Published for SISSA by Springer

Received: *September 11, 2017*
Accepted: *October 17, 2017*
Published: *October 25, 2017*

Classification without labels: learning from mixed samples in high energy physics

Eric M. Metodiev,[a] Benjamin Nachman[b] and Jesse Thaler[a]

[a] *Center for Theoretical Physics, Massachusetts Institute of Technology,*
Cambridge, MA 02139, U.S.A.

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

E-mail: metodiev@mit.edu, bpnachman@lbl.gov, jthaler@mit.edu

Abstract: Modern machine learning techniques can be used to construct powerful models for difficult collider physics problems. In many applications, however, these models are trained on imperfect simulations due to a lack of truth-level information in the data, which risks the model learning artifacts of the simulation. In this paper, we introduce the paradigm of classification without labels (CWoLa) in which a classifier is trained to distinguish statistical mixtures of classes, which are common in collider physics. Crucially, neither individual labels nor class proportions are required, yet we prove that the optimal classifier in the CWoLa paradigm is also the optimal classifier in the traditional fully-supervised case where all label information is available. After demonstrating the power of this method in an analytical toy example, we consider a realistic benchmark for collider physics: distinguishing quark- versus gluon-initiated jets using mixed quark/gluon training samples. More generally, CWoLa can be applied to any classification problem where labels or class proportions are unknown or simulations are unreliable, but statistical mixtures of the classes are available.

Keywords: Jets

ArXiv ePrint: 1708.02949

https://doi.org/10.1007/JHEP10(2017)174

Contents

1 Introduction

In the data-rich environment of the Large Hadron Collider (LHC), machine learning techniques have the potential to significantly improve on many classification, regression, and generation problems in collider physics. There has been a recent surge of interest in applying deep learning and other modern algorithms to a wide variety of problems, such as jet tagging [1–21]. Despite the power of these methods, they all currently rely on significant input from simulations. Existing multivariate approaches for classification used by the LHC experiments all have some degree of mis-modeling by simulations and must be corrected post-hoc using data-driven techniques [22–30]. The existence of these *scale factors* is an indication that the algorithms trained on simulation are sub-optimal when tested on data. Adversarial approaches can be used to mitigate potential mis-modeling effects during training at the cost of algorithmic performance [31]. The only solution that does not compromise performance is to train directly on data. This is often thought to not be possible because data is unlabeled.

In this paper, we introduce *classification without labels* (CWoLa, pronounced “koala”), a paradigm which allows robust classifiers to be trained directly on data in scenarios common in collider physics. Remarkably, the CWoLa method amounts to only a minor variation on well-known machine learning techniques, as one can effectively utilize standard fully-supervised techniques on two mixed samples. As long as the two samples have different compositions of the true classes (even if the label proportions are unknown), we prove that the optimal classifier in the CWoLa framework is the optimal classifier in the fully-supervised case.[1] In practice, after training the classifier on large event samples without using label information, the operating points of the classifier can be determined from a small sample where at least the label proportions are known.

The CWoLa paradigm is part of a broader set of classification frameworks that fall under the umbrella of *weak supervision*. These frameworks go beyond the standard fully-supervised paradigm with the goal of learning from partial, non-standard, or imperfect label information. See ref. [33] for a recent review and comprehensive taxonomy. Weak supervision was first applied in the context of high energy physics in ref. [34] to distinguish jets originating from quarks from those originating from gluons using only class proportions during training; this paradigm is known as *learning from label proportions* (LLP) [35, 36]. For quark versus gluon jet tagging, LLP was an important development because useful quark/gluon discrimination information is often subtle and sensitive to low-energy or wide-angle radiation inside jets, which may not be modeled correctly in parton shower generators [37]. The main drawback of LLP, however, is that there is still uncertainty in the quark/gluon labels themselves, since quark/gluon fractions are determined by matrix element calculations convolved with parton distribution functions, which carry their own uncertainties. The CWoLa paradigm sidesteps the issue of quark/gluon fractions entirely, and only relies on the assumption that the samples used for training are proper mixed samples without contamination or sample-dependent labeling.

The ideas presented below may prove useful for a wide variety of machine learning applications, but for concreteness we focus on *classification*. It is worth emphasizing that the CWoLa framework can be applied to a huge variety of classifiers[2] without modification to the training procedure, by simply training on mixed event samples instead of on pure samples. By contrast, LLP-style weak supervision such as in ref. [34] requires a non-trivial modification to the loss function.[3] For this reason, CWoLa can be applied even for classifiers that are not trained in terms of loss functions at all.

Despite the power and simplicity of the CWoLa approach, there are some important limitations to keep in mind. First, the optimality of CWoLa is only true asymptotically; for a finite training set and a realistic machine learning algorithm, there can be differences, as discussed more below. Second, CWoLa does not apply when one class does not already exist in the data, as may be the case in a search for physics beyond the Standard Model (SM) with an exotic signature. That said, if the new physics can be decomposed into SM-like components, such as different types of jets, then CWoLa may once again be possible. Third, when the CWoLa strategy is employed for training in one event topology and testing in another event topology, there may be systematic uncertainties associated with

[1] After we developed this framework, we learned of a mathematically equivalent (but conceptually different) rephrasing of CWoLa in the language of learning from random noisy labels in ref. [32], where a version of theorem 1 also appears. See the discussion in section 2.3.

