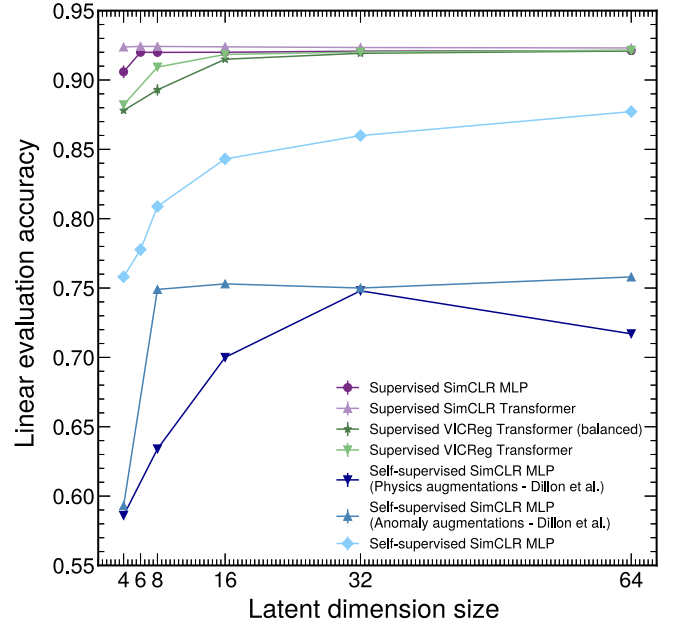


FIG. 2. Linear evaluation accuracy on background classes for the MLP- and transformer-based models trained supervised with a SimCLR loss (dark and light purple, respectively), the transformer-based model trained supervised with a VICReg loss using either balanced background classes (dark green) or background classes weighted to match the composition expected in data (light green), the MLP-based model trained self-supervised with a SimCLR loss with 50% random masking (light blue), physics-inspired augmentations from [26] (medium blue), or physics- and anomaly-inspired augmentations from [26] (dark blue).

positive pairs as the original collision event and its augmented views while the rest of the events within a batch are treated as negative pairs. It is worth noticing that the adopted masking strategy modifies the overall kinematics of the event producing a completely different collision outcome as a view. The latter could be nonphysical or populate a region of the phase space that is not allowed by the original physics process. The augmented view could also resemble events coming from a different background process or an anomaly, with a potentially negative impact on the model performances.

In addition to the augmentations defined for our experiments, we consider the two sets of physics-inspired and anomaly-inspired augmentations introduced in [36] (and defined in [55]). As in [36], we perform two sets of experiments, the first one using only the physics-inspired augmentations and in the second one combining physics-inspired and anomaly-inspired augmentations. The effect of self-supervised approaches on both background classification and anomaly detection performance is discussed in Sec. V D.

C. Baseline comparison of neural embeddings

After pretraining the different neural encoders, the learned latent space is used as an input to a simple linear

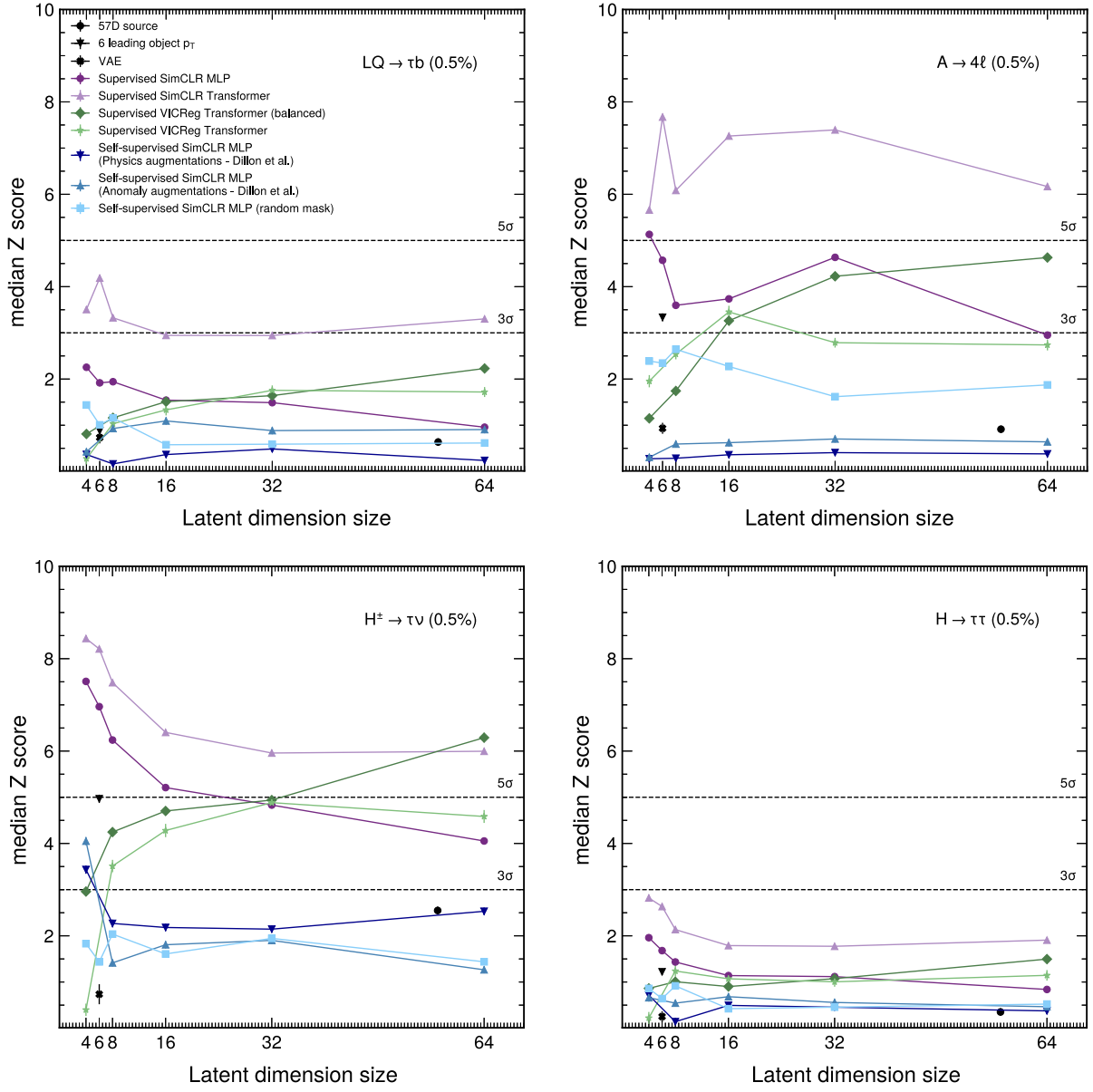


FIG. 3. The median observed Z score as a function of the feature embedding size after injecting a signal into the Standard Model (SM) background pseudodataset, corresponding to 0.5% of the total integral. Results are shown for the $LQ \rightarrow \tau b$ (upper left), $A \rightarrow 4\ell$ (upper right), $H^\pm \rightarrow \tau \nu$ (bottom left), and $H \rightarrow \tau \tau$ (bottom right) signals. The median Z score is presented for different embedding models: the MLP- and transformer-based models trained with supervision and SimCLR loss (dark and light purple, respectively), the transformer-based model trained with supervision and VICReg loss using either balanced background classes (dark green) or background classes weighted to match the composition expected in data (light green), the MLP-based model trained self-supervised with a SimCLR loss with 50% random masking (light blue), physics-inspired augmentations from [26] (medium blue), or physics- and anomaly-inspired augmentations from [26] (dark blue). Additionally, we compare these results with three baseline embeddings: 57D source (black circles), the 6D leading object p_T (black triangles), and the 6D VAE (black crosses). The typical 3σ level for evidence and 5σ level for discovery are reported in dashed lines. Where empirical Z scores cannot be computed ($Z > 3\sigma$) we rely on the asymptotic formula.

evaluation classifier to determine its capability to retain and encode sufficient information for distinguishing between background classes. The linear evaluation is performed using a single linear layer on top of the neural embedding. The layer is trained using a supervised cross-entropy objective for classifying the Standard Model background events. The weights of the neural encoders are frozen

during linear evaluation. We compute the standard accuracy as the mean across the fivefold evaluation datasets, along with its uncertainty (i.e., the standard error of the mean). We use this quantity as a figure of merit to perform a grid search in the space of hyperparameters of the considered models. The values obtained with the selected hyperparameters for different sizes of the latent dimension are

