M. van Beekveld[HEF, Nikhef], S. Caron[HEF, Nikhef], A. De Simone[SISSA, INFN], A. Farbin[UnivArlington], L. Hendriks[HEF], A. Jueid[UnivKonkuk], A. Leinweber[UnivAdelaide], J. Mamuzic[IFIC], E. Merényi[UnivRice], A. Morandini[SISSA, INFN], C. Nellist[HEF, Nikhef], S. Otten[HEF, UnivAmsterdam], M. Pierini[CERN], R. Ruiz de Austri[IFIC], J. Schouwenberg[HEF, Nikhef], S. Sekmen[KNU], R. Vilalta[UnivHouston], M. White[UnivAdelaide]

[HEF] HEF, Radboud University, Nijmegen, the Netherlands

[Nikhef] Nikhef, Amsterdam, the Netherlands

[SISSA] SISSA, Trieste, Italy

[INFN] INFN, Trieste, Italy

[UnivArlington] University of Texas Arlington

[UnivKonkuk] School of Physics, Konkuk University, Seoul, Republic of Korea

[UnivAdelaide] University of Adelaide

[IFIC] Instituto de Física Corpuscular, IFIC-UV/CSIC, Valencia, Spain

[UnivRice] Rice University

[UnivAmsterdam] University of Amsterdam

[CERN] CERN

[KNU] Department of Physics, Kyungpook National University, Daegu, South Korea

[UnivHouston] University of Houston

Acknowledgements MvB acknowledge support from the Dutch NWO-I program 156, "Higgs as Probe and Portal". The work of A.J. is supported by the National Research Foundation of Korea, Grant No. NRF-2019R1A2C1009419. The work of J. Mamuzic is supported in part by the Generalitat Valenciana (GV) through the contract APOSTD/2019/165, and by Spanish and European funds under the project PGC2018-094856-B-I00 (MCIU/AEI/FEDER, EU).

1 Model-Independent Signal Detection: A Challenge using Benchmark Monte Carlo Data and Machine Learning [1]

We discuss model-independent signal detection algorithms, with a particular focus on approaches that are based on unsupervised machine learning. We also offer a set of simulated LHC events, corresponding to $10/,\mathrm{fb}^{-1}$ of data. These events can be used as a benchmark dataset, for example for the comparison of signal detection algorithms. We explain the main features, the data format and describe the use of this data for an upcoming data challenge. The data is available at the webpage https://www.phenoMLdata.org.

1.1 Introduction and Goals

Problem

The Standard Model (SM) has been tremendously successful in describing particle physics phenomena. Nevertheless, many questions still remain unanswered, e.g. the origin of neutrino masses, the nature of dark matter, or the dynamics of electroweak symmetry breaking. Therefore, it is commonly accepted that physics beyond the SM (BSM) is needed in order to provide answers to the questions not addressed in the SM. A key ingredient for the journey towards a new physics discovery is handling the huge amount of complex experimental data collected at LHC. LHC data was initially analyzed for various signals that were predicted by high energy models that extend the SM. Typical examples are supersymmetry (SUSY) or models with extra dimensions. Since these searches did not show any significant deviations from the SM, the LHC search strategy was expanded by using so-called "simplified models" and "effective models". For simplified models, a certain production and a decay of a new hypothetical particle is assumed, and the model is tested using LHC data by optimizing data selection criteria on the energy, momenta and types of particle predicted by the model. For effective models, a new effective interaction is added to the SM Lagrangian and the new interaction is typically constrained with the measurement of SM processes. A sign of new particles typically shows up as an overproduction of events (compared to the SM) in a specific data-selection where the number of events expected from SM processes is compared to the number of measured events in statistical tests. A hint of new physics requires that the "SM-only" hypothesis is highly disfavoured. Often the test is quantified with the help of a p-value defined as the probability that a given result (or a more significant result) occurs under the SM hypothesis. A typical requirement for the discovery of an *expected* signal (such as the Higgs particle) is $p < 3 \times 10^{-7}$ corresponding to 5 standard deviations (5σ).
To date, no signal of BSM physics has been found at the LHC. However, the new physics could look different than generally assumed. This project deals with the question of how to search for a signal in collider data without adopting a specific signal hypothesis.

Attempts

A few attempts have been made to systematically search for new physics without signal assumption by scanning specific observables, such as the sum of the transverse momenta, or the invariant mass. Scans have been done with the help of model-independent (i.e. unsupervised) algorithms to locate anomalies. Such general searches without an explicit BSM signal assumption have been been performed by the DØ Collaboration [1–4] at the Tevatron using an unsupervised multivariate signal detection algorithm termed SLEUTH, by the H1 Collaboration [5,6] at HERA using a 1-dimensional signal detection algorithm, and by the CDF Collaboration [7, 8] at the Tevatron (using again 1-dimensional algorithms). A version of these 1-dimensional signal detection algorithms used in general searches is known as BUMPHUNTER in the HEP community [9]. At the LHC, versions of such searches have been performed by the ATLAS Collaboration at $\sqrt{s} = 13$ TeV [10], and preliminary versions have been performed by the ATLAS and

[1] M. van Beekveld, S. Caron, A. De Simone, A. Farbin, L. Hendriks, A. Jueid, A. Leinweber, J. Mamuzic, E. Merényi, A. Morandini, C. Nellist, S. Otten, M. Pierini, R. Ruiz de Austri, S. Sekmen, J. Schouwenberg, R. Vilalta, M. White

CMS Collaboration at $\sqrt{s} = 7$ and 8 TeV. Here, the ATLAS experiment proposed that the observation of one or more significant deviations in some phase-space region(s) can serve as a trigger to perform dedicated and model-dependent analyses where these 'data-derived' phase-space region(s) can be used as signal region(s). Such an analysis can then determine the level of significance by testing the SM hypothesis in these signal regions in a second dataset. Since the signal region is known also control selection can be defined to determine the background expectations in the signal region(s).
The field of machine learning (ML), sitting at the intersection of computational statistics, optimization, and artificial intelligence, has witnessed unprecedented progress over the past decade. Research in ML has recently led to the development of new and enhanced anomaly detection methods that could be used and extended for applications employing LHC or astroparticle data. Examples of such outlier detection algorithms recently proposed for HEP include density-based methods [11], model-independent searches with multi-layer perceptrons [12] , autoencoders [13–15], variational autoencoders [16, 17] or ML extended bump-hunting algorithms [18, 19].

Methodology

This contribution aims to initiate a comparison of signal detection algorithms. To this end, we supply a benchmark dataset containing simulated high-energy collision data. Furthermore, we provide a (non-exhaustive) list of methods that may be employed to extract a possible signal from this dataset in a model-independent and/or unsupervised way.

