# UC Berkeley
## UC Berkeley Previously Published Works

**Title**
High-dimensional anomaly detection with radiative return in e+e− collisions

**Authors**
Gonski, Julia
Lai, Jerry
Nachman, Benjamin
et al.

# High-dimensional anomaly detection with radiative return in $e^+e^-$ collisions

**Julia Gonski,**[a] **Jerry Lai,**[b] **Benjamin Nachman**[c,d] **and Inês Ochoa**[e]

[a]*Nevis Laboratories, Columbia University,*
*136 S Broadway, Irvington, NY, U.S.A.*

[b]*Department of Electrical Engineering and Computer Sciences, University of California,*
*Berkeley, CA 94720, U.S.A.*

[c]*Physics Division, Lawrence Berkeley National Laboratory,*
*Berkeley, CA 94720, U.S.A.*

[d]*Berkeley Institute for Data Science, University of California,*
*Berkeley, CA 94720, U.S.A.*

[e]*Laboratory of Instrumentation and Experimental Particle Physics,*
*Lisbon, Portugal*

*E-mail:* julia.gonski@cern.ch, thejerrylai@berkeley.edu,
bpnachman@lbl.gov, ines.ochoa@cern.ch

ABSTRACT: Experiments at a future $e^+e^-$ collider will be able to search for new particles with masses below the nominal centre-of-mass energy by analyzing collisions with initial-state radiation (radiative return). We show that machine learning methods that use imperfect or missing training labels can achieve sensitivity to generic new particle production in radiative return events. In addition to presenting an application of the classification without labels (CWoLa) search method in $e^+e^-$ collisions, our study combines weak supervision with variable-dimensional information by deploying a deep sets neural network architecture. We have also investigated some of the experimental aspects of anomaly detection in radiative return events and discuss these in the context of future detector design.

# Contents

## 1 Introduction

Searches for resonant new physics has been a cornerstone of high energy physics since at least the discovery of the $\rho$ meson [1] and up to and including the Higgs boson discovery by the ATLAS and CMS Collaborations at the Large Hadron Collider (LHC) [2, 3]. Since that time, there have been many proposals to widen the sensitivity of bump hunt analyses using machine learning (ML) [4–49] (see also ref. [50]). While searching for a peak in the invariant mass of two or more objects is a generic strategy for searching for the signature of new particles, it is not particularly sensitive to any given new physics model. Recent ML proposals identify patterns in high dimensional data, which can improve discovery potential while maintaining a low degree of signal-model and background-model dependence. The ML-assisted bump hunt has also been extended beyond the LHC to astrophysics, in the pursuit of stellar streams [51].

The goal of the current paper is to explore the potential of resonant anomaly detection at a future $e^+e^-$ collider. The identification of anomalous elements is done through a particular class of algorithms based on weakly supervised learning. Most ML applications in high energy physics are based on supervised learning, whereby training examples are produced from simulations with a known origin. Training ML methods directly on data must be able to cope with the lack of labels for what inputs are background or a potential new signal. The strategy used here leverages noisy labels on high-dimensional hadronic final state inputs.

The choice of studying advanced analysis strategies at future colliders is timely given the ongoing discussions of next generation energy frontier experiments [52–64]. This is particularly true if new techniques require dedicated detector modifications or computing resources. Additionally, an increased sensitivity from ML methods may increase the case for one future machine over another. This result explores the use of machine learning without signal/background labels (weak supervision) in $e^+e^-$ collisions and the input features can be high-dimensional and vary in size from event to event (variable-dimensional[1]).

Unlike at a hadron collider, the particles in a lepton collider are designed to have a fixed center-of-mass energy ($\sqrt{s}$). One can perform resonant anomaly detection by sifting through the collision debris for a range of center-of-mass energies. While beam energy scans are one approach to exploring different $\sqrt{s}$ values, we consider a complementary setting in this paper: *radiative return*. Due to initial state photon radiation that can carry a variable amount of energy away from the collision, collisions at fixed beam energies span a wide range in $e^+e^-$ $\sqrt{\hat{s}} \leq \sqrt{s}$, where $\sqrt{\hat{s}}$ is the effective collision energy. These radiative return events are often considered a nuisance for physics analysis, but they may provide a unique setting for new physics searches. The radiative return final state has been studied extensively for hadron spectroscopy at past $e^+e^-$ colliders [66–69] and was considered in ref. [70] for classical resonance searches at future colliders. We extend this study to the case of an ML-assisted bump hunt.

Machine learning has been studied in the context of future $e^+e^-$ colliders [71, 72] for probing Higgs boson properties. In addition to exploring the use of ML for direct new physics searches at a future $e^+e^-$ collider, we also use a deep neural network architecture that is capable of handling complex hadronic final states. In particular, high energy $e^+e^-$ events can produce a variable number of hadrons, with no inherent order. This is a point cloud, which can be processed with graph neural networks or set-based architectures. We use the latter, employing the Deep Sets architecture [73] implemented as a Particle Flow Network (PFN) [74]. Most of the previous weakly supervised methods used at most a handful of high-level observables, while we will be able to use high-dimensional, low-level inputs.

This paper is organized as follows. The simulations used for our empirical studies are presented in section 2. Section 3 introduces machine learning-based resonant anomaly detection methodology and technical details related to the training setup. Our numerical results are shown in section 4 and a discussion about the impact on future detector design is provided in section 5. The paper ends with conclusions and outlook in section 6.

## 2 Simulated samples and processing

We consider $e^+e^-$ collisions at a nominal center-of-mass (CoM) energy of $\sqrt{s} = 1\,\mathrm{TeV}$ that produce final states with jets and a photon from initial state radiation (ISR). The signal process studied is the production of a BSM heavy scalar $X$ that decays into a pair of scalars $a$, each decaying to two $b$-quarks, in association with an ISR photon: $e^+e^- \to X\gamma \to aa\gamma \to b\bar{b}b\bar{b}\gamma$. The following two sets of values for the invariant masses of

---

[1]Earlier studies of weak supervision with high-dimensional features can be found in ref. [65] and earlier studies of high-dimensional weak supervision for anomaly detection can be found in ref. [26].
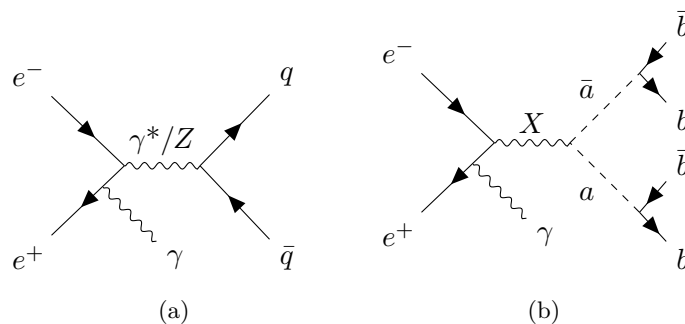
**Figure 1.** Feynman diagrams of the background (a) and signal (b) processes considered.

particles $X$ and $a$ are examined: $m_X, m_a = 350, 40\,\mathrm{GeV}$ and $700, 100\,\mathrm{GeV}$. The background originates from multijet production in association with an ISR photon, with a cross-section that is dominated by the Drell-Yan $\gamma^*/Z$ production and extends to close to the nominal $1\,\mathrm{TeV}$ CoM energy. Leading-order Feynman diagrams of the signal and background processes are shown in figure 1. Note that while we have selected two specific signal models, the methods studied here are largely[2] signal model agnostic.

The generation of background and signal events is done with MAD-GRAPH5_AMC@NLO 2.8.0 [75]. The background is generated at next-to-leading order in the strong coupling constant while the signal is simulated at leading order (MADGRAPH syntax: `e+ e- > j j a [QCD]` and `e+ e- > a h`, respectively, where `a` is the photon). Additional background processes are not included, as they are small, non-resonant in the regions we target, and thus do not change our conclusions. Future work will investigate the impact of more heterogeneous background data. In our study, simulated events are then passed to parton showering and hadronization using PYTHIA 8.244 [76] with its default settings. A minimum $E_T$ threshold of $10\,\mathrm{GeV}$ is placed on the photon, with a pseudo-rapidity that extends to $\pm 5$.

The detector simulation is parameterized with DELPHES 3.4.2 [77], using a card for a generic ILC detector.[3] A particle flow algorithm is used to combine tracking and calorimeter information and define the final reconstructed objects. Photons are built from energy deposits in the electromagnetic calorimeter, using the central and forward calorimeter systems with pseudo-rapidity coverages of $|\eta| < 3.0$ and $3.0 < |\eta| < 4.0$, respectively. Jets are built from particle flow objects (except isolated muons, electrons and photons) measured in the tracker (with an acceptance of up to $|\eta| < 3.0$), electromagnetic and hadronic calorimeters (the latter with an acceptance of up to 2.8 in the central system and up to 3.8 in the forward system in absolute pseudo-rapidity). The jet clustering is performed with the anti-$k_t$ [79] algorithm with a radius $R = 1.0$ implemented in FASTJET 3.3.2 [80, 81].

Events are selected for analysis if they contain at least two jets with a minimum transverse momentum ($p_T$) of $5\,\mathrm{GeV}$. An effective CoM energy can be calculated for all

---

[2]The signal model only plays a role in the choice of the features used for learning and if it does not have a narrow width, in the size of the invariant mass windows.

