# Unsupervised clustering for collider physics

Vinicius Mikuni* and Florencia Canelli
*University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

(Received 7 October 2020; accepted 11 February 2021; published 24 May 2021)

We propose a new method for unsupervised clustering for collider physics named UCluster, where information in the embedding space created by a neural network is used to categorize collision events into different clusters that share similar properties. We show how this method can be developed into an unsupervised multiclass classification of different processes and applied in the anomaly detection of events to search for new physics phenomena at colliders.

DOI: 10.1103/PhysRevD.103.092007

## I. INTRODUCTION

The Standard Model (SM) of particle physics has been successful so far at describing the interaction of fundamental particles in high energy physics (HEP). The ATLAS [1] and CMS [2] Collaborations have tested the SM extensively using particle collision events at the CERN Large Hadron Collider (LHC), while also looking for deviations from the SM that could point to physics beyond the SM (BSM). Since the underlying nature of the new physics is not known, new methods designed to be model independent have proliferated in the recent years. These strategies aim at finding deviations or detecting anomalies where only SM events are used and avoiding any dependence on BSM signals. For a short review of recent approaches, see [3].

For measurements of SM parameters, a fully unsupervised multiclass classification method would be advantageous. This is particularly true for precision measurements of SM parameters. Simulations are often needed to describe the properties of different processes produced in the LHC collisions. However, simulated events are not always precise in all physics process. This can be caused either by a lack of simulated events compared to the data expectation or the need of corrections that are beyond the accuracy of the approximations used in the simulation. Further precision might be computationally prohibitive to achieve or beyond the capability of our current methods. To mitigate these issues, different data-driven methods often replace the event simulations. See [4–7] for recent examples.

When two or more processes are not well modeled, the common approach is to design multiple control regions, often defined using high-level distributions, to create a high purity sample that allows a data-driven estimation and modeling for this process. However, since it is not always straightforward to define each of these regions without relying on simulations, an unsupervised multiclass classification approach could be used instead.

In this paper, we introduce a method for unsupervised clustering (UCluster). The main idea of UCluster is to use a neural network (NN) to reduce the data dimensionality while retaining the main properties of the dataset. In this reduced representation, a clustering objective is added to the training to encourage points embedded in this space to be close together when they share similar properties and far apart otherwise. We test the performance of UCluster in the context of two different tasks: unsupervised multiclass classification of three different SM processes and unsupervised anomaly detection.

## II. RELATED WORKS

Recently, different and innovative strategies have been proposed for unsupervised training in HEP, mostly in the context of event classification. A few examples of methods exploiting anomaly detection signatures as overdensities are [8,9] and, more recently, [3]. In these approaches, anomalous events are identified as localized excesses in some distribution, where machine learning is then used to enhance the local significance of the new physics process.

While many strategies focus on unsupervised anomaly detection, other methods have also been proposed to better understand SM processes without relying on simulation, like the work developed in [10] for quark and gluon classification with jet topics and the methods developed in [11], employing latent Dirichlet allocation to build a data-driven top-quark tagger. In order to create an unsupervised and model independent approach, the majority of

*vinicius.massami.mikuni@cern.ch

the strategies rely on binary classification, where the main goal is to test if an event (or a group of events) resulting from a particle collision is compatible with one out of two competing hypotheses. Approaches applied to mixed samples with more than two components were also studied in [12,13], where prior knowledge of the label proportion for each component in the mixed sample is required to achieve a good performance.

In this work, we propose an unsupervised method for multiclass classification whose only requirement is on the expected number of different components inside a mixed sample. The same method is applied to anomalous event detection, where the data is partitioned into clusters that isolate the anomaly from backgrounds.

## III. METHOD DESCRIPTION

UCluster consists of two components: a classification step to ensure events with similar properties are close in the embedding space created by a NN and a clustering step, where the network learns to cluster embedded events of similar properties. These two tasks are accomplished by means of a combined loss function containing independent components to guarantee each of the described steps.

The classification loss ($L_{\text{focal}}$), applied to the output nodes of a NN, is defined by the focal loss [14]. The focal loss improves the classification performance for unbalanced labels; the case for the classification tasks is to be introduced in the following sections. The expression for the focal loss is

$$L_{\text{focal}} = -\frac{1}{N}\sum_{j}^{N}\sum_{m}^{M} y_{j,m}(1 - p_{\theta,m}(x_j))^{\gamma} \times \log(p_{\theta,m}(x_j)), \tag{1}$$

where $p_{\theta,m}(x_j)$ is the networks confidence, for event $x_j$ with trainable parameters $\theta$, to be classified as class $m$. The term $y_{j,m}$ is 1 if class $m$ is the correct assignment for event $j$ and 0 otherwise. In this work, we fix the hyperparameter $\gamma = 2$ of the focal loss.

The clustering loss ($L_{\text{cluster}}$) is defined similarly as the loss developed in [15]

$$L_{\text{cluster}} = \frac{1}{N}\sum_{k}^{K}\sum_{j}^{n} \|f_\theta(x_j) - \mu_k\|^2 \pi_{jk}, \tag{2}$$

where the distance between each event $j$ and each cluster centroid $\mu_k$ is calculated in the embedding space $f_\theta$ of the neural network with trainable parameters $\theta$. The function $\pi_{jk}$ weighs the importance of each event and takes the form,

$$\pi_{jk} = \frac{e^{-\alpha\|f_\theta(x_j)-\mu_k\|}}{\sum_{k'} e^{-\alpha\|f_\theta(x_j)-\mu_k\|}}, \tag{3}$$

with hyperparameter $\alpha$ identified as an inverse temperature term. Since $L_{\text{cluster}}$ is differentiable, stochastic gradient descent can be used to optimize jointly the trainable parameters $\theta$ and the centroid positions $\mu_k$.

The combined loss to be minimized is

$$L = L_{\text{focal}} + \beta L_{\text{cluster}}. \tag{4}$$

The hyperparameter $\beta$ controls the relative importance between the two losses. For these studies, we fix $\beta = 10$ to ensure that both components have the same order of magnitude.

Since $L_{\text{cluster}}$ requires an initial guess for the centroid positions, we pretrain the model using only $L_{\text{focal}}$ for 10 epochs. After the pretraining, the k-means algorithm [16] is applied to the object embeddings to initialize the cluster centroids. The full training is then carried out with the combined loss defined in Eq. (4). To allow the cluster centers to change, the inverse temperature $\alpha$ has a starting value of 1 and linearly increases by 2 for each following epoch.