[2] CWoLA can be applied to train any classifier with a threshold that can be varied to sweep over operating points. k-nearest neighbors classification, for instance, does not have this property.

[3] The recent study in ref. [38], which was initially inspired by the LLP paradigm, is actually performing weak supervision using the CWoLa approach. We thank Timothy Cohen, Marat Freytsis, and Bryan Ostdiek for clarifications on this point.

the extrapolation. Of course, this is also true for traditional fully-supervised classification, which may introduce residual dependence on simulation; indeed, one could even combine adversarial approaches with CWoLa in this case to mitigate simulation dependence [31]. Finally, the CWoLa approach presented here only applies to mixtures of two categories, and further developments would be needed to disentangle multicategory samples.

The remainder of this paper is organized as follows. In section 2, we explain the theoretical foundations of the CWoLa paradigm and contrast it with LLP-style weak supervision and full supervision. We illustrate the power of CWoLa with a toy example of two gaussian random variables in section 3. We then apply CWoLa to the challenge of quark versus gluon jet tagging in section 4, using a dense network of five standard quark/gluon discriminants to highlight the performance of CWoLa on mixed samples. The paper concludes in section 5 with a summary and future outlook.

2 Machine learning with and without labels

The goal of classification is to distinguish two processes from each other: signal S and background B. Let $\vec{x}$ be a list of observables that are useful for distinguishing signal from background, and define $p_S(\vec{x})$ and $p_B(\vec{x})$ to be the probability distributions of $\vec{x}$ for the signal and background, respectively. A classifier $h : \vec{x} \mapsto \mathbb{R}$ is designed such that higher values of h are more signal-like and lower values are more background-like. A classifier operating point is defined by a threshold cut $h > c$; the signal efficiency is then $\epsilon_S = \int \mathrm{d}\vec{x}\, p_S(\vec{x})\, \Theta(h(\vec{x}) - c)$ and the background efficiency (i.e. mistag rate) is $\epsilon_B = \int \mathrm{d}\vec{x}\, p_B(\vec{x})\, \Theta(h(\vec{x}) - c)$, for the Heaviside step function Θ. The performance of a classifier h can be described by its receiver operating characteristic (ROC) curve which is the function $1 - \epsilon_B^h(\epsilon_S)$. A classifier h is *optimal* if for any other classifier h', $\epsilon_B^{h'}(\epsilon_S) \geq \epsilon_B^h(\epsilon_S)$ for all possible ϵ_S. By the Neyman-Pearson lemma [39], an optimal classifier is the likelihood ratio: $h_{\text{optimal}}(\vec{x}) = p_S(\vec{x})/p_B(\vec{x})$. Therefore, the goal of classification is to learn h_{optimal} or any classifier that is monotonically related to it.

In practice, one learns to approximate $h_{\text{optimal}}(\vec{x})$ from a set of signal and background $\vec{x}$ examples (*training data*). When the dimensionality of $\vec{x}$ is small and the number of examples large, it is often possible to approximate $p_S(\vec{x})$ and $p_B(\vec{x})$ directly by using histograms. When the dimensionality is large, an explicit construction is often not possible. In this case, one constructs a loss function that is minimized using a machine learning algorithm like a boosted decision tree or (deep) neural network. The following section describes three paradigms for learning $h_{\text{optimal}}(\vec{x})$ with different amounts of information available at training time: full supervision, LLP, and CWoLa. The ideas presented here apply to any procedure for constructing $h_{\text{optimal}}(\vec{x})$.

2.1 Full supervision

Fully supervised learning is the standard classification paradigm. Each example $\vec{x}_i$ comes with a label $u_i \in \{S, B\}$. For models trained to minimize loss functions, typical loss functions are the mean squared error:

$$\ell_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\Big(h(\vec{x}_i) - \mathbb{I}(u_i = S)\Big)^2, \tag{2.1}$$

for the indicator function $\mathbb{I}$, or the cross-entropy:

$$\ell_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbb{I}(u_i = S)\log h(\vec{x}_i) + \big(1 - \mathbb{I}(u_i = S)\big)\log\big(1 - h(\vec{x}_i)\big)\Big), \tag{2.2}$$

where N is the size of the subset (*batch*) of the available training data. With large enough training samples, flexible enough model parameterization, and suitable minimization procedure, the learned h should approach the performance of h_{optimal}.

2.2 Learning from label proportions

For weak supervision, one does not have complete and/or accurate label information. Here, we consider the case of accurate labels, but in the context of mixed samples. Consider two processes M_1 and M_2 that are mixtures of the original signal and background processes:

$$p_{M_1}(\vec{x}) = f_1\, p_S(\vec{x}) + (1 - f_1)\, p_B(\vec{x}), \tag{2.3}$$

$$p_{M_2}(\vec{x}) = f_2\, p_S(\vec{x}) + (1 - f_2)\, p_B(\vec{x}), \tag{2.4}$$

with the signal fractions satisfying $0 \le f_2 < f_1 \le 1$.