[3]The code for this detector card is available at https://github.com/iLCSoft/ILCDelphes. The parameterizations are mostly based on the ILD detector design [78].
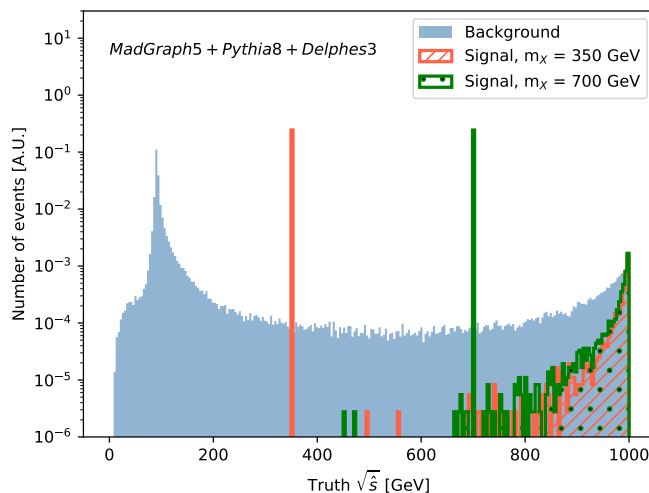
**Figure 2.** Effective center-of-mass energy calculated from truth-level quantities for the background and two signal processes considered.

events based on the energy of the colliding $e^+e^-$ pair after the photon radiation, as shown in figure 2 for all generated samples and based on truth-level quantities. Distributions of the photon transverse energy and pseudo-rapidity at detector-level are shown on figure 3 for the background and signal processes considered. For detector-level $\sqrt{\hat{s}}$ studies, we select the highest $p_T$ photon.

## 3 Methodology

A ML-assisted bump hunt is performed by scanning three different $\sqrt{\hat{s}}$ measures in $e^+e^-$ collisions: truth $\sqrt{\hat{s}}$, where the truth-level energy of the photon is subtracted from the incoming electron and positron beam energies; the $\gamma$-measured $\sqrt{\hat{s}}$, relying on the measured energy of the detected photon; and the hadron-measured $\sqrt{\hat{s}}$, computed using all the measured particles (in this case hadrons) with the exception of the photon. The truth-level analysis is presented as nominal, representing the ideal performance of the method without consideration of detector effects in the construction of the signal regions (detector effects are always included for the classification). Further description of the measured $\sqrt{\hat{s}}$ quantities and their impact on performance is given in section 5.

Signal regions are defined for both mass points as windows of 50 GeV centered at the resonance mass $m_X$. The sideband region then extends 50 GeV in both directions, excluding the signal region. A summary of these definitions can be found in table 1. The training (which is described later in more detail) is performed with a fixed luminosity over events from each of the three $\sqrt{\hat{s}}$ measures. The luminosity is determined from a normalization chosen to give 25000 events in the $m_X = 350$ GeV sideband region. A total of 5000225 background and 90001 signal events were generated. These events are split into three
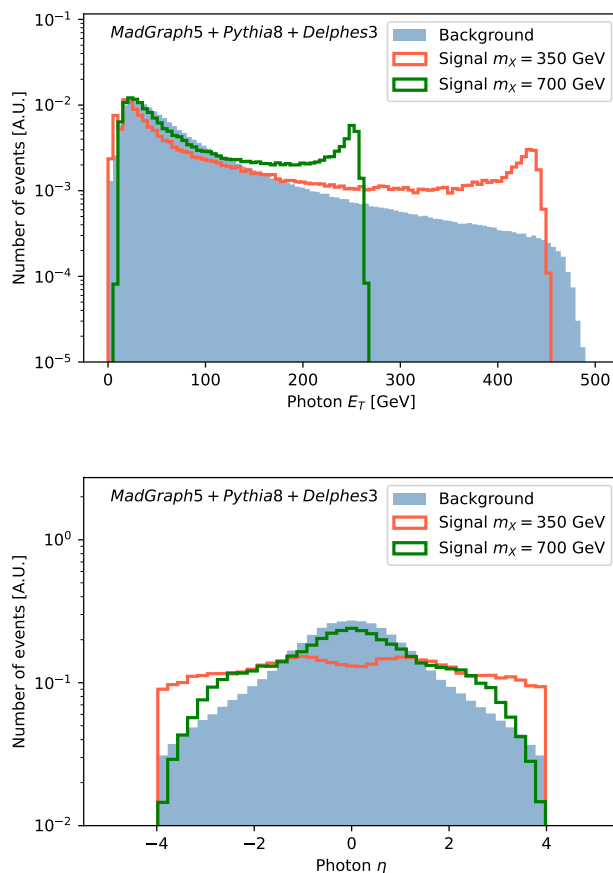
**Figure 3.** Distributions at detector-level of the photon transverse energy and pseudo-rapidity for the background and two signal processes. A minimum cut of $200\,\mathrm{GeV}$ is placed on the truth $\sqrt{\hat{s}}$.

orthogonal categories for training, validation, and testing, which ensures that statistical fluctuations are uncorrelated.

In particular, the unpolarized cross-section for the background process in our fiducial region is about 1 pb (as reported by MADGRAPH5_AMC@NLO). This amount of statistics would correspond to a collected luminosity of approximately $6.5\,\mathrm{ab}^{-1}$, well within the integrated luminosity estimates of the operating scenarios for colliders at $1\,\mathrm{TeV}$ [82]. While we have used this particular setup to provide concrete conclusions, the methodology and qualitative conclusions apply more generally to any future $e^+e^-$ collider.

For a given $\sqrt{\hat{s}}$ distribution, training utilizes one label for background events and another label for pseudodata, which is composed of background and a number of signal contamination events. The number of injected signal events is calculated in terms of the local signal significance $\sigma$ in the signal region, where $\sigma \equiv S/\sqrt{B}$, the approximate statistical significance, where $S$ is the number of signal events and $B$ is the number of background events, both in the pseudodata input class which draws events from the signal region $\sqrt{\hat{s}}$ window. The training for a given configuration is run seven times, with signal con-

|  |  |  | Signal region [GeV] | Sideband region [GeV] |
|---|---|---|---|---|
| $m_X$, $m_a$ = 350 GeV, 40 GeV |  |  | [325, 375) | [275, 325) $\cup$ [375, 425) |
| $m_X$, $m_a$ = 700 GeV, 100 GeV |  |  | [675, 725) | [625, 675) $\cup$ [725, 775) |

**Table 1.** Table of selections on $\sqrt{\hat{s}}$ that define the analysis signal and sideband regions.

|  |  |  | Signal Region | Sideband Region |
|---|---|---|---|---|
| Truth $\sqrt{\hat{s}}$ | $m_X = 350$ GeV | Background | 12133 | 25000 |
|  |  | Signal | 76 | 0 |
|  | $m_X = 700$ GeV | Background | 10832 | 22407 |
|  |  | Signal | 68 | 0 |
| $\gamma$-measured $\sqrt{\hat{s}}$ | $m_X = 350$ GeV | Background | 12121 | 25000 |
|  |  | Signal | 76 | 0 |
|  | $m_X = 700$ GeV | Background | 10441 | 21928 |
|  |  | Signal | 65 | 0 |
| Hadron-measured $\sqrt{\hat{s}}$ | $m_X = 350$ GeV | Background | 12147 | 25000 |
|  |  | Signal | 76 | 2 |
|  | $m_X = 700$ GeV | Background | 11754 | 24459 |
|  |  | Signal | 74 | 1 |

**Table 2.** Training yields for all $\sqrt{\hat{s}}$ measures, using the region definitions given in table 1, for an example signal injection corresponding to a $1.0\sigma$ significance.

taminations that correspond to $\sigma = 0, 0.5, 1, 2, 3, 5$ and $\infty$ (100% signal). The background events are the same in each training, while the signal events added to the pseudodata are randomly selected. Event yields used in training for the background and signal samples in each region can be found in table 2, for all three $\sqrt{\hat{s}}$ measures.

Two different training configurations for classification are used, which vary in how the two input labeled classes are populated.

- A semi-supervised scenario is studied as a benchmark. In this case, a neural network is trained to classify background events from a combination of signal and background events in the signal region, where the amount of signal contamination varies. In a context where real $e^+e^-$ collision data is available, this would correspond to classifying background-only simulation from data (where a potential signal could be included) [4, 7]. The assumption is made that the simulation perfectly describes the data; though this is not achievable with current simulations, future advancements may make such a strategy feasible.

- Weakly supervised learning can be used to enhance the presence of potential localized signals in data, without relying on simulation. Here, the classification without labels (CWoLa) approach is used, where a classifier is trained to distinguish events in the signal region from those in the sidebands. When a signal is present, this method

takes advantage of the different compositions in signal and sideband regions to learn how to distinguish background from signal [5, 6, 23, 83].

The envisioned analysis strategy in real data would be to define a signal-enriched region using the classifier score, and then estimate the background with a completely data-driven bump hunt (e.g. a parametric fit to the sidebands).

In both the semi-supervised and weakly-supervised scenarios, classifiers are parameterized with a Deep Sets model [73], using the Particle Flow Networks (PFN) implementation [74] in the `EnergyFlow` package.[4] A PFN is composed of a per-jet[5] function $\Phi$ and an event-level function $F$, both learnable and each with 3 layers and 20 nodes per layer. The inputs to the PFN are the variable-length set of jets, considering up to 15 jets per event. For each jet, 10 variables are considered: its four-vector, 5 angular radiation moments ($N$-subjettiness [84, 85] for $1 \le N \le 5$) and a flavor-tagging discriminant encoding 4 bins of $b$-tagging efficiency (50%, 70%, 90%, 100%). Plots of the input variables for leading and sub-leading jets in signal and background processes are shown in appendix A.

The neural networks are trained using KERAS [86] with the TENSORFLOW [87, 88] backend. The categorical cross-entropy loss function was minimized, using the Adam optimizer [89] with an initial learning rate of 0.0001. Starting with higher (up to 0.01) and lower (down to $10^{-5}$) learning rates was found to be suboptimal. Adagrad [90] and RM-SProp [91] optimizers were also studied, with no significant impact on the performance. The PFN was trained for 30 epochs with a batch size of 100. A longer training time of 100 epochs was also considered and did not strongly affect the final performance.

