Published for SISSA by Springer

Received: February 29, 2024
Revised: May 24, 2024
Accepted: June 4, 2024
Published: June 25, 2024

# Cluster Scanning: a novel approach to resonance searches

**I. Oleksiyuk**[a,b] **J.A. Raine**[a] **M. Krämer**[c] **S. Voloshynovskiy**[b] **and T. Golling**[a]

[a] *Département de Physique Nucléaire et Corpusculaire, University of Geneva, Quai Ernest-Ansermet 24, 1211 Geneva, Switzerland*

[b] *Department of Computer Science, University of Geneva, Route de Drize 7, 1211 Geneva, Switzerland*

[c] *Institute for Theoretical Particle Physics and Cosmology, RWTH Aachen University, Sommerfeldstrasse 16, 52074 Aachen, Germany*

*E-mail:* ivan.oleksiyuk@unige.ch, john.raine@unige.ch, mkraemer@physik.rwth-aachen.de, svyatoslav.voloshynovskyy@unige.ch, tobias.golling@unige.ch

Abstract: We propose a new model-independent method for new physics searches called Cluster Scanning. It uses the k-means algorithm to perform clustering in the space of low-level event or jet observables, and separates potentially anomalous clusters to construct a signal-enriched region. The spectra of a selected observable (e.g. invariant mass) in these two regions are then used to determine whether a resonant signal is present. A pseudo-analysis on the LHC Olympics dataset with a $Z'$ resonance shows that Cluster Scanning outperforms the widely used 4-parameter functional background fitting procedures, reducing the number of signal events needed to reach a $3\sigma$ significant excess by a factor of 0.61. Emphasis is placed on the speed of the method, which allows the test statistic to be calibrated on synthetic data.

Keywords: Jets and Jet Substructure, Specific BSM Phenomenology

ArXiv ePrint: 2402.17714



https://doi.org/10.1007/JHEP06(2024)163

## Contents



JHEP06(2024)163

## 1 Introduction

The Standard Model (SM) is the current apex of theoretical physics, describing the electromagnetic, weak and strong interactions with unparalleled precision. Unfortunately, it is still far from complete, as several phenomena remain unexplained. In order to create a "theory of everything", one would not only need to combine the SM with general relativity, but also provide an explanation for many other issues, including the existence of neutrino masses, the origin of the matter-antimatter asymmetry, and most importantly, the origin of dark matter. To solve these problems, researchers are collaborating to formalise new theories, design, build and carry out new experiments, as well as simulate and analyze research data.

One of the most renowned experimental facilities, the Large Hadron Collider (LHC) was constructed with the purpose of testing the SM in the high energy regime. The last elementary particle predicted by the SM, the Higgs boson, was discovered in 2012 [1, 2]. Since then, LHC research has shifted towards precision measurements and searches for beyond the Standard Model (BSM) effects.

Many extensions of the SM imply the existence of as yet undiscovered massive particles, often associated with proposed new symmetry groups. If a new particle has a narrow decay width, the straightforward method is to search for a resonant peak in the spectrum of a mass-like observable, such as the invariant mass of a dijet event. However, such a bump hunt is not completely free of assumptions. Often complex analytical functions need to be chosen to model the background distribution, with the possibility to introduce spurious signals and varying sensitivity under the assumption of different functional forms. Furthermore, additional observables or fiducial cuts need to be chosen and optimised to enhance sensitivity in the case where potential signal yields are low, causing searches to become more model-specific. To broaden the range of signal models covered by searches one may employ the model-unspecific search steategies [3–5].

Over the past decade, machine learning-based algorithms have become increasingly popular for solving a multitude of problems. Deep learning, in particular, has gained popularity for various tasks, with large neural networks being utilised. For example, many methods were implemented to perform anomaly detection (AD) tasks in various industries. Some of these AD methods have been repurposed and extended to support BSM searches [6–85] (see refs. [86–89] for a comparison of various ML assisted BSM methods and refs. [36, 90] for a comparison of weakly supervised and unsupervised approaches). The ATLAS collaboration produced the first experimental results for such searches applied to experimental data using weakly supervised methods [91] and unsupervised ML anomaly detection methods [80, 92]. However, these efforts have not observed any significant deviations from the SM expectation.

Many AD approaches rely on the assumption that any new signal would form a set of outliers. However, in a bump hunt the assumption is instead that any new signal would be localised in some feature space, in particular in an invariant mass spectrum. Weakly supervised approaches, on the other hand, aim to enhance the sensitivity by applying a cut on a classifier trained directly on the data. However, in both instances the same bump hunt restrictions apply with either functional forms or input observables impacting the sensitivity to a model.

In this work we introduce a new data-driven method, Cluster Scanning (CS), which builds on the foundations of the bump hunt but addresses several limitations. By leveraging more information from the event CS is able to enhance sensitivity to potential signals without enforcing any model specific assumptions, and can also provide a direct estimate of the background distribution. The proposed approach complements existing techniques and is designed to be computationally efficient.

The paper is structured as follows. In section 2, we briefly describe the LHCO R&D dataset [93], commonly used to benchmark the performance of anomaly detection techniques, and introduce our data preprocessing steps. Section 3 touches on the general topic of bump-hunting strategies in the literature, introduces the novel CS method, and discusses similarities and differences between them. In section 4 we provide the results of applying CS in an anomaly search. Finally, we draw conclusions in section 5.



JHEP06(2024)163

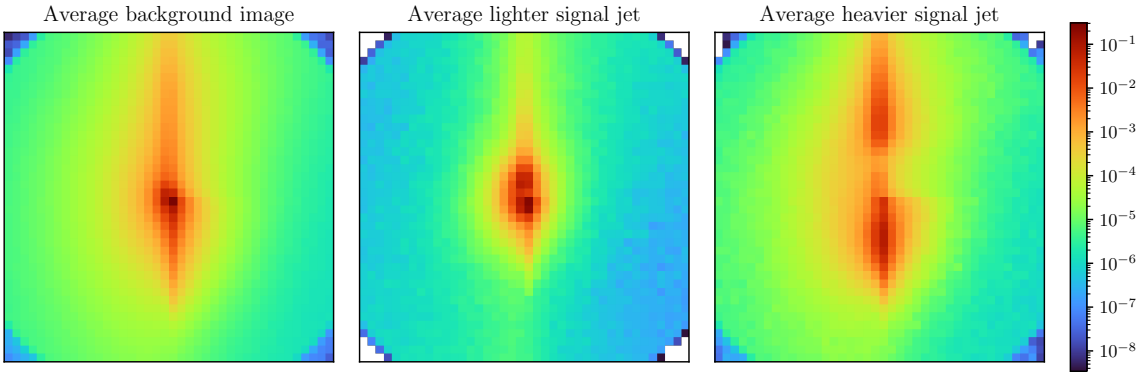



**Figure 1.** From left to right: average of all 2M available QCD jet images, average image of all 100K lighter jets in a $Z'$ event and average image of all 100K heavier jets in a $Z'$ event before smearing and pixel scaling.

## 2 Dataset

The LHCO R&D dataset consists of one million background Standard Model dijet events (also subsequently referred to as QCD) and 100 000 signal BSM $Z' \to XY$ events, where massive particles with $m_X = 500\,\text{GeV}$ and $m_Y = 100\,\text{GeV}$ decay into quark-antiquark pairs. The resonance itself has a mass of $m_{Z'} = 3.5\,\text{GeV}$. This anomaly model is discussed in detail in ref. [94].

All the events were produced using Pythia 8.219 [95] and Delphes 3.4.1 [96–98] using default settings. The jets were clustered using an anti-$k_T$ algorithm [99] with $R = 1$ using FastJet [100] with a `python` interface provided through the `pyjet` library in Scikit-HEP [101]. Jets are required to have $p_T > 1.2\,\text{TeV}$ and fall within $|\eta| < 2.5$.

### 2.1 Jet images

In addition to the di-jet invariant mass ($m_{jj}$) of the event, used in a bump hunt, we extract additional information from the image representations of the two jets. This allows for a more model agnostic approach than selecting specific jet substructure observables. The jet images are processed following a prescription similar to that used in refs. [11, 102–104] from the $\eta$, $\phi$ and $p_\text{T}$ of the jet constituents. Individual jet images are centred, rotated, and flipped in order to provide a consistent input to a convolutional neural network, reducing the number of symmetries the ML method would need to learn.

The jet images are cropped to $[-0.8, 0.8] \times [-0.8, 0.8]$ in $\eta - \phi$ space relative to the jet centre, binned with a $40 \times 40$ pixel grid, and normalised such that the sum of all pixels is equal to one. Figure 1 shows the average jet images for QCD background, and the separate averages of all lighter (mostly $Y$) and heavier (mostly $X$) jets in each $Z'$ event.

Despite being used in many applications, the jet image representation has two main drawbacks, namely the sparsity of non-zero pixels (see appendix B) and the imbalance in the magnitudes of their intensities. This is particularly problematic for approaches that depend on the $L_2$ (Euclidean) distance. We address both of these problems with the solutions introduced in refs. [38, 63].



JHEP06(2024)163

To take the soft constituents into account, which have intensities orders of magnitudes lower than hard constituents, we apply a non-linear scaling to all pixels of $I_{ij} \to I_{ij}^{\gamma}$. To address sparsity we convolve (smear) the whole image with a two-dimensional Gaussian kernel with an isotropic standard deviation $\sigma_k$. We find that using a value of $\gamma = 0.5$ for the pixel scaling alongside $\sigma_k = 1$ for the Gaussian kernel provides a adequate solution to both issues without excessive impact on the structure of the jets.

## 3 Method

### 3.1 Bump hunt

The bump hunt approach is a standard method used to search for excesses over a non-resonant background in HEP (high-energy physics) data. This method usually follows four main steps that we briefly discuss below. Each of these steps is a complex topic in itself with several different approaches in the literature, thus for our study, we choose only simplified and basic approaches.

#### 3.1.1 Signal enrichment

Signal enrichment, in general, refers to the selection of a subset of experimental data in such a manner that the fraction of signal events in it is increased compared to the initial sample. Most often, this is done by cutting out a region of the observable space where the signal is expected to be abundant compared to the rest of the space, typically using a theoretical model of the signal of interest.

These approaches, despite being sensitive to specific signal processes, make the search less model-agnostic and are ill-suited for general anomaly detection searches. Alternatively, one can hope to define a signal-rich region of the experimental data using a plethora of unsupervised ML (machine learning) techniques, which are expected to provide enhanced sensitivity over a wider range of potential signal processes.

In our particular example of LHCO data, we choose to explore a wide, smoothly falling region of the spectrum of dijet events with invariant dijet mass $m_{jj}$ from 3000 GeV to 4600 GeV. We choose this lower bound to avoid the turn on curve of the mass distribution, resulting from the jet trigger, and the upper bound is selected to remain in a region with relatively high statistics, so that we work in the region where the fit functions from subsection 3.1.2 are applicable. This interval contains, in total, around 380,000 QCD events and nearly all $Z'$ events. We divide this region into 16 non-intersecting bins with 100 GeV width each, as in refs. [105, 106].

#### 3.1.2 Background estimation

To perform a hypothesis test, one must first postulate a null hypothesis, which in counting experiments takes form of the expected background coming from the Standard Model processes. Often the background prediction relies on a theoretical basis to calculate the cross sections of the hard process and a simulation to account for detector response and measurement uncertainties. Still there are a number of searches where theory and simulation cannot provide



JHEP06(2024)163

a reliable background estimate. In these cases the background has to be estimated from the data itself in an empiric manner, using some general assumptions.

In dijet-like searches a background is often estimated by fitting a function of the form

$$f(x) = p_1(1-x)^{p_2}x^{p_3+p_4\ln(x)+p_5\ln(x^2)} \tag{3.1}$$

to a smoothly falling part of the dijet mass distribution [107–122], where $x = m_{jj}/\sqrt{s}$. This function is referred to as the "$n$-parameter dijet fit function", where $n$ is the number of nonzero free parameters $p_i$ used in the function. Despite being a good fit to the simulated data, this functional form is still an empirical assumption and thus is subjects to a systematic error. Furthermore, after applying some selection criteria on the events which could be correlated with $m_{jj}$, this function may no longer well describe the resulting distribution.

More advanced methods of fitting, such as the Sliding Window Fit (SWIFT) [123] and the ABCD method used in [124, 125] are other methods that reduce the assumption of a functional form but introduce their own assumptions instead. However, due to the simplicity and wide use of the $n$-parameter fit function, we choose to use global 3-parameter and 4-parameter function fits as the benchmark analysis strategy. Further details of the (pseudo-)analysis on the LHCO R&D data performed using these background estimates are given in appendix A. To access the upper bound on the performance of all background estimation methods, we use the underlying background distribution as an idealised fit, i.e. a fit with no systematic error. The (pseudo-)analysis using this is also described in appendix A.