## IV. GENERAL IMPLEMENTATION

The implementation of UCluster is done using ABCNet [17]. ABCNet is a graph-based neural network where each reconstructed particle is taken as a node in a graph. The importance of each node is then learned by the model by the usage of attention mechanisms. The embedding space for the clustering loss in Eq. (2) is taken as the output of a max-pooling layer. For the following studies, the 10 nearest neighbors from each particle are used to calculate the GAPLayers [18]. The initial distances are calculated in the pseudorapidity-azimuth $(\eta - \phi)$ space using the distance $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$. The second GAPLayer uses the Euclidean distances in the space created by subsequent fully connected layers. The architectures used for multiclass classification and anomaly detection are depicted in Fig. 1. Besides the output classification size, both tasks share almost identical architectures. The model used for anomaly detection uses additional high-level distributions and additional skip connections after the pooling layer to improve the classification performance. In both cases, the batch size is set to 1024, and the training is stopped after for 100 epochs.

ABCNet is implemented in TENSORFLOW v1.14 [19]. An Nvidia GTX 1080 Ti graphics card is used for the training and evaluation steps. For all tasks described in this paper, the Adam optimizer [20] is used. The learning rate starts from 0.001 and decreases by a factor 2 every three epochs, until reaching a minimum of 1e-5.

## V. UNSUPERVISED MULTICLASS CLASSIFICATION

The applicability of UCluster is demonstrated on an important problem in high energy physics: unsupervised

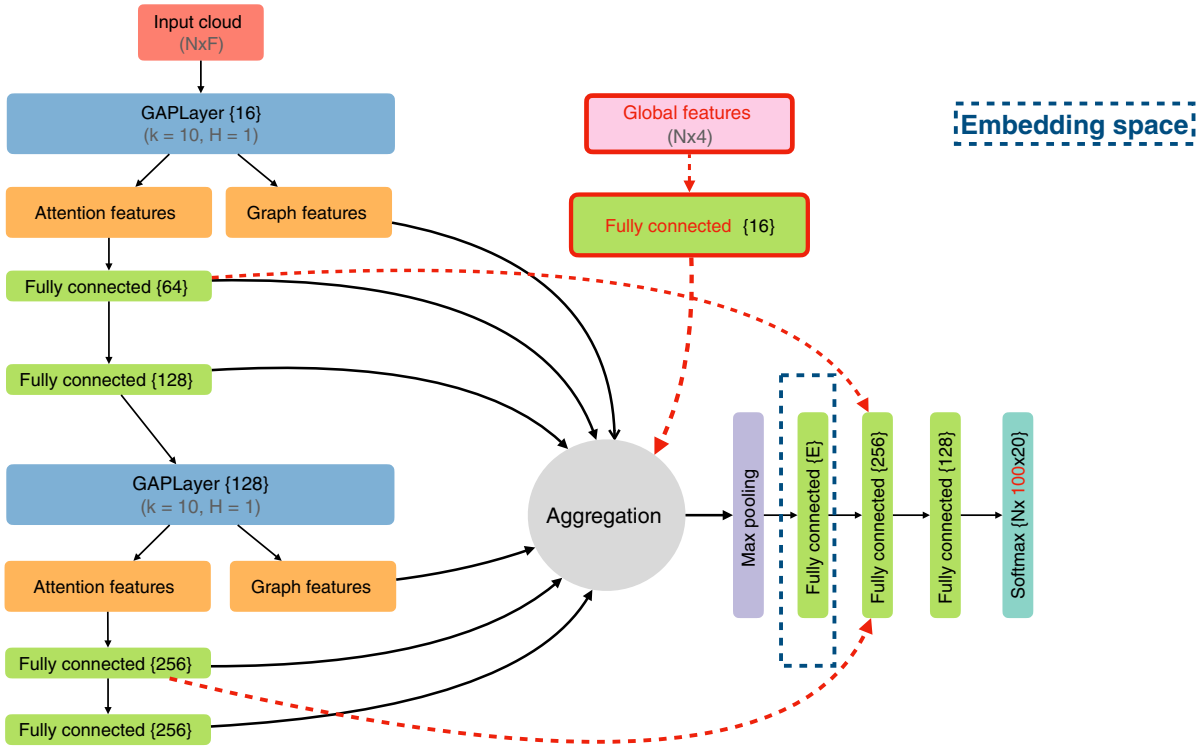


FIG. 1. ABCNet architecture used in UCluster for a batch size N, F input features, and embedding space of size E. Fully connected layers and encoding node sizes are denoted inside "{}". For each GAPLayer, the number of k-nearest neighbors (k) and heads (H) are given. The additional components used only for anomaly detection are shown in red.

multiclass classification. To achieve good performance, we require a task that results in a suitable embedding space. This task should be such that events stemming from the same physics process are found close together in the embedding space as compared to events from different physics processes. Here, a jet mass classification task is chosen in order to provide meaningful event embeddings. Given a set of particles belonging to a jet, we ask our model to correctly identify the invariant mass of the jet. This task chosen is inspired by the correlation of jet substructure observables and the invariant mass of a jet [21,22]. The goal is to have our machine learning method learn to extract relevant information regarding the different jet substructures by first learning how to correctly identify the mass of a jet. The simplest solution to this problem could be achieved by the four-vector sum of all the particle's constituents, leading to an embedding space that does not have separation power for different types of jets. To alleviate this issue, we instead define a jet mass label by taking 20 equidistant steps from 10 to 200 GeV, as shown in Fig. 2. The task is then to identify the correct mass interval a jet belongs to, instead of the specific mass value. The input distributions used for the training are listed in Table I.

For this study, a sample containing simulated jets originating from W bosons, Z bosons, and top quarks produced at $\sqrt{s} = 13$ TeV proton-proton collisions is used. This dataset is created and configured using a parametric description of a generic LHC detector, described in [24,25]. The jets are clustered with the anti-kt algorithm [26] with radius parameter $R = 0.8$, while also requiring that the jet's $p_T$ is around 1 TeV, ensuring that most of the decay

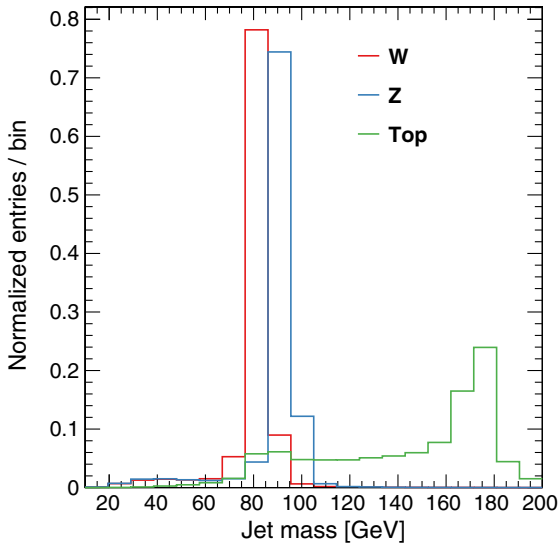


FIG. 2. Normalized distribution of the jet mass of each category used in the unsupervised multiclass classification task. The bin boundaries represent the boundaries used to define the jet mass labels.