In the weakly supervised scenario where training utilizes events from different bins of $\sqrt{\hat{s}}$, care must be taken to ensure that the network output is agnostic to the $\sqrt{\hat{s}}$ of the events. A per-event normalization procedure is implemented to mitigate the $\sqrt{\hat{s}}$ correlation. Each jet's $\eta$ and $\phi$ is centered on the average value for all jets in the event, and its $p_T$ is scaled by the sum of jet transverse momentum in the event. The efficacy of the normalization procedure is verified by training the network to identify background events in the signal region from background events in the sideband. Since these events should only vary in their $\sqrt{\hat{s}}$ values, the normalization procedure can be deemed functional if the classifier is unable to discern these two classes of background events. Figure 4 shows the result of this training, confirming that the chosen normalization is sufficiently able to remove significant correlations of learned information with $\sqrt{\hat{s}}$.

Considerable variance in performance was observed across models with identical training scenarios. An ensemble procedure was developed to mitigate the effect of these fluctuations. Each training result presented here represents the average of 50 trained models, each with a random signal injection. The predicted values on the test set are subject to a non-linear transform known as quantile scaling, after which they follow a normal distribution. Models are then combined by averaging the quantile-scaled results for all 50 models. The results should therefore be interpreted as the expected/average sensitivity.

---

[4]https://energyflow.network/.

[5]We also investigated using all particle-flow candidates directly instead of jets, but the sensitivity was worse than the jet-based learning. Further details are provided in section 4.
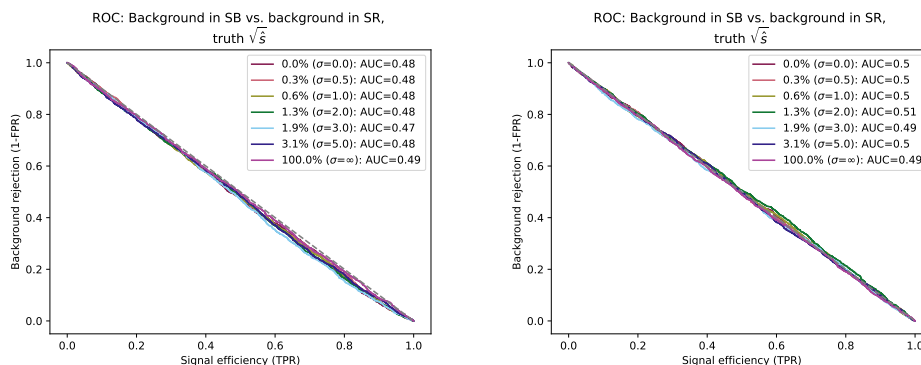
**Figure 4.** Receiver operating characteristic (ROC) curves from the weakly supervised training setup, comparing background in the sideband vs. background in the signal region, for the $m_X = 350\,\text{GeV}$ signal region on the left and the $m_X = 700\,\text{GeV}$ signal region on the right. Each curve represents a different signal contamination in the training set.
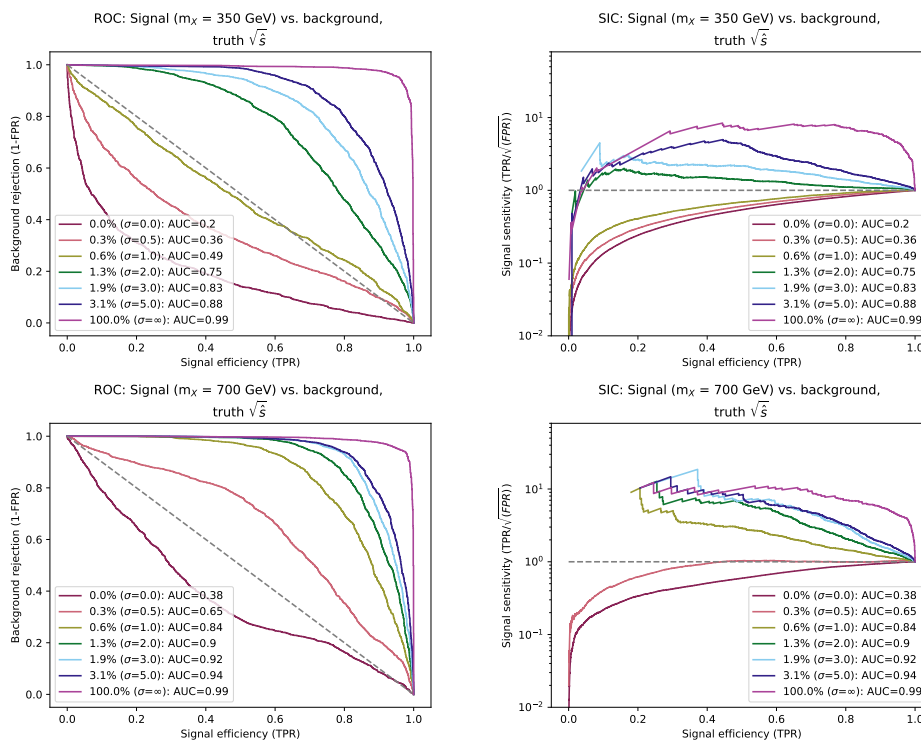
## 4 Results

The results of the training are displayed in two forms. The first is the receiver operating characteristic curve (ROC), which shows background rejection (inverse of the false positive rate, FPR) as a function of the signal efficiency (true positive rate, TPR), and demonstrates the discriminating power of the output net score. Additionally, the significance improvement characteristic (SIC, $\text{TPR}/\sqrt{\text{FPR}}$) is provided, which shows the signal sensitivity as a function of signal efficiency.

The SIC can be used as a proxy of how the output network score can enhance a signal excess in a physics analysis context. In particular, if $1\sigma$ of signal is injected and the SIC value at a particular signal efficiency is $n$, then a cut on the classifier corresponding to that efficiency is expected to increase the amount of signal to about $n\sigma$ (local significance). We do not perform a background fit, so the signal and background yield used to compute the SIC are determined by counting the number of events of each type in the signal region after a particular cut on the classifier output. In practice, the actual significance would slightly degrade from uncertainties in the fit, but as with other recent papers on anomaly detection in HEP, we decouple the network performance and the background fit quality in order to focus on the classifier methodology.

### 4.1 Particle flow networks

The semi-supervised training results are discussed first, as they provide a benchmark of the ideal classifier performance when separating signal from background in the signal region only. Figure 5 gives the ROC and SIC curves for both signal mass hypotheses as a result of the semi-supervised training configuration. The 0% signal contamination curve is shown as a reference. In this figure and in subsequent results, the lack of signal leads to a network that learns arbitrary discriminating information about the training inputs, resulting in small deviations of the area-under-curve (AUC) value from that of a completely random classifier (which can lead to an apparently 'anti-tagging'). A similar trend in the 0.3%
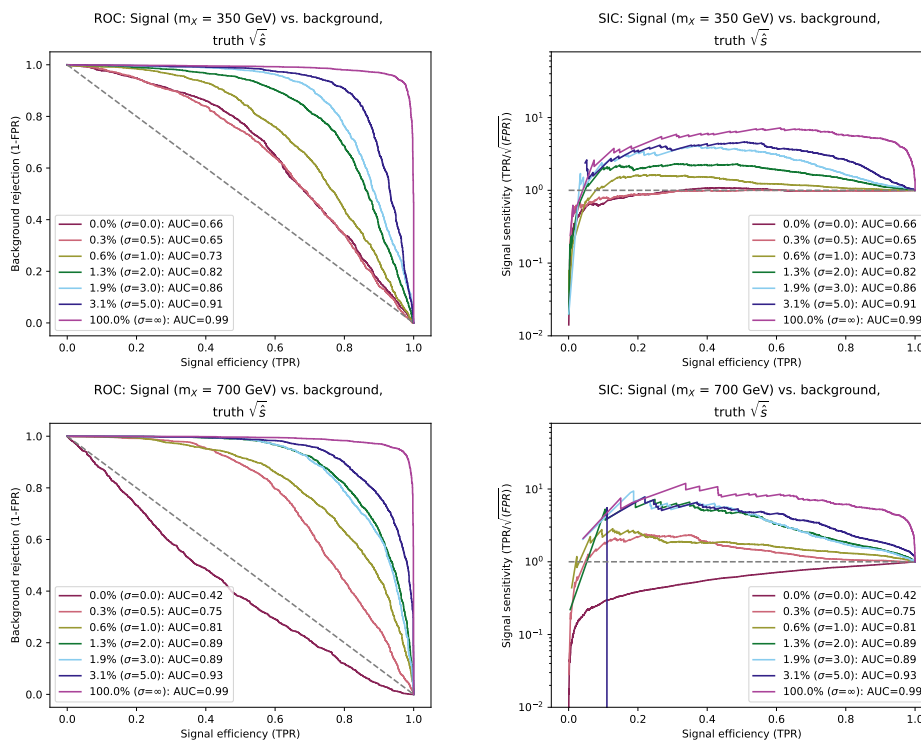
**Figure 5.** Semi-supervised training results in the form of ROC (left) and SIC (right) curves for two signals, $m_X = 350\,\text{GeV}$ (top) and $m_X = 700\,\text{GeV}$ (bottom) vs. background. The dashed grey line indicates no improvement from the use of the neural network score. A significance of $\infty$ indicates the case of pure signal versus pure background.

contamination result indicates that this amount of signal is consistently too small for the network to reliably learn and distinguish it from background.