1.2 Data Description

1.2.1 Data generation procedures

We generate LHC events corresponding to a center-of-mass energy of 13 TeV. Events for the background and signal processes are generated at leading order (LO) with up to two additional partons in the matrix element using the event generator MG5_aMC@NLO v6.3.2 (Madgraph) and versions above [20] with the NNPDF PDF set [21] using 5 flavors in the definition of the proton. Madgraph is interfaced to Pythia 8.2 [22], that handles the showering of the matrix element level generated events. The matching with the parton shower, needed in the case when one or more additional jets are generated in Madgraph, is done using the MLM merging prescription [23]. Then, a quick detector simulation is performed with Delphes 3 [24, 25], using a modified version of the ATLAS detector card. Pileup is not included in this dataset. A repository of the data scripts that are used to generate the events is on GitHub [2].
The final state objects, as described in Table 1, are stored in a one-line-per-event text file (see below Section 1.2.2 for details). An event consists of a variable number of objects. An event is stored when at least one of the following requirements are fulfilled:

- At least one (b)-jet with transverse momentum $p_T > 60$ GeV and pseudorapidity $|\eta| < 2.8$, or
- at least one electron with $p_T > 25$ GeV and $|\eta| < 2.47$, except for $1.37 < |\eta| < 1.52$, or
- at least one muon with $p_T > 25$ GeV and $|\eta| < 2.7$, or
- at least one photon with $p_T > 25$ GeV and $|\eta| < 2.37$.

Of course, these are unrealistic trigger requirements, but we aim to create a flexible data set that allows for different types of studies that might need different selection criteria. The η-restriction on the electrons models a veto in the crack regions as often applied in ATLAS analyses. Such a veto can also be applied to photons by the user. Note that for the processes with the largest cross sections ($W^{\pm}/\gamma/Z + \text{jets}$ and QCD jet production) we have applied cuts on $H_T > 100$ GeV and 600 GeV respectively to make the

[2] https://github.com/melli1992/unsupervised_darkmachines

Symbol ID	Object
j	jet
b	b-jet
e-	electron (e^-)
e+	positron (e^+)
m-	muon (μ^-)
m+	antimuon (μ^+)
g	photon (γ)

Table 1: Definition of symbols used for final-state objects. Only b-quark jets are tagged, no τ- or c-jets have been defined.

data generation manageable. The observable H_T is defined as the scalar sum of the transverse momenta of all jets (with $p_{T,j_i} > 20$ GeV and $|\eta_{j_i}| < 2.8$):

$$H_T = \sum_i |p_{T,j_i}|. \tag{1}$$

Therefore, if one includes any of these processes in their analysis, one must make sure that the same cuts are also applied to the the other processes, which impacts the cross sections that are indicated in Table 2 (and therefore the event weights).
The requirements on the final states objects that are stored are

- (b-)jet: $p_T > 20$ GeV and $|\eta| < 2.8$,
- electron/muon: $p_T > 15$ GeV and $|\eta| < 2.7$,
- photon: $p_T > 20$ GeV and $|\eta| < 2.37$.

This means that, for example, a jet with $p_T = 10$ GeV is not included in the dataset. The detector simulation as performed by Delphes removes any electrons with $|\eta| > 2.5$, as the reconstruction efficiency is set to 0 beyond that point.
The scale choice is set dynamically by Madgraph during the event generation. The resulting cross sections are not reweighted with any of the available higher-order and/or resummed cross sections. All relevant SM (background) processes that have been generated are summarized in Table 2. For each process, the total number of generated events (N_{tot}) is at least the number that is needed for 10 fb^{-1} of data ($N_{10\,\text{fb}^{-1}}$).

For the BSM scenarios (signal events) we have chosen two SUSY channels: gluino ($\tilde{g}$) pair and lightest stop ($\tilde{t}_1$) pair production. The production channels and decays are

$$pp \to \tilde{g}\tilde{g}, \quad \tilde{g} \to t\bar{t}\tilde{\chi}_1^0,$$
$$pp \to \tilde{t}_1\tilde{t}_1, \quad \tilde{t}_1 \to t\tilde{\chi}_1^0.$$

For the gluino events, we used a simplified model in which the lightest SUSY particle is a 1 GeV neutralino. The considered masses of the gluino are indicated in Table 3. All other SUSY particle are set to 4.5 TeV. For the stop production scenarios, we assume a more realistic SUSY scenario with a varying lightest neutralino ($\tilde{\chi}_1^0$) mass. The masses of $\tilde{t}_1$ and $\tilde{\chi}_1^0$ are provided in Table 3. Again for this scenario, all other SUSY masses are set to 4.5 TeV.

SM processes			
Physics process	Process ID	σ (pb)	N_{tot} ($N_{10\,\text{fb}^{-1}}$)
$pp \to jj$	njets	$19718_{H_T>600\,\text{GeV}}$	415331302 (197179140)
$pp \to W^{\pm}(+2j)$	w_jets	$10537_{H_T>100\,\text{GeV}}$	135692164 (105366237)
$pp \to \gamma(+2j)$	gam_jets	$7927_{H_T>100\,\text{GeV}}$	123709226 (79268824)
$pp \to Z(+2j)$	z_jets	$3753_{H_T>100\,\text{GeV}}$	60076409 (37529592)
$pp \to t\bar{t}(+2j)$	ttbar	541	13590811 (5412187)
$pp \to W^{\pm}t(+2j)$	wtop	318	5252172 (3176886)
$pp \to W^{\pm}\bar{t}(+2j)$	wtopbar	318	4723206 (3173834)
$pp \to W^{+}W^{-}(+2j)$	ww	244	17740278 (2441354)
$pp \to t$+jets$(+2j)$	single_top	130	7223883 (1297142)
$pp \to \bar{t}$+jets$(+2j)$	single_topbar	112	7179922 (1116396)
$pp \to \gamma\gamma(+2j)$	2gam	47.1	17464818 (470656)
$pp \to W^{\pm}\gamma(+2j)$	Wgam	45.1	18633683 (450672)
$pp \to ZW^{\pm}(+2j)$	zw	31.6	13847321 (315781)
$pp \to Z\gamma(+2j)$	Zgam	29.9	15909980 (299439)
$pp \to ZZ(+2j)$	zz	9.91	7118820 (99092)
$pp \to h(+2j)$	single_higgs	1.94	2596158 (19383)
$pp \to t\bar{t}\gamma(+1j)$	ttbarGam	1.55	95217 (15471)
$pp \to t\bar{t}Z$	ttbarZ	0.59	300000 (5874)
$pp \to t\bar{t}h(+1j)$	ttbarHiggs	0.46	200476 (4568)
$pp \to \gamma t(+2j)$	atop	0.39	2776166 (3947)
$pp \to t\bar{t}W^{\pm}$	ttbarW	0.35	279365 (3495)
$pp \to \gamma\bar{t}(+2j)$	atopbar	0.27	4770857 (2707)
$pp \to Zt(+2j)$	ztop	0.26	3213475 (2554)
$pp \to Z\bar{t}(+2j)$	ztopbar	0.15	2741276 (1524)
$pp \to t\bar{t}t\bar{t}$	4top	0.0097	399999 (96)
$pp \to t\bar{t}W^{+}W^{-}$	ttbarWW	0.0085	150000 (85)

Table 2: Generated background processes (first column) with the corresponding identification (second column), the LO cross section σ in pb (third column) and the total number of generated events N_{tot} (fourth column). In the last column, we also indicate the number of events corresponding to 10 fb^{-1} of data ($N_{10\,\text{fb}^{-1}}$).

