Learning to isolate muons in data

Edmund Witkowski,[1,*] Benjamin Nachman,[2,3,†] and Daniel Whiteson[1,‡]

[1]*Department of Physics and Astronomy, University of California, Irvine, California 92697, USA*
[2]*Lawrence Berkeley National Laboratory, Physics Division, Berkeley, California 94720, USA*
[3]*Berkeley Institute for Data Science, University of California, Berkeley, California 94720, USA*

(Received 3 August 2023; accepted 20 October 2023; published 17 November 2023)

We use unlabeled collision data and weakly supervised learning to train models that can distinguish prompt muons from nonprompt muons using patterns of low-level particle activity in the vicinity of the muon and interpret the models in the space of energy flow polynomials. Particle activity associated with muons is a valuable tool for identifying prompt muons, those due to heavy boson decay, from muons produced in the decay of heavy flavor jets. The high-dimensional information is typically reduced to a single scalar quantity, isolation, but previous work in simulated samples suggests that valuable discriminating information is lost in this reduction. We extend these studies in LHC collisions recorded by the CMS experiment, where true class labels are not available, requiring the use of the invariant mass spectrum to obtain macroscopic sample information. This allows us to employ classification without labels, a weakly supervised learning technique, to train models. Our results confirm that isolation does not describe events as well as the full low-level calorimeter information, and we are able to identify single energy flow polynomials capable of closing the performance gap. These polynomials are not the same ones derived from simulation, highlighting the importance of training directly on data.

DOI: 10.1103/PhysRevD.108.092008

I. INTRODUCTION

Data collected in hadronic collisions offer a significant opportunity to precisely test the Standard Model (SM) and to search for physics beyond the SM. The identification of muons resulting from electroweak boson decays (called "prompt") is a crucial part of many such studies, as muons are typically well measured and have low rates of background. An important source of background for these events comes from muons produced within jets from decays in flight. This "nonprompt" background is largest at the lower end of the muon transverse momentum spectrum, which has become important in searches for supersymmetry [1–6] as well as for low-mass resonances [7–10].

Prompt muons tend to have less nearby detector activity as compared to muons from jets, which are found near hadrons from the rest of the jet. The concept of "isolation" is therefore important to much of the work involving the discrimination of prompt muons from the nonprompt backgrounds. A complete description of the isolation requires capturing the high-dimensional data in the vicinity of the muon. In practice, high-dimensional data are challenging to analyze and isolation is typically reduced to a scalar quantity [11,12]. However, in the reduction from a high-dimensional (low-level) representation of the data to a lower-dimensional (high-level) one, information can be lost.

Deep learning with low-level inputs has been demonstrated to exceed the performance of engineered high-level observables on a number of tasks in high energy physics, starting with Refs. [13,14] and now including many studies [15]. In the context of prompt muon identification, deep neural networks were able to outperform classical isolation definitions using simulated data—by as much as 50% in nonprompt background rejection at a prompt muon efficiency of 50% [16]. This was achieved by processing all of the calorimeter cells[1] in the vicinity of the muon, corresponding to roughly 1000 dimensions per event. Significant suppression of nonprompt backgrounds with a deep learning approach has the potential to improve the precision and sensitivity of many measurements and searches involving muons at the Large Hadron Collider (LHC).

[*] witkowse@uci.edu
[†] bpnachman@lbl.gov
[‡] daniel@uci.edu

[1] The previous work mentioned here only used calorimeter information, though this study considers both calorimeter and track information.

However, previous studies were based on simulations, with relatively simple detector effects. Hadronic final states are complex and difficult to model, so it is reasonable to be concerned that the performance of a deep-learning-based isolation strategy trained on simulated events may depend on details of the simulation that are not faithful reproductions of collider data. Scale factors derived using standard tag-and-probe methods [17,18] may correct the efficiency, but the performance in data would be suboptimal [19]. Achieving optimal performance in data requires training with data. The limitation is that data are not labeled as prompt or not prompt, so the *supervised* machine learning strategies used in previous studies and which require such labels cannot be applied to data.

We propose to overcome this limitation with *weakly* supervised learning. In contrast to supervised learning, where every event is labeled with certainty as prompt or nonprompt, weakly supervised learning is trained with noisy labels, which describe the overall composition of the sample but not individual events. Specifically, we use the classification without labels (CWoLa) [20] approach to weak supervision where two samples of training events are prepared. One sample is dominated by prompt muons and receives the noisy label of "signal" (and will be called "prompt abundant"); the second sample, while still mostly containing prompt muons, has a relatively higher fraction of nonprompt muons and receives the noisy label of "background" (and will be called "prompt moderate"). Under mild assumptions, training a standard classifier with these noisy labels converges to the same classifier found in a supervised setting. While weak supervision has been used previously for data analysis [21–25], these studies only used 2–18 inputs. Our goal is to approach the muon isolation problem with weak supervision directly on low-level, high-dimensional [$\mathcal{O}(100)$] inputs. While the inputs are high dimensional enough to hold a large number of detected objects, this is only necessary for a small number of events, as on average the inputs have $\mathcal{O}(10)$ nonzero entries.

Even if proven effective in data, deep networks operating on low-level observables can be opaque. To improve the interpretability and compactness of the network, we follow Ref. [16], bridging the performance gap between the low-level observables and classical isolation variables through a small set of additional high-level observables identified by the decisions of a network operating at the low level. We search for new high-level observables among the energy flow polynomials (EFPs) [26], a set of relatively simple combinations of energies and angles of reconstructed objects within the isolation cone. EFP observables are identified automatically using the average decision ordering method [27], which uses the decisions of the low-level network as a guide. While still complex, the resulting EFP is more physically interpretable than the original deep neural network. Interestingly, the first EFP selected through this process was not identified in the previous study as a top candidate for closing the corresponding gap in simulation [16]. This is one more reason why it is essential here to train directly on data. Additionally, it should be noted that, while this study delves into the feasibility of training on data, it does not comprehensively address instrumental effects and systematic uncertainties.

This paper is organized as follows. Section II introduces the dataset, which is from the CMS experiment [28,29]. Then, Sec. III describes the machine learning strategy. Numerical results are presented in Sec. IV. The paper ends with conclusions and outlook in Sec. V.