With semi-supervised training, the network is able to detect an $m_X = 700\,\text{GeV}$ signal contamination of 0.6% from background in a single bin of $\sqrt{\hat{s}}$ and increase its significance by a factor of 11. The setup is considerably less sensitive to the $m_X = 350\,\text{GeV}$ signal, where the 0.6% contamination is indistinguishable, but doubled significance is achieved for a 1.3% contamination. Details on the impact of the neural net output selection can be found in table 3. Both of these signal injections correspond to excesses of $2\sigma$ sensitivity or less, making the role of the neural network substantial in elevating potential new physics to the discovery level of sensitivity.

Figure 6 shows the training results for the weakly supervised training scenario for both signal mass hypotheses. The network is similarly able to enhance a signal contamination down to 0.6% for both signal mass hypotheses, again with specific sensitivity increase factors given in table 3. The comparable results in figures 5 and 6 show that differences in the background-only case between the signal region and sideband region are not reducing the signal sensitivity.

An even higher level of input complexity is possible through utilizing the PFN architecture with all particles in the event. This representation was tried in the training, but
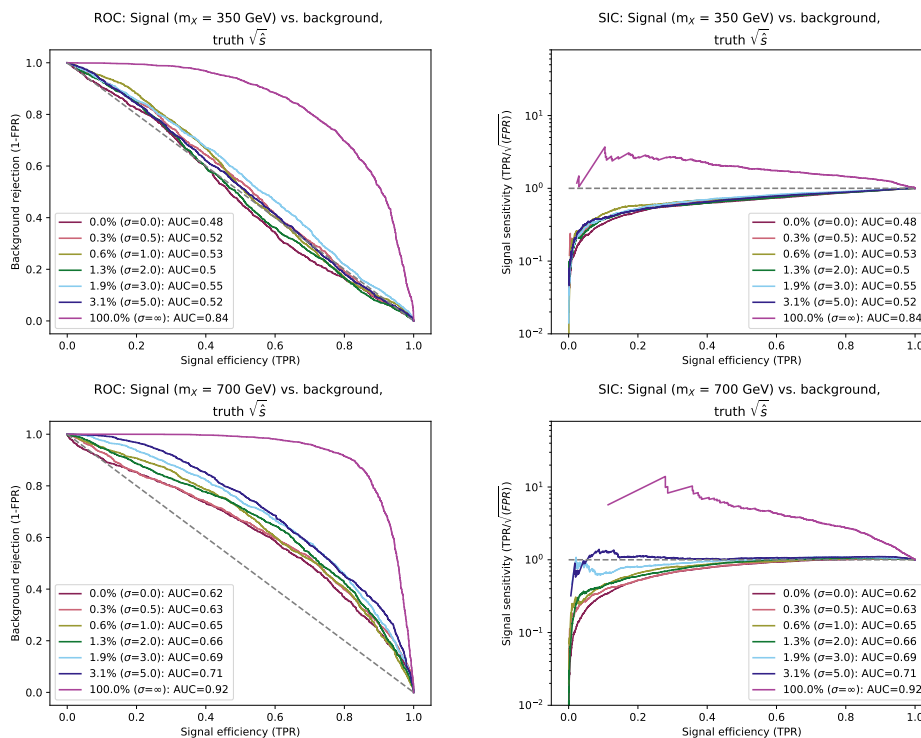
**Figure 6.** Weakly supervised training results in the form of ROC (left) and SIC (right) curves for two signals, $m_X = 350\,\text{GeV}$ (top) and $m_X = 700\,\text{GeV}$ (bottom) vs. background. The dashed grey line indicates no improvement from the use of the neural network score. A significance of $\infty$ indicates the case of pure signal versus pure background.

it was found to be less performant than the jet-level PFN studied here. This indicates a clear dependency of classifier performance on both input dimensionality and accordingly the amount of input statistics available in training.

## 4.2 Event level

To contextualize the results from the PFN training, we perform the same weakly supervised training procedure over input events described solely by event-level kinematic quantities and event shape variables.

The event-level network setup also utilized the `EnergyFlow` package, but replaced the PFN architecture with a dense neural network (DNN) with 15 input dimensions and two dense layers with 100 nodes each. A dropout rate of 0.2 was added to mitigate observed overtraining. As with the PFN, the DNN was trained for 30 epochs with a batch size of 100. Fifteen variables were used to describe input events: the leading and subleading large-radius jet masses and transverse momenta, the leading photon $p_T$, the X particle $p_T$, the particle and jet multiplicity, the ratio of leading jet and X $p_T$ to leading photon pt, $\ln y_{23}$, aplanarity, sphericity, transverse sphericity, and total jet mass. Distributions of these variables for signal and background events can be found in appendix B.

**Figure 7.** Weakly supervised training results using event-level input variables, in the form of ROC (left) and SIC (right) curves for $m_X = 350\,\text{GeV}$ (top) and $m_X = 700\,\text{GeV}$ (bottom). The dashed grey line indicates no improvement from the use of the neural network score. A significance of $\infty$ indicates the case of pure signal versus pure background.

Results are shown in figure 7 for the weakly supervised training scenario. The inability of the DNN to distinguish signal from background, except in the $100\%$ signal contamination scenario, indicates that the event-level variables are suboptimal for the signal of this study. Comparison to figure 6, which gives the analogous result for the PFN training, demonstrates the benefit of using high-dimensional input representations for the task of anomaly detection. Comparable signal sensitivity is delivered by a fully supervised signal vs. background training on event-level variables, and a PFN weakly supervised training with only $3.1\%$ signal contamination.

## 5 Future detector considerations

To extrapolate these results to a search in real collision data, the same method is applied using regions defined with a measured $\sqrt{\hat{s}}$ instead of one computed with truth-level quantities. Two different methods for measuring the total available energy are considered. One assumes that the ISR photon is captured by the detector, and therefore uses the measurement of its energy subtracted from the incoming electron-positron $\sqrt{\hat{s}}$ as a proxy for the amount of energy available in the collision. This is referred to as the *photon-measured* $\sqrt{\hat{s}}$. The second is the *hadron-measured* $\sqrt{\hat{s}}$, which covers the scenario where the photon is lost and the collision CoM must be obtained through measurements of the final-state
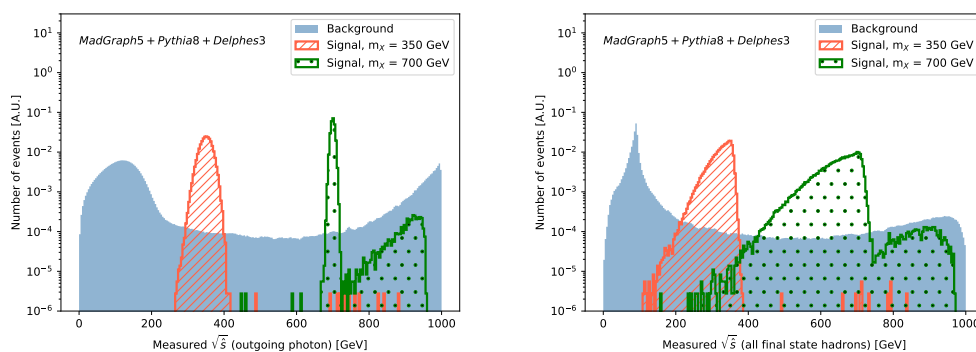
**Figure 8.** Distributions of the measured collision $\sqrt{\hat{s}}$ when the outgoing photon is captured and subtracted from the initial collision energy (left), and computed using only the final state hadrons in the event (right).

hadrons. Note that the highest $p_T$ photon is always used for these calculations. In the photon-measured case, if the true ISR photon is out of acceptance, the predicted $\sqrt{\hat{s}}$ will be significantly different from the true one. In the hadron-measured case, the selected photon is excluded from the calculation of $\sqrt{\hat{s}}$.

Figure 8 shows distributions of these two $\sqrt{\hat{s}}$ measurements for the background and both signal hypotheses. The incorporation of detector information gives each resonance a non-negligible width due to smearing introduced by detector resolution. As a result, the signal-to-noise in the signal region is lower. As seen in table 2, this width can also create some signal contamination in the sideband. Both of these effects make the discrimination task more challenging. In the photon-measured case, the signal and $Z$ peaks are approximately symmetric, with the width dominated by the photon energy resolution. The high-$\sqrt{\hat{s}}$ tail in the 750 GeV case is the result of events where the true ISR photon is out of acceptance and a random photon (the next highest $p_T$ one) is used to compute $\sqrt{\hat{s}}$. In the hadron-measured case, the signal peaks are asymmetric because there are both resolution and acceptance effects playing a role. The $Z$ peak is sharper for the hadron-measured case compared with the photon-measured case because the absolute energy resolution is better at low $p_T$: for the hadron-measured case, all of the particles are $\lesssim m_Z$ while for the photon-measured case, the photon energy is nearly $\sqrt{s}$.

Performance of the method can be found in figures 9 and 10, for the photon-measured and hadron-measured $\sqrt{\hat{s}}$, respectively. The signal significance is calculated using the signal region that is defined by the $\sqrt{\hat{s}}$ measure of interest, and is not normalized to the original truth sensitivity. Although the sensitivity is generally diminished by detector effects, there is still strong enhancement for a variety of signal injections, representing potential for this method in real collision data. Future innovations on hardware (e.g. increased acceptance) and software (e.g. combining photon- and hadron-measurements) may be able to close any remaining gaps between the truth $\sqrt{\hat{s}}$ and the reconstructed version(s).