### 3.1.3 Test statistic definition and calibration

There are several ways to calculate a global test statistic for two spectra. In HEP one of the more popular tests in model agnostic searches, called BumpHunter [126], relies on the maximal local significance (MLS) as the test statistic, where it is computed using a range of different windows over the spectrum. One of the benefits of the MLS test statistic is its simplicity and that it is well suited for signals that give rise to narrow, localised resonances. Here the MLS is applied to the binned $m_{jj}$ distributions of the data. Given a set $\mathcal{B} = \{b_1, \ldots, b_{n_{\text{bins}}}\}$ of non-intersecting bins with $N_{\text{sig+bkg},b}$ events or jets from the signal-rich (experimental) distribution and $N_{\text{bkg},b}$ events or jets from the background estimation, the MLS can be written as

$$\text{MLS} = \max_{b\in\mathcal{B}} Z_b = \max_{b\in\mathcal{B}} (\text{CDF}^{-1}_{\mathcal{N}(0,1)}(\text{CDF}_{Poisson(N_{\text{bkg},b})}(N_{\text{sig+bkg},b}))\,, \tag{3.2}$$

where CDF is the cumulative density function of the respective distribution. In equation (3.2) only overdensities are taken into account, i.e. $Z_b > 0$ only for $N_{\text{sig+bkg},b} > N_{\text{bkg},b}$ as we are searching for a resonance.

For bins with $N_{\text{sig+bkg/bkg},b} \gg 1$ one can approximate the Poisson distribution with a normal distribution $\mathcal{N}(N_{\text{sig+bkg/bkg},b}, \sqrt{N_{\text{sig+bkg/bkg},b}})$. Equation (3.2) then reduces to a much simpler form

$$\text{MLS} = \max_{b\in\mathcal{B}} Z_b = \max_{b\in\mathcal{B}} \frac{N_{\text{sig+bkg},b} - N_{\text{bkg},b}}{\sqrt{N_{\text{bkg},b}}}\,. \tag{3.3}$$

Although some test statistics, like $\chi^2$, have well-known distributions, other more unusual test statistics, like the BumpHunter test statistic, require calibration. This is commonly done by modelling its distribution using Monte Carlo simulation.



JHEP06(2024)163

Moreover, as systematic uncertainties arise from the definition of a signal region selection and the background estimate, this calibration should be performed even in the case where the distribution is known a priori. The calibration for the BumpHunter test statistic is performed in ref. [126] by running pseudo-experiments in which the counts in each bin are varied according to Poisson's law. This can be extended to higher dimensions by resampling the background events with bootstrapping. By calculating the test statistic for each of our bootstrapped background-only pseudo-experiments, we obtain the distribution of the test statistic in the background-only hypothesis. To ensure good modelling of the tail of the test statistic distribution, which corresponds to large significance values in the presence of signal, a large number of pseudo-experiments is required.

#### 3.1.4 Significance evaluation

To obtain a calibrated $p$-value for a given value of the test statistic $t$, one counts the number of background only pseudo-experiments exceeding this value $N_{>t}$ and divides it by the total number of pseudo-experiments done, $N_{\text{tot}}$.

The (one-sided) significance is computed using the inverse cumulative density function of the normal distribution $Z = \text{CDF}^{-1}_{\mathcal{N}(0,1)}(1 - p\text{-value})$.

In the case of $N_{>t} = 0$ arising from the limited number of pseudo-experiments, we instead set a lower bound:

$$p < \frac{1}{N_{\text{tot}}},\ Z > \text{CDF}^{-1}_{\mathcal{N}(0,1)}\left(1 - \frac{1}{N_{\text{tot}}}\right). \tag{3.4}$$

For every experiment with added signal events, we still bootstrap the background (for consistency) and combine it with a given number of signal events chosen at random from 100,000 signal events (around 5% of events fall outside of the evaluation region). Due to statistical fluctuations we also perform several pseudo-experiments in the signal enriched case in order to obtain a robust estimate of the significance for each level of signal doping.

### 3.2 Cluster scanning

In this section we present a novel approach called Cluster Scanning, which follows the same bump hunting scheme, but relies on a distinct set of assumptions than the commonly employed methods and thus has several favorable characteristics. Our approach can be divided into several key steps given below, with the hyperparameters chosen in order to search for narrow resonances in the $m_{jj}$ spectrum of the LHCO R&D data. The motivation for these hyperparameters in each step and the argumentation on how to choose them for a different application case is given in appendix C.

**Training region selection.** We select a narrow $m_{jj}$ window [3000, 3100] GeV for training of the k-means algorithm. This window contains 56,486 original background events. In this publication, we focus on relatively small signal injections that include only 5% or less of the total number of Z' signals available. Therefore the training region is expected to contain 89 signal events or less, which can be regarded as negligible (proof given in appendix G). In appendix G we show an improvement in performance in case the training region matches the resonant peak and thus has a larger portion of signals events involved in clustering. However,



JHEP06(2024)163

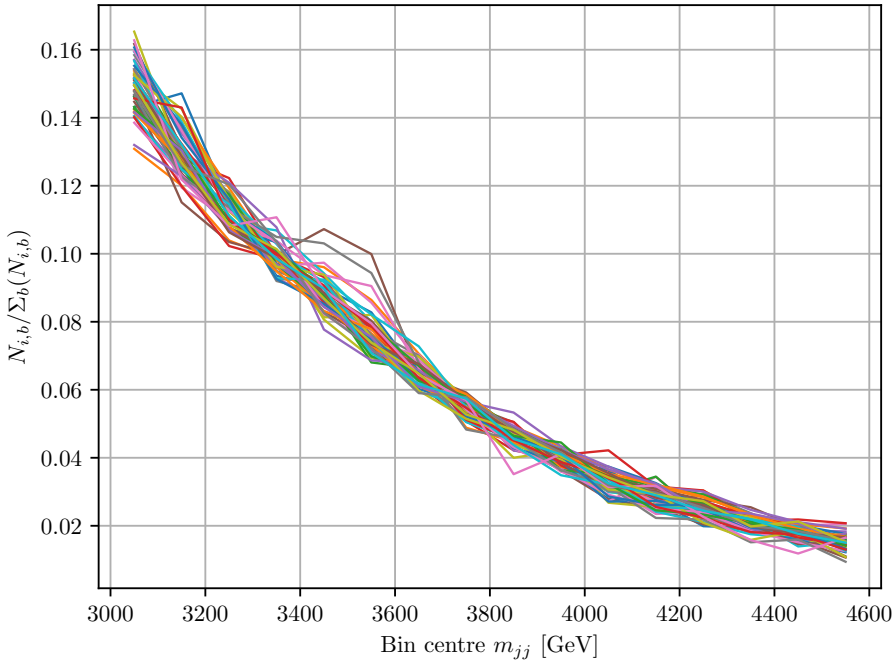



**Figure 2.** The $m_{jj}$ distributions for the jets in each of the 50 clusters, each normalised to unity. Here, 5,000 signal events have been injected into the evaluation dataset, which corresponds to 5% of the total available signal events.

in an actual analysis the position of the peak will be unknown, thus we choose to discuss a more representative case given here, when the training region happens to be in the tail of the signal peak and thus has a negligible number of signal events.

**K-means clustering.** We apply a mini-batch $k$-means clustering algorithm with $k = 50$ implemented in the SCIKIT-LEARN [127] library, with a batch size of 2048 on the set containing jet images of the leading two jets from each event in this $m_{jj}$ window. The mini-batch implementation is chosen due to its computational speed. The seeding of the cluster centroids is performed using the K-MEANS++ prescription described and motivated in ref. [128].

**Cluster spectra.** After performing the fit of $k$-centroids to the data in the training region, we fix the centroid positions and evaluate how many jet images from each of the 16 $m_{jj}$ bins of the evaluation region, defined in subsection 3.1.1 (it includes training region as one of the bins), fall into each of the $k$ clusters $N_{i,b}$ where $i \in \{1 \dots k\}$, $b \in \{1 \dots n_{\text{bins}}\}$. Figure 2 shows the resulting 50 normalised cluster spectra $N_{i,b}/\sum_b(N_{i,b})$ for one pseudo-experiment with signal injection.

**Per bin standardisation.** We note that in each bin the normalised cluster spectra follow an approximately normal distribution with several outliers from the anomalous clusters (see discussion in appendix D). Therefore we standardise the normalised cluster spectra in each bin using outlier robust estimators (described in appendix E) for mean and standard deviation with an outlier factor of 0.2. Here we make the assumption that the majority of the signal is



JHEP06(2024)163

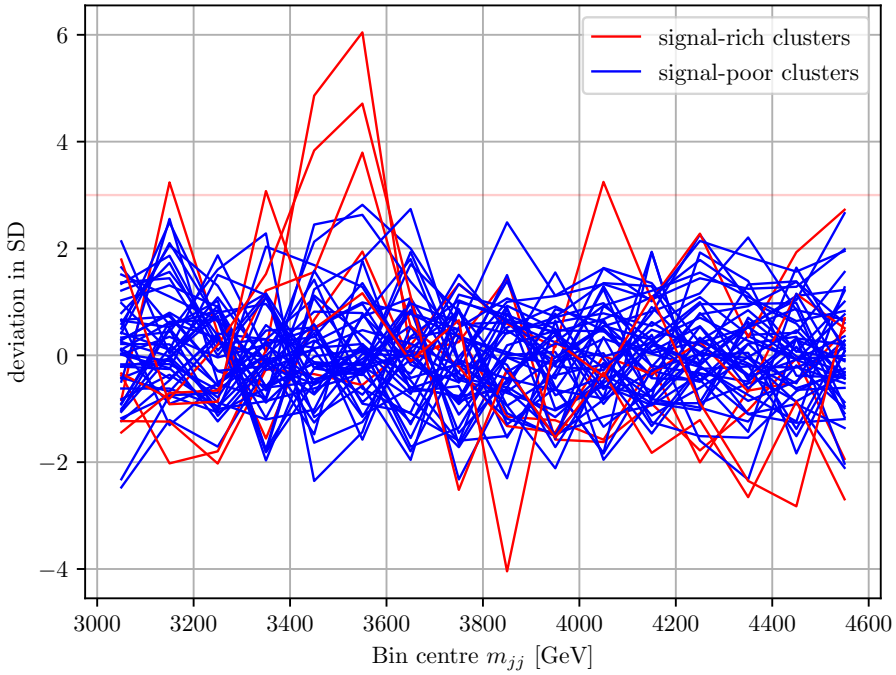



**Figure 3.** Spectra in figure 2 standardised over clusters in each bin. Potentially signal-rich cluster spectra are shown in red.

located in a small number of clusters, and the rest of the clusters are signal depleted. Figure 3 shows the cluster spectra from figure 2 after normalising with the outlier robust estimator.

**Selecting anomalous clusters.** Utilising the assumption that the signal is localised in $m_{jj}$, we select potentially signal-rich cluster spectra as those with a deviation of more than a threshold value of $\theta = 3$ standard deviations from the robust mean in the positive direction as we are only interested in a resonance leading to excess of events. The rest of the clusters are labelled as signal-depleted. The threshold and the selected signal-rich clusters are shown in figure 3 in red.

**Signal-rich and signal-depleted regions.** After the selection, we combine the non-normalised distributions corresponding to our selected signal-rich clusters. This results in a signal-rich spectrum $N_{\text{sig+bkg},b}$ with an example shown in red in figure 4.

The remaining cluster spectra are combined to form a signal-depleted spectrum $N_{\text{poor},b}$. The estimate of the background is then constructed by normalising it to the same total entries as in signal-rich spectrum, namely $N_{\text{bkg},b} = N_{\text{poor},b} \frac{\sum_b N_{\text{sig+bkg},b}}{\sum_b N_{\text{poor},b}}$. It is shown in blue in figure 4.

**Test statistic.** As previously discussed, to test the significance of an observed excess we use the simple maximum local significance, as defined in equation (3.2). It may occur that no cluster is selected as anomalous. In this case we assign a default value of 0, in order to show good agreement with the null-hypothesis expectation. This is similar in motivation to setting the value for an observed deficit in events to zero. Following the discussion in subsection 3.1 for



JHEP06(2024)163

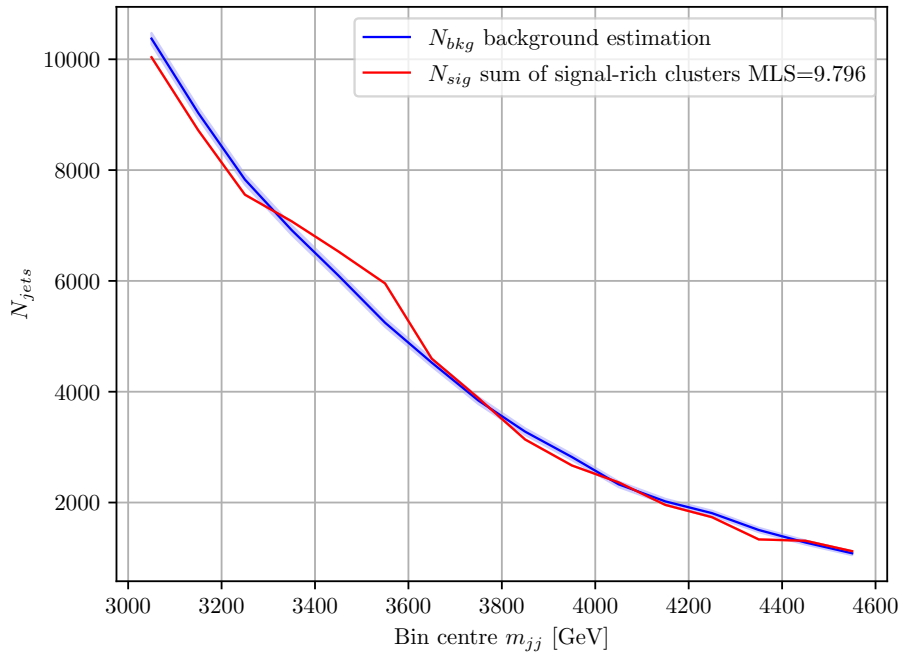



**Figure 4.** Curves corresponding to the sum of signal-rich and signal-poor spectra in figure 3. The blue signal-poor curve is rescaled to have the same total jet number as the signal-rich curve. The coloured region around the blue curve is a $\sigma_{\text{bkg},b} = \sqrt{N_{\text{bkg},b}}$ Poisson deviation after recaling used to compute MLS.

the calibration process, we construct 3,900 pseudo-experiments using bootstrap resampling on 1 million background events. The distribution of the test statistic is discussed in appendix F.