TABLE I. Description of each feature used to define a point in the ABCNet implementation for unsupervised multiclass classification.

| Variable | Description |
|---|---|
| $\Delta\eta$ | Difference between the pseudorapidity of the constituent and the jet |
| $\Delta\phi$ | Difference between the azimuthal angle of the constituent and the jet |
| $\log p_{\mathrm{T}}$ | Logarithm of the constituent's $p_{\mathrm{T}}$ |
| $\log \mathrm{E}$ | Logarithm of the constituent's E |
| $\log \frac{p_{\mathrm{T}}}{p_{\mathrm{T}}(\mathrm{jet})}$ | Logarithm of the ratio between the constituent's $p_{\mathrm{T}}$ and the jet $p_{\mathrm{T}}$ |
| $\log \frac{\mathrm{E}}{\mathrm{E}(\mathrm{jet})}$ | Logarithm of the ratio between the constituent's E and the jet E |
| $\Delta\mathrm{R}$ | Distance in the $\eta-\phi$ space between the constituent and the jet |
| PID | Particle type identifier as described in [23]. |

products of the generated particles are found inside a single jet.

The samples are available at [27]. For each jet, up to 100 particles are stored. If more particles were found inside a jet, the event is truncated, otherwise zero padded up to 100. The training set contains 300,000 jets, while the validation sample consists of 140,000 jets.

To visualize the embedding space, the t-SNE visualization method [28] is used for 1000 jets, taken just after the pretraining with only the classification loss, and compared

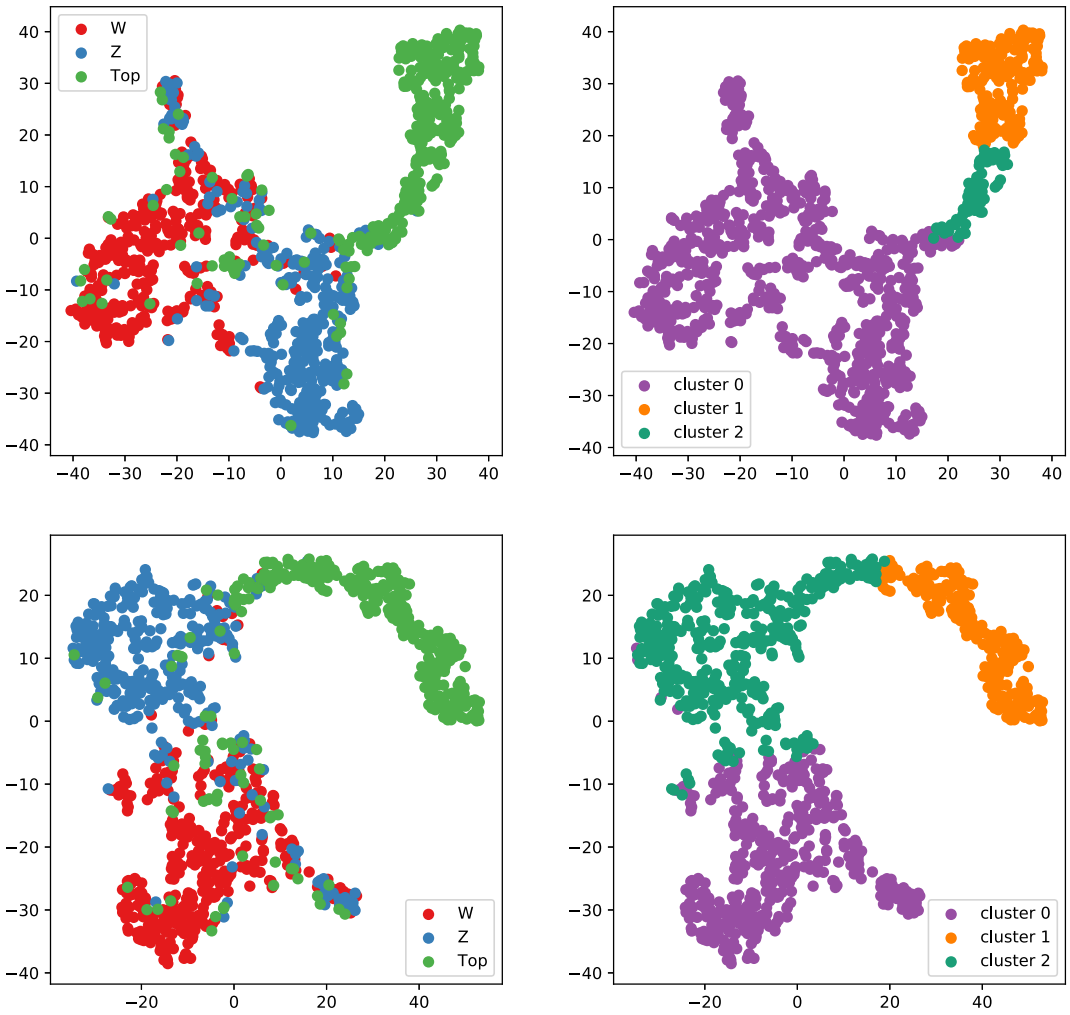


FIG. 3. t-SNE visualization of the embedding space after the pretraining and before the full training (top row) and after the full training (bottom row) for multiclass classification with 1000 jets. The true label information is shown on the left, while the initial cluster labels using a k-means approach is shown on the right.

to the space created after full training is performed. After the pretraining, the initial label assignment is taken from a k-means approach, shown in Fig. 3 (top right), while the true labels are shown in Fig. 3 (top left). At this stage, the clustering accuracy, calculated using the Hungarian algorithm [29], is 51%. After the full training is performed, the trained labels are shown in Fig. 3 (bottom right) with a clustering accuracy of 81% compared to the true label assignment in Fig. 3 (bottom left).

To inspect the quality of the embedding space further, a supervised KNN is trained using only the embedding features as inputs. Its performance is then compared to a separate KNN with the same setup, but using only the jet mass as input. The supervised KNNs are trained to determine class membership given the label of the 30 nearest neighbors. For the training, 35k events are used and tested on an independent sample with 15k events.

The one-vs-all performance is compared using a receiver operating characteristic (ROC) curve in Fig. 4, where one category is considered the signal of interest, while the others are considered a background. The area under curve (AUC) for each process is also shown. The resulting AUC for the supervised training using the event embeddings is higher than the jet mass alone for all categories. Top quark classification shows a particularly large improvement by using the embedding space information. We attribute this improvement to jets containing a top quark showing a broader mass distribution compared to W and Z bosons, resulting in a worse invariant mass separation as seen in Fig. 2. UCluster is able to learn other jet properties beyond the invariant mass, improving the overall performance.