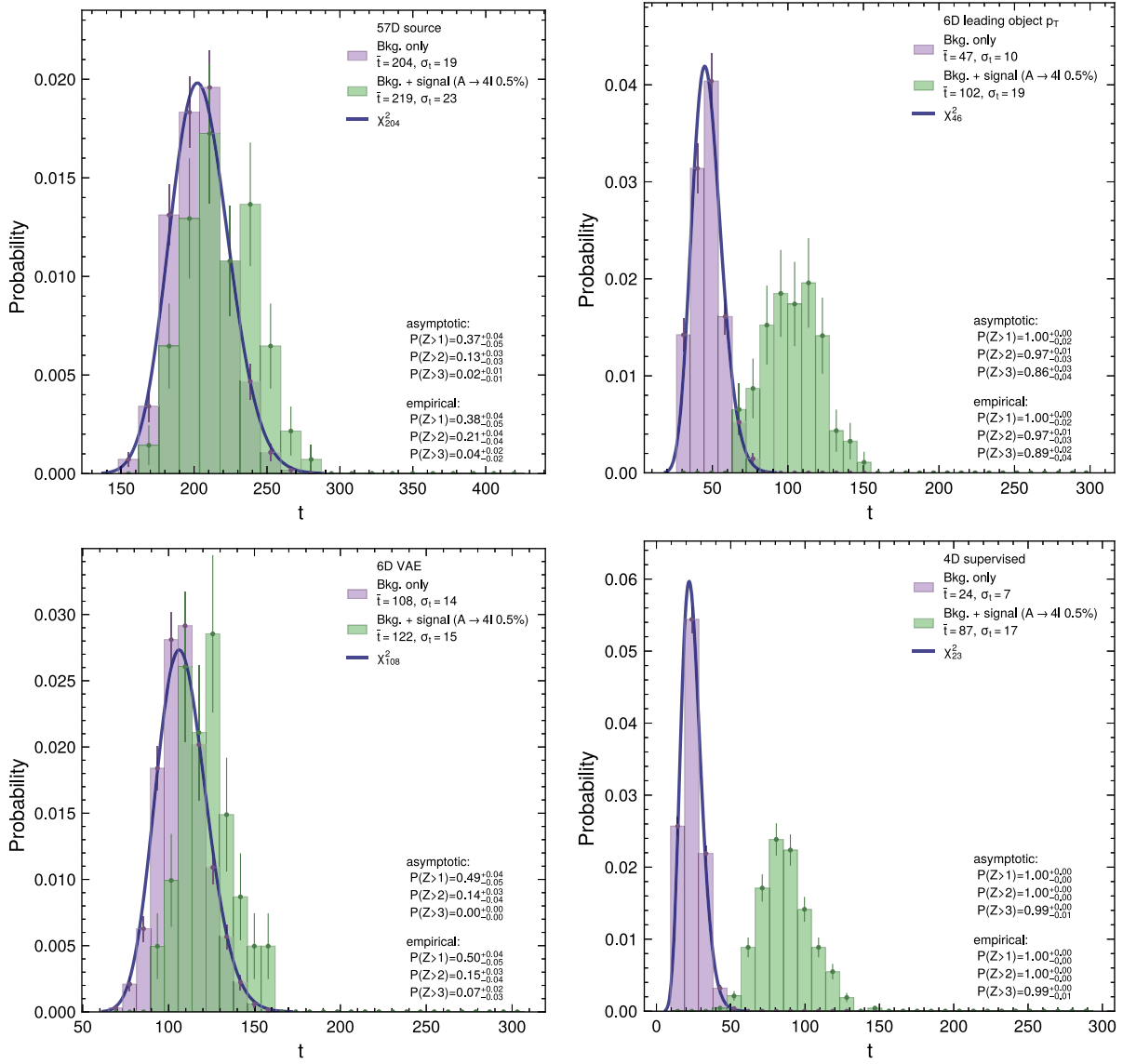


FIG. 4. Anomaly detection applied to $A \rightarrow 4l$ signal benchmark. The outcomes of the NPLM test statistic for 300 toy experiments with 0.5% signal injection (green histogram) are compared to 1000 experiments in absence of signal (purple histogram), representing the empirical null hypothesis. The four panels report the results for different data embeddings: the original 57-dimensional representation (top left); the six-dimensional representation given by the six highest p_T within the 19 objects in the event (top right); the six-dimensional neural embedding given by the variational autoencoder (bottom left); and the four-dimensional neural embedding given by the transformer-based architecture trained with supervised SimCLR loss (bottom right). In each panel we report the power of the test at 1, 2, and 3 σ level of discovery obtained from the empirical distribution or the asymptotic χ^2 showed in solid blue line.

presented in Fig. 2, showcasing the effectiveness of the pretrained latent spaces in the background classification task. As expected, exploiting background supervision in the contrastive learning approach helps reach a high level of accuracy in the background classification task. Moreover, we observe that higher-dimensional latent spaces reach higher levels of accuracy in the classification task. While these results give us an insight on which encoder better captures the discriminative features for an efficient background classification, this does not guarantee that the relevant features for detecting anomalies are preserved.

D. Anomaly detection on neural embeddings

To assess the power of different neural encoders to retain critical information for anomaly detection we apply the NPLM test for anomaly detection outlined in Sec. IV. While fine-tuning the embedding at the stage of testing would be possible, in this work we freeze the embedding and run the NPLM test as a downstream task. To test the sensitivity to previously unseen signals we consider the four signal benchmarks provided in the data challenge. To compute the NPLM test we define a reference sample $\mathcal{R}$ composed of 10 000 000 background data points and a data

TABLE VI. Z scores for each embedding type with 0.5% signal injection across different signals: $LQ \rightarrow \tau b$, $A \rightarrow 4\ell$, $H^\pm \rightarrow \tau\nu$, and $H \rightarrow \tau\tau$.

Embedding type	0.5% signal injection			
	$LQ \rightarrow \tau b$	$A \rightarrow 4\ell$	$H^\pm \rightarrow \tau\nu$	$H \rightarrow \tau\tau$
57D source	$0.57^{0.08}_{0.08}$	$0.93^{0.09}_{0.09}$	$2.47^{0.09}_{0.09}$	$0.41^{0.09}_{0.09}$
6D leading object p_T	$0.96^{0.09}_{0.09}$	$3.42^{0.09}_{0.09}$	$4.95^{0.09}_{0.09}$	$1.19^{0.09}_{0.09}$
6D VAE	$0.7^{0.1}_{0.1}$	$0.9^{0.1}_{0.1}$	$0.7^{0.1}_{0.1}$	$0.2^{0.1}_{0.1}$
4D supervised transformer	$3.66^{0.09}_{0.09}$	$5.61^{0.09}_{0.09}$	$8.51^{0.09}_{0.09}$	$2.81^{0.09}_{0.09}$

sample $\mathcal{D}$ with a number of expected events in the background hypothesis $N(B) = 10000$. To inspect the performances for progressively harder tasks we consider three scenarios of signal injection, $N(S) = 100, 500, 1000$, corresponding to signal fractions of 0.1%, 0.5% and 1%. The implementation of the NPLM test adopted in this work relies on a kernel method, for which the number of Gaussian kernels (M), the kernels' width (σ) and the L2 regularization coefficient (λ) are hyperparameters to be specified. Following the heuristic in Ref. [20], we take a number of kernels equal to square root of the training set size, $M = 1000$; a value of the kernels' width equal to the 90% quantile of the pairwise distribution of the Euclidean distance between reference data points; and a small value of the L2 coefficient that still allows for a stable algorithm convergence, $\lambda = 10^{-6}$. While M and λ are fixed for all experiments, the value of σ varies with the embedding, since different representations are characterized by different typical distance scales. The only exception to the hyperparameter setup just described is made for the 6D variational autoencoder (VAE) baseline that will be introduced in the following. In this case we set $M = 10000$ to obtain a better agreement of the test statistic distribution in absence of signal with the asymptotic χ^2 used as a figure of merit for hyperparameter selection in setting up the NPLM algorithm (see [20] for details).

Figure 3 reports the median Z score obtained using the NPLM test over different signals and different embeddings as a function of the dimensionality of the embedding for a 0.5% signal injection. Similar results on 0.1% and 1% signal injections can be found in Appendix A. In the figure, the neural embeddings based on contrastive learning are compared to three baseline representations: (i) the original embedding, based on 57 kinematic variables describing the 19 physics objects ("57D source"); (ii) a six-dimensional embedding formed of the missing transverse momentum, the transverse momenta of the two highest momentum muons, the two highest momentum electrons and the highest momentum jet ("6D leading object p_T "); and (iii) a six-dimensional neural embedding formed in the latent space of a variational autoencoder ("6D VAE") [56]. Autoencoders are widely used in many fields of data analysis, high-energy

physics included, as a way to perform data compression and anomaly detection. In this work we consider the variational autoencoder developed in [56] as a baseline for anomaly detection on the same dataset. The model learns a probabilistic latent representation of the input data. The encoder maps the 57-dimensional input to a three-dimensional latent space, parametrized by a mean and log variance, resulting in six latent parameters. It consists of two fully connected layers with 32 and 16 units, followed by batch normalization and leaky ReLU activations. The decoder, mirroring the encoder, reconstructs the input using a final dense layer with 57 output units. The model is trained with a loss combining reconstruction error and Kullback-Leibler divergence, weighted by a scaling factor of 0.8, ensuring a structured latent space through the reparametrization trick.