**Ensembling.** Different initialisations lead to a broader distribution over the final test statistic obtained with cluster scanning. In order to obtain a final value for the test statistic, the cluster scanning method is performed 15 times with independent initialisations. The mean of the test statistic from all the runs forms the final ensembled test statistic. The distribution of this statistic is presented in appendix F.

### 3.3 Discussion

As we can see, CS follows the general bump hunt strategy, but introduces novel approaches for the first two steps of this strategy. First of all, CS selects the most anomalous looking clusters to define the signal-enriched region, and constructs a background estimate from the rest of the clusters. Notably though, this selection is completely data-driven and does not target a specific family of signal models. However, CS relies on a set of assumptions that fundamentally differ from those commonly used in other anomaly detection approaches.

**Search for overdensity instead of outliers.** Most anomaly search methods like Autoencoders [38] and SVDDs [88] rely on outlier detection, namely, identifying the data instances that lie in a region of very low probability density or outside the support of the "normal" distribution. Notably, while all normal events share similar characteristics and exhibit easily



JHEP06(2024)163

recognisable trends, anomalous data, such as defects or fraud, can differ in numerous ways and are thus given a wide prior. Although model-agnostic searches should accommodate a wide range of possible anomaly models, it is usually assumed that a signal is produced by only one or a few unknown BSM process. Thus, all anomalous events have many features in common and exhibit some similarity to SM events, as any new particle must radiate and decay into SM particles to be detectable.

Therefore we use the localisation of anomalies in both low-level (e.g. jet images) and high-level variable (e.g. $m_{jj}$) space as the first main assumption of the CS method. Localisation of anomalies in low-level variable space means that only a few out of all clusters contain a fraction of anomalies much higher than the rest of the clusters. This way clustering plays a role of data-driven binning in low-level variable space. Localisation in $m_{jj}$ gives us a possibility to distinguish these anomaly rich clusters from the rest, namely, by searching for an overdensity in $m_{jj}$ in one cluster spectrum compared to all others. Thus, CS is able to select a signal-rich region of events by leveraging the assumption of signal being localised rather than consisting of outliers.

Although semi-supervised methods based on CWoLa (see refs. [105, 106, 129–133]) and density estimation methods are also sensitive to overdensities, they usually require construction of a background template, which until recent developments [132, 133] was preferably constructed for a smooth distribution of low dimensionality, typically using a few high-level observables. In this publication, we show that CS is able to draw significant improvement from a high-dimensional distribution of low-level jet observables. In this way, it can be considered less signal-model dependent than the methods that rely on hand-crafted high-level observables.

**Assume cluster mass independence instead of smoothness.** CS proposes a solution to the second step of the analysis, namely, it estimates the form of the background by combining the signal-depleted clusters. In this way, we do not rely on any assumptions on smoothness or on a particular functional form of the background-only spectrum in $m_{jj}$, which are heavily relied upon by most other bump hunt methods, such as the global functional fit mentioned in subsection 3.1, SWIFT [123] and even Gaussian processes [134].

Instead, the second main assumption of CS is that in the background-only case, the assigned cluster centroid is approximately independent from $m_{jj}$. Ideally, we would want the distribution of background events over $m_{jj}$ in each cluster to be identical within statistical uncertainty, such that the probability of a jet belonging to a cluster and having a specific mass factorises, $p(i, m_{jj}) = p(i)p(m_{jj})$, or at least that the correlation is weak. This would minimise the rate of incorrectly identified signal-enriched clusters. In practice, although figure 2 shows that the distributions all follow a similar trend, there are still some systematic deviations. These are a result of the finite width in $m_{jj}$ of the training window and slight correlations between the distribution of the jet constituents and $m_{jj}$ arising from the transverse momenta of the two non-resonant jets depending on $m_{jj}$. Therefore, for the selection of the clusters, we estimate the uncertainty separately for each experiment and for bin based on the sample of our $k$ cluster spectra values. This uncertainty estimate includes both, statistical Poisson fluctuations and systematical uncertainty from mass dependence (see discussion in appendix D).



JHEP06(2024)163

Unlike in methods with sliding window approach [106, 123], in CS the fit only needs to be performed once.[1] Moreover $k$-means clustering is a simple classical algorithm that typically requires less training than deep learning approaches, making CS a relatively fast analysis method. This is important in the context of the ensembling and calibration, which both require a large number of analysis iterations, and are thus a notable obstacle to incorporating deep learning in HEP analysis under the constraint in computing resources. Fast analysis is also advantageous for testing its efficiency for simulated BSM events in order to produce the exclusion limits (see the RECAST [135] framework). Moreover, CS avoids other disadvantages inherent to sliding window approaches, such as limited search range due to the definition of the sidebands and the need to optimise sideband and signal window widths.

### 3.4 Idealised CS

Despite choosing a narrow $m_{jj}$ window to reduce mass dependence systematics, the variables that we use for clustering are in general not independent of $m_{jj}$. Thus, we observe the background-only spectra of some clusters do not just statistically fluctuate around the expected shape of the background, but exhibit some degree of smooth mass sculpting. This affects the performance of the method by introducing false positives at the cluster selection stage. We expect that this may be partially remedied by a more sophisticated method of selecting anomalous clusters or a better background estimate, both of which would rely on further assumptions. These studies are outside the scope of this publication. However, to give an upper bound on the performance one may achieve with such improvements we propose an idealised version of clusters scanning.

Idealised CS version requires us to modify the distribution of the jets between the clusters. First, we count the numbers of jets that fall into each cluster in the first $m_{jj}$ bin. If no mass dependence were present, the fractions of QCD jets in each cluster $x_{\mathrm{QCD},i,b} = N_{\mathrm{QCD},i,b}/\sum_i N_{\mathrm{QCD},i,b}$ should be independent of bin number $b$ within statistical uncertainties. To simulate this case in all the consecutive bins except the first we distribute the QCD jets in these bins among clusters using a multinomial distribution with weights equal to the fractions obtained in the first bin $x_{\mathrm{QCD},i,b} = x_{\mathrm{QCD},i,1}$, thus generating cluster spectra that follow the original background spectrum with statistical fluctuations, i.e. the case with no mass dependence. The signal jets are distributed as before according to which cluster they belong to, such that the fractions of $Z'$ jets may differ between different bins. This is done because we assume that only the background is distributed roughly proportionally between clusters, which is equivalent to assumption 2, but not the signal.

This distribution of jets creates idealised cluster spectra for each clustering, and the rest of the algorithm remains unchanged.

## 4 Results

As a proof of concept we perform an analysis applying CS and global fit based bump-hunting with the above mentioned hyperparameters to the LHCO R&D dataset with different amounts

[1] Training and evaluating CS using a sliding window approach was considered; however, the resulting spectra exhibited abrupt discontinuities due to relatively low statistics in the high $m_{jj}$ bins, making them unsuitable for further analysis.



JHEP06(2024)163

of signal injection, given in figures either as an absolute number of injected events $\epsilon$ or as a signal to background ratio $S/B$ of events in the considered [3000, 4600] Gev $m_{jj}$ region. For each pseudo-experiment with signal injection we calculate the significance $Z$ as discussed in subsection 3.1.4 using the calibration test statistic distribution. For each signal injection level we run 100 pseudo-experiments with bootstrapped background data and randomly sampled signal events. As a reference for the significance and its statistical variation for each contamination level, we report the median significance of these pseudo-experiments and 0.25 and 0.75 quantiles. We define the ratio between the median significance provided by CS to the significance of a baseline method as the significance improvement (SI). We also quote the relation between the number of events needed to obtain a $3\sigma$ evidence in each analysis strategy.

Figure 5(a) shows how the global significance of CS and the parametric fit-based methods depends on the signal contamination in the pseudo-experiment. It characterises the performance of these realistic analysis strategies, which do not use any truth information for the evaluation of the test statistic, thus including all the systematical uncertainties coming from partially fulfilled assumptions needed for the respective method.

We observe that although 3- and 4-parameter fits give approximately the same results, CS outperforms them by a significant margin in the region from 1500 to 4000 signal events. Beyond 3000 signal events, the significance yield from CS is limited by the number of bootstrap pseudo-experiments in the calibration set, but the lower bound on its significance still remains substantially higher than the significance of the parametric fits. This is the most interesting region as there the transition between non-significant signal (below $1\sigma$) and new physics evidence (above $3\sigma$) takes place. Looking at the lower subplot in figure 5(a) we see that CS gives us an SI of 2 and higher on the majority of regions of interest. We can also see that CS produces a $3\sigma$ evidence for only 61% of the events needed to obtain this evidence with the parametric fit. This shows that although both suffer from fit and assumption induced systematic uncertainties, CS has a clear advantage over parametric fitting procedures.

Above we have also described the idealised version of the cluster scanning method and in appendix A the analysis with an idealised background fit. Both methods rely on event labels to remove systematic uncertainties introduced by the limitations of the assumptions of our methods and to make the background estimate in both cases close to the true background, with only statistical fluctuation taken in consideration. This is done to separate the influence of additional information, namely the low-level observables used in the analysis, from the systematic uncertainties introduced by the fits, and to construct the upper bound on the performance of our methods.

Figure 5(b) shows how the global significance of both idealised methods depend on the signal contamination in the pseudo-experiment. It can be seen that one needs substantially less signal for it to be significant in the idealised methods compared to the realistic methods, as it removes false-positives induced by systematic uncertainties in the fits. Still, we see that the idealised CS outperforms the idealised functional fit on the majority of the interval between 1000 and 2200 signal events. Looking at the lower subplot in figure 5(b) we see that by using CS in the region of interest we gain a significance improvement factor of up to 1.5. We can also see that CS produces a $3\sigma$ evidence with only 69% of the events needed to gain this evidence with the idealised fit. This shows that in the case of negligible



JHEP06(2024)163

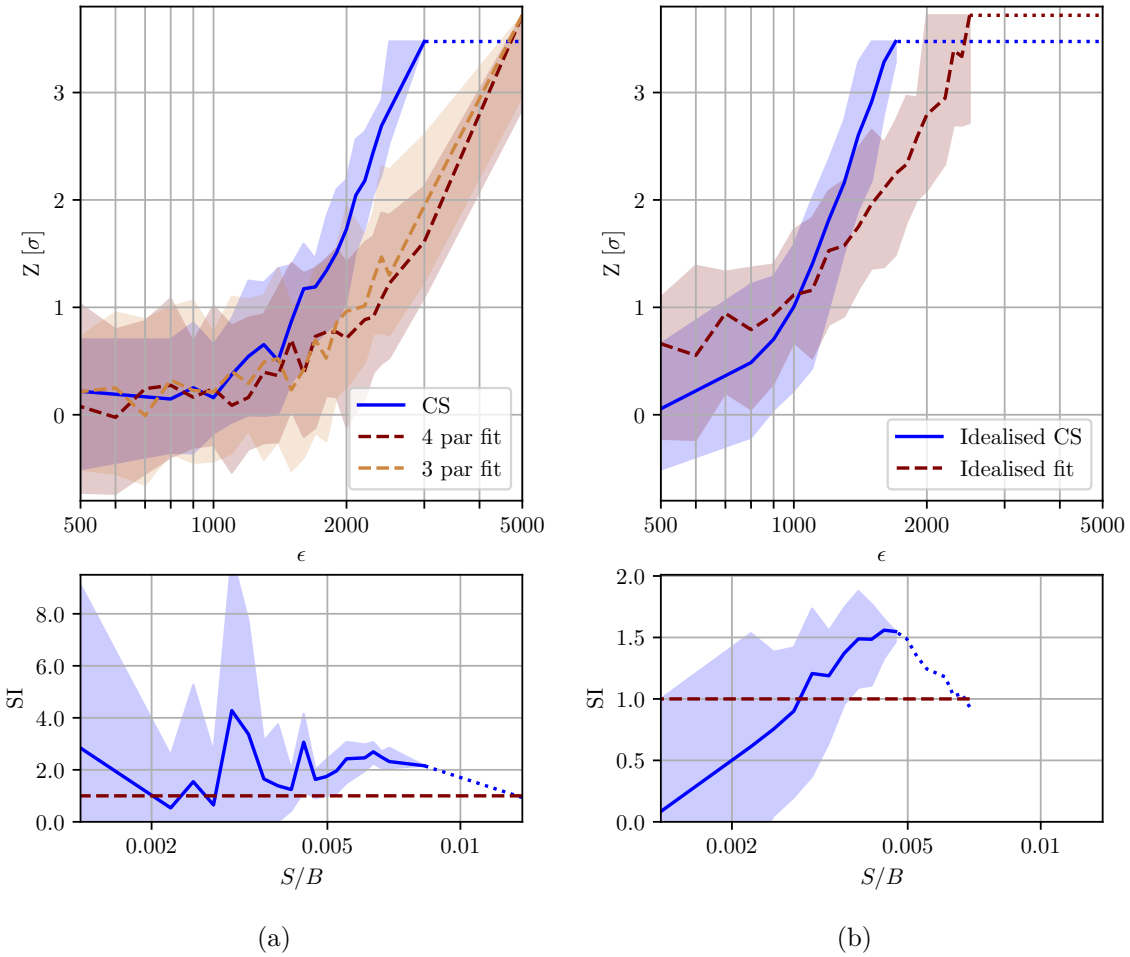



**Figure 5.** Upper figure: Median and the quartile bounds of global significance of the signal contaminated pseudo-experiments as a function of the number of signal events $\epsilon$ injected shown for the (a) realistic and (b) idealised analysis methods. The dotted lines mark lower bounds, as there was not enough statistics to access higher significance levels. Bottom figure: Significance improvement of the CS method compared to the 4-parameter fit for the (a) and the significance improvement of idealised CS compared to idealised fit for the (b) as a function of the signal-to-background ratio $S/B$.

systematic uncertainties, CS gives an improvement over any smooth fit as, in addition to just using information from $m_{jj}$, it also makes use of the low-level event information. From the difference between idealised and non-idealised CS we can see that there is some room for improvement of CS to reduce the false positive rate, and improve the analysis efficiency.