These studies also indicate the effectiveness of the method in a real analysis context where a sliding window would be used to define sets of signal regions and sidebands across
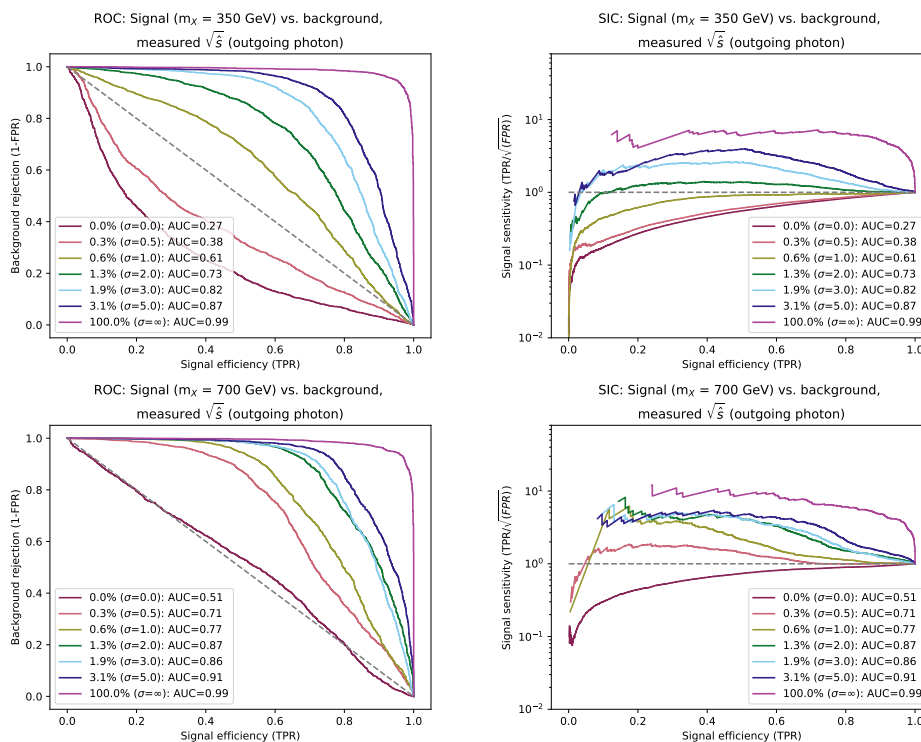
**Figure 9.** Weakly supervised PFN training results using photon-measured $\sqrt{\hat{s}}$, in the form of ROC (left) and SIC (right) curves for $m_X = 350\,\text{GeV}$ (top) and $m_X = 700\,\text{GeV}$ (bottom) vs. background. The dashed grey line indicates no improvement from the use of the neural network score. A significance of $\infty$ indicates the case of pure signal versus pure background.

$\sqrt{\hat{s}}$. In this case, there would necessarily be some sidebands with more signal events than the signal region, leading to suboptimal sensitivity, though at least one such region will exist where the signal is mostly in the signal region. Here a cross-training procedure can be useful to mitigate the effect of a large trials factor,[6] and it is worth noting that this is not accounted for in the significances provided here.

A summary of all training configurations and the impact of the neural net on the achievable signal sensitivity can be found in table 3.

## 6 Conclusions

We present results of high- and variable-dimensional anomaly detection applied to the search for new physics at a future $e^+e^-$ collider. Radiative return events are leveraged to scan $\sqrt{\hat{s}}$ for new resonant particles decaying to hadrons, while remaining agnostic to the mass scale of new physics. The analysis methodology uses different regions of $\sqrt{\hat{s}}$ to obtain training regions with varying signal-to-background ratios, allowing for the construction of a classifier that uses noisy labels for sensitivity to signal characteristics without relying on

---

[6]In the simplest, case one could divide the data in half, train on half and test on the other. This avoids any trials factor, since the two are independent. The two sets can then also be swapped to use all data for training and testing (although never more than half are used for training/testing at a given time).
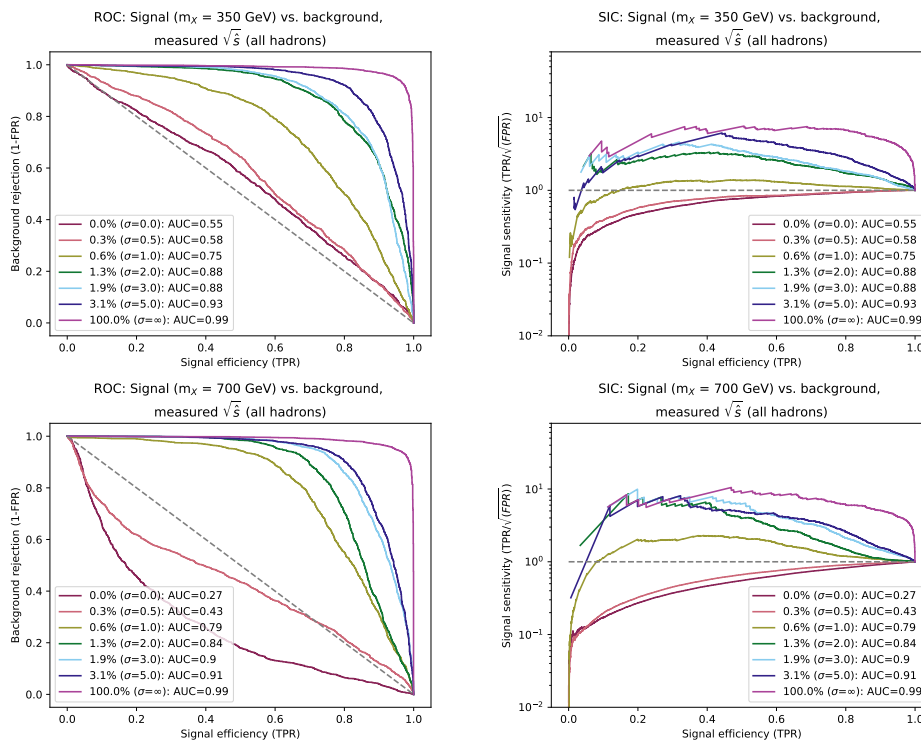
**Figure 10.** Weakly supervised PFN training results using hadron-measured $\sqrt{\hat{s}}$, in the form of ROC (left) and SIC (right) curves for $m_X = 350\,\text{GeV}$ (top) and $m_X = 700\,\text{GeV}$ (bottom) vs. background. The dashed grey line indicates no improvement from the use of the neural network score. A significance of $\infty$ indicates the case of pure signal versus pure background.

an input model. Classifiers are trained over $e^+e^-$ events modeled as particle flow networks, with up to 150 values characterizing each event.

When training over a simulated dataset with a realistic luminosity, signals with an initial signal-to-background ratio of $\lesssim 1\sigma$ are distinguishable by this classifier, and their sensitivity can be enhanced by over an order of magnitude. A normalization procedure is employed that removes correlation of the result with $\sqrt{\hat{s}}$ and ensures that no spurious signals are found in background-only analyses. Together, these results indicate the strong potential for this method to facilitate new physics searches in future $e^+e^-$ datasets. Further studies can be used to inform the design of future detectors by understanding the impact of detector acceptance and resolution on the sensitivity to generic new physics, as well as utilizing full detector simulation. The examination of different input features or even higher input dimensionality could also broaden application of this method beyond hadronically decaying resonant new physics to an even more generic search. These avenues are both closely linked to the development of a robust and exciting physics program at a next-generation $e^+e^-$ collider experiment.

**Code and data.** Our code can be found at https://github.com/bnachman/ILCAnomalies and the data are available at Zenodo.[7]

---

[7]https://zenodo.org/record/5181211.

| | | Signal Signif. [$\sigma$] | 0 | 0.5 | 1.0 | 2.0 | 3.0 | 5.0 | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|
| Semi-supervised, truth $\sqrt{\hat{s}}$ | $\mathrm{m}_X = 350\,\mathrm{GeV}$ | AUC | 0.2 | 0.36 | 0.49 | 0.75 | 0.83 | 0.88 | 0.99 |
| | | Max SIC (new $\sigma$) | 1.0 (0.0) | 1.0 (0.5) | 1.0 (1.0) | 2.0 (4.0) | 4.5 (14) | 5.0 (25) | 8.4 ($\infty$) |
| | $\mathrm{m}_X = 700\,\mathrm{GeV}$ | AUC | 0.38 | 0.65 | 0.84 | 0.9 | 0.92 | 0.94 | 0.99 |
| | | Max SIC (new $\sigma$) | 1.0 (0.0) | 1.0 (0.5) | 10.2 (10.2) | 12.7 (25) | 18.6 (56) | 14.7 (74) | 12.3 ($\infty$) |
| Weakly supervised, truth $\sqrt{\hat{s}}$ | $\mathrm{m}_X = 350\,\mathrm{GeV}$ | AUC | 0.66 | 0.65 | 0.73 | 0.82 | 0.86 | 0.91 | 0.99 |
| | | Max SIC (new $\sigma$) | 1.0 (0.0) | 1.0 (0.5) | 1.6 (1.6) | 2.3 (4.6) | 4.1 (12) | 4.6 (23) | 7.2 ($\infty$) |
| | $\mathrm{m}_X = 700\,\mathrm{GeV}$ | AUC | 0.42 | 0.75 | 0.81 | 0.89 | 0.89 | 0.93 | 0.99 |
| | | Max SIC (new $\sigma$) | 1.0 (0.0) | 2.4 (1.2) | 2.9 (2.9) | 7.2 (14) | 9.4 (28) | 7.8 (39) | 12.0 ($\infty$) |
| Weakly supervised, photon $\sqrt{\hat{s}}$ | $\mathrm{m}_X = 350\,\mathrm{GeV}$ | AUC | 0.27 | 0.38 | 0.61 | 0.73 | 0.82 | 0.87 | 0.99 |
| | | Max SIC (new $\sigma$) | 1.0 (0.0) | 1.0 (0.5) | 1.0 (1.0) | 1.4 (2.8) | 2.6 (7.8) | 4.0 (20) | 7.2 ($\infty$) |
| | $\mathrm{m}_X = 700\,\mathrm{GeV}$ | AUC | 0.51 | 0.71 | 0.77 | 0.87 | 0.86 | 0.91 | 0.99 |
| | | Max SIC (new $\sigma$) | 1.0 (0.0) | 1.8 (0.9) | 6.1 (6.1) | 7.6 (15) | 6.8 (20) | 5.5 (28) | 12.6 ($\infty$) |
| Weakly supervised, hadron $\sqrt{\hat{s}}$ | $\mathrm{m}_X = 350\,\mathrm{GeV}$ | AUC | 0.54 | 0.58 | 0.75 | 0.88 | 0.88 | 0.93 | 0.99 |
| | | Max SIC (new $\sigma$) | 1.0 (0.0) | 1.0 (0.5) | 1.4 (1.4) | 3.3 (6.6) | 4.6 (14) | 6.1 (31) | 7.6 ($\infty$) |
| | $\mathrm{m}_X = 700\,\mathrm{GeV}$ | AUC | 0.27 | 0.43 | 0.79 | 0.84 | 0.9 | 0.91 | 0.99 |
| | | Max SIC (new $\sigma$) | 1.0 (0.0) | 1.0 (0.5) | 2.3 (2.3) | 8.6 (17) | 9.9 (30) | 8.0 (40) | 10.5 ($\infty$) |