Results reported in Fig. 2 showed that, generally, the performance of a linear classifier for background classes tends to be better with larger embeddings. Conversely, results reported in Fig. 3 suggest that anomaly detection benefits from smaller embedding sizes. For two signals, $A \rightarrow 4\ell$ (upper right) and $H^\pm \rightarrow \tau\nu$ (bottom left), SimCLR-based models reach 5σ discovery for embedding dimensionality lower than 16, with a clear increasing trend as the dimensionality is reduced. Moreover, transformers show superior performance, particularly relevant to detect more challenging signals, indicating that they might be better equipped to recognize complex patterns in the data. Methods based on supervised contrastive learning significantly outperform traditional approaches in all benchmarks. Moreover supervised contrastive learning approaches perform generally better than self-supervised ones, suggesting that the physics information carried by the labels, and used to structure the background processes in the latent space, is crucial to enable anomaly detection at the level of new physics processes. In addition, we report in Fig. 4 the distribution of the test statistic for 1000 background-only toys (purple) and for 300 toys of background with 0.5% signal injected (green) using the three baseline representations and the best contrastive approach, given by the 4D supervised SimCLR transformer. In each panel we report the empirical and asymptotic detection power at 1, 2 and 3σ level. The neural embedding based on contrastive learning

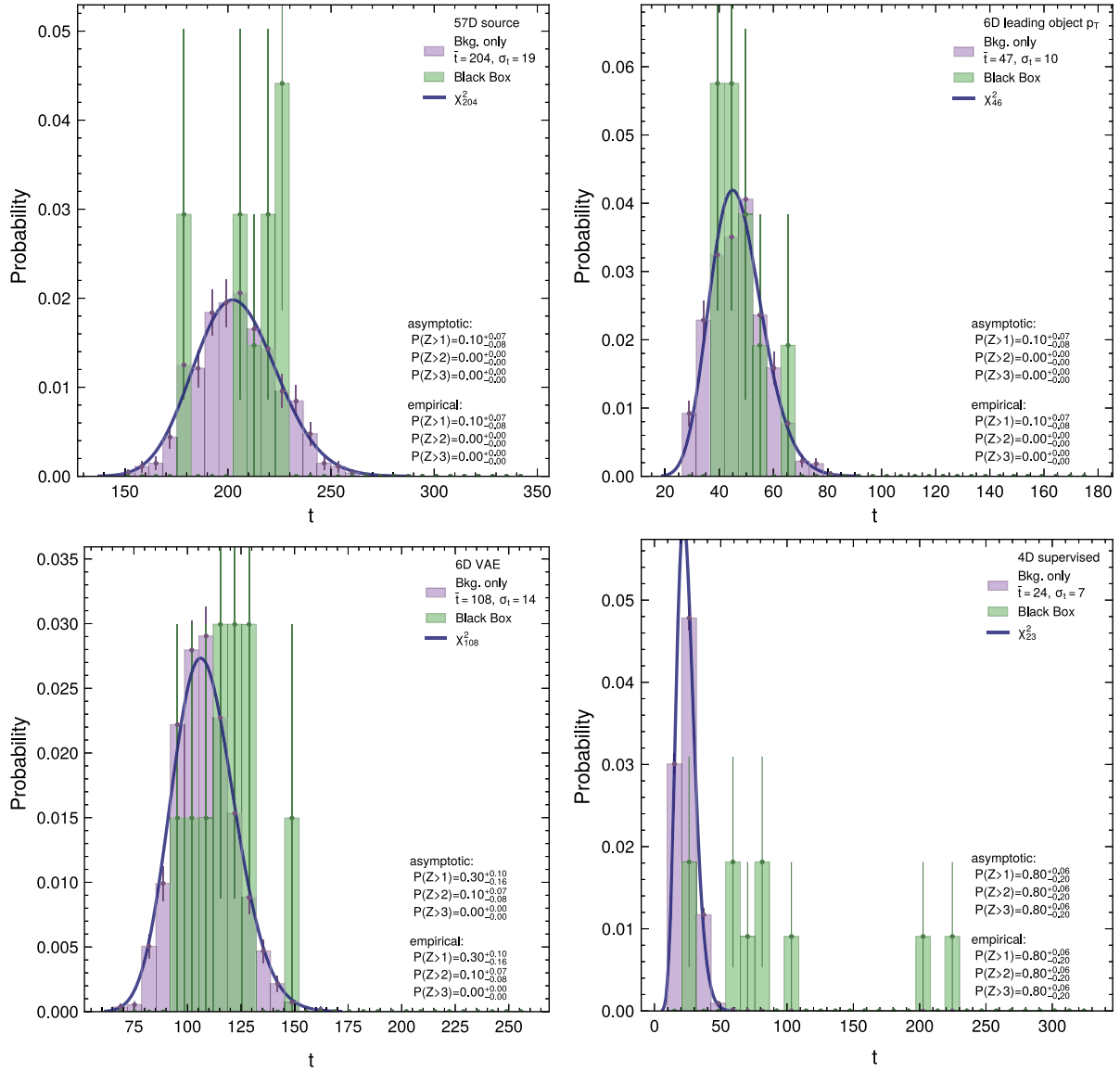


FIG. 5. Anomaly detection applied to the black box dataset. The outcomes of the NPLM test statistic for ten splits of the black box (green histogram) are compared to the test distribution in absence of signals (purple histogram). For each of the splits we compute the NPLM test statistic 10 times and report the average number. The four panels report the results for different data embeddings: the original 57-dimensional representation (top left); the six-dimensional representation given by the six highest p_T within the 19 objects in the event (top right); the six-dimensional neural embedding given by a variational autoencoder from [56] (bottom left); and the four-dimensional neural embedding given by the transformer-based architecture trained with supervised SimCLR loss (bottom right). The neural embedding based on contrastive learning is the only representation that retains discovery power.

is the only representation that achieves almost 100% discovery power at a 3σ level. Numerical results are summarized in Table VI.

Finally, we apply the anomaly detection pipelines studied in this section to a “black box” dataset provided in the AD challenge. The dataset consists of 1 001 000 events of unknown origin, corresponding to 10 times the expected yield of our experiments. The signal injection is expected to be 0.1%. To match the experimental luminosity used in the signal benchmark study we run the NPLM test on ten splits of the dataset separately and we compare the

detection performance of our three baselines (57D source, 6D leading object p_T , and 6D VAE) to that of the best neural embedding, the four-dimensional transformer-based model trained with supervised SimCLR loss. To account for the mild variance due to the finite size of the reference sample, we repeat the NPLM test 10 times for each split, randomizing the sampling of the reference set. Figure 5 compares the NPLM test statistic outcome for each split (green histogram) with the distribution of the test in absence of signal (purple histogram). For each split we report the average observed value of the NPLM test over

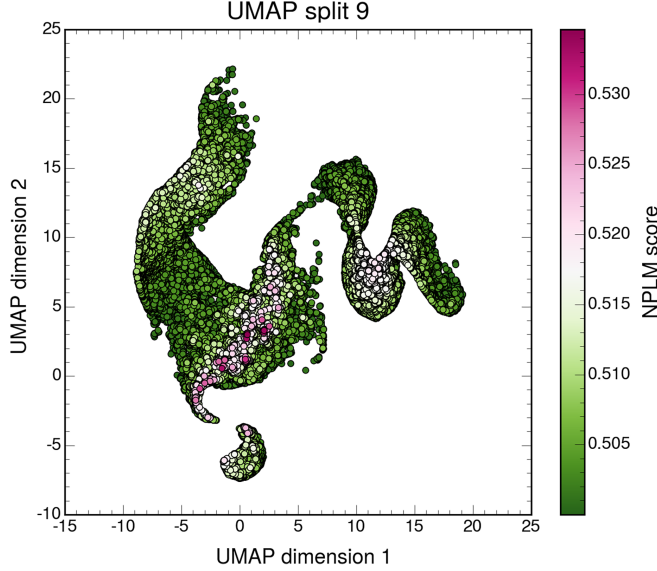


FIG. 6. UMAP visualization of a black box batch. We inspect the most significant batch (split 9) by means of a two-dimensional UMAP, where the color code represents the sigmoid activated score output by NPLM. The plot only shows data with score greater than 0.5. Data points are ordered in score so that higher score data points are on top and always visible.