## 5 Conclusions and outlook

This paper is a first proof of concept for the cluster scanning anomaly search method, which is designed to search for resonant overdensities on the distribution of an observable using clustering techniques in auxiliary observables.

We found that it outperforms the widely used bump-hunting method, which relies on the functional background fits, in several metrics relevant to an analysis. In the transition



JHEP06(2024)163

region, where the benchmark algorithm achieves $1\sigma$ to $3\sigma$ significance, CS improves the result by a factor of 2 or more for the realistic case, or by a factor of 1.5 for the idealised comparison. This reduces the number of signal events required to produce a $3\sigma$ significance by a factor of 0.61 in the realistic case and by a factor of 0.69 in the idealised case. The former factor of improvement should be expected in a real application. We also discuss the comparison of cluster scanning with other anomaly detection algorithms in subsection 3.3, outlining its advantages and limitations.

The CS method should not be seen as a direct competitor to background fitting methods, but rather as a complementary approach that relies on a different set of assumptions about the nature of the anomaly and the background distributions, which are not well known.

There remains a large unexplored field of potential extensions and improvements to this method or synergies with other methods. Straightforward follow-up studies can explore the use of clustering methods other than $k$-means. One can look for other ways of selecting the anomalous clusters, alternatives to the one proposed in subsection 3.2, that would rely on different assumptions. For example, one can require that all anomalous clusters are neighbors in the space of clustered inputs. One can also unify the assumptions of CS and functional fits to produce separate background estimates for each of the clusters separately, greatly reducing the $m_{jj}$ dependent systematic uncertainties.

CS could benefit from using features developed by other algorithms that have already been optimised for other tasks, such as flavour tagging, or even using unsupervised learning for feature extraction.

Furthermore, since many other ML approaches to improve sensitivity in model-independent searches rely on a bump hunt for the final statistical analysis, CS could also be used to further enhance sensitivity. This could be of particular interest when the background distribution is no longer well described by simple, smoothly decreasing functional forms.

## Acknowledgments

The authors would like to acknowledge funding through the SNSF Sinergia grant CRSII5_193716 "Robust Deep Density Models for High-Energy Particle Physics and Solar Flare Analysis (RODEM)" and the SNSF project grant 200020_212127 "At the two upgrade frontiers: machine learning and the ITk Pixel detector". The research of MK is supported by the DFG under grant 396021762 – TRR 257: Particle Physics Phenomenology after the Higgs Discovery.

**Code availability.** The code used to produce all results presented in this paper is available at https://github.com/IvanOleksiyuk/jet_cluster_scanning.

## A Idealised fit and n-parameter fit pseudo-analysis

In general, if a distribution $\mathcal{H}_b$ of a background events is perfectly known a-priory, given a binning for this distribution, one can calculate the expected number of events in each bin. Being provided directly from the true underlying $\mathcal{H}_b$, this background estimate will on average provide the most efficient tests to discriminate samples drawn from $\mathcal{H}_b$ from



JHEP06(2024)163

samples drawn from an alternative hypothesis $\mathcal{H}_{b+s}$ with signal, compared to any other estimate of the expected background for these samples. Hence we call the expectation from $\mathcal{H}_b$ an "idealised fit".

As discussed extensively in section 3.1, estimation of the expected background is only one step of the analysis. To create a benchmark, we do pseudo-analysis on LHCO R&D dataset represented by spectrum $N_{\text{bkg,orig},b}$ using the other choices defined in section 3.1. Namely, we generate pseudo-experiments by bootstrap resampling the events from $N_{\text{bkg,orig},b}$ and add a number of signal events if needed. The "idealised" background estimation for every pseudo-experiment is equal to $N_{\text{bkg,orig},b}$ itself (as the samples were generated with these expected values). Following the discussion in subsection 3.1.3 we use MLS test statistic between this estimate and the generated pseudo-experiments, to generate null-hypothesis test statistic distribution and its value for signal contaminations and thereafter estimate the significance. Depending on the number of doped signal events, the median and quartile region significance given by this test is provided in the main text in figure 5(b).

Unfortunately the background model is usually unknown, so for each experimental sample the background should be estimated in some less precise way relying on weaker assumptions.

As a realistic benchmark to our method we explore how sensitive the analysis is using global n-parameter functional to the kind of signal presented in LHCO R&D dataset. We use the binning with 16 bins defined in subsection 3.1.2 and count the number of background events in each bin to get an original background spectrum $N_{\text{bkg,orig},b}$.

For all the fits in this studies, we use Trust Region Reflective nonlinear least squares fitting algorithm implementation from SCIPY python package [136]. The chosen bins generally contain more than 5000 counts, so the Poisson distributions of these counts can be well approximated by a Gaussian distributions with the variances equal to the bin counts. Using variances to scale the summands in the least squares objective we make it equivalent to the maximum likelihood objective for this setup.

First, we fit our 3- and 4-parameter functions to the spectrum to see if the fit is valid. Resulting fits with 13 and 12 degrees of freedom score $\frac{\chi^2_{3-par}}{n_{\text{dof}}} \approx 1.201$ and $\frac{\chi^2_{4-par}}{n_{\text{dof}}} \approx 1.338$ that correspond to p-values of 0.275 and 0.182 which signify validity of these fits.

Unlike the CS method that doesn't generally rely on the smoothness of the background, global n-parameter takes it as the main assumption, so as $N_{\text{bkg,orig},b}$ already has some statistical fluctuations a distribution resampled from it will have even larger statistical fluctuations than the ones expected for Poisson distribution. To simulate the proper scale Poisson fluctuations in the chosen region for our pseudo-experiments we resample events not from $N_{\text{bkg,orig},b}$ but from the best possible fit. This also negates the systematic error from null-hypothesis not corresponding to the empirical functional form, so these experiments can be viewed as semi-idealised. In a more realistic cases, the space of functions given by all possible parameter values, does not contain the true form of null-hypothesis distribution and can only yield an approximation of it with limited precision. It is usual for fit functions with a small number of parameters, but with increasing number of parameters the function fit problem becomes over-defined and the function can fit the signal bump as well. Experimentally we have observed only insignificant increase in performance when comparing sampling from $N_{\text{bkg,orig},b}$ or from the best fit distributions. On top of



JHEP06(2024)163

the resampled background events we add a number of signal events from signal's original distribution when needed.

The initial parameters of the fit in each experiment are chosen to be equal to the optimal parameters of the initial fit discussed above, so that one gets an "idealised" background fit if no optimisation is done. However, because of the statistical fluctuations and/or added signal contamination, the maximisation of likelihood results in a different set of parameters for this functional form. This error of background mismodeling under its statistical fluctuations and addition of the signal is exactly the type of error we want to demonstrate with this pseudo-analysis.

The results of such analysis for different signal contamination is given in figure 5(a). We can see that the 3-parameter fit provides a slightly better result than 4-parameter fit as the latter has more flexibility to overfit the signal and the statistical fluctuations. This is so because the samples are drawn from 3- and 4-parameter functions with fixed parameters themselves. If we were to sample from other distribution the error coming from mismatch in true end expected functional forms may switch this ordering but it will reduce both performances. Therefore, the curves shown in figure 5(a) are upper limits of these realistic n-parameter fit analyses achievable only when the true distribution is described by one of the functions in the chosen parameterised space.

## B Sparsity of the jet images

Figure 6 show that the jet images are very sparsely populated ususally having less tha 100 non-zero pixels per 1600 pixels total.

## C Hyperparameter selection and motivation

In this appendix we give motivation for every not yet discussed choice of hyperparameter in our pseudo-analysis. All the hyperparameter suggestions are done in an unsupervised way coming from general assumptions about signal and background and are not optimised using the truth information from LHCO R&D data. As such the levels of significance improvements may be further increased by performing a dedicated parameter scan for a specific application, however, we recommend to follow the same reasoning when applying CS in other analyses.

**Training region.** Training on the full spectrum would likely result in each cluster corresponding to a specific mass region, thus the background spectrum for each cluster would not be close to the original mass spectrum. Therefore we perform clustering in a narrow mass window [3000, 3100] GeV. We choose this window as it lies in the studied region defined in subsection 3.1.1 and has the largest statistic of all other 100 GeV windows.

**Number of clusters.** The most important parameter we had to choose is the number of clusters $k$. Two factors play the key role in this choice. On one hand, the number of clusters has to be as large as possible to better narrow down the anomaly-rich region. On the other hand, for a given number of events in the evaluation region and the binning of this region one has to take the number of clusters sufficiently small so that the least populated clusters



JHEP06(2024)163

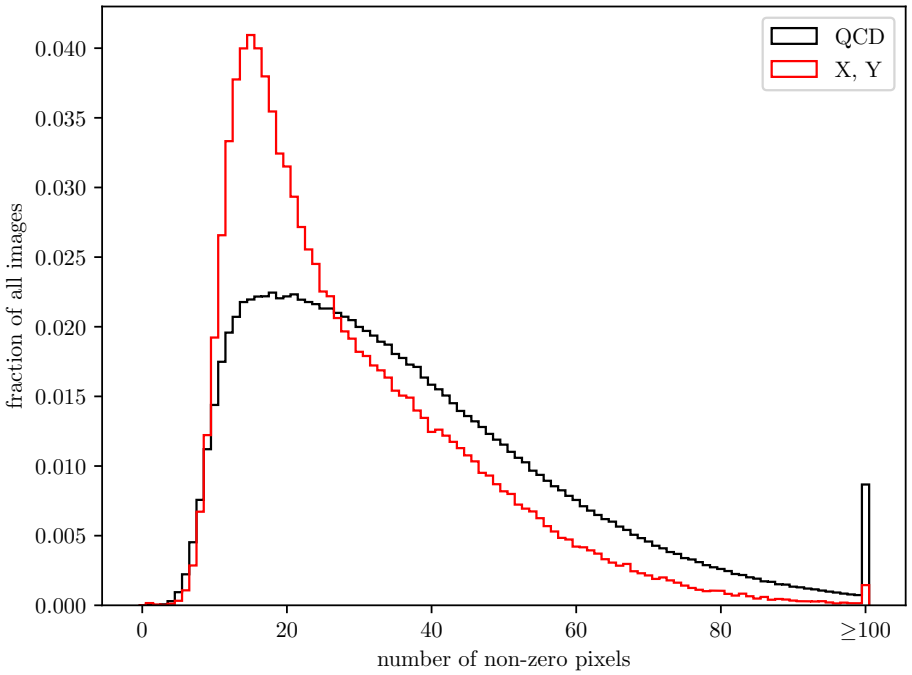



**Figure 6.** Distribution of images in QCD and top datasets from top-tagging task vs the number of non-0 pixels in them.

in the smallest $m_{jj}$ bin $\underset{i,m_{jj}}{min}(N_i(m_{jj}))$ still has enough statistics for a meaningful statistical analysis. We assume that $\underset{k,m_{jj}}{min}(N_i(m_{jj})) = O(50)$ should be sufficient. Using a coarse search, we determine, that for our choice of binning and overall statistic at hand choosing $k = 50$ gives a good trade-off as it has a median of 55 events in smallest cluster-bin and it goes below 20 only 1 time in 1000 pseudo-experiment runs, as can be seen in figure 7.

**Batch size.** Scikit-learn [127] documentation states that the parallelisation is performed on all available $N_{\text{cores}}$ computing cores if the batch size is $N_{\text{cores}} \cdot 256$ or larger. We performed all computations with 8 core parallelisation, thus the natural choice of a batch size was 2048. It is also important to maintain the batch size much larger than the number of clusters to ensure faster convergence.