We include a second BSM model corresponding to a leptophobic topcolor Z' model [26], where an on-shell Z' boson is produced that subsequently decays into a pair of top quarks:

$$pp \to Z' \to t\bar{t}. \tag{2}$$

The masses of the Z' are provided in Table 3. In Table 3, one may find the process ID, cross sections σ, and total number of generated events N_{tot} of the BSM processes mentioned above.
Generally, the processes with lower cross sections are harder to extract out of the background events, as such processes result in a lower number of signal events. A notable exception that is present in the BSM dataset is the scenario where the lightest stop mass is 220 GeV (process ID stop_01). Although the cross section of this production scenario is relatively high, the signal events are nearly indistinguishable from

BSM processes			
Physics process	Process ID	σ (pb)	N_{tot} ($N_{10\,\text{fb}^{-1}}$)
$pp \to \tilde{g}\tilde{g}$ (1 TeV)	gluino_01	0.20	50000 (2013)
$pp \to \tilde{g}\tilde{g}$ (1.2 TeV)	gluino_02	0.05	50000 (508)
$pp \to \tilde{g}\tilde{g}$ (1.4 TeV)	gluino_03	0.014	50000 (144)
$pp \to \tilde{g}\tilde{g}$ (1.6 TeV)	gluino_04	0.004	50000 (44)
$pp \to \tilde{g}\tilde{g}$ (1.8 TeV)	gluino_05	0.001	50000 (14)
$pp \to \tilde{g}\tilde{g}$ (2 TeV)	gluino_06	$4.8{\cdot}10^{-4}$	50000 (5)
$pp \to \tilde{g}\tilde{g}$ (2.2 TeV)	gluino_07	$1.7{\cdot}10^{-4}$	50000 (2)
$pp \to \tilde{t}_1\tilde{t}_1$ (220 GeV), $m_{\tilde{\chi}_1^0} = 20$ GeV	stop_01	26.7	500000 (267494)
$pp \to \tilde{t}_1\tilde{t}_1$ (300 GeV), $m_{\tilde{\chi}_1^0} = 100$ GeV	stop_02	5.7	500000 (56977)
$pp \to \tilde{t}_1\tilde{t}_1$ (400 GeV), $m_{\tilde{\chi}_1^0} = 100$ GeV	stop_03	1.25	250000 (12483)
$pp \to \tilde{t}_1\tilde{t}_1$ (800 GeV), $m_{\tilde{\chi}_1^0} = 100$ GeV	stop_04	0.02	250000 (200)
$pp \to Z'$ (2 TeV)	Zp_01	0.38	50000 (3865)
$pp \to Z'$ (2.5 TeV)	Zp_02	0.12	50000 (1220)
$pp \to Z'$ (3 TeV)	Zp_03	0.044	50000 (442)
$pp \to Z'$ (3.5 TeV)	Zp_04	0.018	50000 (179)
$pp \to Z'$ (4 TeV)	Zp_05	0.008	50000 (80)

Table 3: Generated signal processes (first column) with the corresponding identification (second column), the LO cross section σ in pb (third column) and the total number of generated events N_{tot} (fourth column). In the last column, we also indicate the number of events corresponding to 10 fb^{-1} of data ($N_{10\,\text{fb}^{-1}}$).

the background events due to their topology, making it extremely difficult to separate the signal events from the background events.

1.2.2 Description of the data format

The data are provided in a one-line-per-event text format (CSV file), where each line has variable length and contains 3 event-specifiers, followed by the kinematic features for each object in the event. The format of CSV files are:

```
event ID; process ID; event weight; MET; METphi; obj1, E1, pt1, eta1, phi1;
                      obj2, E2, pt2, eta2, phi2; ...
```

The `event ID` is an event specifier. It is an integer to identify the generation of that particular event, included for debugging purposes only. The `process ID` is a string referring to the process that generated the event, as mentioned in Tables 2 and 3. The event weight w is defined as

$$w = \frac{\sigma}{N_{\text{lines}}} \times \left(10\,\text{fb}^{-1}\right), \tag{3}$$

with σ the cross section for a particular process, and N_{lines} the number of events in a single CSV file. With the release of this contribution we provide files for $N_{\text{lines}} = N_{10\,\text{fb}^{-1}}$ (with $N_{10\,\text{fb}^{-1}}$ in Table 2), such that all weights are 1. Additionally, when $N_{10\,\text{fb}^{-1}} < 20000$, we provide a second CSV file with $N_{\text{lines}} = 20000$. These conclude the event specifiers of each line in the CSV file.

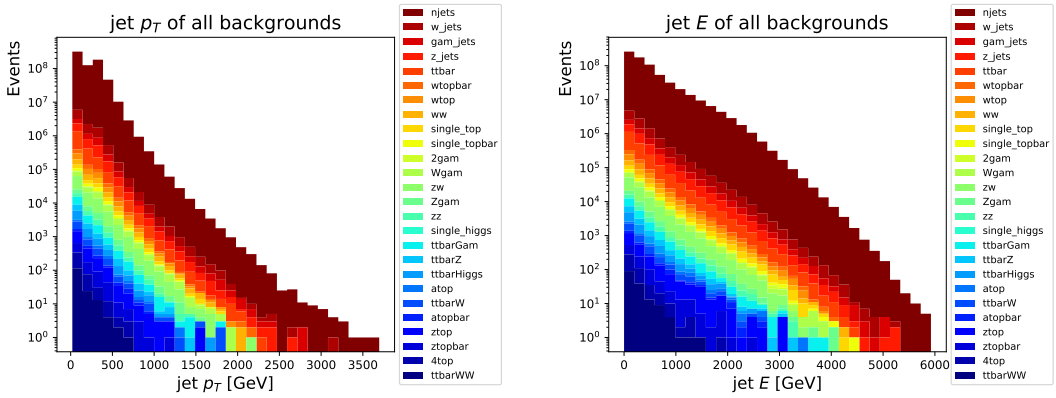

Fig. 1: Transverse momentum p_T (left) and energy E (right) in GeV of the jets for all backgrounds.