II. DATASET

Proton-proton collisions at $\sqrt{s} = 8$ TeV were recorded in 2012 and curated by the CMS Collaboration and made available through the CERN Open Data Portal [29]. The number of collisions corresponds to 19.5 fb^{-1}. Reconstruction was performed with the particle flow (PF) algorithm [11], which integrates calorimeter and tracker information to approximate individual particle four-vectors. The PF algorithm also assigns a particle identification (PID) from one of the following types: muon, charged hadron, neutral hadron, photon, or pileup. For the charged PF objects, the sign of the charge is reconstructed. PF object momenta are represented by their transverse momentum (p_T), pseudorapidity (η), and azimuthal angle (ϕ).

We select events with exactly two muons, both with $p_\mathrm{T} \geq 25$ GeV, $|\eta| \leq 2.1$, and with a dimuon invariant mass between 70 and 110 GeV to accommodate the Z boson mass of 90 GeV [30]. Events are separated into two samples that have different mixtures of prompt and nonprompt muon events, as is required by the CWoLa method. One sample, with a higher fraction of nonprompt muons, consists of all events in which the muons have identical electric charge, as well as events with muon pairs of opposite electric charge but reconstructed invariant mass far from the Z boson invariant mass, below 84 GeV or above 96 GeV. This sample is referred to as the prompt muon moderate sample. The remaining events, which are almost entirely prompt muons, form the complementary sample and are referred to as the prompt muon abundant sample. These regions are illustrated in Fig. 1. The opposite sign sample is almost entirely from Z boson decays and so is peaked at the Z boson mass. The same sign sample is mostly from decays in flight and has a nearly smooth and steeply falling spectrum.

In order to ensure that the two samples have similar kinematic distributions, event weights are computed so that the muon p_T and η spectra are the same between the prompt- and nonprompt-enriched samples. The unbinned likelihood ratio is estimated using a two-dimensional kernel density estimator with Gaussian kernels. The preweighted spectra are displayed in Fig. 2. The p_T spectrum is peaked near $m_Z/2$ and the sharp features in the muon histogram are

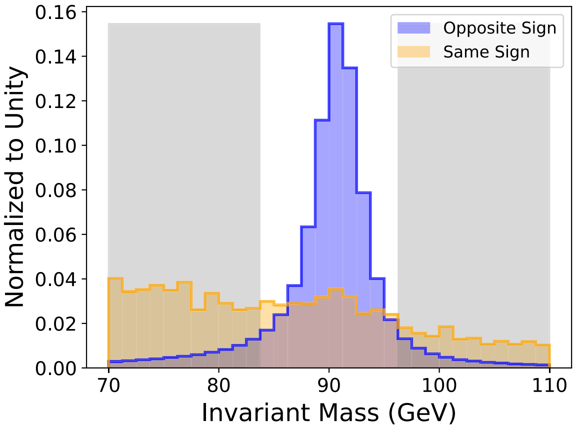

FIG. 1. Histogram of the dimuon invariant mass near the Z boson peak, for events in data with identical (yellow) or opposite (blue) electric charges. The unshaded area indicates the region for the oppositely charged pairs that comprises our prompt muon abundant sample. The gray shaded region for the oppositely charged pairs, as well as the entire region for identically charged pairs, comprise our prompt muon moderate sample.

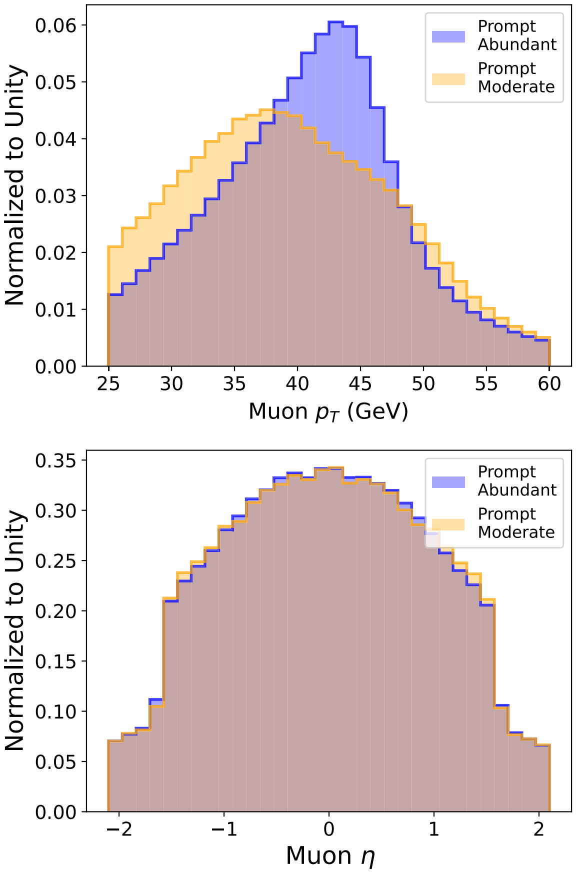

FIG. 2. Histograms of muon p_{T} and pseudorapidity η in the two samples with varying fractions of prompt muons, as defined in text and Fig. 1.

due to detector acceptance effects. We additionally validate the core assumption of CWoLa (see Sec. III)—that the (non)prompt muons look the same in both samples—using samples of simulated muons; see the Appendix.

Once events are selected, they are formatted to be used as inputs to the neural networks. The low-level inputs are composed of the p_{T}, η, ϕ, and PID for each constituent within a 0.45 radius around a given muon. We additionally preprocess the low-level input by centering on the muon and dividing the momenta by the muon transverse momentum. A visualization of the momentum in the vicinity of the muon, not including the muon itself, for both samples is shown in Fig. 3. We see that the sample means per pixel have distinct distributions, with the more prompt sample being more uniform.