Instead of having training data labeled as being from p_S or p_B, we are now only given examples drawn from p_{M_1} and p_{M_2} with the corresponding M_1 and M_2 labels. We are however told f_1 and f_2 ahead of time. The resulting optimization problems are much less constrained than those in section 2.1, but learning is still possible. The key is to use several different mixed samples with sufficiently different fractions in order to avoid trivial failure modes, as discussed in ref. [34]. One possible loss function is given by:

$$\ell_{\text{LLP}} = \left|\sum_{i=1}^{N_{M_1}}\frac{h(\vec{x}_i)}{N_{M_1}} - f_1\right| + \left|\sum_{j=1}^{N_{M_2}}\frac{h(\vec{x}_j)}{N_{M_2}} - f_2\right|, \tag{2.5}$$

where N_{M_1} and N_{M_2} are the number of M_1 and M_2 examples in the batch. One could extend (and improve) this paradigm by adding in more samples with different fractions, but we consider only two here for simplicity.

2.3 Classification without labels

CWoLa is an alternative strategy for weak supervision in the context of mixed samples. Rather than modifying the loss function to accommodate the limited information as in section 2.2, the CWoLa approach is to simply train the model to discriminate the mixed samples M_1 and M_2 from one another. The classifier h trained to distinguish M_1 from M_2 (using full supervision) is then directly applied to distinguish S from B. An illustration of this technique is shown in figure 1. Remarkably, this procedure results in an optimal classifier (as defined in the beginning of section 2) for the S versus B classification problem:

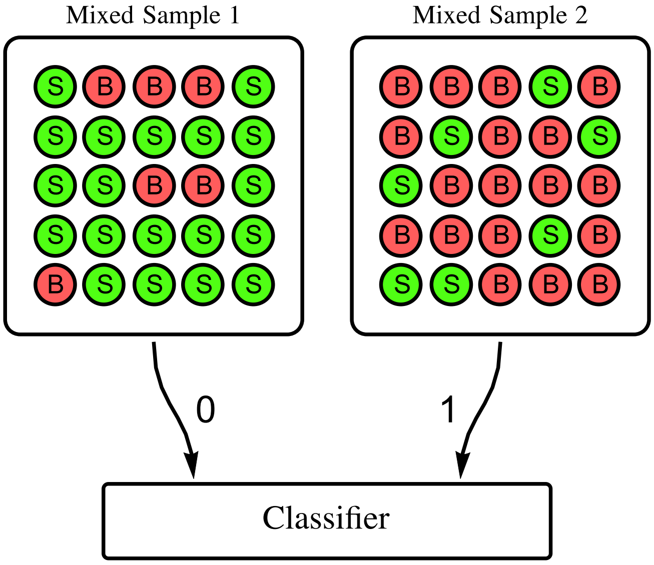


Figure 1. An illustration of the CWoLa framework. Rather than being trained to directly classify signal (S) from background (B), the classifier is trained by standard techniques to distinguish data as coming either from the first or second mixed sample, labeled as 0 and 1 respectively. No information about the signal/background labels or class proportions in the mixed samples is used during training.

Theorem 1. *Given mixed samples M_1 and M_2 defined in terms of pure samples S and B using eqs. (2.3) and (2.4) with signal fractions $f_1 > f_2$, an optimal classifier trained to distinguish M_1 from M_2 is also optimal for distinguishing S from B.*

Proof. The optimal classifier to distinguish examples drawn from p_{M_1} and p_{M_2} is the likelihood ratio $L_{M_1/M_2}(\vec{x}) = p_{M_1}(\vec{x})/p_{M_2}(\vec{x})$. Similarly, the optimal classifier to distinguish examples drawn from p_S and p_B is the likelihood ratio $L_{S/B}(\vec{x}) = p_S(\vec{x})/p_B(\vec{x})$. Where p_B has support, we can relate these two likelihood ratios algebraically:

$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1-f_1)\, p_B}{f_2\, p_S + (1-f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1-f_1)}{f_2\, L_{S/B} + (1-f_2)}, \tag{2.6}$$

which is a monotonically increasing rescaling of the likelihood $L_{S/B}$ as long as $f_1 > f_2$, since $\partial_{L_{S/B}} L_{M_1/M_2} = (f_1 - f_2)/(f_2 L_{S/B} - f_2 + 1)^2 > 0$. If $f_1 < f_2$, then one obtains the reversed classifier. Therefore, $L_{S/B}$ and L_{M_1/M_2} define the same classifier. □

An important feature of CWoLa is that, unlike the LLP-style weak supervision in section 2.2, the label proportions f_1 and f_2 are not required for training. Of course, this proof only guarantees that the optimal classifier from CWoLa is the same as the optimal classifier from fully-supervised learning. We explore the practical performance of CWoLa in sections 3 and 4.

The problem of learning from unknown mixed samples can be shown to be mathematically equivalent to the problem of learning with asymmetric random label noise, where there have been recent advances [32, 40]. The equivalence of these frameworks follows from the fact that randomly flipping the labels of pure samples, possibly with different flip probabilities for signal and background, produces mixed samples. In the language of noisy labels, ref. [32] argues that even unknown class proportions can be estimated from mixed samples under certain conditions using mixture proportion estimation [41], which may have interesting applications in collider physics. There are also connections between learning from unknown mixed samples and the *calibrated classifiers* approach in ref. [42], where measurement of the class proportions from unknown mixtures is also shown to be possible.