To estimate an upper bound on the UCluster performance, a fully supervised model using the full ABCNet architecture is also trained. The ABCNet architecture is used to train a classifier containing the real class labels as targets, achieving an accuracy of 92%. The comparable results between the fully supervised approach and the KNN trained on the event embeddings demonstrate how the method is able to reduce the dimensionality of the input data while retaining relevant information.

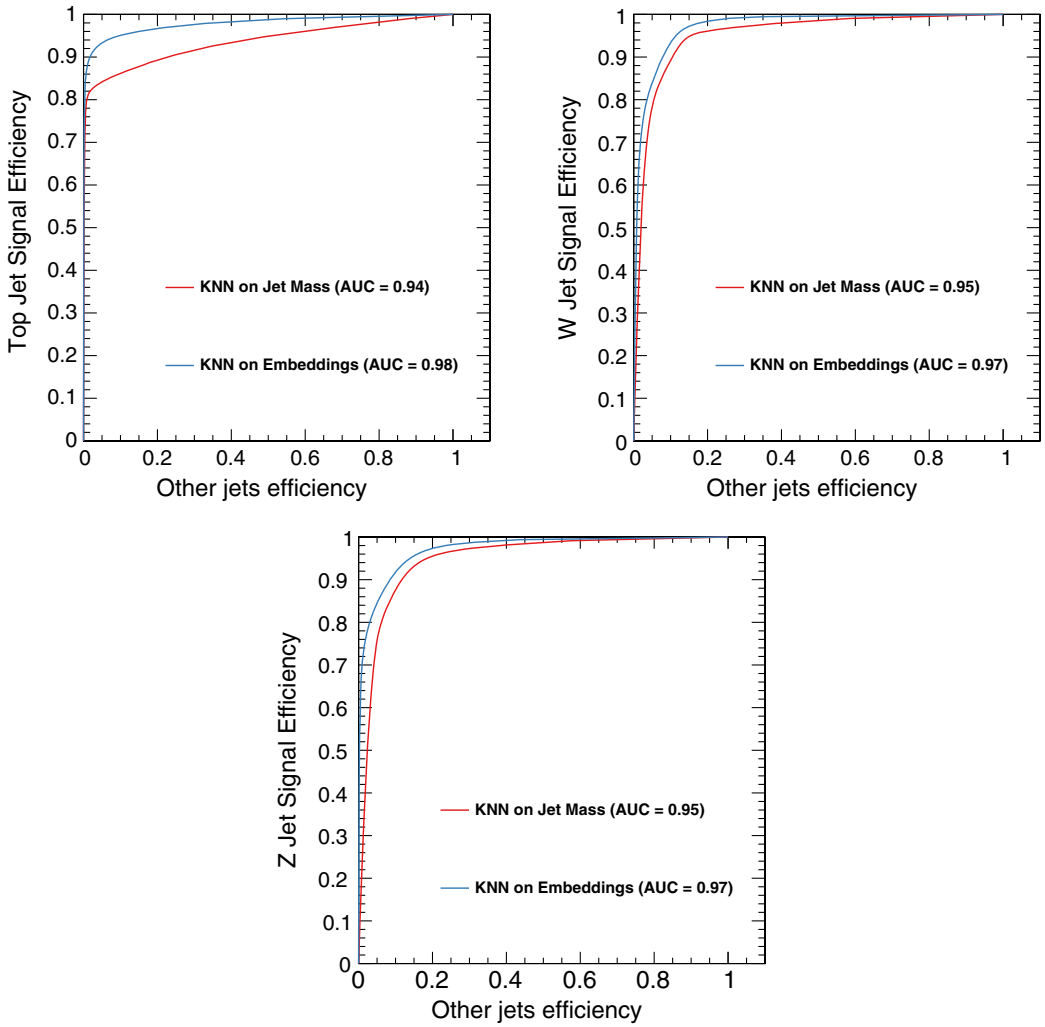


FIG. 4. ROC curves for each jet category when considering the other jet categories as a background.

TABLE II. Supervised and unsupervised clustering accuracy of UCluster when using only the embedding space features.

| Algorithm | Accuracy |
|---|---|
| Pretraining k-means | 51% |
| UCluster | 81% |
| Supervised KNN | 89% |
| Supervised training | 92% |

The accuracies achieved with the full supervision and the other approaches are summarized in Table II.

## VI. ANOMALY DETECTION

UCluster can also be applied to anomaly detection. Here, we show an example where anomalous events, created from an unknown physics process, are found to be close in the embedding space created from a suitable classification task. This technique is motivated by the fact that, irrespective to the underlying physics model, events created by the same physics process carry similar event signatures.

To create a suitable embedding space, we modify the approach described in Sec. V to take into account all the particles created in a collision event rather than a single jet. To do so, the classification task is instead changed to a part segmentation task. We consider all particles associated to a clustered jet. Each particle then receives a label proportional to the mass of the jet that it was clustered into. For this task, we require the model to learn not only the mass of the associated jet the particle belongs to, but also to learn which particles should belong to the same jet. This approach is motivated by the fact that jet substructure often contains useful information for distinguishing different physics processes, as studied in the previous section.

The mass labels are then created by defining 20 equidistant intervals from 10 to 1000 GeV. For simplicity, only the two heaviest jets are considered per event. A simplified example of the label definition is shown in Fig. 5.

To perform these studies, we use the R&D dataset created for the LHC Olympics 2020 [30]. The dataset

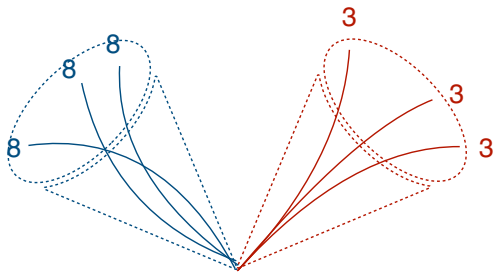


FIG. 5. Schematic of the labels for anomaly detection. Each particle associated to a clustered jet receives a mass label proportional to the respective jet mass. The larger the number, the more massive the associated jet.

consists of a million quantum chromodynamic (QCD) dijet events simulated with PYTHIA 8 [31] without pileup or multiple parton interactions. The BSM signal consists of a hypothetical W' boson with mass $m_W = 3.5$ TeV that decays into an X and Y bosons with masses $m_X = 500$ GeV and $m_Y = 100$ GeV, respectively. The X and Y bosons, on the other hand, decay promptly into quarks.The detector simulation is performed with DELPHES 3.4.1 [32], and particle flow objects are clustered into jets using the FASTJET [33] implementation of the anti-kt algorithm with R = 1.0 for the jet radius. Events are required to have at least one jet with $p_T > 1.3$ TeV. The number of signal events generated is set as 1% of the total number of events. From this dataset, 300k events are randomly selected for training, 150k for testing and 300k events, are used to evaluate the clustering performance.