**Table 3.** Summary of all considered training scenarios, the resulting area-under-curve (AUC), and maximum value of the significance improvement characteristic (SIC) curve. A significance of $\infty$ indicates the case of pure signal versus pure background. The "new $\sigma$" refers to the sensitivity of the signal excess after a neural net selection, which is dictated by the max SIC value.

## Acknowledgments
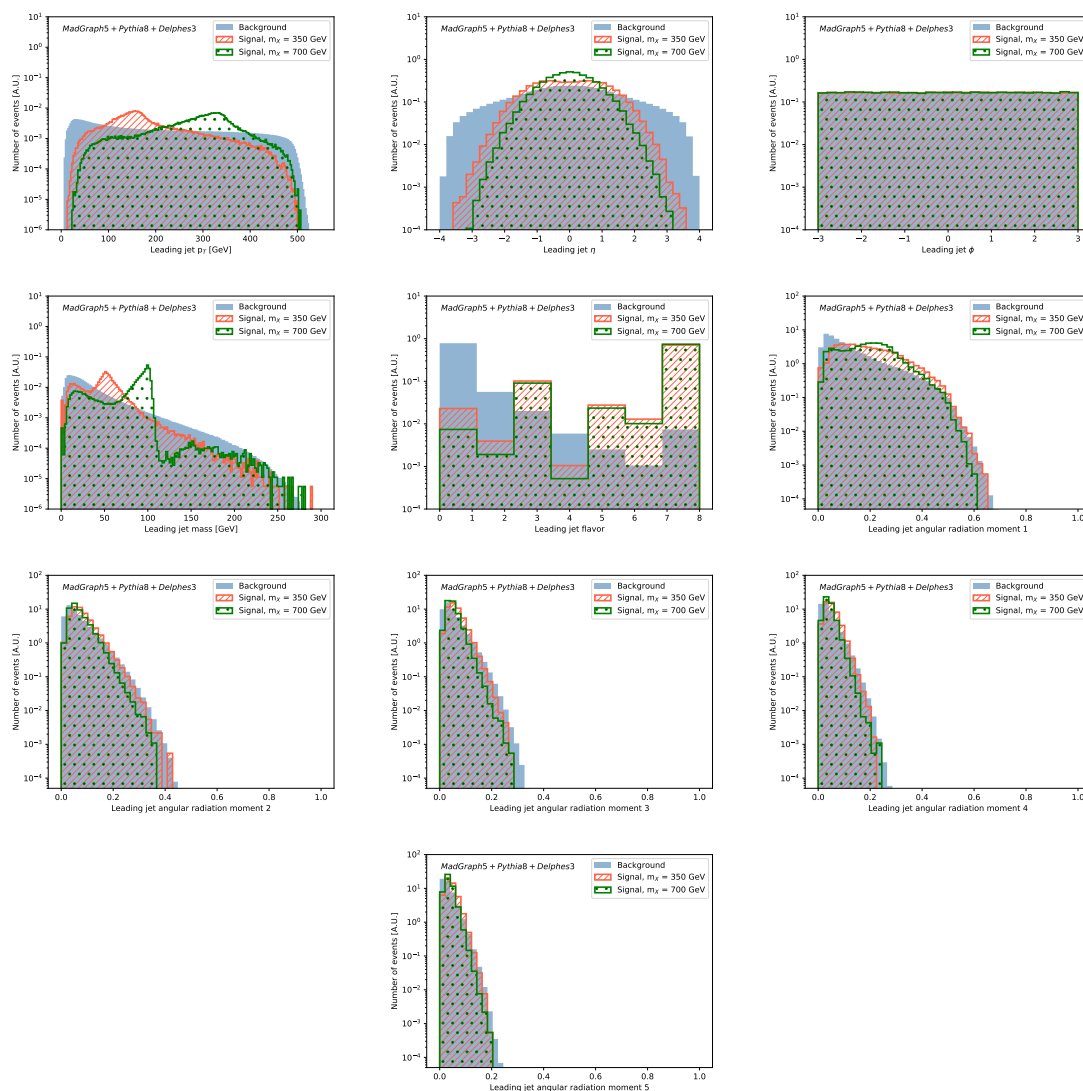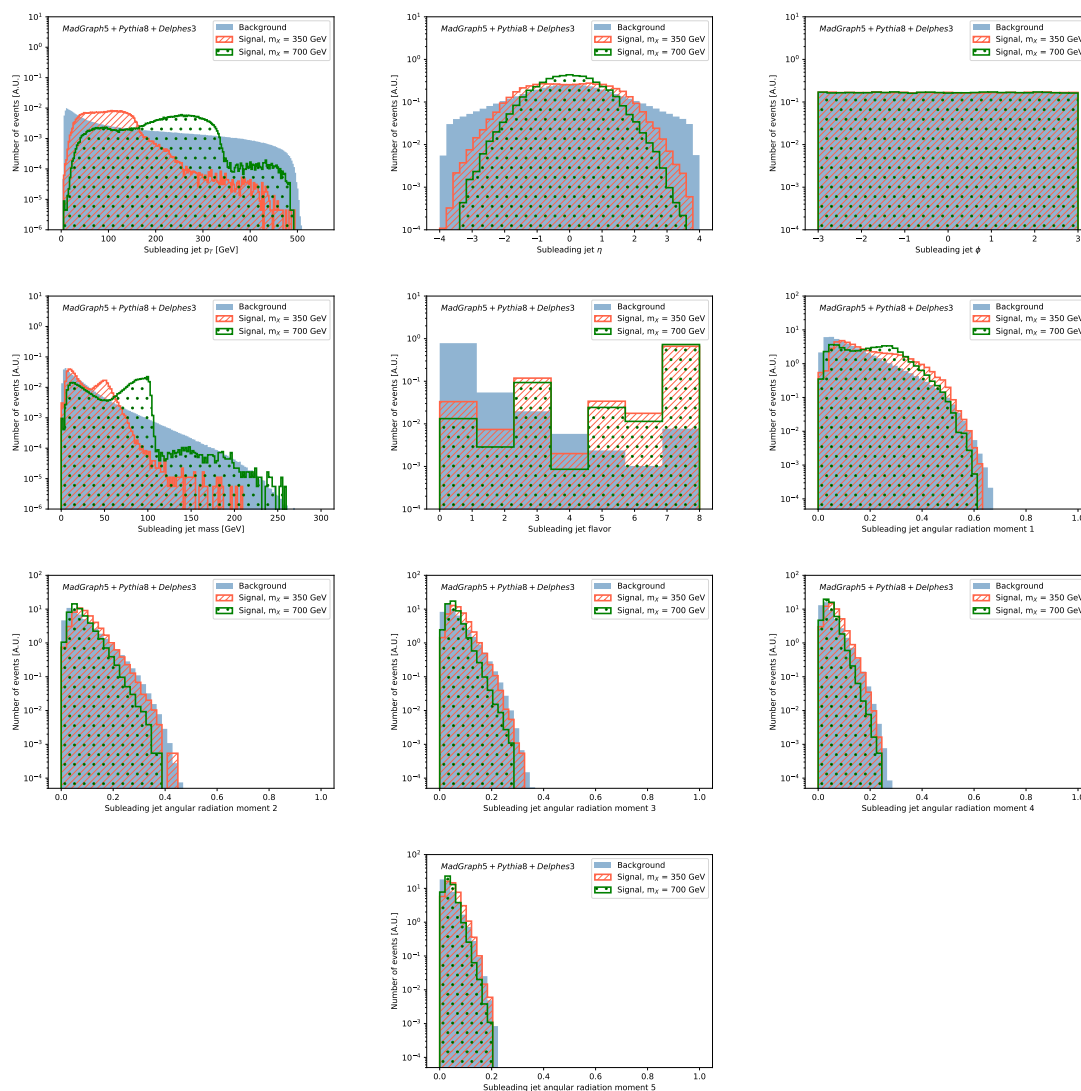
**Figure 11.** Leading jet input PFN variables.