the ten runs obtained resampling the reference. The comprehensive set of average values and corresponding standard deviations translated in Z scores are reported in Table VII. In the table we also provide an overall assessment of the statistical significance across all runs by computing a global significance. The latter is obtained by selecting the most significant observation among the runs and adjust it via Bonferroni correction. The four-dimensional transformer is the only neural embedding that allows the discovery of a hidden signal in the black box sample, significantly surpassing the detection performances of all the baselines considered in this study. The 4D contrastive embedding reaches a global Z score 12.1σ , opposed to 0, 0.6 and 1.4σ for the source, leading p_T objects and variational autoencoder, respectively. To further inspect the properties of the black box dataset we produce

a 2D uniform manifold approximation and projection (UMAP) [57] for each split. UMAP is a dimensionality reduction technique that helps visualize high-dimensional data in 2D or 3D while preserving its essential structure. It first constructs a graph by connecting each data point to its nearest neighbors and then tries to find a 2D representation that preserves this local neighborhood structure as much as possible. Two points that are close to each other in the UMAP visualization are therefore strongly connected in the original four-dimensional space. Figure 6 shows the 2D UMAP of the black box data in split 9. The color code corresponds to the score output by NPLM, and only data points with NPLM score higher than 0.5 are considered. The most anomaly like data points in mesh of pink and purple fall in localized regions of the domain, suggesting that the anomalies are part of the same local structure in the neural embedding. Additional plots for the remaining black box splits are given in Appendix C.

VI. DISCUSSION

The results presented in Sec. V show that contrastive-learning-based neural embeddings retain significantly higher anomaly detection power compared to the original 57-dimensional representation, or to a subset of variables, like the six highest p_T values, that only allows for a partial view of the event. The contrastive learning scheme based on background supervision with SimCLR objective outperforms all other neural embedding strategies studied in this work, maintaining or even enhancing the sensitivity when the compression is more aggressive. As a function of the embedding dimensionality, we observe opposite trends in the accuracy of the background-only classification objective pursued as training task (Fig. 2) and the anomalous signal detection objective pursued as downstream task (Fig. 3). This experimental result remarks that the contrastive learning objective is not necessarily aligned with the anomaly detection objective, which is unknown until real data are analyzed. Due to this reason, despite the increased computational demand, fine-tuning the neural embedding while optimizing the NPLM objective could potentially improve

TABLE VII. Anomaly detection applied to the black box dataset. For each black box split we report the outcome of the NPLM test applied to different data embeddings. The results are reported in terms of average observed Z score, where the average is computed over ten runs of the NPLM test with randomized resampling of the reference set. Error bands represent the 68% confidence interval, computed as the standard deviation over the ten runs. In the last column we report the global significance across all runs. The latter is obtained by selecting the most significant observation among the runs and adjusting it via Bonferroni correction.

Embedding type	Split 0	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8	Split 9	Global
57D source	$0.9^{0.7}_{0.6}$	$0^{0.7}_{0.6}$	$0.6^{0.7}_{0.6}$	$1.1^{0.6}_{0.6}$	$-1.5^{0.8}_{0.7}$	$-0.0^{0.4}_{0.4}$	$-1.3^{0.7}_{0.7}$	$0.3^{0.5}_{0.5}$	$0.7^{0.5}_{0.5}$	$0.9^{0.5}_{0.5}$	0
6D leading object p_T	$-0.2^{0.4}_{0.4}$	$-0.5^{0.3}_{0.3}$	$0.4^{0.3}_{0.3}$	$-0.5^{0.4}_{0.4}$	$-0.6^{0.3}_{0.3}$	$0.7^{0.5}_{0.5}$	$0.6^{0.4}_{0.4}$	$-0.2^{0.6}_{0.6}$	$-0.2^{0.4}_{0.4}$	$1.9^{0.5}_{0.5}$	$0.6^{0.8}_{0.6}$
6D VAE	$1.4^{0.4}_{0.4}$	$0.6^{0.3}_{0.3}$	$0.7^{0.4}_{0.3}$	$-0.1^{0.3}_{0.3}$	$0.3^{0.3}_{0.3}$	$-0.7^{0.3}_{0.3}$	$-0.5^{0.4}_{0.4}$	$0.9^{0.2}_{0.2}$	$1.2^{0.5}_{0.5}$	$2.4^{0.3}_{0.3}$	$1.4^{0.4}_{0.5}$
4D SimCLR sup.	$4.2^{0.2}_{0.2}$	$5.3^{0.2}_{0.2}$	$5.0^{0.1}_{0.1}$	$11.3^{0.2}_{0.2}$	$4.6^{0.3}_{0.2}$	$0.6^{0.2}_{0.2}$	$6.8^{0.2}_{0.2}$	$0.9^{0.3}_{0.2}$	$4.3^{0.2}_{0.2}$	$12.3^{0.1}_{0.1}$	$12.1^{0.1}_{0.1}$

the detection performances. This is an interesting research direction that we leave to future work.

Finally, our study shows that neural embeddings based on supervised contrastive learning—specifically, contrastive learning between background classes—significantly outperform self-supervised methods. While a substantial portion of this improvement can be attributed to the additional information provided by labels, our findings also underscore the need for greater attention to training details in self-supervised approaches, particularly in designing effective data augmentations. Furthermore, due to the absence of label information it is likely that self-supervised methods could benefit from larger training datasets. This could be readily achieved using real experimental data, which are produced in abundance at LHC experiments. We leave to future work a thorough exploration of this avenue. Additionally, supervised contrastive learning could be further expanded to enhance sensitivity to specific families of signal hypotheses by adding a broad collection of labeled signal simulations to the training set. The resulting contrastive space could be used to perform weakly supervised anomaly detection strategies as demonstrated in [58].

VII. CONCLUSION AND OUTLOOK

We presented an end-to-end, signal-agnostic search strategy for inclusive final states at the LHC. By leveraging contrastive learning to build neural embeddings for anomaly detection, we significantly improve sensitivity to several unseen signals. Applying this method to the anomaly detection challenge black box dataset [46] with an injected unknown signal in an unspecified final state, we demonstrate an eightfold improvement in signal sensitivity. This approach enables robust statistical anomaly detection searches for new physics beyond the Standard Model. The promising results obtained in our studies motivate further research aimed at making this pipeline viable for real experimental analyses. In the view of scaling the current study to real experimental scenarios characterized by large datasets, efficient self-supervised contrastive learning strategies leveraging real data for pretraining could represent a promising direction to reduce reliance on computationally expensive labeled simulations. Investigating the dependence of self-supervised performance on training sample size and designing optimized data augmentation techniques are critical directions for future research. Another key challenge of real analyses is the presence of systematic uncertainties affecting the modeling of the background processes. Previous work has demonstrated methods to incorporate such uncertainties into the NPLM test procedure using dedicated machine-learning models [22]. Additionally, contrastive training schemes have shown promise in mitigating systematic effects at the embedding stage [25]. However, understanding the extent to which systematic uncertainties propagate through contrastive-based neural embeddings and their impact on discovery

sensitivity remains an open question for future exploration. Due to the ability to handle inclusive final states while preserving signal independence, the analysis strategy presented in this work shows great potential for the statistical analysis of data collected by event-level anomaly detection triggers, such as the one used by the CMS experiment [59], or for running quasi-online in real-time detector data acquisition systems [60].

ACKNOWLEDGMENTS

G. G. and P. H. acknowledge the financial support of the National Science Foundation (NSF) under Cooperative Agreement No. PHY-2019786 (The National Science Foundation Artificial Intelligence Institute for Artificial Intelligence and Fundamental Interactions). K. G. and P. H. acknowledge the financial support of the National Science Foundation under Cooperative Agreement No. 2117997 (AI-Accelerated Algorithms for Data-Driven Discovery, A3D3) and the National Science Foundation Cyberinfrastructure for Sustained Scientific Innovation (CSSI) program. Computations in this paper were partially run on the Faculty of Arts and Sciences Research Computing (FASRC) Cannon cluster, supported by the Faculty of Arts and Sciences Division of Science Research Computing Group at Harvard University. T. K. Å. is supported by the Swiss National Science Foundation (SNSF) Grant No. PZ00P2_201594. L. X. is supported by the John Reed Fund from the Massachusetts Institute of Technology Undergraduate Research Opportunities Program (MIT UROP) office, and additionally, she was supported by the MIT International Science and Technology Initiatives (MISTI) France program while performing research at the European Organization for Nuclear Research (CERN).

DATA AVAILABILITY

The data that support the findings of this article are openly available [46].