**Outlier fraction.** To quantify the performance of our method we introduce the signal fraction improvement score (SFI) that characterises a subset $\mathcal{S}$ of the events in evaluation set $\mathcal{E}$ by the relative increase in the signal to background ratio

$$\text{SFI}(\mathcal{S}) = \frac{N_{\text{sig}}(\mathcal{S})}{N_{\text{QCD}}(\mathcal{S})} \frac{N_{\text{QCD}}(\mathcal{E})}{N_{\text{sig}}(\mathcal{E})}. \tag{C.1}$$

Our main assumption is that the signal is distributed in clusters unevenly and there only several clusters have a significantly large SFI. To put it in numbers, we assume that not more than 20% of clusters have SFI of 2 or more. Following this assumption we choose the outlier fraction of 0.2 for outlier robust estimators. This is an ad. hoc prior assumption about the



JHEP06(2024)163

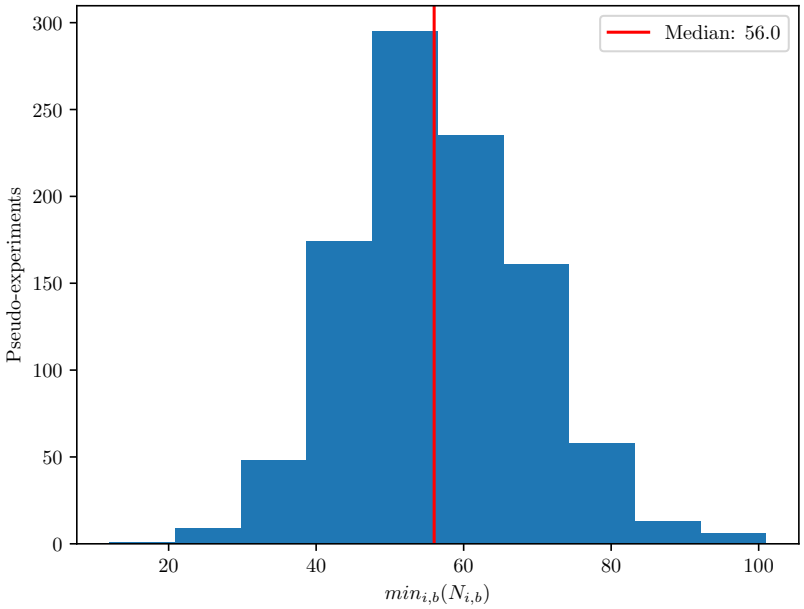



**Figure 7.** Distribution of the number of jet image counts in the least populated bin of the least populated cluster in each of the 1000 random background only runs of the CS algorithm on backround only data.

data at hand, and it has to be made prior to analysis and has no way to be validated without knowing the truth lables. Still we can show that this assumption is satisfied in our case with a margin for the pseudo-experiment shown on all the figures of section 3. 5000 signal events were giving an overdensity on the original spectrum that was not identifiable as a deviation from smooth background by human eye (without knowing the background truth), but in figure 2 one can easily notice two spectra with a significant bump around 3.5 TeV that stand out of the crowd of other spectra. Unsurprisingly these two spectra have SFIs of 9.1 and 8.9. Three more clusters also have a visible overdensity at this position possessing SFIs of 6.3, 5.6, 4.4. In total, exactly 8 clusters have $SFI > 2$. Still as we will see later only 3 of these clusters have a signal significant enough to be selected as anomalous, showing that our assumption is quite conservative in its limit and either the threshold $SFI$ can be increased or the percentage of clusters to path the threshold reduced for it to still remain a valid assumption. Runs of the analysis on other (pseudo-)experiments behave in the similar manner.

**Cluster selection threshold $\theta$.** First of all, we use the threshold only for positive deviations as we only search for excesses of events. Apart from the signal-rich outlier clusters the threshold can be passed by signal poor clusters, but only with an expected false positive rate of $1 - (1 - \text{p-value}_{\mathcal{N}(0,1)}(\theta))^{n_{\text{bins}}}$. Then for large enough thresholds the average number of false positives can be estimated as $k \cdot n_{\text{bins}} \cdot \text{p-value}_{\mathcal{N}(0,1)}$. Higher thresholds result in lower false positive and lower true positive rates. To retain the sensitivity for statistically small signal we choose to use $\theta = 3$ that will result in approximately $50 \cdot 16 \cdot 0.00135 = 1.08$ signal poor cluster being assigned a false positive label on average. Figure 3 shows 4 clusters being chosen using this threshold. Three of them have an overdensity at 3.5 TeV and one does not, implying that it is a likely false posive.



JHEP06(2024)163

| Parameter | value | motivation |
|---|---|---|
| k | 50 | $\min_{i,b}(N_{i,b}) = O(50)$ with binning below |
| mini-batch | 2048 | $N_{\text{cores}} \cdot 256$, must be $\gg k$ |
| Training region | [3, 3.1] TeV | narrow mass window with high statistic |
| Evaluation region | [3, 4.6] TeV | the n-parameter fit is applicable<br>excluding low statistic regions |
| Bin width | 100 Gev | broad enough to have sufficient statistics in each bin |
| outlier<br>fraction $f$ | 0.2 | consistent with assumption<br>on the maximum number of signal clusters |
| Cluster selection<br>threshold $\theta$ | 3 $\sigma$ | low enough to let trough many true positives<br>but high enough to filter most false positives |
| Test statistic | MLS | simple and specialised for local excesses |
| Default TS | 0 | minimal test statistic possible |
| Ensemble size | 15 | As large as possible realistic compute limitations |

**Table 1.** Summary of the hyperparameters used in cluster scanning.

**Ensemble size.** We recommend to take the ensemble size as high as possible, for given computation resource constrains to reduce the width of the test statistic distribution (see appendix F).

## D Distribution of cluster scanning bin entries

The assumption on the Gaussianity of cluster spectra in each bin can be shown to be valid by robustly standardizing the cluster counts in each bin (see appendix E) and checking if these distributions match $\mathcal{N}(0,1)$. The upper part of figure 8 visually demonstrates that the distribution of cluster spectra in each bin in figure 3 matches the Gaussian model, except in bins $3.4\,\text{TeV} < m_{jj} < 3.5\,\text{TeV}$ and $3.5\,\text{TeV} < m_{jj} < 3.6\,\text{TeV}$, where many outlier clusters are found due to the presence of signal. Fifty samples are usually not sufficient to determine whether the distribution is Gaussian or not, but by marginalizing (summing up) over 16 bins, we obtain 800 samples in total. The lower plot in figure 8 shows the said distribution. If we consider only signal-poor clusters, the distribution fits the $\mathcal{N}(0,1)$ distribution well visually and according to a consensus of three normality tests: Shapiro-Wilk, Kolmogorov-Smirnov, and Jarque-Bera (note that the p-values for all the tests are high). Including signal-rich clusters adds outliers, which is reflected in lower p-values for the normality tests; however, apart from these, it still can be well approximated by a unit Gaussian.

The variance among clusters in each bin depends on both statistical fluctuations and mass-dependent systematic effects. In the case of infinite statistics, all cluster spectra would appear smooth but would vary from one another due to differences in mass sculpting. In the absence of mass dependence, the variation in each bin would be caused solely by Poisson fluctuations (as mentioned in subsection 3.4, dedicated to idealised CS). In a realistic scenario, these two factors cannot be separated but can be jointly estimated using a robust standard deviation for each bin (see appendix E).



JHEP06(2024)163

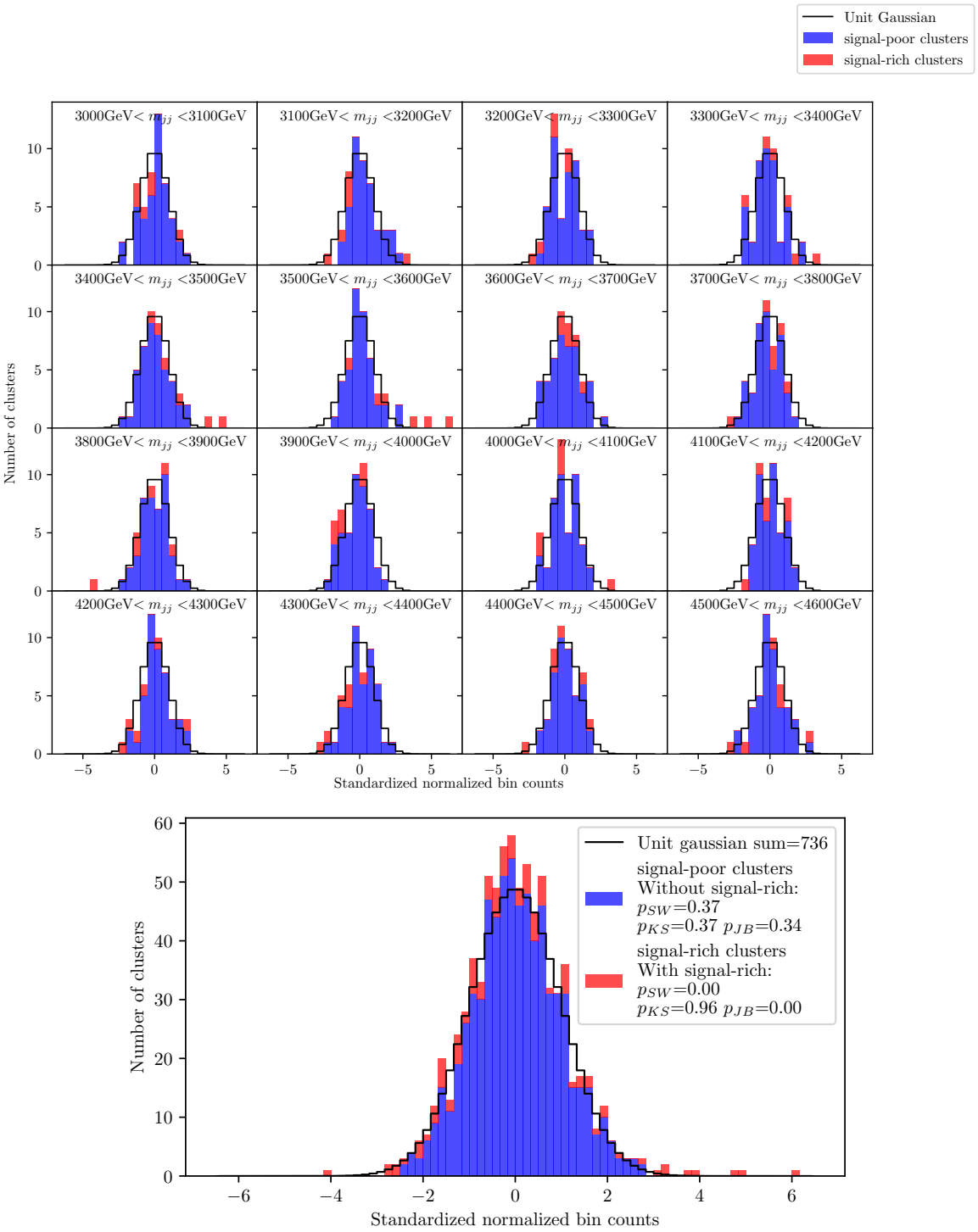



**Figure 8.** Top: a histograms of cluster spectra for each bin in figure 3, depending on the deviation from the robust mean measured in robust standard deviations. Bottom: sum (marginal) of all the histograms at the top compared to expected bin counts of a Gaussian with a sum of 736 count (all the signal poor counts).

JHEP06(2024)163

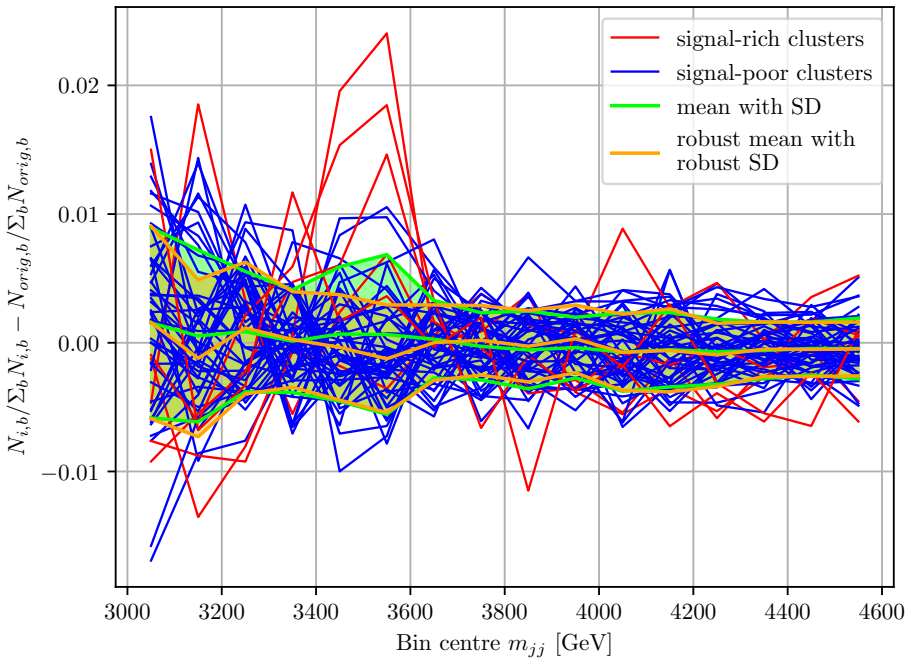



**Figure 9.** Normalised spectra with subtracted normalised original $m_{jj}$ spectrum. Amount of signal is 5000. The selection of the anomalous clusters is taken from figure 4.