The distributions used as an input for ABCNet are described in Table III. To improve the clustering performance, a set of high-level variables is added to the network. The goal of the additional distributions is to parameterize the model performance as described in [34].

Here, we would also like to point out that, even if a proxy of jet masses is given as an input, the trivial solution is still not achieved, since the model also has to identify which particles belong to which jets. To quantify the performance of UCluster, we start by considering only two clusters with an embedding space of same dimension. Figure 6 shows the resulting embedding space without any transformation for 1000 random events.

Most of the BSM events are found in the same trained cluster, confirming the assumption that the signal events would end up close together in the embedding space. However, because of the large QCD background contamination present in the same cluster, the signal-to-background (SB) ratio remains low, increasing only from 1% to 2.5%. If the proximity assumption holds, then the cluster SB ratio can be further enhanced by partitioning the events into more clusters. Indeed, if the classification loss favors an embedding space where signal events remain close together, increasing the number of clusters will decrease the QCD contamination in the signal clusters whose properties differ from the signal events. To test this assumption, the cluster size is varied while keeping all the other network parameters fixed. The maximum SB ratio found in a cluster for different clusters sizes is shown in Fig. 7 left. The SB ratio steadily increases with cluster size, reaching an average of around 28%. To test how the performance changes with the number of events, different training sample sizes were used while keeping the model fixed, the signal fraction fixed to 1% and number of clusters fixed to 30. The result of each training is then evaluated in an independent sample, which is the same size as the training sample. The result of the approximate significance ($S\sqrt{B}$) is shown in Fig. 7 on the right. For initial

TABLE III. Descriptions of each feature used to define a point in the point cloud implementation for anomaly detection. The last two lines are the global information added to parameterize the network.

| Variable | Description |
|---|---|
| $\Delta\eta$ | Pseudorapidity difference between the constituent and the associated jet |
| $\Delta\phi$ | Azimuthal angle difference between the constituent and the associated jet |
| $\log p_T$ | Logarithm of the constituent's $p_T$ |
| $\log E$ | Logarithm of the constituent's E |
| $\log \frac{p_T}{p_T(\text{jet})}$ | Logarithm of the ratio between the constituent's $p_T$ and the associated jet $p_T$ |
| $\log \frac{E}{E(\text{jet})}$ | Logarithm of the ratio between the constituent's E and the associated jet E |
| $\Delta R$ | Distance in the $\eta-\phi$ space between the constituent and the associated jet |
| $\log m_{J\{1,2\}}$ | Logarithm of the masses of the two heaviest jets in the event |
| $\tau_{21}^{\{1,2\}}$ | Ratio of $\tau_1$ to $\tau_2$ for the two heaviest jets in the event, with $\tau_N$ defined in [35] |

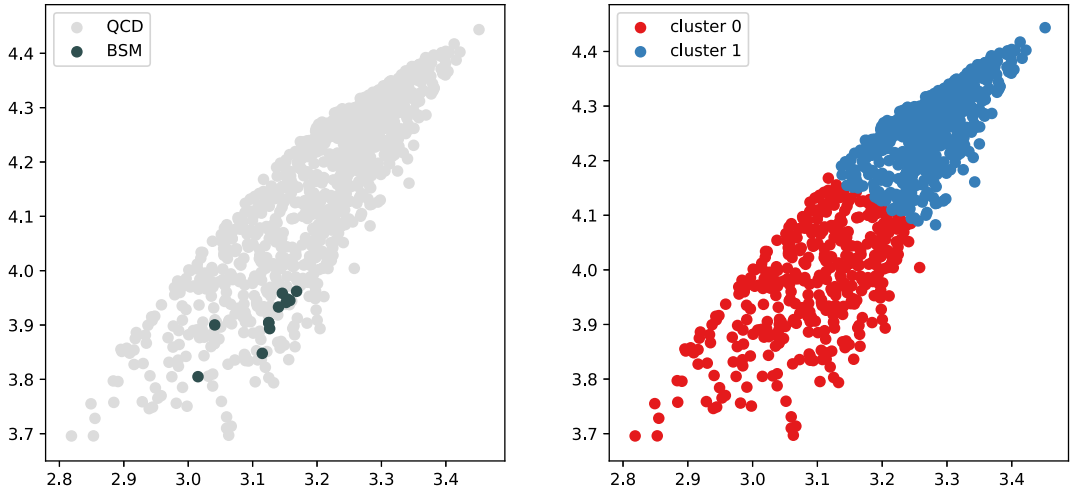


FIG. 6. Visualization of the embedding space created for anomaly detection using 1000 events. Since the embedding space is already two-dimensional, no additional transformation is applied. The true labels are shown on the left, while the clusters created by UCluster are shown on the right.

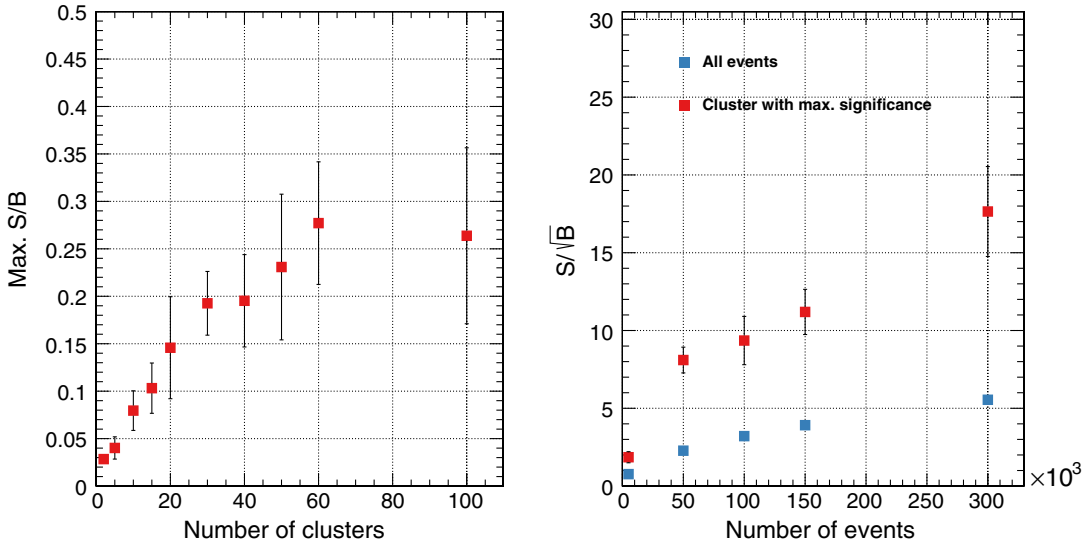


FIG. 7. Maximum signal-to-background ratio found for different clustering sizes (left) and maximum significance found for UCluster trained and evaluated on a different number of events with cluster size fixed to 30 (right). The uncertainty shows the standard deviation of the results from five trainings with different random weight initialization.

significance in the range 2–6, we observe enhancements by factors 3–4.