APPENDIX A: MEDIAN Z SCORE FOR DIFFERENT SIGNAL INJECTIONS

In this appendix we report additional examples of observed median Z score as the ones shown in Fig. 3 in the main body of this paper. Figure 7 is obtained from a 0.1% signal injection whereas Fig. 8 from 1% signal injection.

APPENDIX B: DISCOVERY POWERS FOR 0.5% AND 1% SIGNAL INJECTIONS

Figures 9 and 10 report the statistical power for signal discovery at 3σ level obtained using the NPLM test over 0.5% and 1% signal injections, respectively. The four panels in each figure report the power curves for the four

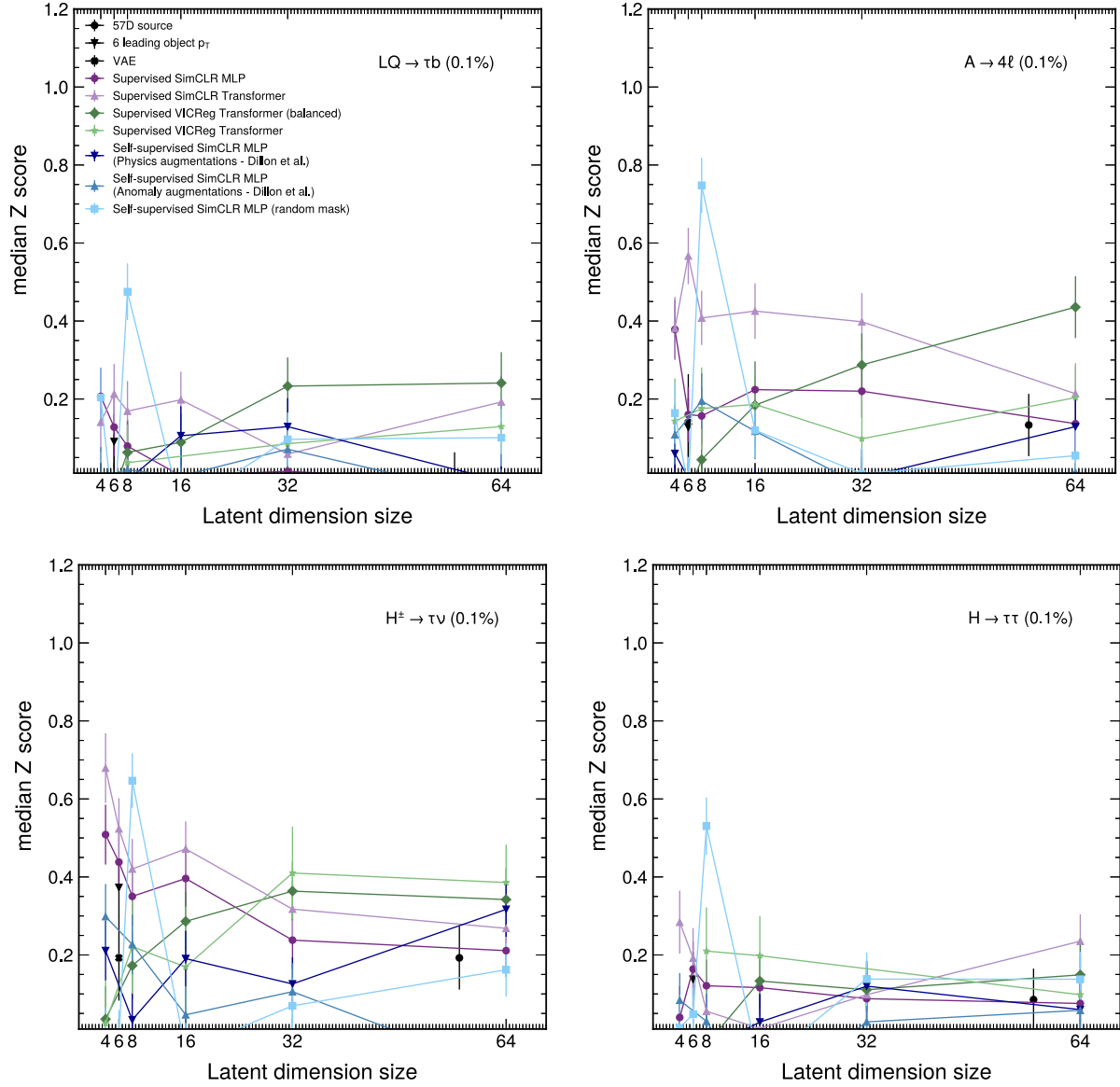


FIG. 7. The median observed Z score as a function of the feature embedding size after injecting a signal into the SM background pseudodataset, corresponding to 0.1% of the total integral. Results are shown for the $LQ \rightarrow \tau b$ (upper left), $A \rightarrow 4\ell$ (upper right), $H^\pm \rightarrow \tau\nu$ (bottom left), and $H \rightarrow \tau\tau$ (bottom right) signals. The median Z score is presented for different embedding models: the MLP- and transformer-based models trained with supervision and SimCLR loss (dark and light purple, respectively), the transformer-based model trained with supervision and VICReg loss using either balanced background classes (dark green) or background classes weighted to match the composition expected in data (light green), the MLP-based model trained self-supervised with a SimCLR loss with 50% random masking (light blue), physics-inspired augmentations from [26] (medium blue), or physics- and anomaly-inspired augmentations from [26] (dark blue). Additionally, we compare these results with three baseline embeddings: 57D source (black circles), the 6D leading object p_T (black triangles), and the 6D VAE (black crosses). The typical 3σ level for evidence and 5σ level for discovery are reported in dashed lines. Where empirical Z scores cannot be computed ($Z > 3\sigma$) we rely on the asymptotic formula.

signal benchmarks considered in this work: $LQ \rightarrow \tau b$ (upper left), $A \rightarrow 4\ell$ (upper right), $H^\pm \rightarrow \tau\nu$ (bottom left), and $H \rightarrow \tau\tau$ (bottom right) signals. Different curves represent the power of different embeddings as a function of

the dimensionality of the embedding. The statistical power at 3σ represent the probability of observing a p value lower or equal to 1.35×10^{-3} , namely a Z score greater or equal to 3.