2.4 Operating points

While the optimal classifier from CWoLa is independent of the mixed sample compositions, some minimal input is needed in order to establish classification operating points. Specifically, to define a cut on the classifier h at a value c to achieve signal efficiency ϵ_S, one requires some degree of label information.

One practical strategy is to use CWoLa to train on two large mixed samples without label or class proportion information, and then benchmark it on two smaller samples where the class proportions f_1 and f_2 are precisely known. In that case, one can solve a simple system of equations on the smaller samples:

$$\Pr(h(x) > c \,|\, M_1) = \epsilon_S\, f_1 + \epsilon_B\,(1 - f_1) \tag{2.7}$$

$$\Pr(h(x) > c \,|\, M_2) = \epsilon_S\, f_2 + \epsilon_B\,(1 - f_2), \tag{2.8}$$

where the probabilities can be estimated numerically by counting the number of events that pass the classifier cut in some sample, e.g. $\Pr(h(x) > c \,|\, M_1) \approx \sum_{x\in\mathcal{M}_1} \mathbb{I}[h(x) > c]/|\mathcal{M}_1|$, where $\mathcal{M}_1$ is the mixed sample data. Thus with class proportions only, the ROC curve of a classifier can be determined.[4]

For the purpose of establishing working points, one might need to rely on simulations to determine the label proportions of the test samples. In many cases, though, label proportions are better known than the details of the observables used to train the classifier. For instance, in jet tagging, the label proportions of kinematically-selected samples are largely determined by the hard scattering process, with only mild sensitivity to effects such as shower mismodeling. In this way, one is sensitive only to simulation uncertainties associated with sample composition, which in most cases are largely uncorrelated with uncertainties associated with tagging performance.

To summarize, the CWoLa paradigm does not need class proportions during training, and it only requires a small sample of test data where class proportions are known in order to determine the classifier performance and operating points, with minimal input from simulation.

3 Illustrative example: two gaussian random variables

Before demonstrating the combination of CWoLa with a modern neural network, we first illustrate the various forms of learning discussed in section 2 through a simplified example where the optimal classifier can be obtained analytically. Consider a single observable x for

[4] We are grateful to Francesco Rubbo for bringing this to our attention.

distinguishing a signal S from a background B. For simplicity, suppose that the probability distribution of x is a Gaussian with mean μ_S and standard deviation σ_S for the signal and a Gaussian with mean μ_B and standard deviation σ_B for the background. We then consider the mixed samples M_1 and M_2 from eqs. (2.3) and (2.4) with signal fractions f_1 and f_2.

In this one-dimensional case, the optimal fully-supervised classifier can be constructed analytically:

$$h_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}. \tag{3.1}$$

Of course, non-parameterically estimating eq. (3.1) numerically requires a choice of binning which can introduce numerical fluctuations. To avoid this effect, we discretize x into 50 bins between -40 and 40 (under/overflow is added to the first/last bins). There are then a finite number of possibilities for the likelihood ratio in eq. (3.1).

Using a calligraphic font to denote explicit training samples, we test the following classifiers on signal ($\mathcal{S}$), background ($\mathcal{B}$), and mixed ($\mathcal{M}_{1,2}$) training samples of the same size:

1. *Full Supervision* (section 2.1): by construction, every example in the signal training dataset $\mathcal{S}$ is a signal event and every example in the background training set $\mathcal{B}$ is a background event. The classifier is the numerical approximation to eq. (3.1):

$$h_{\text{full}}(x) = \frac{\sum_{y\in\mathcal{S}} \mathbb{I}[y=x]}{\sum_{y\in\mathcal{B}} \mathbb{I}[y=x]}. \tag{3.2}$$

2. *LLP* (section 2.2): the events in the mixed training samples $\mathcal{M}_1$ and $\mathcal{M}_2$ are a mixture of signal and background events. Weak supervision proceeds by solving the system of equations in eqs. (2.3) and (2.4) and using numerical estimates for p_{M_1} and p_{M_2}:

$$h_{\text{LLP}}(x) = \frac{(1-f_2)\sum_{y\in\mathcal{M}_1} \mathbb{I}[y=x] - (1-f_1)\sum_{y\in\mathcal{M}_2} \mathbb{I}[y=x]}{f_1\sum_{y\in\mathcal{M}_2} \mathbb{I}[y=x] - f_2\sum_{y\in\mathcal{M}_1} \mathbb{I}[y=x]}. \tag{3.3}$$