Concerning the kinematic features, the `MET` and `METphi` entries are the magnitude E_T^{miss} and the azimuthal angle $\phi_{E_T^{\text{miss}}}$ of the missing transverse energy vector of the event. The E_T^{miss} is based on the truth E_T^{miss}, meaning the transverse energy of those objects that genuinely escape detection. The object identifiers (`obj1`, `obj2`,...) are strings identifying each object in the event, using the identifiers of Table 1. Each object identifier is followed by 4 comma-separated values fully specifying the 4-vector of the object: `E1`, `pt1`, `eta1`, `phi1`. The quantities `E1` and `pt1` respectively refer to the full energy E and transverse momentum p_T of `obj1` in units of MeV. The quantities `eta1` and `phi1` refer to the pseudo-rapidity η and azimuthal angle ϕ of `obj1`.
As an example, an event corresponding to the final state of the $t\bar{t} + 2j$ process with two b-jets (with $E = 33.2$ GeV and $E = 55.8$ GeV) and one jet (with $E = 100.4$ GeV) reads:

94;ttbar;1;112288;1.74766;b,331927,147558,-1.44969,-1.76399;j,100406,85589,-0.568259,-1.17144;b,55808.8,54391.4,-0.198215,1.726

In Figures 1-4 we show the (stacked) distributions of the kinematic variables E, p_T, η, and ϕ of the jets and leptons in all of the generated background processes. In Figure 6 we show the number of jets N_{jet} and leptons (N_{lepton}) for the generated backgrounds. The E_T^{miss} and $\phi_{E_T^{\text{miss}}}$ distributions are shown in Figure 5, and the H_T distribution is shown in Figure 7. Note that only for Figure 7, we have filtered out the events with $H_T < 600$ GeV. For the other Figures, we show the events for all values of H_T for most backgrounds, except for the ones with tags njets ($H_T > 600$ GeV), w_jets, gam_jets and z_jets ($H_T > 100$ GeV). We stress again that for any analysis, the same kinematic cuts on *all* the background and signal events should be made.

1.2.3 Data storage

The generated MC data is stored in the form of ROOT files (including all stable hadrons) and in CSV files including only the information as described above. The CSV files corresponding to 10 fb^{-1} of data per process are available in `https://www.PhenoMLdata.org` for further validation. We encourage the community to explore the data, and report any issue to the authors of the proceedings. In the near future we plan to extend the dataset and to make the full set of ROOT files available, which currently take about 150 TB of disk space.

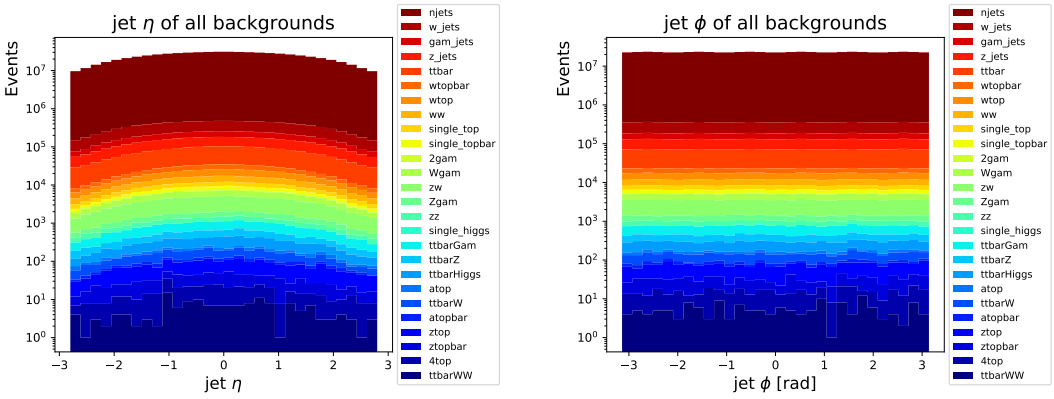

Fig. 2: Pseudorapidity η (left) and azimuthal angle ϕ (right) of the jets for all backgrounds.

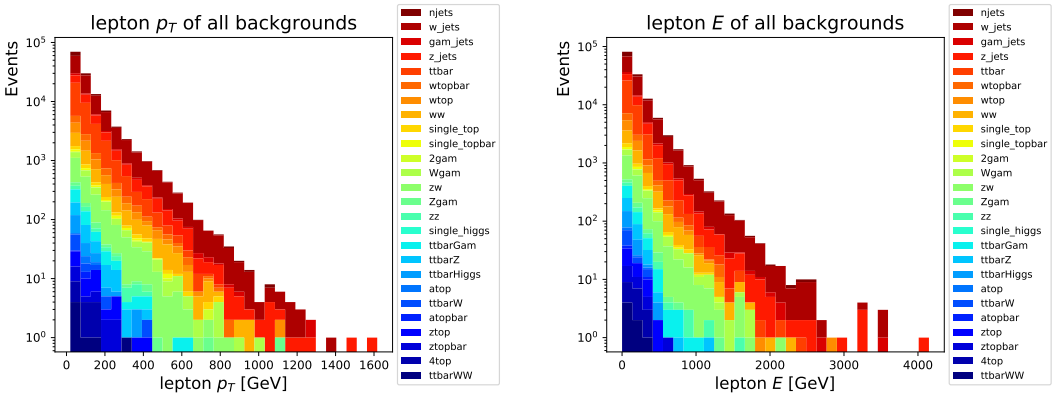

Fig. 3: Transverse momentum p_T (left) and energy E (right) in GeV of the leptons (e^+, e^-, μ^+, μ^-) for all backgrounds.

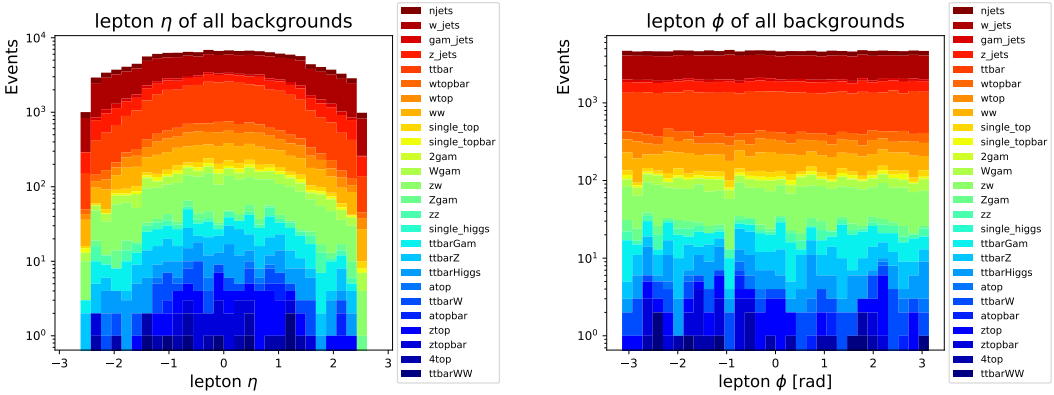

Fig. 4: Pseudorapidity η (left) and azimuthal angle ϕ (right) of the leptons (e^+, e^-, μ^+, μ^-) for all backgrounds.