Traditional high-level scalar observables are calculated from the low-level data. These observables include the summed p_{T} of nonmuon objects in an event, isolation, and EFP observables. We calculate isolation as defined in

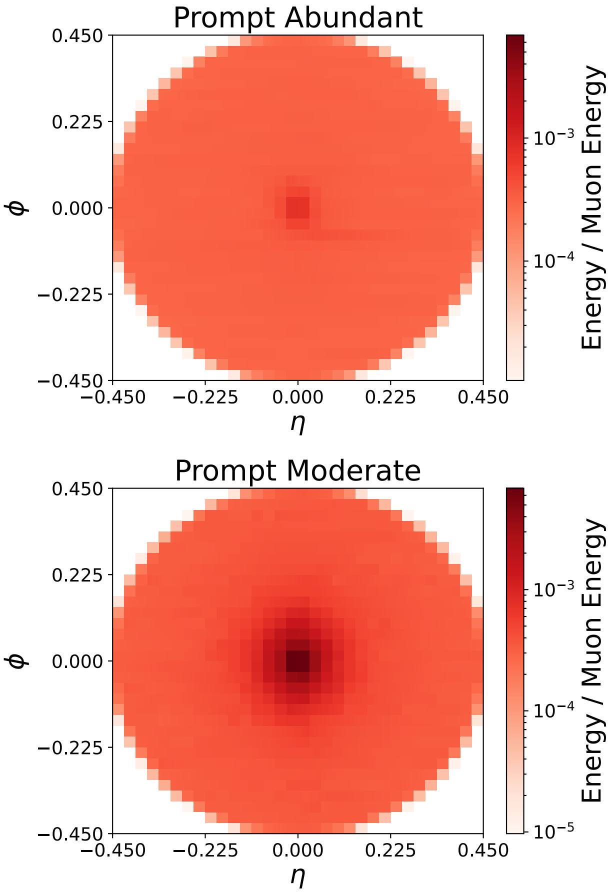

FIG. 3. The average image of hadronic activity in the vicinity of an identified muon, in angular coordinates of azimuthal angle ϕ and pseudorapidity η, for our two training samples, one which is dominated by prompt muons (top) and a second which has a more moderate mixture of prompt and nonprompt muons (bottom). The muon itself is excluded from these visualizations, but the energies are normalized by that of the muon.

Eq. (1), where $h_\pm$ and h_0 denote charged and neutral hadrons, respectively. This definition quantifies the activity around a muon within a given radius strictly in terms of particle flow objects and treats the objects differently according to their particle flow ID. The expression is composed of terms that sum over the transverse momenta of the nonmuon particle flow candidates within the chosen radius, and the result is normalized by the muon momentum. Pileup is mitigated by subtracting half of its sum from the neutral hadron and photon sums, and clamping the result of this subtraction at 0. Distributions of the isolation for two choices of cone radius are shown in Fig. 4. The larger of the two choices of radius tends to yield larger isolation values, as one might expect,

$$I_\mu(R_0) = \left[\sum_{i,R<R_0}^{N_{\mathrm{h}_\pm}} p^{\mathrm{i}}_{\mathrm{T},\mathrm{h}_\pm} + \max\left(0, \sum_{i,R<R_0}^{N_{\mathrm{h}_0}} p^{\mathrm{i}}_{\mathrm{T},\mathrm{h}_0} + \sum_{i,R<R_0}^{N_\gamma} p^{\mathrm{i}}_{\mathrm{T},\gamma} - \frac{1}{2}\sum_{i,R<R_0}^{N_{\mathrm{pileup}}} p^{\mathrm{i}}_{\mathrm{T},\mathrm{pileup}}\right)\right] / p_{\mathrm{T},\mathrm{muon}}. \qquad (1)$$

We calculate isolation quantities for a set of radii from 0.025 to 0.45 in steps of 0.025. CMS has previously studied isolation at radius of 0.3 [11], which is included in our generated set.

While, in principle, the demonstration of weak supervision as a technique for learning to improve muon isolation beyond cone-based quantities could use simulation instead of data, we have chosen to use collider data for a number of reasons. First, realistic simulation of muon isolation is very challenging, for both the prompt and nonprompt categories; see the Appendix. Second, a demonstration in data can confirm (or refute) the results of earlier studies in simulation, which showed a significant gap between the power of isolation cones and full use of the lower-level data. If such a gap exists in collider data, it would indicate that additional information is available in nature; the interpretation of that gap in terms of EFP observables will provide clues as to the physical processes involved, and the size of the gap can motivate a further study in a complete experimental context. For this reason, we also do not estimate systematic uncertainties, which would be required before application to searches and measurements. As a data-driven method, there are no simulation-based uncertainties, but there would be method closure uncertainties related to the underlying assumptions of CWoLa and sPlots.

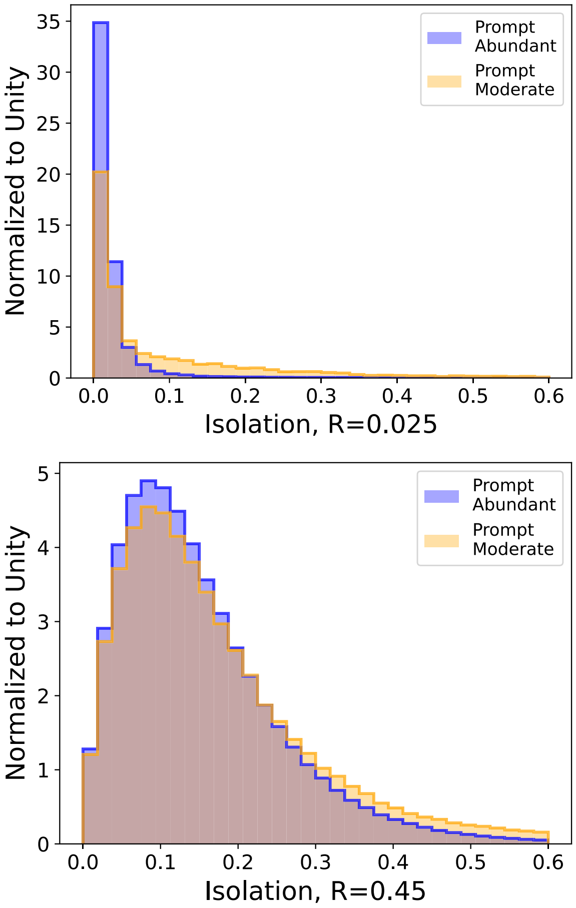

FIG. 4. Histograms of the muon isolation [defined in Eq. (1)] for each of our training samples, one of which is dominated by prompt muons, for two choices of isolation cone radius parameter $R_0 = 0.025$ (top) and $R_0 = 0.45$ (bottom).