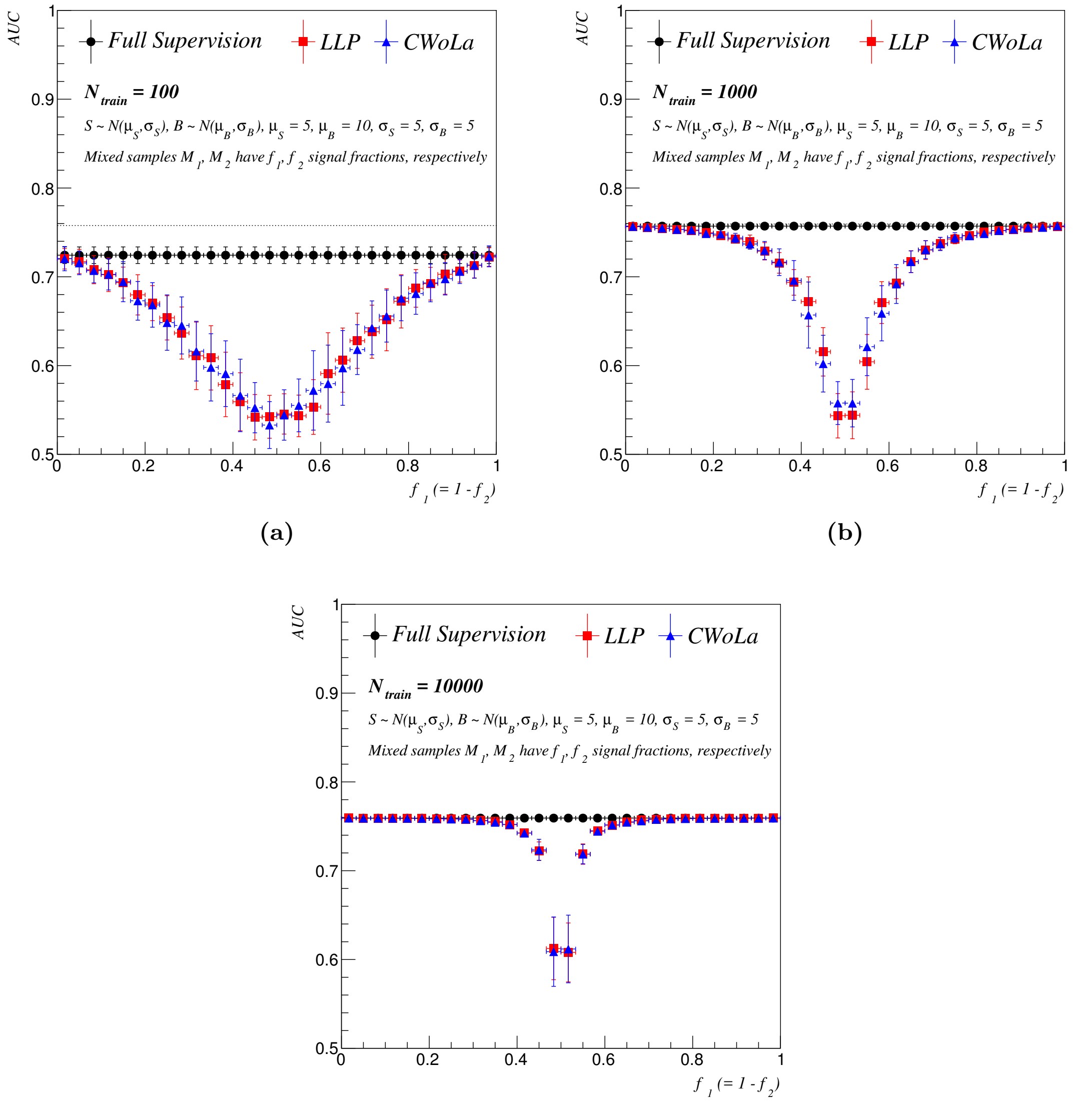
3. *CWoLa* (section 2.3): the input is the same as for the LLP case, though the fractions f_1 and f_2 are not needed as input. The CWoLa classifier is the same as in eq. (3.2), only now signal and background distributions are replaced by the available mixed examples:

$$h_{\text{CWoLa}}(x) = \frac{\sum_{y\in\mathcal{M}_1} \mathbb{I}[y=x]}{\sum_{y\in\mathcal{M}_2} \mathbb{I}[y=x]}. \tag{3.4}$$

The performance of the classifiers trained in this way is evaluated on a holdout set of signal and background examples that is large enough such that statistical fluctuations are negligible. We use the area under the curve (AUC) metric to quantify performance. For continuous random variables, the AUC can be defined as $\Pr(h(x|S) > h(x|B))$. This notion extends well to discrete random variables (indexed by integers):

$$\text{AUC} = \sum_{i=1}\sum_{j=i+1} \Pr(x=i\,|\,S)\,\Pr(x=j\,|\,B) + \frac{1}{2}\sum_{i=1} \Pr(x=i\,|\,S)\,\Pr(x=i\,|\,B). \tag{3.5}$$

For a properly constructed classifier, the AUC ≥ 0.5. In all of the numerical examples shown below, the classifier is inverted if necessary so that by construction, AUC ≥ 0.5.



(a)

(b)

(c)

Figure 2. The AUC for the LLP and CWoLa methods as a function of the signal fraction f_1, for training sizes $N_{\rm train}$ of (a) 100 events, (b) 1k events, and (c) 10k events. Here, the complementary signal fraction is $f_2 = 1 - f_1$. By construction, the AUC for full supervision is independent of f_1. The horizontal dashed line indicates the fully-supervised AUC with infinite training statistics. For $N_{\rm train}$ sufficiently large and f_1 sufficient far from 0.5, all three methods converge to the optimal case.