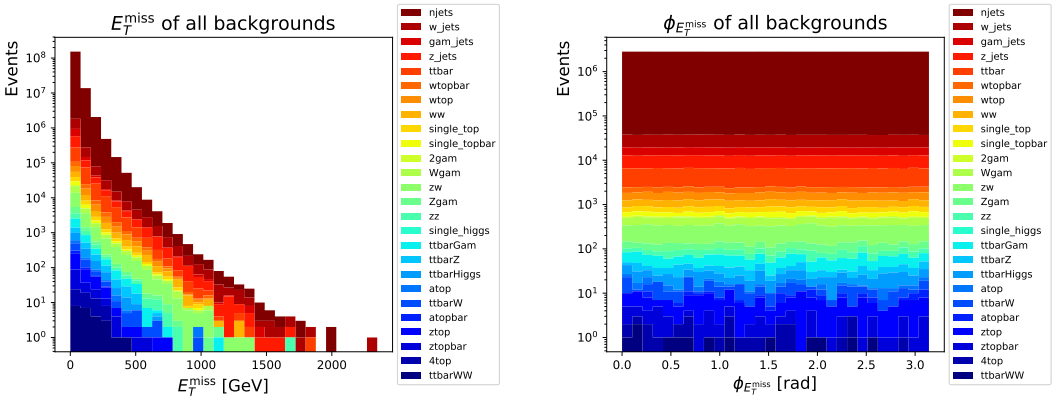

Fig. 5: Missing transverse energy E_T^{miss} in GeV and azimuthal angle $\phi_{E_T^{\text{miss}}}$ for all backgrounds.

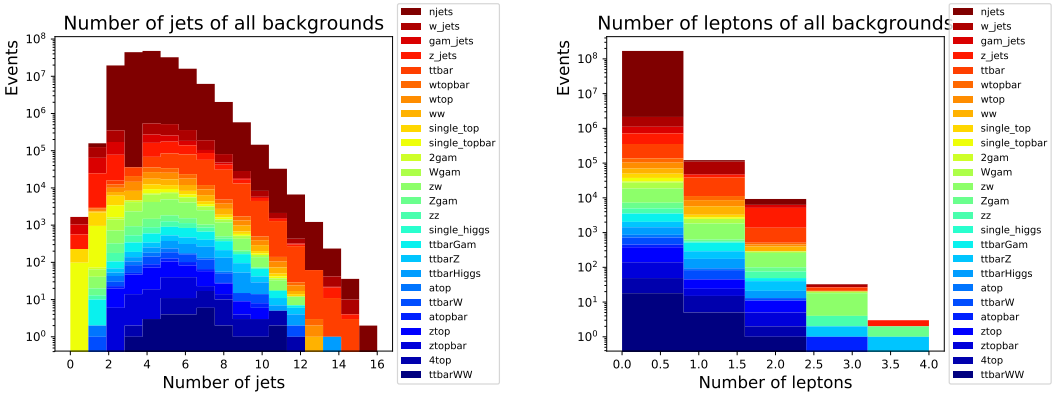

Fig. 6: Number of jets (left) and leptons (right).

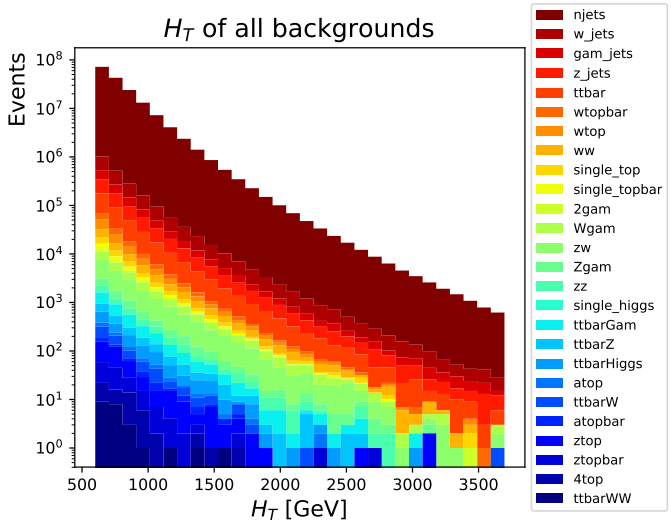

Fig. 7: The scalar sum of the jet transverse momenta H_T in GeV (see Eq. (1)) for the all backgrounds with $H_T > 600$ GeV imposed.

1.2.4 Benchmarking

The dataset presented in this paper is the result of an effort started back in 2018 in one of the working groups of the DarkMachines initiative [3]. We plan to use this dataset as a benchmark dataset to open a challenge, addressed to both particle physics and computer science communities. The challenge will aim at stimulating these communities to design and employ new methods/algorithms for detecting and characterizing signals in datasets featuring degenaracy, high dimensionality, and low signal-to-noise ratios, such as those faced when searching for new physics at the LHC.
Anomaly detection datasets used in ML are e.g. credit card fraud detection data [27]. Other challenges similar in spirit have been previously ran, e.g. in 2006 teams of theorists compared LHC data analysis approaches with mock datasets [4] and a QCD-oriented LHC Olympics 2020 [28] [5] On the other hand, we provide data set with very high statistics and including event-level features as well as full 4-vectors features, with several potential use cases.
The dataset in this paper (corresponding to an integrated luminosity of 10 fb^{-1} of data) can be used by the interested readers for training and validation, using the BSM signal samples provided. For the challenge, we will provide a statistically independent dataset, where signal events are included. These signal events are generated e.g. by one or more (undisclosed) types of BSM processes. The goal of the challenge will be to identify and characterize such signals. The submitted solutions will be judged and ranked according to specified metrics, based on the classification performance of the proposed algorithm with respect to the dataset with true labels assigned. More details will be provided when officially opening the challenge, and in a follow-up paper.

1.3 Approaches to the problem

The task at hand is to distinguish background from signal events. Since signal events are very similar to background events in terms of their topology, it is usually impossible to identify them by looking at individual events. Therefore, one needs to take into account effects that only appear when examining distributions in a collection of events. Since the signal and background events can be viewed as samples drawn from an (unknown in the case of signal) multi-dimensional probability distribution, and we only have a finite amount of data, we are restricted to statistical investigations, e.g. in the form of a hypothesis test against the null hypothesis that the given dataset does not contain any signal. In this section, we aim to give some examples for a signal detection algorithm.
In order to maximize the power of the test, it may be very helpful to transform the low-level (raw) features of the events into high-level ones. This crucial step of feature selection/engineering can be performed by using unsupervised learning techniques [29], e.g. by creating low-dimensional (latent) model of the data. There are at least four different approaches to design the signal detection algorithm and train it on data.