The uncertainties in Fig. 7 show the standard deviation of five independent trainings with different random initial weights. When many clusters are used, the clustering stability starts to decrease, as evidenced by the larger error bars. This behavior is expected, since a large cluster multiplicity requires clusters to target more specific event properties that might differ in between different trainings.

To qualitatively verify the cluster composition, the dijet mass distributions for all events (left) and for the cluster with the highest SB ratio (right) are shown in Fig. 8.

### A. Background estimation

In the previous section, the sensitivity to an anomalous signal was shown to improve with the number of clusters required by UCluster. However, requiring a larger number of clusters also requires a method to select interesting partitions for further inspection. A local p value for each cluster can be determined for a background-only hypothesis, where the cluster with the lowest p value is selected for further investigation. We also note that a global p value can be derived by taking into account the look-elsewhere effect [36], which is already mitigated by the usage of independent samples during training, testing, and evaluation of UCluster. The main difficulty to estimate the p value is to

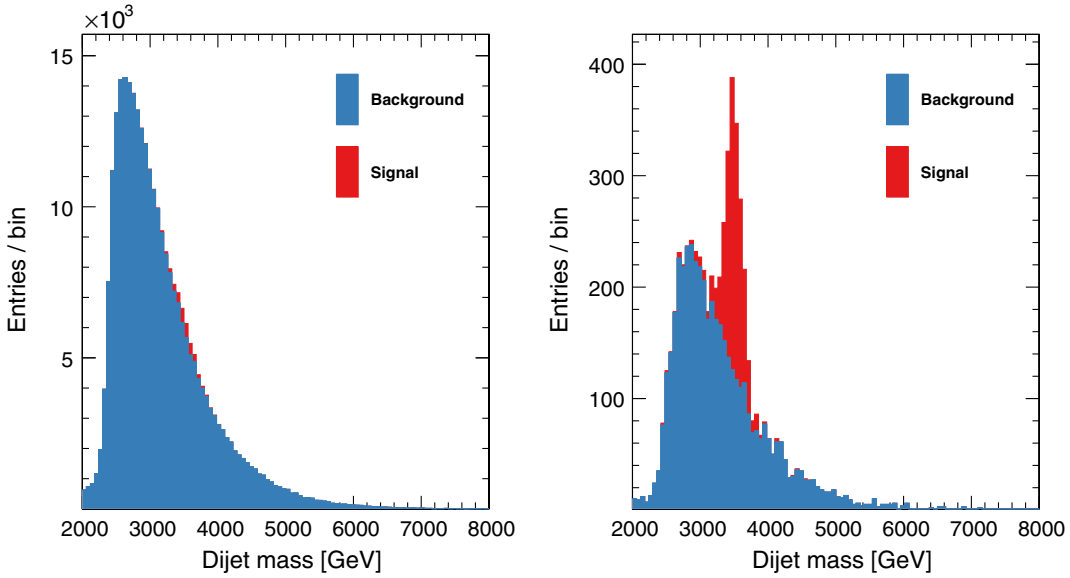


FIG. 8. Dijet mass distributions of the events prior to clustering (left) and for the cluster with the highest SB ratio (right), found when the data are partitioned into 60 clusters.

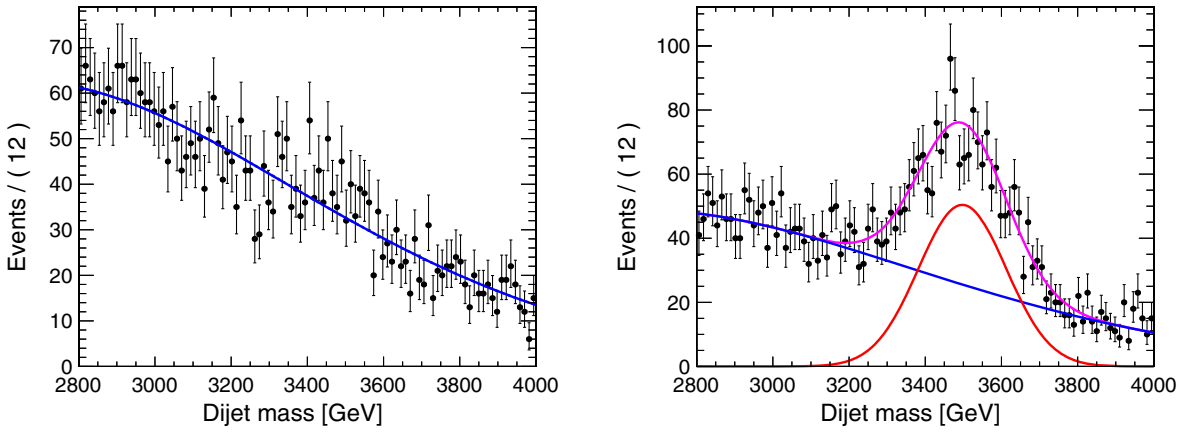


FIG. 9. Dijet mass distribution for events in the cluster closest to the cluster with the highest SB ratio (left) and the events in the cluster with the highest SB ratio (right). The background component (blue) is determined in the closest cluster and extrapolated to the highest SB ratio cluster. The signal contribution is shown in red, while the sum of signal and background contributions are shown in magenta.

have a reliable background estimation for each cluster. Given that UCluster is encouraged to create clusters with more specific properties, the background shape for a given partition might not have a trivial description. A possible way to mitigate this issue is to use the nearest cluster, in embedding space, as a background model for the cluster under study. Given that the anomalous signal remains localized in a particular cluster, the nearest clusters have the benefit to be signal free while still retaining similar properties to the cluster under consideration. To exemplify this idea, the cluster with the highest SB ratio shown in Fig. 8 is used. To model the data distribution in the closest cluster, a smooth falling distribution with four free parameters, commonly used in dijet resonance searches, is used [37–39], described as

$$\frac{dN}{dm_{jj}} = p_0 \frac{(1-x)^{p_1}}{x^{p_2+p_3 \ln(x)}}, \qquad x = m_{jj}/1 \text{ TeV}. \tag{5}$$

After the fit is performed, all background parameters, besides the overall normalization, are kept frozen. This function is then used to model the background in the cluster with the highest SB ratio. The signal modeling is done with a Gaussian function. The results of both fits are shown in Fig. 9.