In figure 2, we illustrate the performance of the three classification paradigms described above with 100, 1k, and 10k training examples each of S and B, or M_1 and M_2 in the LLP and CWoLa cases, taking $f_1 = 1 - f_2$ for concreteness. Testing is performed on 100k S and B examples in all cases. The LLP and CWoLa paradigms have nearly the same dependence on the number of training events and the signal fraction f_1. The full supervision does not depend on the signal composition of M_1 and M_2 as it is trained directly on labeled signal and background examples. As expected, the performance is poor when the number of

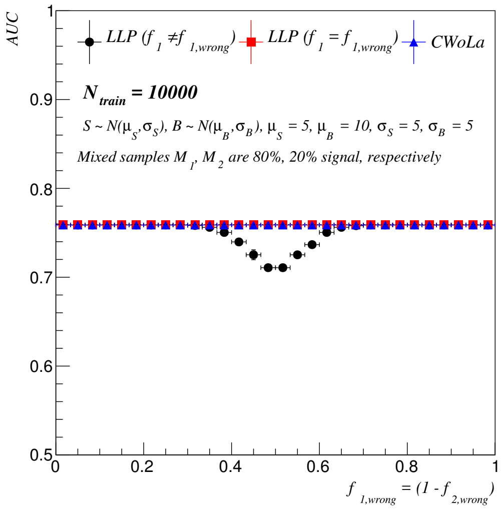


Figure 3. The AUC for LLP and CWoLa as a function of the (possibly incorrect) signal fraction provided for training. By construction, CWoLa does not depend on the input fraction and LLP is only sensitive to provided signal fraction information when that fraction is near 50%.

training examples is small or f_1 is close to f_2 (so the effective number of useful events is small). As $f_1 \to f_2$, the two mixtures become identical and there is thus no way to distinguish M_1 and M_2; in the context of LLP, this corresponds to attempting to solve a degenerate system of equations. With sufficiently many training examples and/or well-separated fractions f_1 and f_2, the techniques trained with M_1 and M_2 converge to the fully supervised case, as expected from Theorem 1.

One advantage of CWoLa over the LLP approach is that the fractions f_1 and f_2 are not required for training. In figure 3, we demonstrate the impact on the AUC for LLP when the wrong fractions are provided at training time. Here, the true fractions are $f_1 = 80\%$ and $f_2 = 20\%$, but different fractions $f_{1,\text{wrong}} = 1 - f_{2,\text{wrong}}$ are used to calculate eq. (3.3). For $f_{1,\text{wrong}}$ far from 50%, there is little dependence on the fraction used for training. This insensitivity is likely due to the preservation of monotonicity to the full likelihood with small perturbations in f, as discussed in detail in ref. [38].

With this one-dimensional example, the estimate for the optimal classifier under each of the three schemes is computable directly. It is often the case that $\vec{x}$ is highly multi-dimensional, though, in which case a more sophisticated learning scheme may be required. We investigate the performance of CWoLa in a five-dimensional space in the next section.

4 Realistic example: quark/gluon jet discrimination

Quark- versus gluon-initaited jet tagging [43–51] is a particularly important classification problem in high energy physics where training on data would be beneficial. This is because correlations between key observables known to be useful for tagging are not always well-modeled by simulations as they depend on the detailed structure of a jet's radiation pattern [24, 52]. Furthermore, even the LLP paradigm proposed in ref. [34] can be sensitive to the input fractions which are themselves dependent on non-perturbative information from parton distribution functions. In this section, we test the performance of CWoLa in a realistic context where a small number of quark/gluon discriminants are combined into one classifier, similar to the CMS quark/gluon likelihood [25, 26].

A key limitation of this study is that we artificially construct mixed samples $\mathcal{M}_1$ and $\mathcal{M}_2$ from pure "quark" ($\mathcal{S}$) and pure "gluon" ($\mathcal{B}$) samples.[5] In the practical case of interest at the LHC, one would measure a quark-enriched sample in Z plus jet events and a gluon-enriched sample in dijet events, with more sophisticated selections possible as well [53]. However, the "quark" jet in $pp \to Z + j$ event is not the same as the "quark" jet in $pp \to 2j$, since there are soft color correlations with the rest of the event. Jet grooming techniques [54–59] can mitigate the impact of soft effects to provide a more universal "quark" jet definition [60, 61]. Still, one needs to validate the robustness of quark/gluon classifiers to the possibility of sample-dependent labels, and we leave a detailed study of this effect to future work.