III. METHODS

CWoLa defines a weakly supervised setting that relies on the principle that, given two classes, an optimal classifier may be obtained by training to discriminate between two samples composed of different mixtures of the classes, rather than training directly on two pure class samples. This technique only requires that the two samples have different class mixtures, and these mixtures do not need to be known in order for training to proceed. The essential assumption is that class fraction is the only feature that determines the different properties of the two samples. This means that the spectrum of radiation around the muon for prompt leptons is identical for the prompt muon abundant and the prompt muon moderate samples. Similarly, the probability density for hadrons around the muon for nonprompt leptons should be the same within the prompt muon abundant and the prompt muon moderate samples. We expect this to be the case here, since the invariant mass of the muons and their relative electric charges should be statistically independent from the radiation pattern around the muons given the

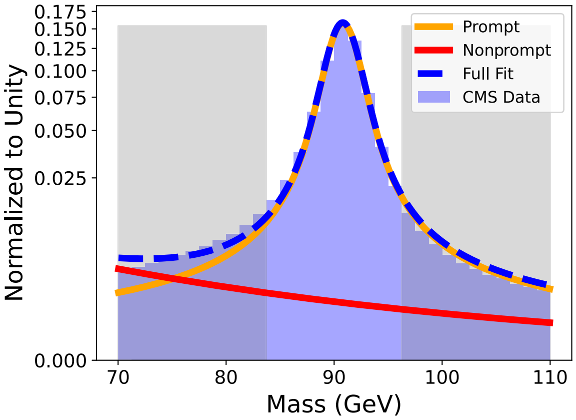

FIG. 5. A visualization of the masses overlaid with the fit and its prompt/nonprompt components. The shaded regions indicate events that are included in the relatively less prompt sample. Here we fit the full CMS sample used in the study, finding that it is 95.6 ± 0.6% prompt overall.

prompt status. This expectation is validated in simulation in the Appendix.

While CWoLa does not need class labels to derive a classifier, some class information is required to determine the performance of the method. The only information needed is the proportion of prompt muons in each sample; from this information, it is possible to characterize the full trade-off between signal efficiency and background rejection. The prompt muon fraction is measured directly from the data in each sample by modeling the invariant mass distribution as a mixture model with two components: one peaking component of Z bosons, which decay to two prompt muons, and a second, nonpeaking component. The invariant mass spectrum is fit using a Voigt profile and an exponential function for the respective components. Fitting is done with SciPy v1.7.3 [31] and visually demonstrated in Fig. 5, where the fit is applied to the full dataset, finding an overall prompt fraction of $95.6 \pm 0.6\%$, where the error bar corresponds to 1σ statistical. In the prompt muon abundant sample, the prompt fraction is measured to be 98.9%; in the prompt muon moderate sample, the prompt fraction is measured to be 56.0%. This is the first application of weak supervision in particle physics where the relative proportions have also been extracted directly from data.

Characterizing the network performance is nontrivial without pure samples. To measure the efficiency of a varying network threshold in the prompt and nonprompt samples, one could fit the distribution of the invariant mass of events surpassing each threshold. Measurement of the efficiencies of each class allows calculation of performance metrics, such as the standard receiver operating characteristic (ROC) and its associated statistics. However, fits are expensive and stochastic. Fitting the mass spectrum for each threshold output can be avoided using the sPlots technique [32], which can decompose the prompt and nonprompt contributions to distributions of the network output given weights from the single invariant mass fit into the full sample. sPlots assumes that the variable being weighted is statistically independent of the invariant mass, within the individual classes. The correlation coefficient may be evaluated to assess how well this assumption holds. While we lack a pure background sample, the coefficient for the signal case may be well estimated using the nearly pure prompt muon abundant sample. Evaluating the average of this coefficient between the outputs of the networks trained and the invariant masses yields a small value of 0.0664, suggesting that the method may be applied. Once the variable has been separated by the components, the resulting histograms may be integrated to calculate true and false positive rates and construct a ROC curve. Performance is evaluated through the area under the curve (AUC) and the signal efficiency at 50% background efficiency. While we do not perform a full determination of the uncertainty, we do consider statistical sources of uncertainty from the training and from the fit.[2] While not an uncertainty *per se* [33], the statistical variation from the finite size of the training dataset[3] gives a sense for the stability and optimality of the result. This effect is estimated using bootstrapping [34] with 100 event ensembles with a new classifier trained per ensemble. Additionally, we propagate the statistical uncertainty from the fit in each ensemble by sampling 100 times from the fitted parameter covariance matrix. Metrics are recomputed and averaged across each ensemble, and we report the 1σ confidence intervals according to the resulting set of values.

We consider two types of neural networks: high-level networks with an increasing list of engineered observables (such as isolation) and low-level networks that process the full muon image. For the high-level networks, one of our goals is to determine the minimal set of isolation observables that will saturate the performance. To do this, we start by training a network using the single isolation cone corresponding to the largest radius in our set and subsequently train networks with incrementally smaller cones included as inputs. The summed event p_T is included as an input in all of these sets, in order to be sensitive to overall normalization effects.

The low-level networks take the full high-dimensional representations of the events as inputs. We use the deep set architecture [35] implemented as particle flow networks (PFNs) [36] to process these data. This architecture was chosen because the inputs are a permutation invariant, variable-length set of four-vectors and so a point-cloud model is the natural choice for processing them. Deep set models are composed of two fully connected networks. The first network embeds each particle flow object [represented by $(p_T, \eta, \phi, \mathrm{PID})$] into a latent space. The second network

[2] While these are the only sources of uncertainty quantified in Table I, other sources are present, such as a bias due to imperfect description of the mass distribution by the fit function.

[3] The random initialization of the network is also folded into this estimation.

processes the sum of these latent space vectors across all inputs. The sum operation is permutation invariant and can readily process variable-length inputs.

Additionally, we strive to close the gap in performance between low- and high-level networks using relatively simple variables. Energy flow polynomials [26] serve as a set of potential variables for this purpose. EFPs are a set of parametrized functions that sum over objects within an event, were each term is weighted using the angular relations between these objects. EFPs can be represented using graphs, where

$$\text{each node} \Rightarrow \sum_{i=1}^{N} z_i, \tag{2}$$

$$\text{each } k\text{-fold edge} \Rightarrow (\theta_{ij})^k. \tag{3}$$

$$(z_i)^\kappa = \left(\frac{p_{\mathrm{T}i}}{\sum_j p_{\mathrm{T}j}}\right)^\kappa, \tag{4}$$

$$\theta_{ij}^\beta = (\Delta\eta_{ij}^2 + \Delta\phi_{ij}^2)^{\beta/2}. \tag{5}$$

When $\kappa = 1$ the EFPs form a basis for infrared-and-collinear- (IRC) safe observables [26]. We compute a set of EFPs that contains IRC-safe, as well as unsafe, information, using the same parametrizations as in Ref. [16]: $\kappa \in [-1, 0, \frac{1}{4}, \frac{1}{2}, 1, 2]$ and $\beta \in [\frac{1}{4}, \frac{1}{2}, 1, 2, 3, 4]$, for graphs with up to $n = 7$ nodes and up to $d = 7$ edges.