### B. Global distribution effects on clusters

In order to relate the clusters in embedding space to physical observables, four high-level features were added

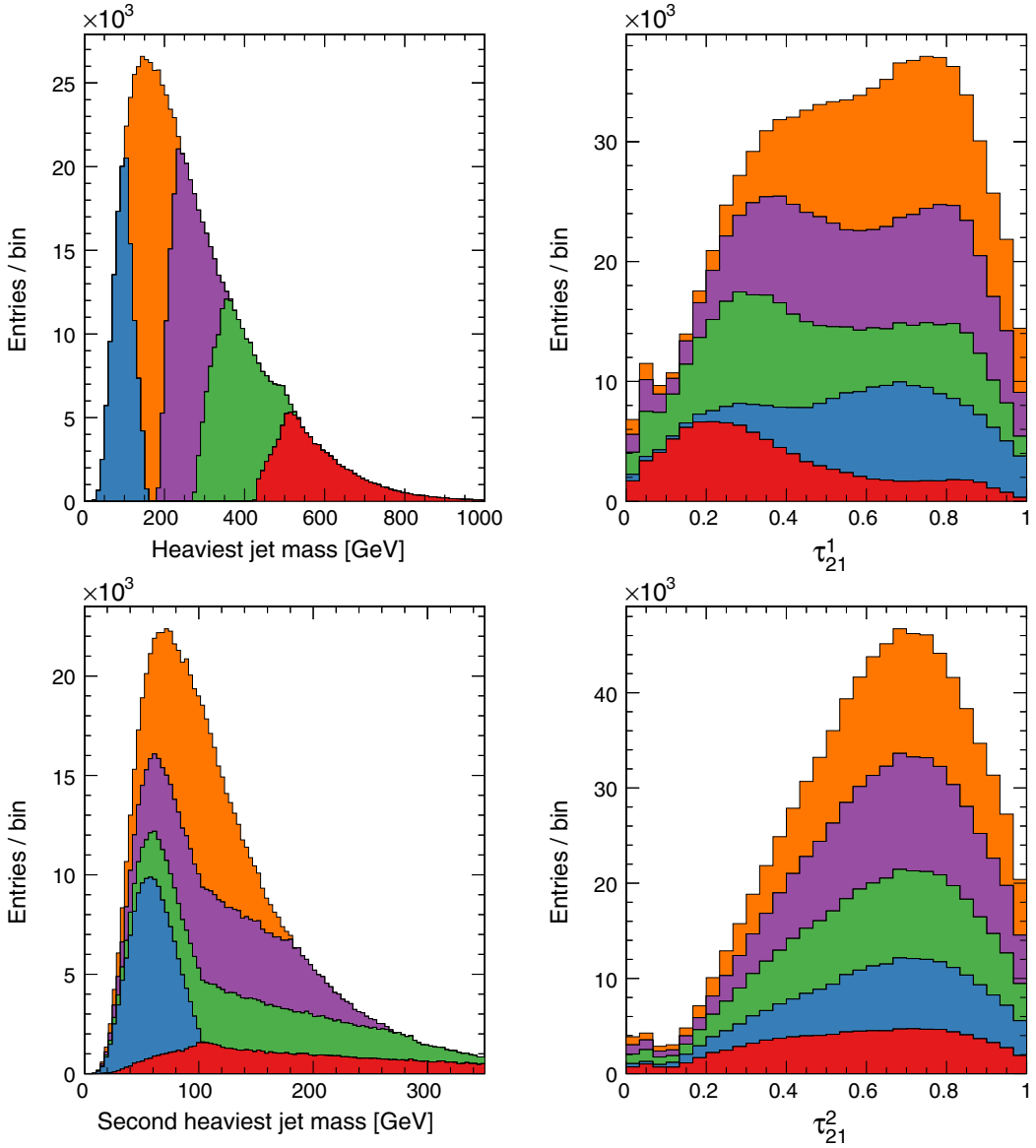


FIG. 10. Distributions for the four high level features used to parameterize the performance of UCluster trained with five clusters. Events belonging to the same clusters receive the same color. The stacked contribution of all clusters is then shown.

to the anomaly detection model: the invariant mass and $\tau_{21}$ of the two heaviest jets in the event.

To visualize the physical properties of the clusters, histograms of these four observables are shown in Fig. 10 with the stacked contributions of each individual cluster shown for UCluster with five clusters. From these distributions, there is a sharp separation between the cluster boundaries for the mass of heaviest jet in the dijet event. The sharp separation in jet mass is also related to the separation that is observed in the heaviest jet $\tau_{21}$. As pointed out in [22], QCD jets show a more distinctive two-prong structure when they have a larger mass. Therefore, heavier jets tend to have lower values of $\tau_{21}$. This correlation between jet mass and jet substructure is why the jet mass classification task leads to clusters where jets within a cluster have similar substructure.

## VII. CONCLUSION AND FUTURE PROSPECTS

In this work, we presented UCluster, a new method to perform unsupervised clustering for collision events in high energy physics. We explored two potential applications for this method: unsupervised multiclass classification and anomaly detection.

The ability of the embedding space to separate different processes is directly connected to the secondary task used in conjunction with the clustering objective. We proposed a classification task, which was motivated by the observations of the correlation between the jet mass and jet substructure observables, which is often useful for jet tagging. By learning to classify the mass of a jet, UCluster created an embedding space that was shown to have a better separation power for all the class components in the dataset compared to the jet mass alone.

UCluster was also studied for unsupervised anomaly detection. In this context, the classification task on jet masses was expanded to cover the entire event topology. Using this method, we were able to increase the signal-to-background ratio in a given cluster from an initial value of 1% up to 28%, while also observing a stable performance even for a large cluster multiplicity. A data-driven background estimation is also possible by using the closest cluster in embedding space to the cluster under investigation. This data-driven method allows for the selection of interesting clusters by comparing the background compatibility with the nearest cluster. Clusters of interest can be further investigated by a dedicated analysis.

We remark that different tasks than the ones proposed in this work can also be used to create meaningful embeddings. In particular, recent advances in autoencoders applied to particle physics [40] are strong candidates for a summary statistic that can encapsulate the event information in a lower dimensional representation, suitable for clustering.