This study is based on five key jet substructure observables which are known to be useful quark/gluon discriminants [37]. The discriminants are combined using a modern neural network employing either CWoLa or fully-supervised learning. We do not show a benchmark curve for LLP since it is difficult to ensure a fair comparison. By contrast, CWoLa and full supervision use the same loss function with the same training strategy, so a direct comparison is meaningful. All of the observables can be written in terms of the generalized angularities [51] (see also [62–64]):

$$\lambda^\kappa_\beta = \sum_{i \in \text{jet}} z_i^\kappa \theta_i^\beta, \quad \text{with} \quad z_i = \frac{p_{\mathrm{T},i}}{\sum_{j \in \text{jet}} p_{\mathrm{T},j}}, \quad \theta_i = \frac{\Delta R_i}{R}, \tag{4.1}$$

where ΔR_i is the rapidity/azimuth distance to the E-scheme jet axis,[6] $p_{\mathrm{T},i}$ is the particle transverse momentum, and R is the jet radius. The observables used to train the network use (κ, β) values of:

$$\begin{array}{ccccc} (0,0) & (2,0) & (1,0.5) & (1,1) & (1,2) \\ \text{multiplicity} & p_\mathrm{T}^\mathrm{D} & \text{LHA} & \text{width} & \text{mass} \end{array} \tag{4.2}$$

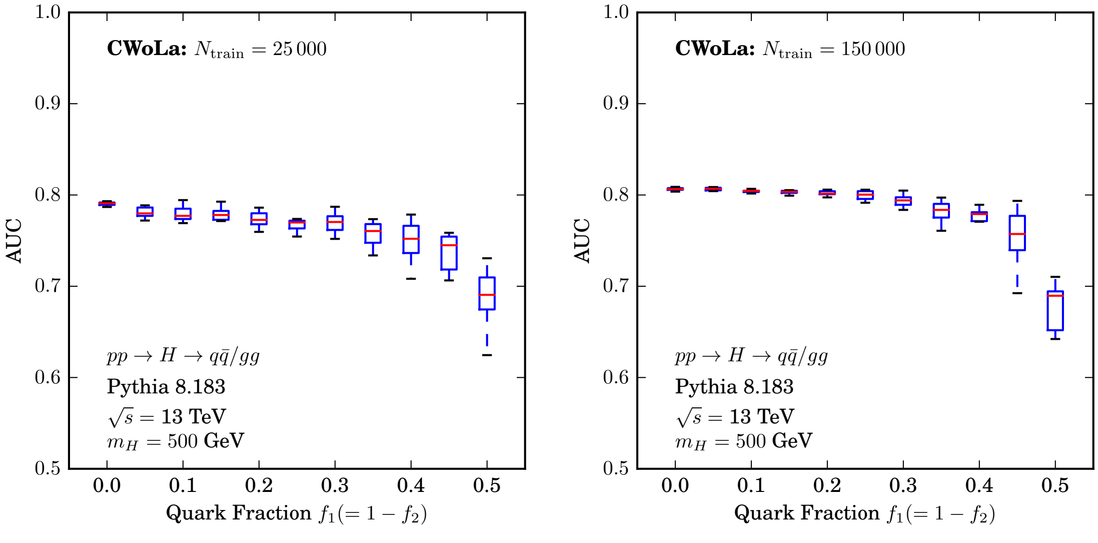
where the names map onto the well-known discriminants in the quark/gluon literature.[7]

Quark and gluon jets are simulated from the decay of a heavy scalar particle H with $m_H = 500\,\text{GeV}$ in either the $pp \to H \to q\bar{q}$ or $pp \to H \to gg$ channel. Production, decay, and fragmentation are modeled with Pythia 8.183 [70]. Jets are clustered using the anti-k_t algorithm [71] with radius $R = 0.6$ implemented in FastJet 3.1.3 [72]. Only detector-stable hadrons are used for jet finding. Since the gluon color factor C_A is larger than the quark color factor C_F by about a factor of two, gluon jets have more particles and are "wider" on average as measured by the angularities listed above.

[5] The reason for the scare quotes is discussed at length in ref. [37], as the definition of a quark or gluon jet is fundamentally ambiguous.

[6] This is in contrast to ref. [37], which uses the winner-take-all axis [65–67].

[7] Strictly speaking $(2,0)$ is the square of p_T^D [68], and $(1,2)$ is mass-squared over energy-squared in the soft-collinear limit. For this study, we use the angularity definition of the five observables. Note that the first observable is infrared and collinear (IRC) unsafe, the second observable is IR safe but C unsafe, and the last three observables with $\kappa = 1$ are all IRC safe. LHA refers to the Les Houches Angularity from the eponymous study in refs. [37, 69].