We use the average decision ordering (ADO) [27] metric to determine which EFPs from this generated set might bridge the performance gap to the PFN. ADO compares two classifiers on signal and background input pairs, measuring how often the classifiers rank the inputs in the same way. This is quantified with a Heaviside step function on many different pairs, and the results are averaged to obtain the ADO. The ADO can be interpreted as the probability that a given pair will be ordered in the same way by the two classifiers. This is intuitively similar to the AUC metric, which measures the probability that a given signal example will be ranked higher than a given background example. While AUC can be seen as comparing a classifier to the truth, the ADO compares two classifiers to one another without regard for correct ordering. To avoid training a large set of new high-level networks, one for each EFP being considered as an additional observable, we follow the strategy of Ref. [27] and search for EFPs that have a high ADO with our PFN for the subset of events where the PFN and the high-level network disagree. In general, this process can be iterated, selecting new observables until the ADO no longer improves.

IV. RESULTS

The performance of each network is measured through ROC AUC as well as the signal efficiency at a fixed background efficiency of 50%. Figure 6 illustrates the effects of including additional isolation cones as network input features. Adding cones tends to increase performance up until nine cones are used, after which there is no clear further gain in AUC. There is a significant performance gap between the network that uses nine cones and the PFN, which, respectively, yield AUCs of 0.848(1)[4] and 0.874(1), as well as signal efficiencies of 0.939(1) and 0.957(1). This suggests that isolation cones alone do not capture all discriminating information available in the low-level data. This is consistent with previous results shown on simulation [16], and it is notable that it holds for real collider data. While the improvement over isolation is subtle, any performance gain is valuable given the importance of muons. For example, in multilepton final states, the event efficiency depends on the lepton efficiency raised to the number of leptons and so even a modest change may be significant.

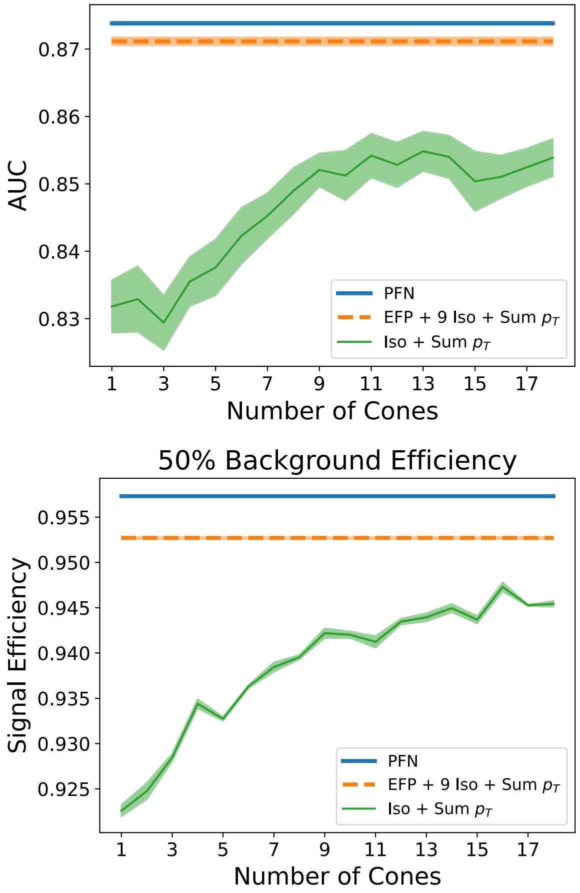

FIG. 6. Isolation network performance shown as a function of number of input cones. Performance of the PFN and best performing high-level network are shown as benchmarks. ROC AUC is shown for each model (top) as well as the signal efficiency at a fixed background efficiency (bottom).

We use the ADO metric to search among the EFP observables for ways to close the gap with the PFN

[4] The reported error values should be understood as rounded to 1×10^{-3} from values calculated to be $\lesssim 1 \times 10^{-3}$.

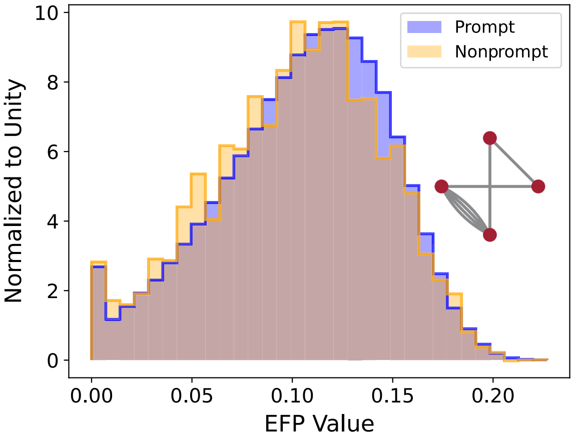

FIG. 7. Distribution of the EFP observable identified in the search described by the text. Samples shown are separated by class using the sPlots weighting technique after applying a 50% background efficiency cut according to the outputs of the 9 isolation cone network. Also shown is the graph representation of the EFP.

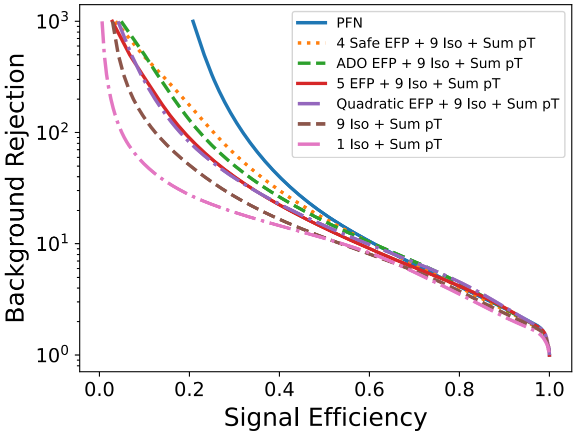

FIG. 8. Comparison of the performance of the networks described in Table I, via ROC curves. Shown is background rejection (inverse of efficiency) versus signal efficiency.