## A   Particle flow network variable distributions

Figures 11 and 12 show example distributions of the PFN variables used in training for both signals and the background. Only the leading and subleading jet inputs are shown, though training is done over the same variables for up to 15 jets per event.
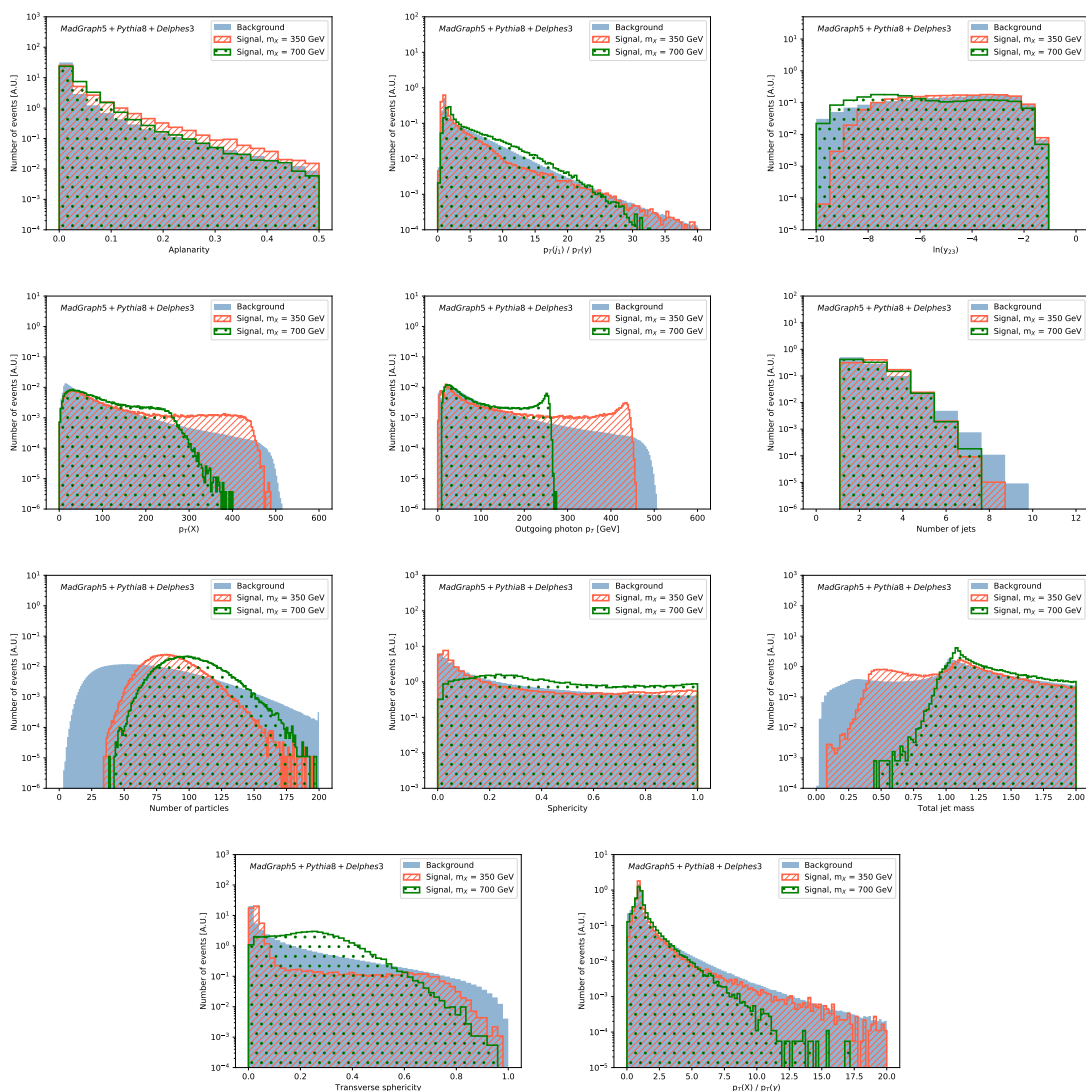
## B   Event level variable distributions

The following 15 variables were used as input to the event-level DNN approach.

- Leading and subleading large-R jet masses

- Leading and subleading large-R jet transverse momenta

**Figure 12.** Subleading jet input PFN variables.

- Leading photon $p_T$

- Measured $X$ $p_T$, defined by the two leading jets

- Number of particles per event

- Number of jets per event

- Ratio of leading jet $p_T$ to leading photon $p_T$

- Ratio of measured $X$ $p_T$ to leading photon $p_T$

- $\ln y_{23}$, calculated with all reconstructed jets in the event [92, 93].

**Figure 13.** Event-level variables used in training, overlaying background and both signal mass hypothesis distributions.

- Aplanarity, sphericity, and transverse sphericity [94, 95]

- Total jet mass [96].

Figure 13 show the distributions of the event-level variables used in training for the background, $m_X = 350\,\mathrm{GeV}$ signal, and $m_X = 700\,\mathrm{GeV}$ signal. The two leading jet masses and $p_T$ distributions are also used in the PFN training, and can be seen in appendix A.

# References

[1] J. Button, G.R. Kalbfleisch, G.R. Lynch, B.C. Maglić, A.H. Rosenfeld and M.L. Stevenson, *Pion-pion interaction in the reaction $\bar{p} + p \to 2\pi^+ + 2\pi^- + n\pi^0$*, *Phys. Rev.* **126** (1962) 1858 [INSPIRE].

[2] 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].

[3] 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].

[4] R.T. D'Agnolo and A. Wulzer, *Learning new physics from a machine*, *Phys. Rev. D* **99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].

[5] 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].

[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] R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer and M. Zanetti, *Learning multivariate new physics*, *Eur. Phys. J. C* **81** (2021) 89 [arXiv:1912.12155] [INSPIRE].

[8] M. Farina, Y. Nakai and D. Shih, *Searching for new physics with deep autoencoders*, *Phys. Rev. D* **101** (2020) 075021 [arXiv:1808.08992] [INSPIRE].

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

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

[11] 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] [INSPIRE].

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

[13] 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].

[14] A. De Simone and T. Jacques, *Guiding new physics searches with unsupervised learning*, *Eur. Phys. J. C* **79** (2019) 289 [arXiv:1807.06038] [INSPIRE].

[15] A. Casa and G. Menardi, *Nonparametric semisupervised classification for signal detection in high energy physics*, arXiv:1809.02977 [INSPIRE].

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

[17] A. Andreassen, B. Nachman and D. Shih, *Simulation assisted likelihood-free anomaly detection*, *Phys. Rev. D* **101** (2020) 095004 [arXiv:2001.05001] [INSPIRE].

[18] B. Nachman and D. Shih, *Anomaly detection with density estimation*, *Phys. Rev. D* **101** (2020) 075042 [arXiv:2001.04990] [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, N.F. Castro, J.G. Milhano, R. Pedro and T. Vale, *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. Knapp, O. Cerri, G. Dissertori, T.Q. Nguyen, M. Pierini and J.-R. Vlimant, *Adversarially learned anomaly detection on CMS open data: re-discovering the top quark*, *Eur. Phys. J. Plus* **136** (2021) 236 [arXiv:2005.01598] [inSPIRE].

[23] 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].

[24] B.M. Dillon, D.A. Faroughy, J.F. Kamenik and M. Szewc, *Learning the latent structure of collider events*, *JHEP* **10** (2020) 206 [arXiv:2005.12319] [inSPIRE].

[25] M. Crispim Romão, N.F. Castro and R. Pedro, *Finding new physics without learning about it: anomaly detection as a tool for searches at colliders*, *Eur. Phys. J. C* **81** (2021) 27 [*Erratum ibid.* **81** (2021) 1020] [arXiv:2006.05432] [inSPIRE].

[26] 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].

[27] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette and T. Golling, *Variational autoencoders for anomalous jet tagging*, arXiv:2007.01850 [inSPIRE].

[28] C.K. Khosa and V. Sanz, *Anomaly awareness*, arXiv:2007.14462 [inSPIRE].

[29] J.A. Aguilar-Saavedra, F.R. Joaquim and J.F. Seabra, *Mass Unspecific Supervised Tagging (MUST) for boosted jets*, *JHEP* **03** (2021) 012 [*Erratum ibid.* **04** (2021) 133] [arXiv:2008.12792] [inSPIRE].

[30] K. Benkendorfer, L.L. Pottier and B. Nachman, *Simulation-assisted decorrelation for resonant anomaly detection*, *Phys. Rev. D* **104** (2021) 035003 [arXiv:2009.02205] [inSPIRE].

[31] A.A. Pol, V. Berger, G. Cerminara, C. Germain and M. Pierini, *Anomaly detection with conditional variational autoencoders*, in *Eighteenth international conference on machine learning and applications*, (2020) [arXiv:2010.05531] [inSPIRE].

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

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

[34] S.E. Park, D. Rankin, S.-M. Udrescu, M. Yunus and P. Harris, *Quasi anomalous knowledge: searching for new physics with embedded knowledge*, *JHEP* **06** (2021) 030 [arXiv:2011.03550] [inSPIRE].

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

[36] G. Stein, U. Seljak and B. Dai, *Unsupervised in-distribution anomaly detection of new physics through conditional density estimation*, in 34th *conference on neural information processing systems*, (2020) [arXiv:2012.11638] [InSPIRE].

[37] 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].

[38] A. Blance and M. Spannowsky, *Unsupervised event classification with graphs on classical and photonic quantum computers*, *JHEP* **08** (2021) 170 [arXiv:2103.03897] [InSPIRE].

[39] B. Bortolato, B.M. Dillon, J.F. Kamenik and A. Smolkovič, *Bump hunting in latent space*, arXiv:2103.06595 [InSPIRE].

[40] 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].

[41] 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].

[42] T. Finke, M. Krämer, A. Morandini, A. Mück and I. Oleksiyuk, *Autoencoders for unsupervised anomaly detection in high energy physics*, *JHEP* **06** (2021) 161 [arXiv:2104.09051] [InSPIRE].