## E Outlier robust estimators

While searching for outliers, it is preferred to use outlier robust estimators for standard deviation (SD) and mean. We define them as follows: given a sample of observations $S = \{x_1, x_2, \dots x_n\}$ we find a median $med(S)$ (which is itself an outlier robust estimator) of this sample and take a subsample $\tilde{S}_f$ that is constructed from $S$ by discarding a fraction $0 < f < 1$ of all samples that have largest absolute distance to this median. In this way we have discarded the outliers. After that we construct estimators $\tilde{\mu}_f = mean(\tilde{S}_f)$ and $\tilde{\sigma}_f = SD(\tilde{S}_f) \cdot g(f)$. If $S$ is a sample from $\mathcal{N}(\mu, \sigma)$ it is obvious that with $\lim\limits_{n\to\infty} \tilde{\mu}_f = \lim\limits_{n\to\infty} mean(S) = \mu$. If one takes $S$ from $\mathcal{N}(0, 1)$ and rescales $x_i \to \sigma x_i$, then both estimators transform as $\tilde{\sigma}_f \to \sigma\tilde{\sigma}_f$ and $SD(S) \to \sigma SD(S)$ by definition, so both estimators $\tilde{\sigma}_f$ and $SD(S)$ are proportional to a true $\sigma$ of the Gaussian distribution.

Both $\tilde{\sigma}_f$ and $SD(S)$ are independent of $\mu$ and there are no other parameters of the normal distribution for estimators to depend on, therefore for a family of Gaussian distribution estimators $\tilde{\sigma}_f$ and $SD(S)$ are proportional to each other by some constant factor $g(f)$ in the limit of infinite sample. In other words, adjusting numerically $g(f) = \frac{SD(\mathcal{N}(0,1))}{\tilde{\sigma}_f(\mathcal{N}(0,1))} = \frac{1}{\tilde{\sigma}_f(\mathcal{N}(0,1))}$ is sufficient to sattisfy $\lim\limits_{n\to\infty} \tilde{\sigma}_f = \lim\limits_{n\to\infty} SD(S) = \sigma$. So $\tilde{\mu}_f$ and $\tilde{\sigma}_f$ are unbiased estimators of $\mu$ and $\sigma$ of a normal distribution, although depending on $f$ they are less efficient than usual non-robust mean and SD.

Figure 9 shows us the cluster spectra from figure 2 with subtracted normalised original spectrum (which is only needed for better visualisation as this step has no effect on the standardisation).



JHEP06(2024)163

Figure 9 also shows the conventional and the outlier robust estimations of mean and SD of the cluster spectra values in each bin. As expected for lower $m_{jj}$ the SD is higher as these deviations is partially caused by the Poisson fluctuations which are proportional to $\sqrt{N_{i,b}}$. We can also see the conventional estimators have a bump around 3.5 TeV that is induced by our outlier signal-rich clusters, while the robust estimators are unaffected by the outliers.

## F Calibration distributions

The distribution of the test statistics given by CS without ensembling for all background only pseudo-experiments is shown in figure 10(a) as a histogram. We see that around 300 of those were assigned test statistic of 0 as they had no clusters selected as anomalous. Other cases where one or more anomalous clusters were selected form a smooth continuous distribution.

The median CS test statistic for 100 signal contaminated pseudo-experiments is represented in figure 10(a) by a vertical line, and the vertical band represent the region between the quartiles of such a test statistic sample. For each signal-doped pseudo-experiment we calculate significance as it is described in subsection 3.1.4. The median significance is quoted in the legend of the figure.

Figure 10(b) shows the distribution that is analogous to the one in figure 10(a), but with an ensemble of 15 runs of CS algorithm for each pseudo-experiment. We notice that the distribution in figure 10(b) is significantly narrower than in 10(a) which reduces the frequency of background only experiment having large test statistic, thus increasing the sensitivity to signal injection. An additional benefit is that the uncertainty region (between two quarterlies) for each signal doping have significantly decreased which is important for lower uncertainty in the analysis on experimental data on the excess significance or on the exclusion limits.

This motivates, that in general the ensemble size should be taken as large as reasonably possible. Our choice of ensemble size 15 together with the number of pseudo-experiments 3900 were dictated by the computing time and storage memory limits as the amount of full CS algorithm iterations is the product of those numbers (excluding the pseudo-experiments with signal injection)

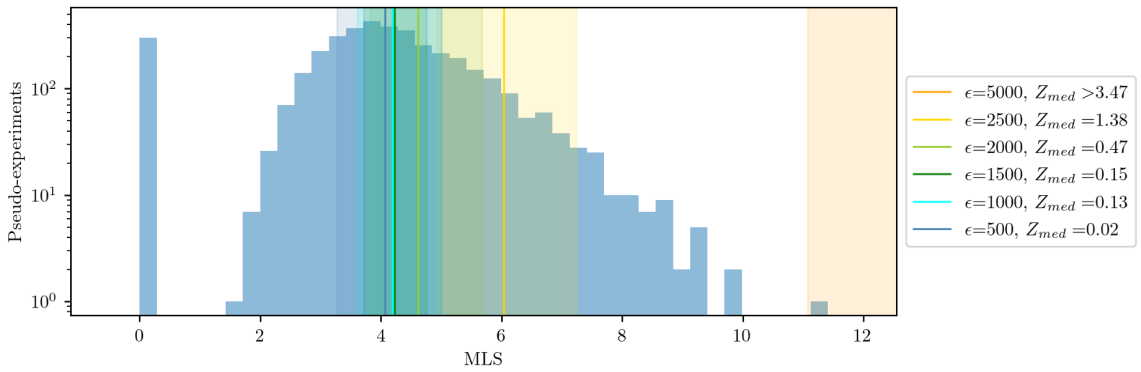
Finally, figure 10(c) shows that for idealised CS without systematics introduced by mass correlations the MLS between our signal spectrum and background estimate is lower. Moreover, as expected, it improved the sensitivity of the method to the signal. Obviously this technique cannot be utilised in an actual analysis as jet labels are needed to distribute signal and background jets in a different manner.

## G Impact of signal in the training region

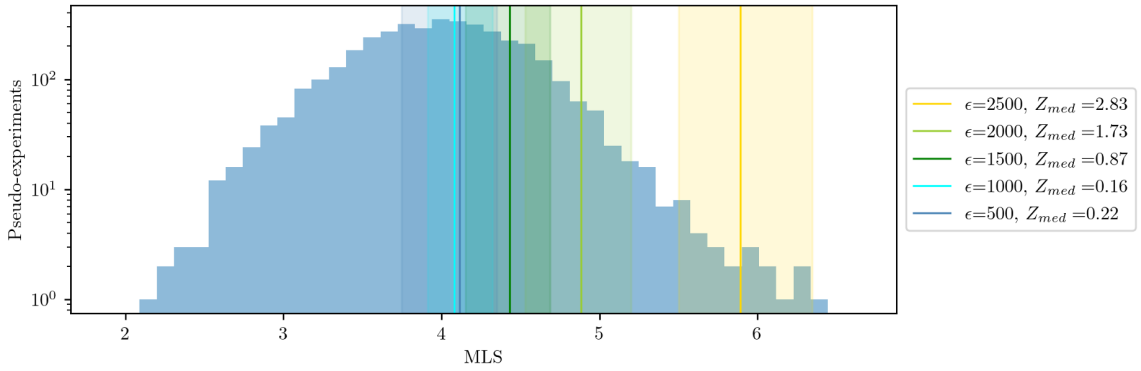
First of all, we retrain all the signal-contaminated experiments such that we do not include the O(100) or fewer signal events while performing k-means fit. We still include the correct number of signal events when evaluating. The performance of this version of CS is demonstrated with the curve labeled as "ignore signal" in figure 11. It is evident that the two versions have a statistically insignificant difference. We conclude that indeed the original CS version well describes a general and realistic case of having negligible to no signal contamination in the training region.



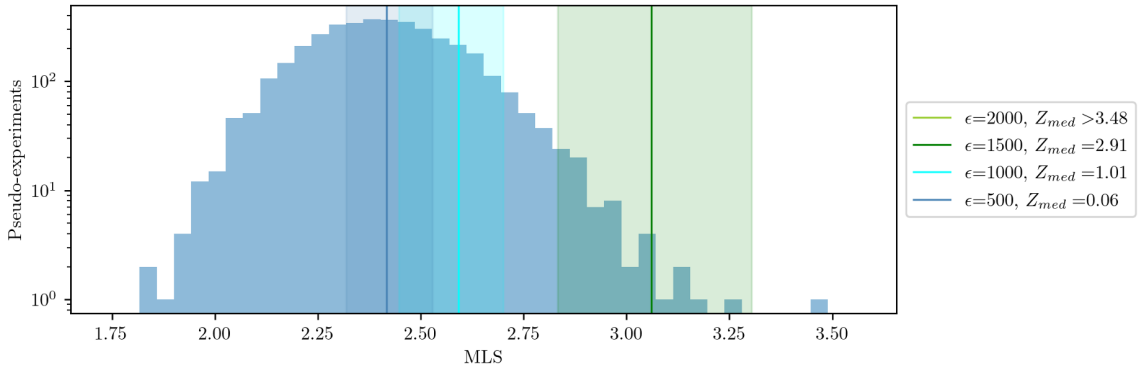
JHEP06(2024)163




(a)




(b)




(c)

**Figure 10.** Histogram of the CS test statistic for pseudo-experiments with bootstrapped background only samples. Vertical lines and vertical bands show median and region between lower and higher quartiles of test statistics for pseudo-experiments with signal injection. Several signal injection levels are represented by different colours. Panel (a) shows a case with only 1 initialisation of clusters in CS per pseudo-experiment, panel (b) shows a case for ensembling 15 runs of CS with different intialisations per pseudo-experiment and panel (c) shows a case for ensembling 15 runs of idealised CS with different initialisations per pseudo-experiment.

JHEP06(2024)163

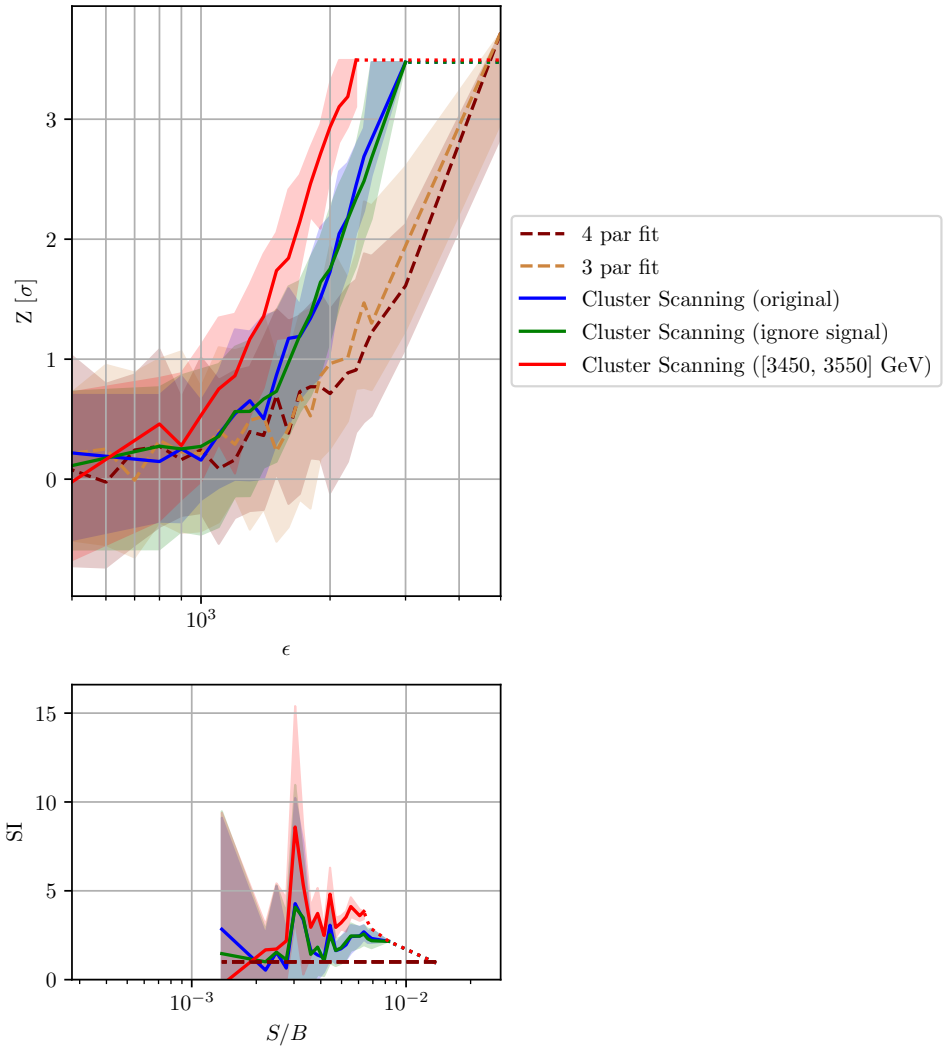



**Figure 11.** Analogous to figure 5(a), with two curves added for the experiments of training without signal and for training in the most signal-rich region of [3450, 3550] GeV.