performance. Note that the EFPs lack the built-in radial symmetry of the isolation cones and so may contain additional useful information. The networks using EFP features are also provided the nine largest isolation cones and the summed event p_T. Remarkably, the ADO search method is able to identify a *single* IRC-safe EFP that obtains an AUC of 0.871(1) and signal efficiency of 0.953(1), almost fully closing the gap in AUC to the PFN from 0.026 to 0.003. The graph representation of this EFP, as well as class distributions separated through the sPlots technique, are illustrated in Fig. 7. This EFP corresponds to parameters $\kappa = 1$ and $\beta = 0.25$, and the full expression is provided in Eq. (6),

$$\sum_{a,b,c,d=1}^{N} z_a z_b z_c z_d (\theta_{ab}\theta_{ac}\theta_{bd}\theta_{cd}^4)^{1/4}. \qquad (6)$$

TABLE I. Comparison of the performance of the various networks discussed in the text. Performance is measured through ROC AUC, as well as signal efficiency [the true positive rate (TPR), or proportion of actual positives correctly identified] at 50% background efficiency. Standard error is evaluated to be $\lesssim 1 \times 10^{-3}$ for both metrics over a 1σ confidence interval (see Sec. III for details on calculation). While the reported performance values refer only to testing done on CMS data, the "EFP scan" column indicates whether the EFP inputs used were identified as useful by a scan over CMS or simulated (Sim) data. These results correspond to the ROC curves in Fig. 8.

Input features	AUC	TPR	EFP scan
Single isocone + $\sum p_T$	0.835	0.922	
9 Iso, $\sum p_T$	0.848	0.939	
9 Iso, $\sum p_T$, ADO EFP	0.871	0.953	CMS
9 Iso, $\sum p_T$, Quadratic EFP	0.870	0.956	CMS
9 Iso, $\sum p_T$, 4 IRC-safe EFP	0.868	0.949	Sim
9 Iso, $\sum p_T$, 5 EFP	0.865	0.954	Sim
Full details PFN	0.874	0.957	

An additional scan is done over the quadratic EFPs included in our full set of calculated EFPs, as these are simple in structure and are therefore more interpretable. This identifies another single EFP with $\kappa = 1$ and $\beta = 0.25$ which yields performance close to that of the one identified by the first ADO search, at an AUC of 0.870(1) and signal efficiency of 0.956(1). We further check the performance of sets of EFPs identified as useful by previous work done on simulation [16], which selected an IRC-safe set of EFPs, as well as a set not restricted to be safe. The IRC-safe set yields an AUC of 0.868(1) with a signal efficiency of 0.949(1), while the unsafe set yields an AUC of 0.865(1) with a signal efficiency of 0.954(1). While these sets identified in simulation close much of the performance gap, they require more features and are outperformed by the EFPs identified directly on the CMS data, underscoring the importance of training in data.

A full summary of performance across the methods used is presented in Table I, as well as depicted in Fig. 8. Our results indicate that we are able to construct a minimal set of high-level observables that perform comparably to the low-level inputs, allowing for the use of more physically intuitive features and less complex networks without making concessions regarding performance.

V. CONCLUSIONS

On collision data from the LHC, we apply neural networks to the problem of prompt muon discrimination. We investigate how much information is present in high- and low-level representations of the data, finding that the traditionally used scalar isolation does not capture all useful classification information present at the low level.

Furthermore, we find that another high-level set of observables, the EFPs, may be used to create a network that performs almost as well as one operating at the low level, while providing the advantage of being less complex and more human interpretable. In addition to being notable for using real rather than simulated data, this study demonstrates the use of weakly supervised training methods with CWoLa on low-level features, as well as performance evaluation without having access to individual class labels. Future work may include investigating the interpretation of the observables selected here, exploring how much information might be captured by other types of high-level observables, and the generalizability of these results. While our study indicates that additional information is available beyond the use of simple cones, and the identification of a single EFP observable that captures that information allows for simple application and interpretation, further work would be required before implementation within an experimental context. A robust estimate of the systematic uncertainties involved has not been done, which would be necessary to establish the optimal observables. Our result does not replace work by the experimental collaborations, but motivates further study.

The code for this paper can be found at [37] to isolate muons in data. The datasets will be provided upon reasonable request to the authors.

ACKNOWLEDGMENTS

The authors are grateful to Troels Petersen, David Shih, Gregor Kasieczka, Jesse Thaler, Michael Fenton, and Aishik Ghosh for useful comments and suggestions. B. N. is supported by the U.S. Department of Energy, Office of Science under Award No. DE-AC02-05CH11231. D. W. is supported by the Department of Energy, Office of Science.

APPENDIX: ASSUMPTIONS AND SUPPLEMENTARY FIGURES

CWoLa assumes that the mixed samples are generated in such a way that a given component feature is distributed the same way in one sample as it is in the other. While we cannot explicitly demonstrate this on an unlabeled dataset, we can use a simulated dataset similar to the experimental data to probe whether we can reasonably expect this assumption to hold.

We simulate events where prompt muons are generated by the process $pp \to Z \to \mu^+\mu^-$ and nonprompt muons by $pp \to b\bar{b}$. A center of mass energy of $s = \sqrt(13)$ TeV is used. MadGraph5, PYTHIA, and DELPHES are used, respectively, for collision and heavy boson decay simulation, showering and hadronization, and the detector simulation, with pileup included. In total we generate 22766 events, where half are prompt and the other half are nonprompt events. The muon transverse momentum and pseudorapidity distributions for this dataset are shown in Fig. 9, and the average event images are shown in Fig. 10, where quantities are separated between the prompt and nonprompt distributions.