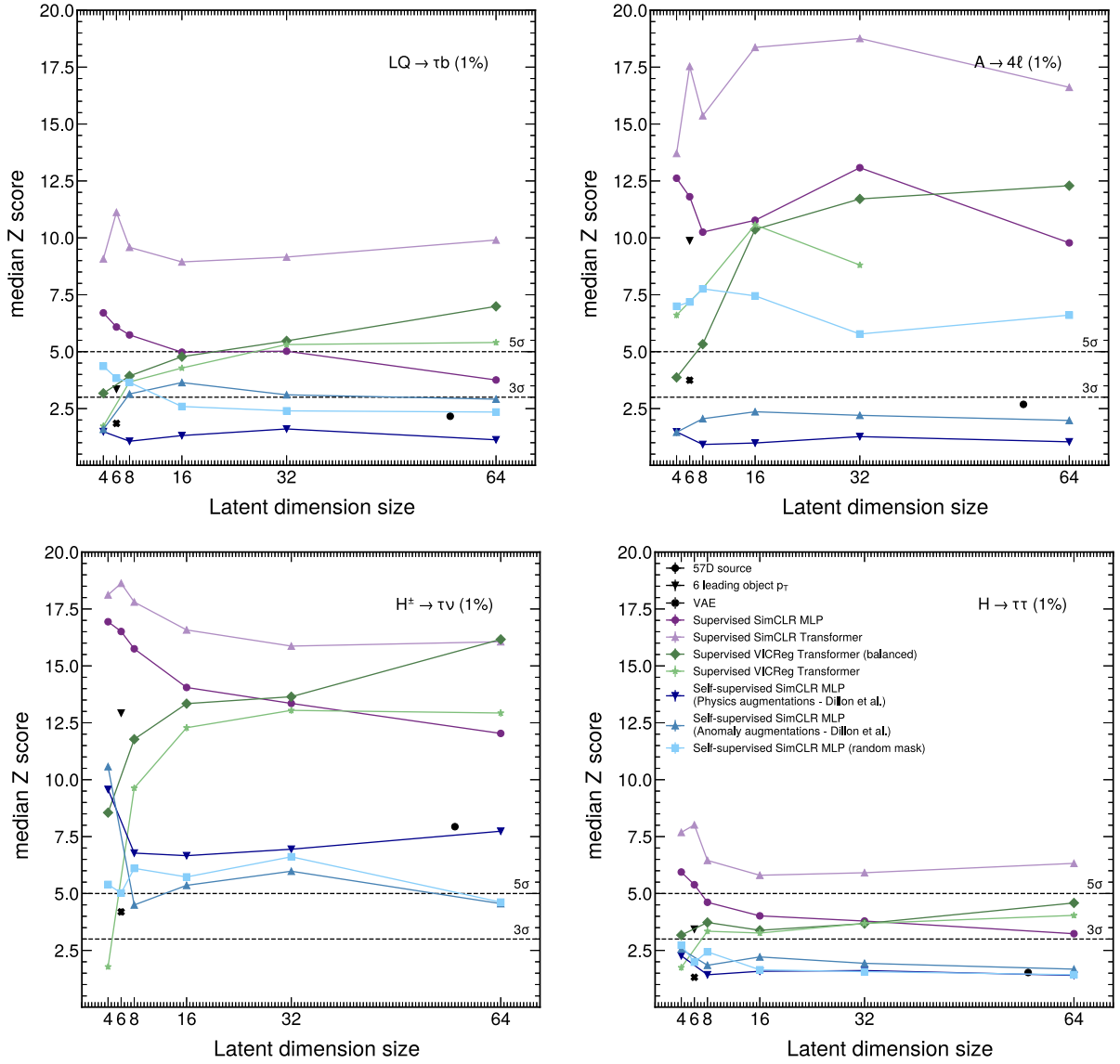


FIG. 8. The median observed Z score as a function of the feature embedding size after injecting a signal into the SM background pseudodataset, corresponding to 1% of the total integral. Results are shown for the $LQ \rightarrow \tau b$ (upper left), $A \rightarrow 4\ell$ (upper right), $H^\pm \rightarrow \tau \nu$ (bottom left), and $H \rightarrow \tau \tau$ (bottom right) signals. The median Z score is presented for different embedding models: the MLP- and transformer-based models trained with supervision and SimCLR loss (dark and light purple, respectively), the transformer-based model trained with supervision and VICReg loss using either balanced background classes (dark green) or background classes weighted to match the composition expected in data (light green), the MLP-based model trained self-supervised with a SimCLR loss with 50% random masking (light blue), physics-inspired augmentations from [26] (medium blue), or physics- and anomaly-inspired augmentations from [26] (dark blue). Additionally, we compare these results with three baseline embeddings: 57D source (black circles), the 6D leading object p_T (black triangles), and the 6D VAE (black crosses). The typical 3σ level for evidence and 5σ level for discovery are reported in dashed lines. Where empirical Z scores cannot be computed ($Z > 3\sigma$) we rely on the asymptotic formula.

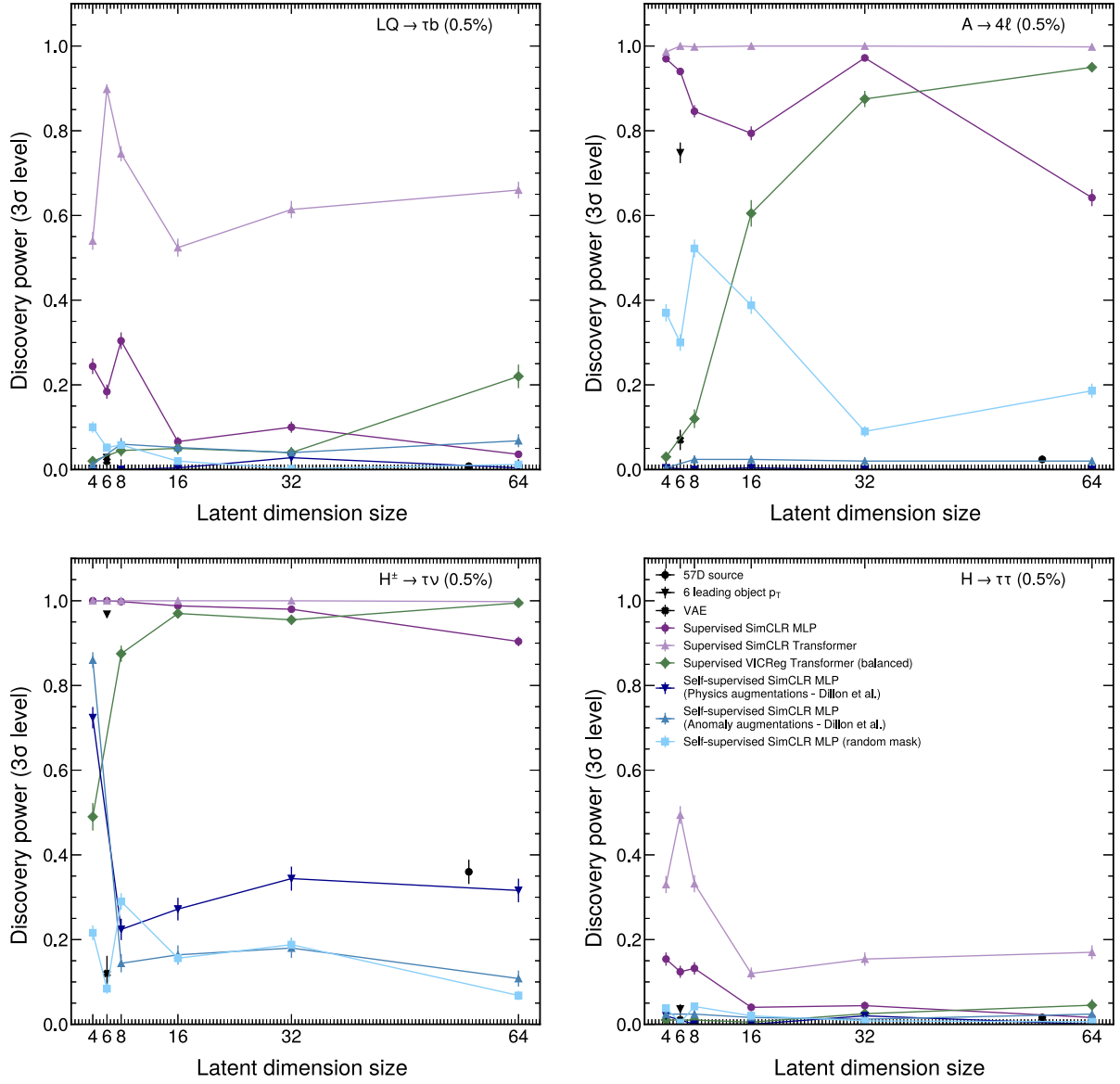


FIG. 9. The empirical discovery power (3σ) as a function of the feature embedding size after injecting a signal into the SM background pseudodataset, corresponding to 0.5% of the total integral. Results are shown for the $LQ \rightarrow \tau b$ (upper left), $A \rightarrow 4\ell$ (upper right), $H^\pm \rightarrow \tau\nu$ (bottom left), and $H \rightarrow \tau\tau$ (bottom right) signals. The discovery power is presented for different models: the MLP- and transformer-based models trained with supervision and SimCLR loss (dark and light purple, respectively), the transformer-based model trained with supervision and VICReg loss using balanced background classes (dark green), the MLP-based model trained self-supervised with a SimCLR loss with 50% random masking (light blue), physics-inspired augmentations from [26] (medium blue), or physics- and anomaly-inspired augmentations from [26] (dark blue). Additionally, we compare these results with three baseline embeddings: 57D source (black circles), the 6D leading object p_T (black triangles), and the 6D VAE (black crosses). The typical 3σ level for evidence and 5σ level for discovery are reported in dashed lines.