Compared to [12,13], we relax the requirements on the label proportion for each different component in a mixed sample. One interesting point to notice is that, as presented in [41], the clustering assignment problem can instead be interpreted as an optimal transport problem. This insight is particularly interesting when the label proportions are known *a priori*. In this case, the additional knowledge of the label proportions can be directly added to the model as a regularization term of the form,

$$L_{\text{reg. cluster}} = \min \sum_{k}^{K} \sum_{j}^{n} \|f_\theta(x_j) - \mu_k\|^2 \pi_{jk} + \alpha \pi_{jk}(\log(\pi_{jk}) - 1). \tag{6}$$

This approach requires the term $\pi_{jk}$ to be numerically solved, subject to

$$\pi \mathbf{1}_K = \frac{1}{n} \mathbf{1}_N, \qquad \pi^{\mathrm{T}} \mathbf{1}_N = w, \tag{7}$$

where $w$ represents the vector of label proportions.

Furthermore, we considered an application where the initial number of mixed components was known. This condition was necessary to select a suitable number of clusters. However, this requirement could also be relaxed, as shown in [42,43], for example, where the clustering model is able to identify the optimal number of partitions given the properties of a dataset.

Finally, UCluster can also be used in conjunction with other anomaly detection approaches, where first a set of interesting clusters is identified and then further inspected by other methods.

## ACKNOWLEDGMENTS

The authors would like to thank Kyle James Read Cormier for the valuable suggestions regarding the development and clarity of this document. This research was supported in part by the Swiss National Science Foundation (SNF) under Contract No. 200020-182037.

[1] G. Aad *et al.* (ATLAS Collaboration), The ATLAS Experiment at the CERN large hadron collider, J. Instrum. **3**, S08003 (2008).

[2] S. Chatrchyan *et al.* (CMS Collaboration), The CMS Experiment at the CERN LHC, J. Instrum. **3**, S08004 (2008).

[3] B. Nachman and D. Shih, Anomaly detection with density estimation, Phys. Rev. D **101**, 075042 (2020).

[4] M. Aaboud *et al.* (ATLAS Collaboration), Search for new phenomena with large jet multiplicities and missing transverse momentum using large-radius jets and flavour-tagging at ATLAS in 13 TeV $pp$ collisions, J. High Energy Phys. 12 (2017) 034.

[5] A. M. Sirunyan *et al.* (CMS Collaboration), Measurement of the $t\bar{t}b\bar{b}$ production cross section in the all-jet final state in pp collisions at $\sqrt{s} = 13$ TeV, Phys. Lett. B **803**, 135285 (2020).

[6] A. M. Sirunyan *et al.* (CMS Collaboration), Search for high mass dijet resonances with a new background prediction method in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 05 (**2020**) 033.

[7] G. Aad *et al.* (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).

[8] E. M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.

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

[10] E. M. Metodiev and J. Thaler, Jet Topics: Disentangling Quarks and Gluons at Colliders, Phys. Rev. Lett. **120**, 241602 (2018).

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

[12] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, Estimating labels from label proportions, J. Mach. Learn. Res. **10**, 2349 (2009).

[13] G. Patrini, R. Nock, P. Rivera, and T. Caetano, (Almost) no label no cry, in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., Cambridge, MA, USA, 2014), pp. 190–198.

[14] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, Focal loss for dense object detection, arXiv:1708.02002.

[15] M. M. Fard, T. Thonet, and É. Gaussier, Deep k-means: Jointly clustering with k-means and learning representations, arXiv:1806.10069.

[16] J. A. Hartigan and M. A. Wong, Algorithm as 136: A k-means clustering algorithm, J. R. Stat. Soc. Ser. C **28**, 100 (1979).

[17] V. Mikuni and F. Canelli, ABCNet: An attention-based method for particle tagging, Eur. Phys. J. Plus **135**, 463 (2020).

[18] C. Chen, L. Zanotti Fragonara, and A. Tsourdos, GAPNet: Graph attention based point neural network for exploiting local feature of point cloud, arXiv:1905.08705.

[19] M. Abadi *et al.*, TENSORFLOW: Large-scale machine learning on heterogeneous systems (2015), software available from tensorflow.org.

[20] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[21] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure, J. High Energy Phys. 05 (2016) 156.

[22] P. T. Komiske, E. M. Metodiev, and J. Thaler, Metric Space of Collider Events, Phys. Rev. Lett. **123**, 041801 (2019).

[23] M. Tanabashi *et al.*, Review of particle physics, Phys. Rev. D **98**, 030001 (2018).

[24] E. Coleman, M. Freytsis, A. Hinzmann, M. Narain, J. Thaler, N. Tran, and C. Vernieri, The importance of calorimetry for highly-boosted jet substructure, J. Instrum. **13**, T01003 (2017).

[25] J. Duarte *et al.*, Fast inference of deep neural networks in FPGAs for particle physics, J. Instrum. **13**, P07027 (2018).

[26] M. Cacciari, G. P. Salam, and G. Soyez, The anti-$k_T$ jet clustering algorithm, J. High Energy Phys. 04 (2008) 063.

[27] M. Pierini, J. M. Duarte, N. Tran, and M. Freytsis, Hls4 ml lhc jet dataset (100 particles) (2020).

[28] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. **9**, 2579 (2008).

[29] H. W. Kuhn, The hungarian method for the assignment problem, Naval research logistics quarterly **2**, 83 (1955).

[30] G. Kasieczka, B. Nachman, and D. Shih, R&D dataset for LHC Olympics 2020 anomaly detection challenge, 10.5281/zenodo.2629073 (2019).

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

[32] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lematre, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES 3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.

[33] M. Cacciari, G. P. Salam, and G. Soyez, FASTJET user manual, Eur. Phys. J. C **72**, 1896 (2012).

[34] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, Parameterized neural networks for high-energy physics, Eur. Phys. J. C **76**, 235 (2016).

[35] J. Thaler and K. Van Tilburg, Identifying boosted objects with n-subjettiness, J. High Energy Phys. 11 (2011) 015.

[36] E. Gross and O. Vitells, Trial factors for the look elsewhere effect in high energy physics, Eur. Phys. J. C **70**, 525 (2010).

[37] T. Aaltonen *et al.* (CDF Collaboration), Search for new particles decaying into dijets in proton-antiproton collisions at s(1/2) = 1.96-TeV, Phys. Rev. D **79**, 112002 (2009).

[38] G. Aad *et al.* (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**, 302 (2016).

[39] V. Khachatryan *et al.* (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**, 031802 (2016).

[40] Deep generative models for fast shower simulation in ATLAS, Technical Report No. ATL-SOFT-PUB-2018-001, CERN, Geneva, 2018.
[41] A. Genevay, G. Dulac-Arnold, and J.-P. Vert, Differentiable deep clustering with cluster size constraints, arXiv:1910.09036.
[42] Y. Ren, N. Wang, M. Li, and Z. Xu, Deep density-based image clustering, arXiv:1812.04287.
[43] C. Patil and I. Baidari, Estimating the optimal number of clusters k in a dataset using data depth, Data Sci. Eng. **4**, 132 (2019).