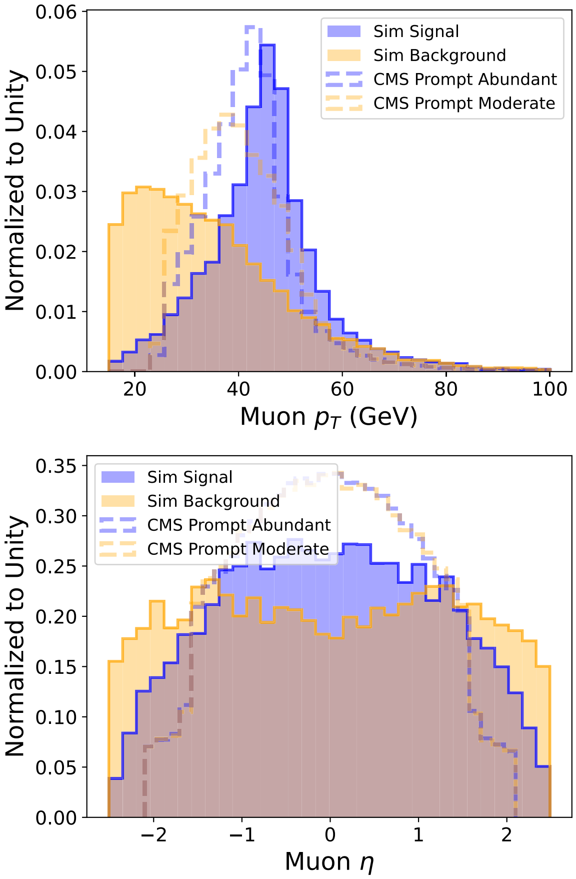

FIG. 9. (MG5+PYTHIA+DELPHES) Distributions of the simulated muon transverse momentum and pseudorapidity.

Using the simulated dataset, we compute one of the features included in our models that use the CMS dataset, the summed transverse momentum of the objects in an event. We see in Fig. 11 that the component distributions do approximately match across the samples for the simulated dataset. Similarly, the class components of a network classifier should be distributed the same way, regardless of which mixed sample the events were drawn from. We check this by training a PFN using the simulated dataset and looking at the distributions of the outputs, as shown in Fig. 12. Once again we see that the distributions depend on the class rather than the mixed sample to which events belong.

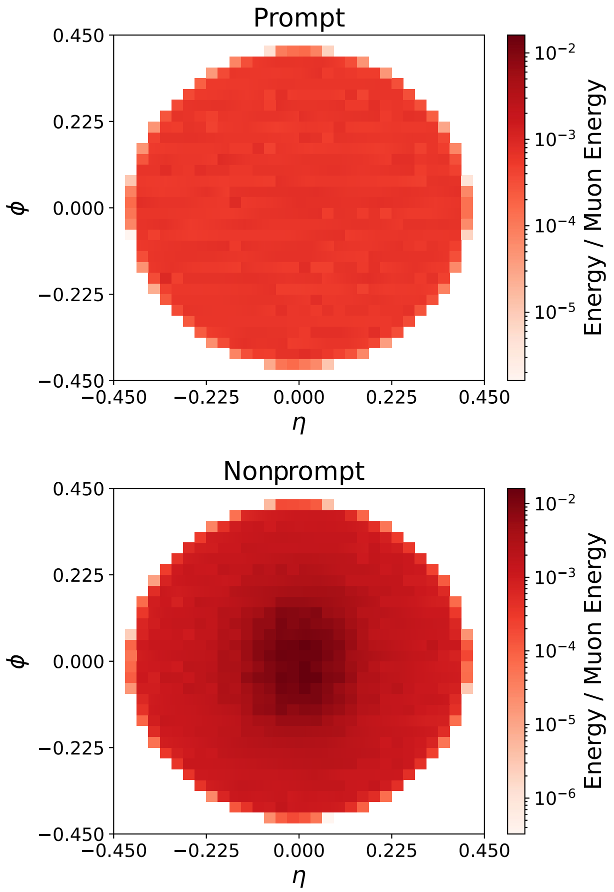

FIG. 10. (MG5+PYTHIA+DELPHES) Average event images similar to Fig. 3, but for the simulated dataset and separated by prompt and nonprompt events.

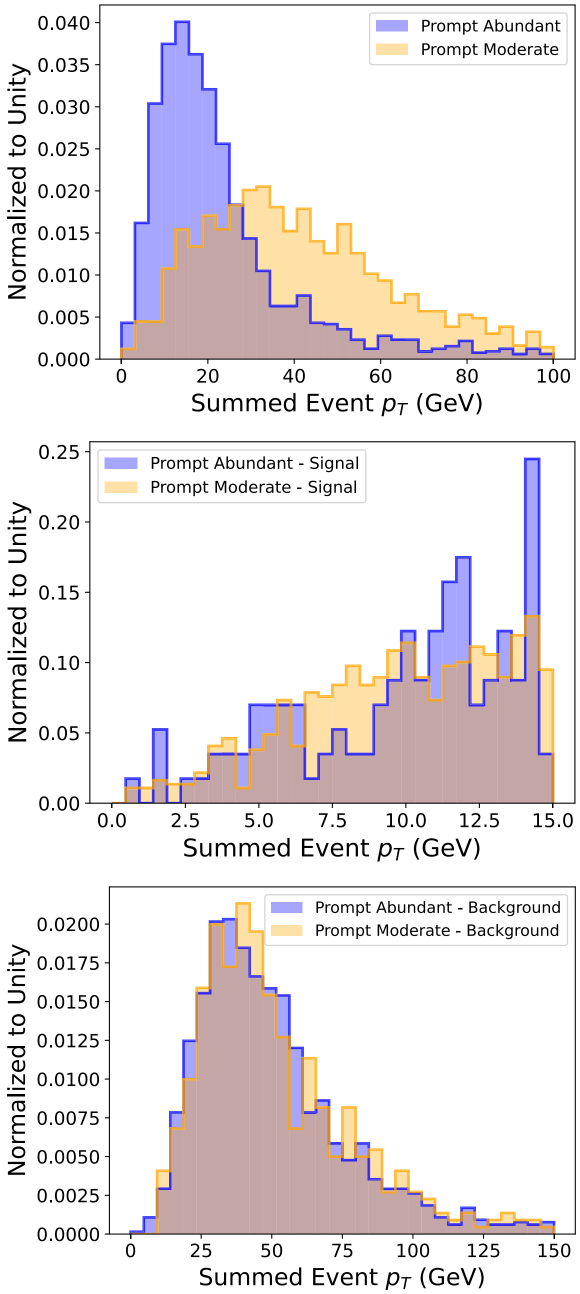

FIG. 11. (MG5+PYTHIA+DELPHES) Top: the total summed event p_T distributions for two simulated mixed samples. Middle: only the signal components of the two simulated mixed samples. Bottom: only the background components of the two simulated mixed samples. We see that, while the class proportions are different, the signal and background distributions are approximately the same across the samples.