For the next experiment, we train clustering in the region with the highest signal event fraction, namely [3450, 3550] GeV. We run background-only and signal-contaminated pseudo-experiments in this region with all other hyperparameters equal to the values used in main studies. In the case of a non-negligible signal fraction in the training region, the cluster centroids will be attracted to the regions of signal event concentration. We observe this effect, as the resulting signal clusters have a much higher signal fraction as the signal events "pull" the corresponding cluster centers closer, leading to a large increase in the performance of the CS method that is visible in figure 11. Although in the general case the position of the signal-rich region is unknown, these studies prove that figure 5(a) and figure 5(b) only show the lower bound for the discovery potential of CS.



JHEP06(2024)163

## References

[1] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1 [arXiv:1207.7214] [INSPIRE].

[2] CMS collaboration, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30 [arXiv:1207.7235] [INSPIRE].

[3] D0 collaboration, *Sleuth: A Quasimodel Independent Search Strategy for New Physics*, in the proceedings of the *36th Rencontres de Moriond on QCD and Hadronic Interactions*, Les Arcs, France, March 17–24 (2001) [hep-ex/0105027] [INSPIRE].

[4] CDF collaboration, *Model-Independent Global Search for New High-p(T) Physics at CDF*, arXiv:0712.2534 [DOI:10.2172/922303] [INSPIRE].

[5] CMS collaboration, *MUSiC: a model-unspecific search for new physics in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, *Eur. Phys. J. C* **81** (2021) 629 [arXiv:2010.02984] [INSPIRE].

[6] J.H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, *Phys. Rev. D* **99** (2019) 014038 [arXiv:1902.02634] [INSPIRE].

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

[8] R.T. D'Agnolo and A. Wulzer, *Learning New Physics from a Machine*, *Phys. Rev. D* **99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].

[9] R.T. D'Agnolo et al., *Learning multivariate new physics*, *Eur. Phys. J. C* **81** (2021) 89 [arXiv:1912.12155] [INSPIRE].

[10] M. Farina, Y. Nakai and D. Shih, *Searching for New Physics with Deep Autoencoders*, *Phys. Rev. D* **101** (2020) 075021 [arXiv:1808.08992] [INSPIRE].

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

[12] T.S. Roy and A.H. Vijay, *A robust anomaly finder based on autoencoders*, arXiv:1903.02032 [INSPIRE].

[13] O. Cerri et al., *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, *JHEP* **05** (2019) 036 [arXiv:1811.10276] [INSPIRE].

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

[15] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty Detection Meets Collider Physics*, *Phys. Rev. D* **101** (2020) 076015 [arXiv:1807.10261] [INSPIRE].

[16] A. De Simone and T. Jacques, *Guiding New Physics Searches with Unsupervised Learning*, *Eur. Phys. J. C* **79** (2019) 289 [arXiv:1807.06038] [INSPIRE].

[17] A. Mullin et al., *Does SUSY have friends? A new approach for LHC event analysis*, *JHEP* **02** (2021) 160 [arXiv:1912.10625] [INSPIRE].



JHEP06(2024)163

[18] B.M. Dillon, D.A. Faroughy and J.F. Kamenik, *Uncovering latent jet substructure*, *Phys. Rev. D* **100** (2019) 056002 [arXiv:1904.04200] [INSPIRE].

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

[20] M. Romão Crispim, N.F. Castro, R. Pedro and T. Vale, *Transferability of Deep Learning Models in Searches for New Physics at Colliders*, *Phys. Rev. D* **101** (2020) 035042 [arXiv:1912.04220] [INSPIRE].

[21] M. Crispim Romão et al., *Use of a generalized energy Mover's distance in the search for rare phenomena at colliders*, *Eur. Phys. J. C* **81** (2021) 192 [arXiv:2004.09360] [INSPIRE].

[22] O. Amram and C.M. Suarez, *Tag N' Train: a technique to train improved classifiers on unlabeled data*, *JHEP* **01** (2021) 153 [arXiv:2002.12376] [INSPIRE].

[23] T. Cheng et al., *Variational autoencoders for anomalous jet tagging*, *Phys. Rev. D* **107** (2023) 016002 [arXiv:2007.01850] [INSPIRE].

[24] C.K. Khosa and V. Sanz, *Anomaly Awareness*, *SciPost Phys.* **15** (2023) 053 [arXiv:2007.14462] [INSPIRE].

[25] P. Thaprasop, K. Zhou, J. Steinheimer and C. Herold, *Unsupervised Outlier Detection in Heavy-Ion Collisions*, *Phys. Scripta* **96** (2021) 064003 [arXiv:2007.15830] [INSPIRE].

[26] S. Alexander et al., *Decoding Dark Matter Substructure without Supervision*, arXiv:2008.12731 [INSPIRE].

[27] V. Mikuni and F. Canelli, *Unsupervised clustering for collider physics*, *Phys. Rev. D* **103** (2021) 092007 [arXiv:2010.07106] [INSPIRE].

[28] M. van Beekveld et al., *Combining outlier analysis algorithms to identify new physics at the LHC*, *JHEP* **09** (2021) 024 [arXiv:2010.07940] [INSPIRE].

[29] S.E. Park et al., *Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge*, *JHEP* **06** (2020) 030 [arXiv:2011.03550] [INSPIRE].

[30] D.A. Faroughy, *Uncovering hidden new physics patterns in collider events using Bayesian probabilistic models*, *PoS* **ICHEP2020** (2021) 238 [arXiv:2012.08579] [INSPIRE].

[31] T. Golling et al., *The Mass-ive Issue: Anomaly Detection in Jet Physics*, in the proceedings of the *34th Conference on Neural Information Processing Systems*, Online Conference, Canada, December 06–12 (2020) [arXiv:2303.14134] [INSPIRE].

[32] P. Chakravarti, M. Kuusela, J. Lei and L. Wasserman, *Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests*, arXiv:2102.07679 [INSPIRE].

[33] J. Batson, C.G. Haaf, Y. Kahn and D.A. Roberts, *Topological Obstructions to Autoencoding*, *JHEP* **04** (2021) 280 [arXiv:2102.08380] [INSPIRE].

[34] A. Blance and M. Spannowsky, *Unsupervised event classification with graphs on classical and photonic quantum computers*, *JHEP* **08** (2020) 170 [arXiv:2103.03897] [INSPIRE].

[35] B. Bortolato, A. Smolkovič, B.M. Dillon and J.F. Kamenik, *Bump hunting in latent space*, *Phys. Rev. D* **105** (2022) 115009 [arXiv:2103.06595] [INSPIRE].

[36] J.H. Collins, P. Martín-Ramiro, B. Nachman and D. Shih, *Comparing weak- and unsupervised methods for resonant anomaly detection*, *Eur. Phys. J. C* **81** (2021) 617 [arXiv:2104.02092] [INSPIRE].



JHEP06(2024)163

[37] B.M. Dillon, T. Plehn, C. Sauer and P. Sorrenson, *Better Latent Spaces for Better Autoencoders*, *SciPost Phys.* **11** (2021) 061 [arXiv:2104.08291] [INSPIRE].

[38] T. Finke et al., *Autoencoders for unsupervised anomaly detection in high energy physics*, *JHEP* **06** (2021) 161 [arXiv:2104.09051] [INSPIRE].

[39] D. Shih, M.R. Buckley, L. Necib and J. Tamanas, *via machinae: Searching for stellar streams using unsupervised machine learning*, *Mon. Not. Roy. Astron. Soc.* **509** (2021) 5992 [arXiv:2104.12789] [INSPIRE].

[40] O. Atkinson et al., *Anomaly detection with convolutional Graph Neural Networks*, *JHEP* **08** (2021) 080 [arXiv:2105.07988] [INSPIRE].

[41] A. Kahn et al., *Anomalous jet identification via sequence modeling*, 2021 *JINST* **16** P08012 [arXiv:2105.09274] [INSPIRE].

[42] T. Dorigo et al., *RanBox: anomaly detection in the copula space*, *JHEP* **01** (2023) 008 [arXiv:2106.05747] [INSPIRE].

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

[44] E. Govorkova et al., *LHC physics dataset for unsupervised New Physics detection at 40 MHz*, *Sci. Data* **9** (2022) 118 [arXiv:2107.02157] [INSPIRE].

[45] G. Kasieczka, B. Nachman and D. Shih, *New Methods and Datasets for Group Anomaly Detection From Fundamental Physics*, in the proceedings of the *Conference on Knowledge Discovery and Data Mining*, Online Conference, Singapore, August 14–18 (2021) [arXiv:2107.02821] [INSPIRE].

[46] S. Volkovich, F. De Vito Halevy and S. Bressler, *A data-directed paradigm for BSM searches: the bump-hunting example*, *Eur. Phys. J. C* **82** (2022) 265 [arXiv:2107.11573] [INSPIRE].

[47] 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*, *Nature Mach. Intell.* **4** (2022) 154 [arXiv:2108.03986] [INSPIRE].

[48] B. Ostdiek, *Deep Set Auto Encoders for Anomaly Detection in Particle Physics*, *SciPost Phys.* **12** (2022) 045 [arXiv:2109.01695] [INSPIRE].

[49] K. Fraser et al., *Challenges for unsupervised anomaly detection in particle physics*, *JHEP* **03** (2022) 066 [arXiv:2110.06948] [INSPIRE].

[50] P. Jawahar et al., *Improving Variational Autoencoders for New Physics Detection at the LHC With Normalizing Flows*, *Front. Big Data* **5** (2022) 803685 [arXiv:2110.08508] [INSPIRE].

[51] J. Herrero-Garcia, R. Patrick and A. Scaffidi, *A semi-supervised approach to dark matter searches in direct detection data with machine learning*, *JCAP* **02** (2022) 039 [arXiv:2110.12248] [INSPIRE].

[52] J.A. Aguilar-Saavedra, *Anomaly detection from mass unspecific jet tagging*, *Eur. Phys. J. C* **82** (2022) 130 [arXiv:2111.02647] [INSPIRE].

[53] R. Tombs and C.G. Lester, *A method to challenge symmetries in data with self-supervised learning*, 2022 *JINST* **17** P08024 [arXiv:2111.05442] [INSPIRE].

[54] C.G. Lester and R. Tombs, *Using unsupervised learning to detect broken symmetries, with relevance to searches for parity violation in nature. (Previously: "Stressed GANs snag desserts")*, arXiv:2111.00616 [INSPIRE].



JHEP06(2024)163

[55] V. Mikuni, B. Nachman and D. Shih, *Online-compatible unsupervised nonresonant anomaly detection*, *Phys. Rev. D* **105** (2022) 055006 [arXiv:2111.06417] [INSPIRE].

[56] S. Chekanov and W. Hopkins, *Event-Based Anomaly Detection for Searches for New Physics*, *Universe* **8** (2022) 494 [arXiv:2111.12119] [INSPIRE].

[57] R.T. d'Agnolo et al., *Learning new physics from an imperfect machine*, *Eur. Phys. J. C* **82** (2022) 275 [arXiv:2111.13633] [INSPIRE].

[58] F. Canelli et al., *Autoencoders for semivisible jet detection*, *JHEP* **02** (2022) 074 [arXiv:2112.02864] [INSPIRE].

[59] V.S. Ngairangbam, M. Spannowsky and M. Takeuchi, *Anomaly detection in high-energy physics using a quantum autoencoder*, *Phys. Rev. D* **105** (2022) 095004 [arXiv:2112.04958] [INSPIRE].

[60] L. Bradshaw, S. Chang and B. Ostdiek, *Creating simple, interpretable anomaly detectors for new physics in jet substructure*, *Phys. Rev. D* **106** (2022) 035014 [arXiv:2203.01343] [INSPIRE].

[61] J.A. Aguilar-Saavedra, *Taming modeling uncertainties with mass unspecific supervised tagging*, *Eur. Phys. J. C* **82** (2022) 270 [arXiv:2201.11143] [INSPIRE].

[62] J.A. Aguilar-Saavedra et al., *Gradient Boosting MUST taggers for highly-boosted jets*, arXiv:2305.04957 [INSPIRE].

[63] T. Buss et al., *What's anomalous in LHC jets?*, *SciPost Phys.* **15** (2023) 168 [arXiv:2202.00686] [INSPIRE].

[64] S. Alvi, C.W. Bauer and B. Nachman, *Quantum anomaly detection for collider physics*, *JHEP* **02** (2023) 220 [arXiv:2206.08391] [INSPIRE].

[65] B.M. Dillon, R. Mastandrea and B. Nachman, *Self-supervised anomaly detection for new physics*, *Phys. Rev. D* **106** (2022) 056005 [arXiv:2205.10380] [INSPIRE].

[66] M. Birman et al., *Data-directed search for new physics based on symmetries of the SM*, *Eur. Phys. J. C* **82** (2022) 508 [arXiv:2203.07529] [INSPIRE].

[67] M. Letizia et al., *Learning new physics efficiently with nonparametric methods*, *Eur. Phys. J. C* **82** (2022) 879 [arXiv:2204.02317] [INSPIRE].

[68] C. Fanelli, J. Giroux and Z. Papandreou, *'Flux+Mutability': a conditional generative approach to one-class classification and anomaly detection*, *Mach. Learn. Sci. Tech.* **3** (2022) 045012 [arXiv:2204.08609] [INSPIRE].

[69] T. Finke, M. Krämer, M. Lipp and A. Mück, *Boosting mono-jet searches with model-agnostic machine learning*, *JHEP* **08** (2022) 015 [arXiv:2204.11889] [INSPIRE].