(a) Training the algorithm on real data, possibly being a mixture of signal and background. This is necessary when a reliable or accurate model for the background is not available. It will then be tested on another independent sample of real data.

(b) Training the algorithm on computer-generated backgrounds. It will then be tested on real data.

(c) Training the algorithm by two-sample comparison of background data and real data (e.g. [11, 12])

(d) Training the algorithm on a specific signal and background. This is what is typically done at LHC. Another possibility would be to train the algorithm on a large number of possible signals with a large variety.

[3] https://www.darkmachines.org/

[4] see e.g. http://public-archive.web.cern.ch/en/Spotlight/SpotlightOlympics-en.html

[5] The LHC Olympics 2020 is a low statistics dataset and a challenge to study anomalies in jets, i.e. to build an "anomaly jet detection algorithm" with inputs being the kinematical features of stable hadrons.

In all cases, the outcome can be reduced to constructing one or more variables which maximize the power to discriminate signal from background (e.g. the probability of being an outlier, see Fig 8).
In the cases where the training involves a real dataset, it is possible to replace it by a mock dataset where signals of various kinds are injected. This is done to validate the algorithm and assess its performance to spot outliers.
In the approaches involving the background-only data, one should keep in mind that the simulated data are not a perfect description of the LHC data, and mismodeling may show up as fake new physics signals.
Several traditional ML techniques and various deep learning techniques enable the design of an algorithm that promise to serve our purpose of identifying new physics from LHC data: Kernel Density Estimation [30], Gaussian Mixture Models [31], Flow models [32], Variational Autoencoders [33] and GANs [34]. In the subsequent subsections we have listed four overlapping approaches to the problem of finding new physics: anomaly detection, clustering, dimensional reduction and density estimation, all of whom could potentially be supported by the above-mentioned ML techniques.

1.3.1 Anomaly detection

Anomaly detection generally describes the process of identifying unexpected events in a dataset. With the aid of ML tools, this can be achieved in a supervised, semi-supervised or unsupervised manner. Since, in a model-independent scenario, we do not have labels for new physics signals, we are in principle only interested in the unsupervised approach. Nevertheless, one could label the SM expectation of some observable. A special feature of this label is that it does not correspond to an individual event, but to a collection of events. This allows us to employ all the possible approaches mentioned above, while still being unsupervised with respect to the signal. A successful anomaly detection algorithm would then be able to tag the signal events as outliers. A potential problem of the approaches is that very rare SM events may also be part of the collection of outliers.

We present an instructive toy example in Figure 8. We simulated data from a background expectation distributed exponentially and we combined it with a narrow Gaussian signal anomaly. In order to give an anomaly score to the points we trained the Local Outlier Factor (LOF) [35] on a background-only simulation, and subsequently used it on the dataset containing both inliers and outliers (this would correspond to approach (b) mentioned above). Despite its simplicity, this example shows two interesting characteristics. First, it is clear that feature selection is important, since the variable on the x-axis is discriminating, while the variable on the y-axis is less discriminating. This is because the exponential distribution of the background has a different variance in the two directions. Second, the example has the characteristics that it is difficult to separate an anomaly from the background with a simple selection on one of the two plotted variables. The purpose of anomaly detection in this context is not to find *all* anomalous points, but to be able to reliably state when a point (or a set of points) is anomalous and worth studying. The LOF gives a score to all points in order to assess how much they differ from the background. On the right-hand side of Figure 8, we see that most outliers have a high probability of being part of the signal, and not belong to the background.

Once all points are assigned an anomaly score, one may compare the distribution of such scores to a validation set containing only SM events. Therefore, we use the framework of a two-sample test, aimed at detecting statistically significant differences in the score distributions of inliers and outliers.

1.3.2 Clustering

We expect data amenable for analysis to lack in class labels (e.g. it is not known if the data is a signal event); it will then be necessary to extract information in an unsupervised fashion. A solution is to invoke clustering techniques [36, 37], where the goal is to group the data into clusters, each cluster bearing certain unique properties. Specifically, the goal is to partition the data such that the average distance

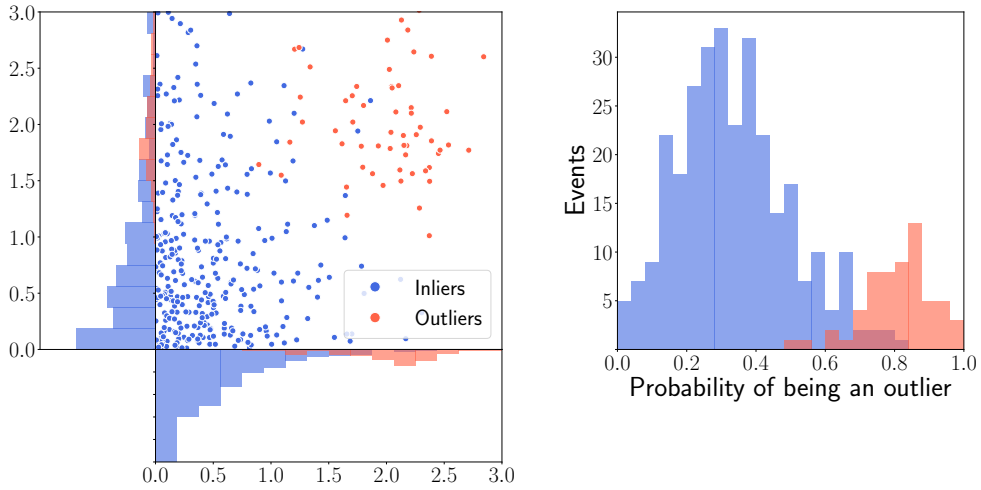

Fig. 8: *Left:* A narrow Gaussian anomaly centered around $(2, 2)$ (in red) is added to an exponentially-distributed background (in blue). *Right:* The probability of belonging to the signal events (outliers) is assigned to each point of the dataset and we can perform a counting. In this case, higher probabilities are correctly assigned to the outliers.

between objects in the same cluster (the average intra-distance) is significantly less than the distance between objects in different clusters (the average inter-distance). Several approaches have been developed to cluster data based on diverse criteria, such as the cluster representation (e.g. flat, hierarchical), the criterion function to identify sensible clusters (e.g. sum-of-squared errors, minimum variance), and the proximity measure that quantifies the degree of similarity between data objects (e.g. Euclidean distance, Manhattan norm, inner product). Our goal is to experiment with a variety of clustering approaches to gain a better understanding of the type of patterns emerging from clustering structures.
In order to analyze clusters to identify novel groupings that may point to new physics, one approach is to use what is known as *cluster validation* [38], where the idea is to assess the value of the output of a clustering algorithm by computing statistics over the clustering structure. Clusters with high degree of *cohesiveness*, where events within the group are sampled from regions of high probability density, are particularly relevant for analysis. In addition, one could carry out a form of *external cluster validation* [39], where the idea is to compare the output clusters to existing, known classes of particles. While finding clusters resembling existing classes may serve to confirm existing theories, clusters bearing no resemblance to known classes can potentially drive the search for new physics models.