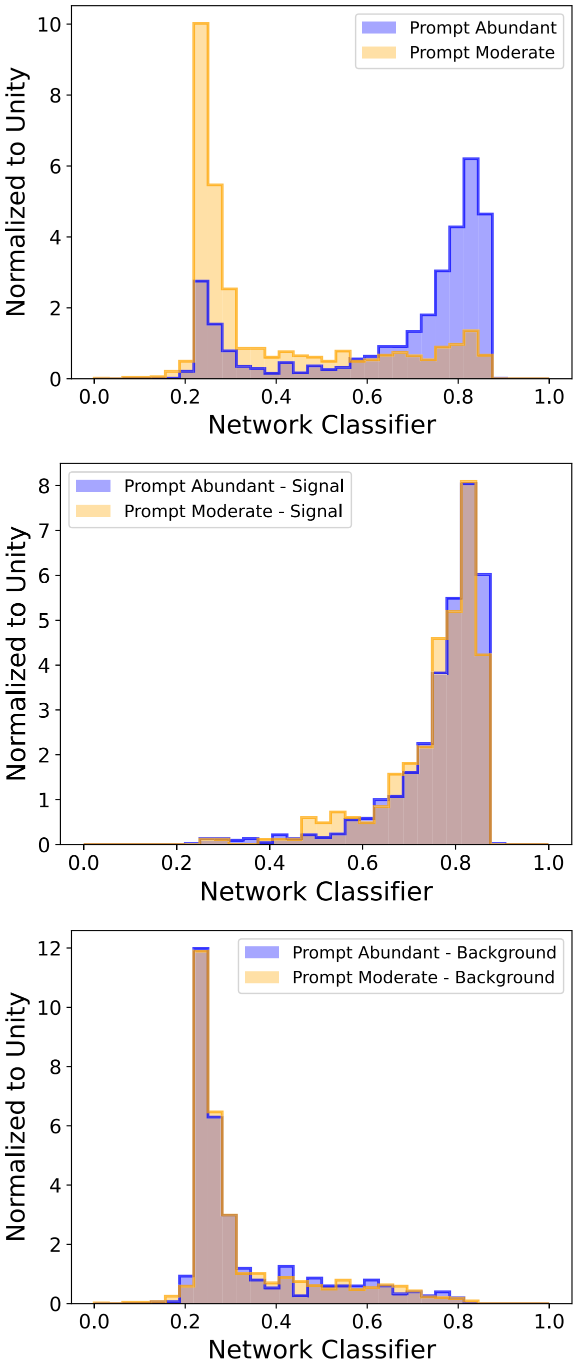

FIG. 12. (MG5+PYTHIA+DELPHES) Similar to Fig. 11, but demonstrating that network output distributions for each class match across mixed samples.

[1] M. Aaboud *et al.* (ATLAS Collaboration), Phys. Rev. D **97**, 052010 (2018).

[2] R. Schoefbeck, Nucl. Part. Phys. Proc. **273–275**, 631 (2016).

[3] V. Khachatryan *et al.* (CMS Collaboration), J. High Energy Phys. 11 (2015) 189.

[4] ATLAS Collaboration, arXiv:2209.13935.

[5] A. Tumasyan *et al.* (CMS Collaboration), J. High Energy Phys. 04 (2022) 091.

[6] G. Aad *et al.* (ATLAS Collaboration), Phys. Rev. D **101**, 052005 (2020).

[7] I. Hoenig, G. Samach, and D. Tucker-Smith, Phys. Rev. D **90**, 075016 (2014).

[8] ATLAS Collaboration, arXiv:2301.09342.

[9] R. Aaij *et al.* (LHCb Collaboration), J. High Energy Phys. 10 (2020) 156.

[10] A. Tumasyan *et al.* (CMS Collaboration), Eur. Phys. J. C **82**, 290 (2022).

[11] A. M. Sirunyan *et al.* (CMS Collaboration), J. Instrum. **12**, P10003 (2017).

[12] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **81**, 578 (2021).

[13] P. Baldi, P. Sadowski, and D. Whiteson, Nat. Commun. **5**, 4308 (2014).

[14] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, J. High Energy Phys. 07 (2016) 069.

[15] M. Feickert and B. Nachman, arXiv:2102.02770.

[16] J. Collado, K. Bauer, E. Witkowski, T. Faucett, D. Whiteson, and P. Baldi, J. High Energy Phys. 10 (2020) 200.

[17] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **76**, 292 (2016).

[18] A. M. Sirunyan *et al.* (CMS Collaboration), J. Instrum. **13**, P06015 (2018).

[19] C. Cesarotti, Y. Soreq, M. J. Strassler, J. Thaler, and W. Xue, Phys. Rev. D **100**, 015021 (2019).

[20] E. M. Metodiev, B. Nachman, and J. Thaler, J. High Energy Phys. 10 (2017) 174.

[21] A. M. Sirunyan *et al.* (CMS Collaboration), Phys. Lett. B **803**, 135285 (2020).

[22] V. M. Mikuni, Collider physics measurements in high jet multiplicity final states (2021), https://cds.cern.ch/record/2781479.

[23] J. H. Collins, K. Howe, and B. Nachman, Phys. Rev. D **99**, 014038 (2019).

[24] J. H. Collins, K. Howe, and B. Nachman, Phys. Rev. Lett. **121**, 241803 (2018).

[25] G. Aad *et al.* (ATLAS Collaboration), Phys. Rev. Lett. **125**, 131801 (2020).

[26] P. T. Komiske, E. M. Metodiev, and J. Thaler, J. High Energy Phys. 04 (2018) 013.

[27] T. Faucett, J. Thaler, and D. Whiteson, Phys. Rev. D **103**, 036020 (2021).

[28] S. Chatrchyan *et al.* (CMS Collaboration), J. Instrum. **3**, S08004 (2008).

[29] CERN Open Data Portal, http://opendata.cern.ch.

[30] R. L. Workman *et al.* (Particle Data Group), Prog. Theor. Exp. Phys. **2022**, 083C01 (2022).

[31] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, Nat. Methods **17**, 261 (2020).

[32] M. Pivk and F. R. Le Diberder, Nucl. Instrum. Methods Phys. Res., Sect. A **555**, 356 (2005).

[33] B. Nachman, SciPost Phys. **8**, 090 (2020).

[34] B. Efron, Ann. Stat. **7**, 1 (1979).

[35] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola, arXiv:1703.06114.

[36] P. T. Komiske, E. M. Metodiev, and J. Thaler, J. High Energy Phys. 01 (2019) 121.

[37] https://github.com/Edwit4/learning_to_isolate_muons_in_data.