[70] R. Verheyen, *Event Generation and Density Estimation with Surjective Normalizing Flows*, *SciPost Phys.* **13** (2022) 047 [arXiv:2205.01697] [INSPIRE].

[71] B.M. Dillon et al., *A normalized autoencoder for LHC triggers*, *SciPost Phys. Core* **6** (2023) 074 [arXiv:2206.14225] [INSPIRE].

[72] S. Caron, R.R. de Austri and Z. Zhang, *Mixture-of-Theories training: can we find new physics and anomalies better by mixing physical theories?*, *JHEP* **03** (2023) 004 [arXiv:2207.07631] [INSPIRE].

[73] S.E. Park, P. Harris and B. Ostdiek, *Neural embedding: learning the embedding of the manifold of physics data*, *JHEP* **07** (2023) 108 [arXiv:2208.05484] [INSPIRE].



JHEP06(2024)163

[74] J.F. Kamenik and M. Szewc, *Null hypothesis test for anomaly detection*, *Phys. Lett. B* **840** (2023) 137836 [arXiv:2210.02226] [INSPIRE].

[75] G. Kasieczka et al., *Anomaly detection under coordinate transformations*, *Phys. Rev. D* **107** (2023) 015009 [arXiv:2209.06225] [INSPIRE].

[76] J.Y. Araz and M. Spannowsky, *Quantum-probabilistic Hamiltonian learning for generative modeling and anomaly detection*, *Phys. Rev. A* **108** (2023) 062422 [arXiv:2211.03803] [INSPIRE].

[77] J. Schuhmacher et al., *Unravelling physics beyond the standard model with classical and quantum anomaly detection*, *Mach. Learn. Sci. Tech.* **4** (2023) 045031 [arXiv:2301.10787] [INSPIRE].

[78] S. Roche et al., *Nanosecond anomaly detection with decision trees and real-time application to exotic Higgs decays*, *Nature Commun.* **15** (2024) 3527 [arXiv:2304.03836] [INSPIRE].

[79] L. Vaslin, V. Barra and J. Donini, *GAN-AE: an anomaly detection algorithm for New Physics search in LHC data*, *Eur. Phys. J. C* **83** (2023) 1008 [arXiv:2305.15179] [INSPIRE].

[80] ATLAS collaboration, *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** (2023) 052009 [arXiv:2306.03637] [INSPIRE].

[81] S.V. Chekanov and R. Zhang, *Enhancing the hunt for new phenomena in dijet final states using anomaly detection filters at the high-luminosity large Hadron Collider*, *Eur. Phys. J. Plus* **139** (2024) 237 [arXiv:2308.02671] [INSPIRE].

[82] CMS ECAL collaboration, *Autoencoder-based Anomaly Detection System for Online Data Quality Monitoring of the CMS Electromagnetic Calorimeter*, arXiv:2309.10157 [INSPIRE].

[83] G. Bickendorf et al., *Combining resonant and tail-based anomaly detection*, *Phys. Rev. D* **109** (2024) 096031 [arXiv:2309.12918] [INSPIRE].

[84] M. Freytsis, M. Perelstein and Y.C. San, *Anomaly detection in the presence of irrelevant features*, *JHEP* **02** (2024) 220 [arXiv:2310.13057] [INSPIRE].

[85] E.M. Metodiev, J. Thaler and R. Wynne, *Anomaly Detection in Collider Physics via Factorized Observables*, arXiv:2312.00119 [INSPIRE].

[86] G. Karagiorgi et al., *Machine learning in the search for new fundamental physics*, *Nature Rev. Phys.* **4** (2022) 399 [arXiv:2112.03769] [INSPIRE].

[87] G. Kasieczka et al., *The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics*, *Rept. Prog. Phys.* **84** (2021) 124201 [arXiv:2101.08320] [INSPIRE].

[88] T. Aarrestad et al., *The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider*, *SciPost Phys.* **12** (2022) 043 [arXiv:2105.14027] [INSPIRE].

[89] V. Belis, P. Odagiu and T.K. Aarrestad, *Machine learning for anomaly detection in particle physics*, *Rev. Phys.* **12** (2024) 100091 [arXiv:2312.14190] [INSPIRE].

[90] T. Golling et al., *The interplay of machine learning-based resonant anomaly detection methods*, *Eur. Phys. J. C* **84** (2024) 241 [arXiv:2307.11157] [INSPIRE].

[91] ATLAS collaboration, *Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector*, *Phys. Rev. Lett.* **125** (2020) 131801 [arXiv:2005.02983] [INSPIRE].

JHEP06(2024)163

[92] ATLAS collaboration, *Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. Lett.* **132** (2024) 081801 [arXiv:2307.01612] [INSPIRE].

[93] G. Kasieczka, B. Nachman and D. Shih, *Official Datasets for LHC Olympics 2020 Anomaly Detection Challenge (Version v6)*, DOI:10.5281/zenodo.4536624.

[94] J.H. Kim, K. Kong, B. Nachman and D. Whiteson, *The motivation and status of two-body resonance decays after the LHC Run 2 and beyond*, *JHEP* **04** (2020) 030 [arXiv:1907.06659] [INSPIRE].

[95] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

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

[97] A. Mertens, *New features in Delphes 3*, *J. Phys. Conf. Ser.* **608** (2015) 012045 [INSPIRE].

[98] M. Selvaggi, *DELPHES 3: A modular framework for fast-simulation of generic collider experiments*, *J. Phys. Conf. Ser.* **523** (2014) 012033 [INSPIRE].

[99] M. Cacciari, G.P. Salam and G. Soyez, *The anti-$k_t$ jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[100] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J. C* **72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[101] E. Rodrigues et al., *The Scikit HEP Project — overview and prospects*, *EPJ Web Conf.* **245** (2020) 06028 [arXiv:2007.03577] [INSPIRE].

[102] L. de Oliveira et al., *Jet-images — deep learning edition*, *JHEP* **07** (2016) 069 [arXiv:1511.05190] [INSPIRE].

[103] A. Butter et al., *The Machine Learning landscape of top taggers*, *SciPost Phys.* **7** (2019) 014 [arXiv:1902.09914] [INSPIRE].

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

[105] A. Hallin et al., *Classifying anomalies through outer density estimation*, *Phys. Rev. D* **106** (2022) 055006 [arXiv:2109.00546] [INSPIRE].

[106] J.A. Raine, S. Klein, D. Sengupta and T. Golling, *CURTAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals*, *Front. Big Data* **6** (2023) 899345 [arXiv:2203.09470] [INSPIRE].

[107] CMS collaboration, *Search for narrow resonances in the b-tagged dijet mass spectrum in proton-proton collisions at s=13 TeV*, *Phys. Rev. D* **108** (2023) 012009 [arXiv:2205.01835] [INSPIRE].

[108] ATLAS collaboration, *Search for new phenomena in the dijet mass distribution using $p-p$ collision data at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *Phys. Rev. D* **91** (2015) 052007 [arXiv:1407.1376] [INSPIRE].

[109] ATLAS collaboration, *Search for new phenomena in dijet mass and angular distributions from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Lett. B* **754** (2016) 302 [arXiv:1512.01530] [INSPIRE].



JHEP06(2024)163

[110] CMS collaboration, *Search for dijet resonances in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on dark matter and other models*, *Phys. Lett. B* **769** (2017) 520 [arXiv:1611.03568] [INSPIRE].

[111] CMS collaboration, *Search for narrow resonances decaying to dijets in proton-proton collisions at $\sqrt{s} = 13$ tev*, *Phys. Rev. Lett.* **116** (2016) 071801 [arXiv:1512.01224] [INSPIRE].

[112] CMS collaboration, *Search for narrow resonances in dijet final states at $\sqrt{s} = 8$ tev with the novel cms technique of data scouting*, *Phys. Rev. Lett.* **117** (2016) 031802 [arXiv:1604.08907] [INSPIRE].

[113] ATLAS collaboration, *Search for New Physics in Dijet Mass and Angular Distributions in pp Collisions at $\sqrt{s} = 7$ TeV Measured with the ATLAS Detector*, *New J. Phys.* **13** (2011) 053044 [arXiv:1103.3864] [INSPIRE].

[114] CMS collaboration, *Search for resonances and quantum black holes using dijet mass spectra in proton-proton collisions at $\sqrt{s} = 8$ TeV*, *Phys. Rev. D* **91** (2015) 052009 [arXiv:1501.04198] [INSPIRE].

[115] CMS collaboration, *Search for Narrow Resonances Using the Dijet Mass Spectrum in pp Collisions at $\sqrt{s} = 8$ TeV*, *Phys. Rev. D* **87** (2013) 114015 [arXiv:1302.4794] [INSPIRE].

[116] CMS collaboration, *Search for Narrow Resonances and Quantum Black Holes in Inclusive and b-Tagged Dijet Mass Spectra from pp Collisions at $\sqrt{s} = 7$ TeV*, *JHEP* **01** (2013) 013 [arXiv:1210.2387] [INSPIRE].

[117] ATLAS collaboration, *Search for new physics in the dijet mass distribution using 1 $fb^{-1}$ of pp collision data at $\sqrt{s} = 7$ tev collected by the atlas detector*, *Phys. Lett. B* **708** (2012) 37 [arXiv:1108.6311] [INSPIRE].

[118] ATLAS collaboration, *ATLAS search for new phenomena in dijet mass and angular distributions using pp collisions at $\sqrt{s} = 7$ TeV*, *JHEP* **01** (2013) 029 [arXiv:1210.1718] [INSPIRE].

[119] CMS collaboration, *Search for Resonances in the Dijet Mass Spectrum from 7 TeV pp Collisions at CMS*, *Phys. Lett. B* **704** (2011) 123 [arXiv:1107.4771] [INSPIRE].

[120] CDF collaboration, *Search for new particles decaying into dijets in proton-antiproton collisions at $\sqrt{s} = 1.96$-TeV*, *Phys. Rev. D* **79** (2009) 112002 [arXiv:0812.4036] [INSPIRE].

[121] CMS collaboration, *Search for dijet resonances in 7 tev pp collisions at cms*, *Phys. Rev. Lett.* **105** (2010) 211801 [arXiv:1010.0203] [INSPIRE].

[122] ATLAS collaboration, *Search for low-mass dijet resonances using trigger-level jets with the atlas detector in pp collisions at $\sqrt{s} = 13$ TeV*, *Phys. Rev. Lett.* **121** (2018) 081801 [arXiv:1804.03496] [INSPIRE].

[123] K. Sekhon, R.C. Edgar and D. Amidei, *SWiFt: Sliding Window Fit Method for Resonance Searches*, Tech. Rep. ATL-COM-PHYS-2018-161, CERN, Geneva (2018).

[124] J. Alison et al., *Search for resonances in the di-jet mass distribution with one or two jets identified as b-jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Tech. Rep. ATL-COM-PHYS-2016-1561, CERN, Geneva (2016).

[125] G. Kasieczka, B. Nachman, M.D. Schwartz and D. Shih, *Automating the ABCD method with machine learning*, *Phys. Rev. D* **103** (2021) 035021 [arXiv:2007.14400] [INSPIRE].



JHEP06(2024)163

[126] L. Vaslin, S. Calvet, V. Barra and J. Donini, *pyBumpHunter: A model independent bump hunting tool in Python for High Energy Physics analyses*, *SciPost Phys. Codeb.* **2023** (2023) 15 [arXiv:2208.14760] [INSPIRE].

[127] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*, *J. Mach. Learn. Res.* **12** (2011) 2825 [arXiv:1201.0490] [INSPIRE].

[128] D. Arthur and S. Vassilvitskii, *K-means++: The advantages of careful seeding*, in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, (U.S.A.), p. 1027-1035, Society for Industrial and Applied Mathematics (2007).

[129] A. Andreassen, B. Nachman and D. Shih, *Simulation Assisted Likelihood-free Anomaly Detection*, *Phys. Rev. D* **101** (2020) 095004 [arXiv:2001.05001] [INSPIRE].

[130] T. Golling, S. Klein, R. Mastandrea and B. Nachman, *Flow-enhanced transportation for anomaly detection*, *Phys. Rev. D* **107** (2023) 096025 [arXiv:2212.11285] [INSPIRE].

[131] D. Sengupta, S. Klein, J.A. Raine and T. Golling, *CURTAINs Flows For Flows: Constructing Unobserved Regions with Maximum Likelihood Estimation*, arXiv:2305.04646 [INSPIRE].

[132] E. Buhmann et al., *Full phase space resonant anomaly detection*, *Phys. Rev. D* **109** (2024) 055015 [arXiv:2310.06897] [INSPIRE].

[133] D. Sengupta et al., *Improving new physics searches with diffusion models for event observables and jet constituents*, *JHEP* **04** (2024) 109 [arXiv:2312.10130] [INSPIRE].

[134] M. Frate et al., *Modeling Smooth Backgrounds and Generic Localized Signals with Gaussian Processes*, arXiv:1709.05681 [INSPIRE].

[135] K. Cranmer and I. Yavin, *RECAST: Extending the Impact of Existing Analyses*, *JHEP* **04** (2011) 038 [arXiv:1010.2506] [INSPIRE].

[136] P. Virtanen et al., *SciPy 1.0 — Fundamental Algorithms for Scientific Computing in Python*, *Nature Meth.* **17** (2020) 261 [arXiv:1907.10121] [INSPIRE].

JHEP06(2024)163