1.3.3 Dimensionality Reduction

Data stemming from LHC arrive in copious amounts, are highly dimensional, and lack class labels; clustering can be useful to find patterns hidden in the data, a task whose importance has been highlighted in the previous section. Unfortunately, highly dimensional data create a plethora of complications during the data analysis process. Two possible solutions exist: we can either pre-process the data through dimensionality reduction techniques [40], or we can make use of specialized approaches [41].
Dimensionality reduction can be done through feature selection, by determining which features are most relevant, i.e. those that possess a high power to discriminate signal from background. This may come with some information loss, but it is commonly the case at the LHC that only a subset of information is needed to distinguish among different types of data. Another approach is to invoke principal component analysis: the data is transformed while eliminating cross correlations among the new features; the resulting subset can be further analyzed to filter out irrelevant features.
Another promising direction is to use ML to attain a reduced representation of the data by performing non-linear transformations [42,43]. This approach can have a strong impact in the search for new physics since it implements data transformations that can unveil hidden patterns corresponding to new particle

signals.

1.3.4 Density estimation

Events produced at the LHC (either real or simulated) can be thought as samples drawn from an unknown probability density function (PDF) that characterizes the complex physical processes leading to the generation of the events themselves. The PDF of a new physics signal might be different from the PDF of the SM. However, also the estimated PDFs of the SM, and the one from real experimental data may be different. Spotting and analyzing the differences in these two densities can provide a great deal of information about the underlying process (i.e. the true physical model) that generates the signal events.

However, estimating the PDF reliably starting from the raw data is far from trivial, especially if the number of features is high. This constitutes an active field of research in data science and, depending on the specific task, different approaches may be suitable [30, 44]. One such approach is kernel density estimation, which estimates the PDF by a sum of kernel functions (e.g. multivariate Gaussians) centered around each data point [45].

Assuming density estimation can be performed accurately, there are several ways to use it for model independent unsupervised analysis. For instance, one can compare the PDFs of real and simulated data from the LHC to detect differences. They point towards interesting signal regions, which can be used in order to guide further scrutiny. Furthermore, one could also perform clustering and anomaly detection in a way independent from the approaches mentioned before, see e.g. [46, 47].

One difficulty in applying density estimation on the dataset described in this work is the fact that the events change in dimensionality because the number of objects is not the same in every event. Additionally, there are both continuous data (for example energy and angles) and categorical data (object symbol). To circumvent these issues, one might try to map events to a different parameter space, a potential methodology is described in Ref. [17].

1.4 Conclusions

In this paper we have described a dataset aimed at constituting a benchmark for future model-independent studies of new physics detection at the LHC. We described the details of the data generation and the data format, which allow the user to easily handle the data with any programming language. We encourage the community to acquire familiarity with this dataset, which will also form the basis for a signal detection challenge to be announced soon. The challenge will be addressed to both computer scientists and particle physicists, fostering fruitful collaborations between them. Furthermore, we outlined some approaches, inspired by machine learning, to the problem of signal identification in background-dominated situations, like the ones commonly faced in high-energy physics.

With a benchmark dataset such as the one described in this paper it is possible to test and compare different techniques and algorithms for signal detection. We believe the effort of designing and comparing new algorithms tailored to the needs of high-energy physics will prove very useful for the future of the field.

References

[1] D0 Collaboration, B. Abbott et al., *Search for new physics in $e\mu X$ data at DØ using Sherlock: A quasi model independent search strategy for new physics*, Phys. Rev. **D62** (2000) 092004, arXiv:hep-ex/0006011 [hep-ex].

[2] D0 Collaboration, D0 Collaboration, *Quasi-model-independent search for new physics at large transverse momentum*, Phys. Rev. D **64** (2001) 012004, arXiv:hep-ex/0011067.

[3] D0 Collaboration, D0 Collaboration, *Quasi-Model-Independent Search for New High p_T Physics at DO*, Phys. Rev. Lett. **86** (2001) 3712–3717, arXiv:hep-ex/0011071.

[4] D0 Collaboration, D0 Collaboration, *Model independent search for new phenomena in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV*, Phys. Rev. D **85** (2012) 092015, arXiv:1108.5362 [hep-ex].

[5] H1 Collaboration, A. Aktas et al., *A General search for new phenomena in ep scattering at HERA*, Phys. Lett. **B602** (2004) 14–30, arXiv:hep-ex/0408044 [hep-ex].

[6] H1 Collaboration, F. D. Aaron et al., *A General Search for New Phenomena at HERA*, Phys. Lett. **B674** (2009) 257–268, arXiv:0901.0507 [hep-ex].

[7] CDF Collaboration, T. Aaltonen et al., *Model-Independent and Quasi-Model-Independent Search for New Physics at CDF*, Phys. Rev. **D78** (2008) 012002, arXiv:0712.1311 [hep-ex].

[8] CDF Collaboration, T. Aaltonen et al., *Global Search for New Physics with 2.0 fb^{-1} at CDF*, Phys. Rev. **D79** (2009) 011101, arXiv:0809.3781 [hep-ex].

[9] G. Choudalakis, *On hypothesis testing, trials factor, hypertests and the BumpHunter*, in *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011*. 2011. arXiv:1101.0390 [physics.data-an].

[10] ATLAS Collaboration, M. Aaboud et al., *A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment*, Eur. Phys. J. **C79** (2019) no. 2, 120, arXiv:1807.07447 [hep-ex].

[11] A. De Simone and T. Jacques, *Guiding New Physics Searches with Unsupervised Learning*, Eur. Phys. J. **C79** (2019) no. 4, 289, arXiv:1807.06038 [hep-ph].

[12] R. T. D'Agnolo and A. Wulzer, *Learning New Physics from a Machine*, Phys. Rev. **D99** (2019) no. 1, 015014, arXiv:1806.02350 [hep-ph].

[13] M. Farina, Y. Nakai, and D. Shih, *Searching for New Physics with Deep Autoencoders*, arXiv:1808.08992 [hep-ph].

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

[15] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, *Novelty Detection Meets Collider Physics*, arXiv:1807.10261 [hep-ph].