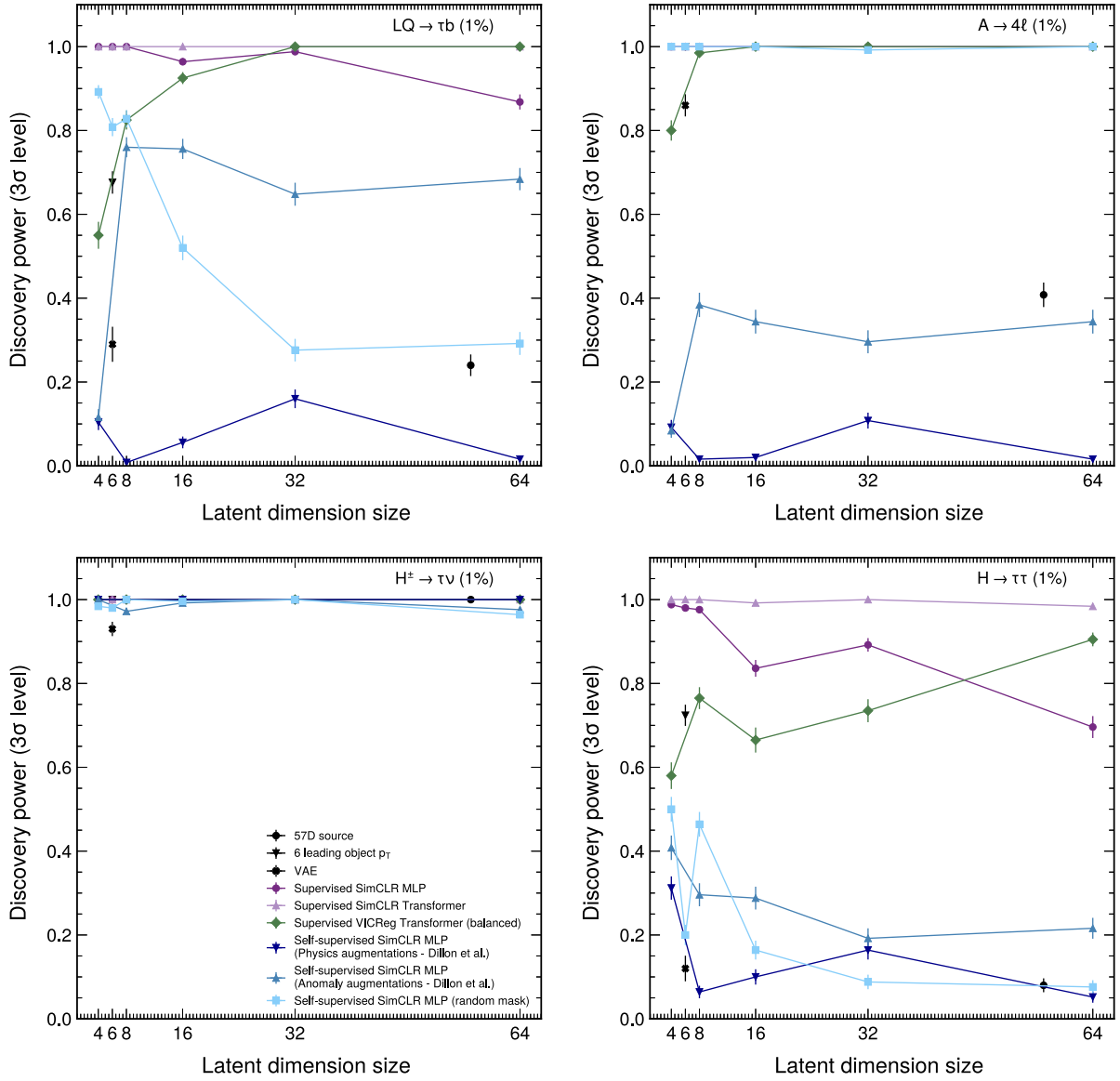


FIG. 10. The empirical discovery power (3σ) as a function of the feature embedding size after injecting a signal into the SM background pseudodataset, corresponding to 1% of the total integral. Results are shown for the $LQ \rightarrow \tau b$ (upper left), $A \rightarrow 4\ell$ (upper right), $H^\pm \rightarrow \tau\nu$ (bottom left), and $H \rightarrow \tau\tau$ (bottom right) signals. The discovery power is presented for different models: the MLP- and transformer-based models trained with supervision and SimCLR loss (dark and light purple, respectively), the transformer-based model trained with supervision and VICReg loss using balanced background classes (dark green), the MLP-based model trained self-supervised with a SimCLR loss with 50% random masking (light blue), physics-inspired augmentations from [26] (medium blue), or physics- and anomaly-inspired augmentations from [26] (dark blue). Additionally, we compare these results with three baseline embeddings: 57D source (black circles), the 6D leading object p_T (black triangles), and the 6D VAE (black crosses). The typical 3σ level for evidence and 5σ level for discovery are reported in dashed lines.

APPENDIX C: UMAPS

To complement Fig. 6, we show in Fig. 11 the two-dimensional UMAP of the various black box data splits embedded in the 4D contrastive space used in the studies

described in Sec. V. The plots show data with NPLM score greater than 0.5. Data points are ordered by score so that higher score data points are on top and always well visible.

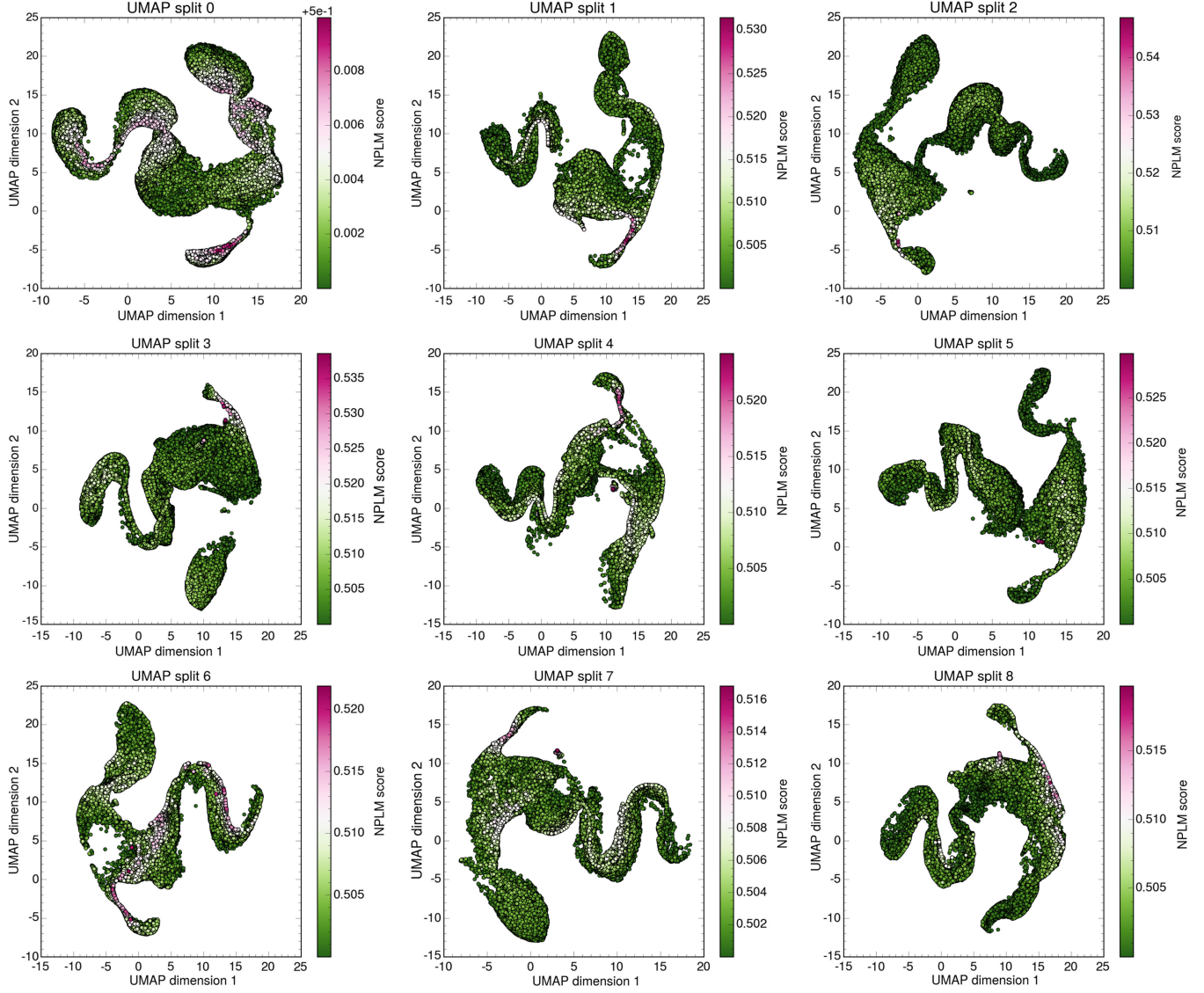


FIG. 11. UMAP visualization of a black box batch. We inspect each split by means of a two-dimensional UMAP, where the color code represents the sigmoid activated score output by NPLM. The plots only show data with score greater than 0.5. Data points are ordered in score so that higher score data points are on top and always visible.

-
- [1] OpenAI, GPT-4 technical report, 2024, [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
 - [2] Y. Yang, S. Mandt, and L. Theis, An introduction to neural data compression. *Found. Trends® Comput. Graph. Vision* **15**, 113 (2023).
 - [3] V. M. Abazov *et al.*, A quasi model independent search for new physics at large transverse momentum, *Phys. Rev. D* **64**, 012004 (2001).
 - [4] B. Abbott *et al.*, A quasi model independent search for new high p_T physics at D0, *Phys. Rev. Lett.* **86**, 3712 (2001).
 - [5] A. Aktas *et al.*, A general search for new phenomena in ep scattering at HERA, *Phys. Lett. B* **602**, 14 (2004).
 - [6] T. Aaltonen *et al.*, Model-independent and quasi-model-independent search for new physics at CDF, *Phys. Rev. D* **78**, 012002 (2008).
 - [7] F. D. Aaron *et al.*, A general search for new phenomena at HERA, *Phys. Lett. B* **674**, 257 (2009).
 - [8] T. Aaltonen *et al.*, Global search for new physics with 2.0 fb⁻¹ at CDF, *Phys. Rev. D* **79**, 011101 (2009).