[43] O. Atkinson, A. Bhardwaj, C. Englert, V.S. Ngairangbam and M. Spannowsky, *Anomaly detection with convolutional graph neural networks*, *JHEP* **08** (2021) 080 [arXiv:2105.07988] [InSPIRE].

[44] A. Kahn, J. Gonski, I. Ochoa, D. Williams and G. Brooijmans, *Anomalous jet identification via sequence modeling*, 2021 *JINST* **16** P08012 [arXiv:2105.09274] [InSPIRE].

[45] 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].

[46] T. Dorigo, M. Fumanelli, C. Maccani, M. Mojsovska, G.C. Strong and B. Scarpa, *RanBox: anomaly detection in the Copula space*, arXiv:2106.05747 [InSPIRE].

[47] 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].

[48] E. Govorkova, E. Puljak, T. Aarrestad, M. Pierini, K.A. Woźniak and J. Ngadiuba, *LHC physics dataset for unsupervised new physics detection at* 40 *MHz*, arXiv:2107.02157 [InSPIRE].

[49] G. Kasieczka, B. Nachman and D. Shih, *New methods and datasets for group anomaly detection from fundamental physics*, in *Conference on knowledge discovery and data mining*, (2021) [arXiv:2107.02821] [InSPIRE].

[50] M. Feickert and B. Nachman, *A living review of machine learning for particle physics*, arXiv:2102.02770 [InSPIRE].

[51] 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].

[52] FCC collaboration, *FCC physics opportunities: Future Circular Collider conceptual design report volume* 1, *Eur. Phys. J. C* **79** (2019) 474 [ɪɴSPIRE].

[53] FCC collaboration, *FCC-ee: the lepton collider. Future Circular Collider conceptual design report volume* 2, *Eur. Phys. J. ST* **228** (2019) 261 [ɪɴSPIRE].

[54] FCC collaboration, *FCC-hh: the hadron collider. Future Circular Collider conceptual design report volume* 3, *Eur. Phys. J. ST* **228** (2019) 755 [ɪɴSPIRE].

[55] T. Behnke et al. eds., *The International Linear Collider technical design report — volume* 1: *executive summary*, arXiv:1306.6327 [ɪɴSPIRE].

[56] H. Baer et al. eds., *The International Linear Collider technical design report — volume* 2: *physics*, arXiv:1306.6352 [ɪɴSPIRE].

[57] C. Adolphsen et al. eds., *The International Linear Collider technical design report — volume* 3.*I: accelerator* & *in the technical design phase*, arXiv:1306.6353 [ɪɴSPIRE].

[58] C. Adolphsen et al. eds., *The International Linear Collider technical design report — volume* 3.*II: accelerator baseline design*, arXiv:1306.6328 [ɪɴSPIRE].

[59] H. Abramowicz et al., *The International Linear Collider technical design report — volume* 4: *detectors*, arXiv:1306.6329 [ɪɴSPIRE].

[60] Iɴᴛᴇʀɴᴀᴛɪᴏɴᴀʟ Lɪɴᴇᴀʀ Cᴏʟʟɪᴅᴇʀ Iɴᴛᴇʀɴᴀᴛɪᴏɴᴀʟ Dᴇᴠᴇʟᴏᴘᴍᴇɴᴛ Tᴇᴀᴍ collaboration, *Proposal for the ILC preparatory laboratory (pre-lab)*, arXiv:2106.00602 [ɪɴSPIRE].

[61] CEPC Sᴛᴜᴅʏ Gʀᴏᴜᴘ collaboration, *CEPC conceptual design report: volume* 1 — *accelerator*, arXiv:1809.00285 [ɪɴSPIRE].

[62] CEPC Sᴛᴜᴅʏ Gʀᴏᴜᴘ collaboration, *CEPC conceptual design report: volume* 2 — *physics* & *detector*, arXiv:1811.10545 [ɪɴSPIRE].

[63] L. Linssen, A. Miyamoto, M. Stanitzki and H. Weerts eds., *Physics and detectors at CLIC: CLIC conceptual design report*, CERN yellow report CERN-2012-003, CERN, Geneva, Switzerland (2012) [arXiv:1202.5940] [ɪɴSPIRE].

[64] CLICᴅᴘ and CLIC collaborations, *The Compact Linear Collider (CLIC)* — 2018 *summary report*, arXiv:1812.06018 [ɪɴSPIRE].

[65] P.T. Komiske, E.M. Metodiev, B. Nachman and M.D. Schwartz, *Learning to classify from impure samples with high-dimensional data*, *Phys. Rev. D* **98** (2018) 011502 [arXiv:1801.10158] [ɪɴSPIRE].

[66] A. Denig, *The radiative return: a review of experimental results*, *Nucl. Phys. B Proc. Suppl.* **162** (2006) 81 [hep-ex/0611024] [ɪɴSPIRE].

[67] W. Kluge, *Initial state radiation: a success story*, *Nucl. Phys. B Proc. Suppl.* **181-182** (2008) 280 [arXiv:0805.4708] [ɪɴSPIRE].

[68] H. Czyz, A. Grzelinska and J.H. Kühn, *Narrow resonances studies with the radiative return method*, *Phys. Rev. D* **81** (2010) 094014 [arXiv:1002.0279] [ɪɴSPIRE].

[69] V.P. Druzhinin, S.I. Eidelman, S.I. Serednyakov and E.P. Solodov, *Hadron production via $e^+e^-$ collisions with initial state radiation*, *Rev. Mod. Phys.* **83** (2011) 1545 [arXiv:1105.4975] [ɪɴSPIRE].

[70] M. Karliner, M. Low, J.L. Rosner and L.-T. Wang, *Radiative return capabilities of a high-energy, high-luminosity $e^+e^-$ collider*, *Phys. Rev. D* **92** (2015) 035010 [arXiv:1503.07209] [ɪɴSPIRE].

[71] G. Li, Z. Li, Y. Wang and Y. Wang, *Improving the measurement of the Higgs boson-gluon coupling using convolutional neural networks at $e^+e^-$ colliders*, *Phys. Rev. D* **100** (2019) 116013 [arXiv:1901.09391] [InSPIRE].

[72] L. Li, Y.-Y. Li, T. Liu and S.-J. Xu, *Learning physics at future $e^-e^+$ colliders with machine*, *JHEP* **10** (2020) 018 [arXiv:2004.15013] [InSPIRE].

[73] M. Zaheer et al., *Deep sets*, in *Proceedings of the* 31st *international conference on neural information processing systems, NIPS'17, Red Hook, NY, U.S.A.*, Curran Associates Inc., U.S.A. (2017), p. 3394.

[74] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy flow networks: deep sets for particle jets*, *JHEP* **01** (2019) 121 [arXiv:1810.05165] [InSPIRE].

[75] J. Alwall et al., *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] [InSPIRE].

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

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

[78] ILD CONCEPT GROUP collaboration, *International Large Detector: interim design report*, arXiv:2003.01116 [InSPIRE].

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

[80] M. Cacciari and G.P. Salam, *Dispelling the $N^3$ myth for the $k_t$ jet-finder*, *Phys. Lett. B* **641** (2006) 57 [hep-ph/0512210] [InSPIRE].

[81] M. Cacciari, G.P. Salam and G. Soyez, *FastJet user manual*, *Eur. Phys. J. C* **72** (2012) 1896 [arXiv:1111.6097] [InSPIRE].

[82] T. Barklow et al., *ILC operating scenarios*, arXiv:1506.07830 [InSPIRE].

[83] 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] [InSPIRE].

[84] J. Thaler and K. Van Tilburg, *Identifying boosted objects with $N$-subjettiness*, *JHEP* **03** (2011) 015 [arXiv:1011.2268] [InSPIRE].

[85] J. Thaler and K. Van Tilburg, *Maximizing boosted top identification by minimizing $N$-subjettiness*, *JHEP* **02** (2012) 093 [arXiv:1108.2701] [InSPIRE].

[86] F. Chollet et al., *Keras*, https://github.com/fchollet/keras, (2015).

[87] M. Abadi et al., *TensorFlow: large-scale machine learning on heterogeneous distributed systems*, arXiv:1603.04467 [InSPIRE].

[88] *TensorFlow webpage*, https://www.tensorflow.org/.

[89] D.P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, 12, 2014 [arXiv:1412.6980] [InSPIRE].

[90] J. Duchi, E. Hazan and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, *J. Mach. Learn. Res.* **12** (2011) 2121.

[91] G. Hinton, N. Srivastava and K. Swersky, *Neural networks for machine learning: lecture* 6a, http://www.cs.toronto.edu/∼tijmen/csc321/slides/lecture_slides_lec6.pdf.

[92] A. Banfi, G.P. Salam and G. Zanderighi, *Phenomenology of event shapes at hadron colliders*, *JHEP* **06** (2010) 038 [arXiv:1001.4082] [INSPIRE].

[93] ALEPH collaboration, *Studies of quantum chromodynamics with the ALEPH detector*, *Phys. Rept.* **294** (1998) 1 [INSPIRE].

[94] OPAL collaboration, *QCD studies with $e^+e^-$ annihilation data at* 130 *GeV and* 136 *GeV*, *Z. Phys. C* **72** (1996) 191 [INSPIRE].

[95] ATLAS collaboration, *Measurement of event shapes at large momentum transfer with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV*, *Eur. Phys. J. C* **72** (2012) 2211 [arXiv:1206.2135] [INSPIRE].

[96] CMS collaboration, *Event shape variables measured using multijet final states in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **12** (2018) 117 [arXiv:1811.00588] [INSPIRE].