[16] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J.-R. Vlimant, *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, JHEP **05** (2019) 036, arXiv:1811.10276 [hep-ex].

[17] S. Otten, S. Caron, W. de Swart, M. van Beekveld, L. Hendriks, C. van Leeuwen, D. Podareanu, R. Ruiz de Austri, and R. Verheyen, *Event Generation and Statistical Sampling for Physics with Deep Generative Models and a Density Information Buffer*, arXiv:1901.00875 [hep-ph].

[18] E. M. Metodiev, B. Nachman, and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, JHEP **10** (2017) 174, arXiv:1708.02949 [hep-ph].

[19] J. H. Collins, K. Howe, and B. Nachman, *Anomaly Detection for Resonant New Physics with Machine Learning*, Phys. Rev. Lett. **121** (2018) no. 24, 241803, arXiv:1805.02664 [hep-ph].

[20] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *The automated computation of tree-level and next-to-leading order*

differential cross sections, and their matching to parton shower simulations, JHEP **07** (2014) 079, arXiv:1405.0301 [hep-ph].

[21] NNPDF Collaboration, R. D. Ball et al., *Parton distributions from high-precision collider data*, Eur. Phys. J. **C77** (2017) no. 10, 663, arXiv:1706.00428 [hep-ph].

[22] 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** (2015) 159–177, arXiv:1410.3012 [hep-ph].

[23] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, *ALPGEN, a generator for hard multiparton processes in hadronic collisions*, JHEP **07** (2003) 001, arXiv:hep-ph/0206293 [hep-ph].

[24] DELPHES 3 Collaboration, J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. LemaÃ®tre, A. Mertens, and M. Selvaggi, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, JHEP **02** (2014) 057, arXiv:1307.6346 [hep-ex].

[25] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, Eur. Phys. J. **C72** (2012) 1896, arXiv:1111.6097 [hep-ph].

[26] R. M. Harris and S. Jain, *Cross Sections for Leptophobic Topcolor Z' Decaying to Top-Antitop*, Eur. Phys. J. **C72** (2012) 2072, arXiv:1112.4928 [hep-ph].

[27] A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi, *Learned lessons in credit card fraud detection from a practitioner perspective*, Expert Systems with Applications **41** (08, 2014) 4915–4928. https://www.kaggle.com/mlg-ulb/creditcardfraud.

[28] G. Kasieczka, B. Nachman, and D. Shih, *R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge*, Apr, 2019. https://indico.cern.ch/event/809820/page/16782-lhcolympics2020.

[29] K. Albertsson, P. Altoe, D. Anderson, J. Anderson, M. Andrews, J. P. A. Espinosa, A. Aurisano, L. Basara, A. Bevan, W. Bhimji, D. Bonacorsi, B. Burkle, P. Calafiura, M. Campanelli, L. Capps, F. Carminati, S. Carrazza, Y. fan Chen, T. Childers, Y. Coadou, E. Coniavitis, K. Cranmer, C. David, D. Davis, A. D. Simone, J. Duarte, M. Erdmann, J. Eschle, A. Farbin, M. Feickert, N. F. Castro, C. Fitzpatrick, M. Floris, A. Forti, J. Garra-Tico, J. Gemmler, M. Girone, P. Glaysher, S. Gleyzer, V. Gligorov, T. Golling, J. Graw, L. Gray, D. Greenwood, T. Hacker, J. Harvey, B. Hegner, L. Heinrich, U. Heintz, B. Hooberman, J. Junggeburth, M. Kagan, M. Kane, K. Kanishchev, P. Karpiński, Z. Kassabov, G. Kaul, D. Kcira, T. Keck, A. Klimentov, J. Kowalkowski, L. Kreczko, A. Kurepin, R. Kutschke, V. Kuznetsov, N. Köhler, I. Lakomov, K. Lannon, M. Lassnig, A. Limosani, G. Louppe, A. Mangu, P. Mato, N. Meenakshi, H. Meinhard, D. Menasce, L. Moneta, S. Moortgat, M. Neubauer, H. Newman, S. Otten, H. Pabst, M. Paganini, M. Paulini, G. Perdue, U. Perez, A. Picazio, J. Pivarski, H. Prosper, F. Psihas, A. Radovic, R. Reece, A. Rinkevicius, E. Rodrigues, J. Rorie, D. Rousseau, A. Sauers, S. Schramm, A. Schwartzman, H. Severini, P. Seyfert, F. Siroky, K. Skazytkin, M. Sokoloff, G. Stewart, B. Stienen, I. Stockdale, G. Strong, W. Sun, S. Thais, K. Tomko, E. Upfal, E. Usai, A. Ustyuzhanin, M. Vala, J. Vasel, S. Vallecorsa, M. Verzetti, X. Vilasís-Cardona, J.-R. Vlimant, I. Vukotic, S.-J. Wang, G. Watts, M. Williams, W. Wu, S. Wunsch, K. Yang, and O. Zapata, *Machine Learning in High Energy Physics Community White Paper*, 2018. arXiv:1807.02876 [physics.comp-ph].

[30] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 2015. https://books.google.it/books?id=pIAZBwAAQBAJ.

[31] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*, vol. 38. M. Dekker New York, 1988.

[32] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, *Improved variational inference with inverse autoregressive flow*, in *Advances in neural information*

processing systems, pp. 4743–4751. 2016.

[33] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, arXiv:1312.6114 [stat.ML].

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Nets*, in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., pp. 2672–2680. Curran Associates, Inc., 2014. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[35] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, *LOF: identifying density-based local outliers*, in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104. 2000.

[36] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011. https://books.google.com/books?id=pQws07tdpjoC.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 ed., 2009. http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[38] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*. Academic Press, Inc., USA, 4th ed., 2008.

[39] B. E. Dom, *An Information-Theoretic External Cluster-Validity Measure*, in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pp. 137–145. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

[40] T. Kohonen, *Self-organized formation of topologically correct feature maps*, Biological Cybernetics **43** (Jan, 1982) 59–69. https://doi.org/10.1007/BF00337288.

[41] I. Assent, *Clustering high dimensional data*, WIREs Data Mining and Knowledge Discovery **2** (2012) no. 4, 340–350, https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1062.

[42] G. E. Hinton and R. R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, Science **313** (July, 2006) 504–507.

[43] Y. Bengio, A. Courville, and P. Vincent, *Representation Learning: A Review and New Perspectives*, IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013) no. 8, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

[44] Z. Wang and D. W. Scott, *Nonparametric density estimation for high-dimensional data, Algorithms and applications*, WIREs Computational Statistics **11** (2019) no. 4, e1461, https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1461.

[45] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[46] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, *Density-based clustering*, WIREs Data Mining and Knowledge Discovery **1** (2011) no. 3, 231–240, https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.30.

[47] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*, 2020. arXiv:2001.04990 [hep-ph].