- [9] V.M. Abazov *et al.*, Model independent search for new phenomena in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV, *Phys. Rev. D* **85**, 092015 (2012).
- [10] ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV, ATLAS Report No. ATLAS-CONF-2014-006, 2014, <https://cds.cern.ch/record/1666536>.
- [11] A.M. Sirunyan *et al.*, MUSiC: A model-unspecific search for new physics in proton–proton collisions at $\sqrt{s} = 13$ TeV, *Eur. Phys. J. C* **81**, 629 (2021).
- [12] ATLAS Collaboration, Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector, *Phys. Rev. Lett.* **125**, 131801 (2020).
- [13] CMS Collaboration, Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV, *Rep. Prog. Phys.* **88**, 067802 (2025).
- [14] G. Aad *et al.*, Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector, *Phys. Rev. D* **108**, 052009 (2023).
- [15] G. Aad *et al.*, Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $s = 13$ TeV with the ATLAS detector, *Phys. Rev. Lett.* **132**, 081801 (2024).
- [16] G. Aad *et al.*, Weakly supervised anomaly detection for resonant new physics in the dijet final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, [arXiv:2502.09770](https://arxiv.org/abs/2502.09770).
- [17] V. Belis, P. Odagiu, and T. K. Aarrestad, Machine learning for anomaly detection in particle physics, *Rev. Phys.* **12**, 100091 (2024).
- [18] T. Finke *et al.*, Back to the roots: Tree-based algorithms for weakly supervised anomaly detection, *Phys. Rev. D* **109**, 034033 (2024).
- [19] M. Freytsis, M. Perelstein, and Y. C. San, Anomaly detection in the presence of irrelevant features, *J. High Energy Phys.* **02** (2024) 220.
- [20] M. Letizia *et al.*, Learning new physics efficiently with nonparametric methods, *Eur. Phys. J. C* **82**, 879 (2022).
- [21] G. Grosso, Anomaly-aware summary statistic from data batches, *J. High Energy Phys.* **12** (2024) 093.
- [22] R. T. d’Agnolo, G. Grosso, M. Pierini, A. Wulzer, and M. Zanetti, Learning new physics from an imperfect machine, *Eur. Phys. J. C* **82**, 275 (2022).
- [23] A. Khot, X. Wang, A. Roy, V. Kindratenko, and M. S. Neubauer, Evidential deep learning for uncertainty quantification and out-of-distribution detection in jet identification using deep neural networks, *Mach. Learn. Sci. Tech.* **6**, 035003 (2025).
- [24] S. E. Park, P. Harris, and B. Ostdiek, Neural embedding: Learning the embedding of the manifold of physics data, *J. High Energy Phys.* **07** (2023) 108.
- [25] P. Harris, M. Kagan, J. Krupa, B. Maier, and N. Woodward, Re-simulation-based self-supervised learning for pre-training foundation models, *Phys. Rev. D* **111**, 032010 (2025).
- [26] B. M. Dillon *et al.*, Symmetries, safety, and self-supervision, *SciPost Phys.* **12**, 188 (2022).
- [27] V. Mikuni and B. Nachman, OmniLearn: A method to simultaneously facilitate all jet physics tasks, *Phys. Rev. D* **111**, L051504 (2025).
- [28] M. Leigh *et al.*, Is tokenization needed for masked particle modelling?, [arXiv:2409.12589](https://arxiv.org/abs/2409.12589).
- [29] J. Birk, A. Hallin, and G. Kasieczka, OmniJet- α : The first cross-task foundation model for particle physics, *Mach. Learn. Sci. Tech.* **5**, 035031 (2024).
- [30] A. J. Wildridge *et al.*, Bumblebee: Foundation model for particle physics discovery, in *Proceedings of the 38th Conference on Neural Information Processing Systems* (2024), [arXiv:2412.07867](https://arxiv.org/abs/2412.07867).
- [31] J. Bardhan, R. Agrawal, A. Tilak, C. Neeraj, and S. Mitra, HEP-JEPA: A foundation model for collider physics using joint embedding predictive architecture, [arXiv:2502.03933](https://arxiv.org/abs/2502.03933).
- [32] T. Golling *et al.*, Masked particle modeling on sets: Towards self-supervised high energy physics foundation models, *Mach. Learn. Sci. Tech.* **5**, 035074 (2024).
- [33] C. Li *et al.*, Accelerating resonance searches via signature-oriented pre-training, [arXiv:2405.12972](https://arxiv.org/abs/2405.12972).
- [34] B. M. Dillon, R. Mastandrea, and B. Nachman, Self-supervised anomaly detection for new physics, *Phys. Rev. D* **106**, 056005 (2022).
- [35] L. Favaro, M. Krämer, T. Modak, T. Plehn, and J. Rüsckamp, Semi-visible jets, energy-based models, and self-supervision, *SciPost Phys.* **18**, 042 (2025).
- [36] B. M. Dillon, L. Favaro, F. Feiden, T. Modak, and T. Plehn, Anomalies, representations, and self-supervision, *SciPost Phys. Core* **7**, 056 (2024).
- [37] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798 (2013).
- [38] X. Chen and K. He, Exploring simple siamese representation learning, in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York, 2021), pp. 15745–15753.
- [39] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, Supervised Contrastive Learning, in *Advances in Neural Information Processing Systems*, Vol. 33 (Curran Associates, Inc., Red Hook, NY, 2020), pp. 18661–18673.
- [40] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in *Proceedings of the 37th International Conference on Machine Learning (PMLR, 2020)*, Vol. 119, pp. 1597–1607.
- [41] A. Bardes, J. Ponce, and Y. LeCun, VICReg: Variance-invariance-covariance regularization for self-supervised learning, [arXiv:2105.04906](https://arxiv.org/abs/2105.04906).
- [42] G. Grosso, M. Letizia, M. Pierini, and A. Wulzer, Goodness of fit by Neyman-Pearson testing, *SciPost Phys.* **16**, 123 (2024).
- [43] R. T. D’Agnolo and A. Wulzer, Learning new physics from a machine, *Phys. Rev. D* **99**, 015014 (2019).
- [44] J. Neyman and E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, *Phil. Trans. R. Soc. A* **231**, 289 (1933).

- [45] G. Meanti, L. Carratino, L. Rosasco, and A. Rudi, Kernel methods through the roof: Handling billions of points efficiently, in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (Curran Associates Inc., Red Hook, NY, USA, 2020), pp. 14410–14422.
- [46] E. Govorkova *et al.*, LHC physics dataset for unsupervised new physics detection at 40 MHz, *Nat. Sci. Data* **9**, 118 (2022).
- [47] T. Sjöstrand *et al.*, An introduction to PYTHIA8.2, *Comput. Phys. Commun.* **191**, 159 (2015).
- [48] J. de Favereau *et al.*, DELPHES3, A modular framework for fast simulation of a generic collider experiment, *J. High Energy Phys.* **02** (2014) 057.
- [49] <https://www.drawio.com/>.
- [50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* **30**, 5998 (2017).
- [51] Pytorch transformer, <https://pytorch.org/docs/stable/generated/torch.nn.Transformer.html>.
- [52] M. Caron *et al.*, Self-supervised vision transformers with dino, <https://github.com/facebookresearch/dino/tree/main>.
- [53] Y. You, I. Gitman, and B. Ginsburg, Large batch training of convolutional networks (2017).
- [54] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *Proceedings of the 7th International Conference on Learning Representations (ICLR, New Orleans, LA, 2019)*.
- [55] B. M. Dillon, L. Favaro, F. Feiden, T. Modak, and T. Plehn, Anomalies, representations, and self-supervision, *SciPost Phys.* **7**, 056 (2024).
- [56] E. Govorkova *et al.*, 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 (2022).
- [57] L. McInnes, J. Healy, N. Saul, and L. Großberger, UMAP: Uniform manifold approximation and projection, *J. Open Source Software* **3**, 861 (2018).
- [58] S. E. Park, D. Rankin, S.-M. Udrescu, M. Yunus, and P. Harris, Quasi anomalous knowledge: Searching for new physics with embedded knowledge, *J. High Energy Phys.* **06** (2021) 030.
- [59] 2024 Data Collected with AXOLITL anomaly detection at the CMS level-1 trigger (2024), <https://cds.cern.ch/record/2904695>.
- [60] R. Ardino *et al.*, A 40 MHz level-1 trigger scouting system for the CMS phase-2 upgrade. *Nucl. Instrum. Methods Phys. Res., Sect. A* **1047**, 167805 (2023).