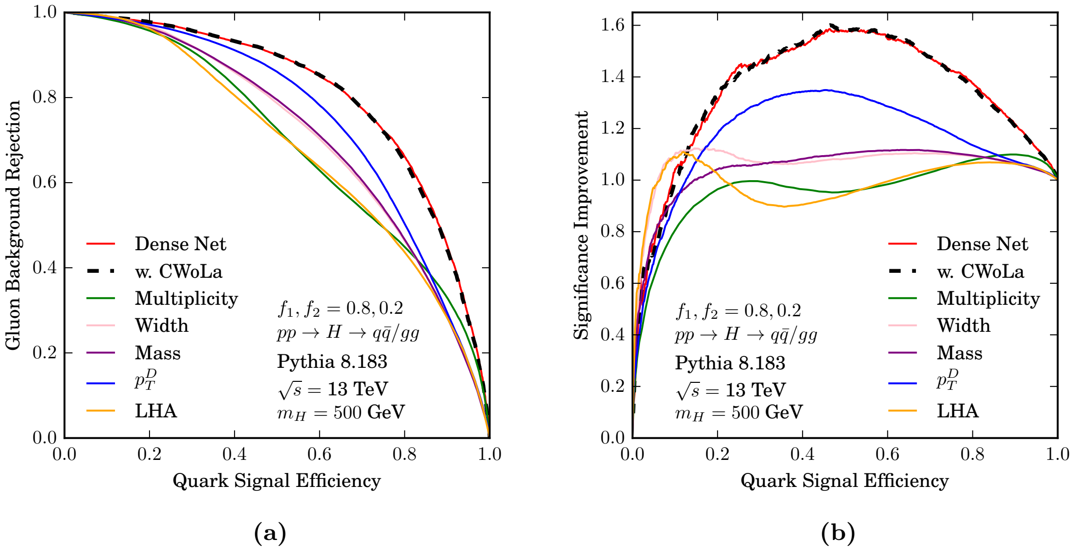
(a) **(b)**

Figure 4. Training performance of the CWoLa method on two mixed samples with $f_1 = 1 - f_2$ quark fraction. Shown are the range of AUC values obtained from 10 repetitions of training the neural network on (a) 25k events and (b) 150k events for 10 epochs.

To classify quarks and gluons with either the CWoLa or fully-supervised method, we use a simple neural network consisting of two dense layers of 30 nodes with rectified linear unit (ReLU) activation functions connected to a 2-node output with a softmax activation function. All neural network training was performed with the Python deep learning library Keras [73] with a Tensorflow [74] backend. The data consisted of 200k quark/gluon events, partitioned into 20k validation event, 20k test events, and the remainder used as training event samples of various sizes. He-uniform weight initialization [75] was used for the model weights. The network was trained with the categorical cross-entropy loss function using the Adam algorithm [76] with a learning rate of 0.001 and a batch size of 128.

In figure 4, we show the performance of CWoLa training for quark/gluon classification using mixed samples of different purities. These mixed samples of 25k and 150k training events were generated by shuffling the pure samples into two sets in different proportions. Performance is measured in terms of the classifier AUC. The behavior resembles that found in the toy model of figure 2, with more training data resulting in increased robustness to sample impurity. It is remarkable that such good performance can be obtained even when the signal/background events are so heavily mixed.

In figure 5, we show ROC and significance improvement (SI) curves for 150k training events, where SI is a curve of $\epsilon_q/\sqrt{\epsilon_g}$ at different ϵ_q values [50]. Results are given for the fully-supervised classifier trained on pure samples and the CWoLa classifier trained on mixed samples with $f_1 = 80\%$ and $f_2 = 20\%$, along with the curves of the input observables. Both the fully-supervised and CWoLa dense networks achieve similar performance, with the expected improvement over the individual input observables. This suggests that the proof of CWoLa optimality in theorem 1 is achievable in practice, though many more studies are needed to demonstrate this in a wider range of contexts.



(a) **(b)**

Figure 5. Quark/gluon discrimination performance in terms of (a) ROC curves and (b) SI curves. Shown are results for the dense net trained on 150k pure samples, and then with CWoLa on $f_1 = 80\%$ versus $f_2 = 20\%$ mixed samples, as well as the input observables individually. The classifier trained on the mixed samples achieves similar performance to the classifier trained on the pure samples, with improvement in performance over the input observables.

5 Conclusions

We introduced the CWoLa framework for training classifiers on different mixed samples of signal and background events, without using true labels or class proportions. The observation that the optimal classifier for mixed samples of signal and background is also optimal for pure samples of signal and background, proven in theorem 1, could be of tremendous practical use at the LHC for learning directly from data whenever truth information is unknown or uncertain and whenever detailed and reliable simulations are unavailable. We highlight that no new specific code, loss function, or model architecture is needed to implement CWoLa. Any tools for training a classifier using truth information can be directly applied to discriminate mixed samples and thus to train in the CWoLa framework directly on data.

Using a toy example, we found that CWoLa performs as well as LLP (which requires knowledge of the class proportions), suggesting that CWoLa is a robust paradigm for weak supervision. Of course, to determine operating points and classification power for the CWoLa method, some label information is needed, but it can be furnished by a smaller sample of testing data that can be separate from the larger mixed samples used for training. It is also worth remembering that CWoLa assumes that the mixed samples are not subject to contamination or sample-dependent labeling, though one could imagine using data-driven cross-validation with more than two mixed samples to identify and mitigate such effects. More ambitiously, one could try to apply CWoLa to event samples that otherwise look identical, to try to tease out potential subpopulations of events.

As a realistic example, we applied the CWoLa framework to the important case of quark/gluon discrimination, a classification task for which simulations are typically unreliable and true labels are unknown. We showed that the CWoLa method can be successfully used to train a dense neural network for quark/gluon classification on mixed samples with five jet substructure observables as input. Though the realistic example made use of a neural network, the CWoLa paradigm can be used to train many other types of classifiers. While in this study we considered a relatively small network on a small (but important) number of inputs, the same principles apply for any type of model or input. In future work, we plan to study CWoLa in the context of deeper architectures and larger inputs.

Acknowledgments

The authors are grateful to Timothy Cohen, Kyle Cranmer, Marat Freytsis, Patrick Komiske, Bryan Ostdiek, Francesco Rubbo, Matthew Schwartz, and Clayton Scott for helpful discussions and suggestions. Cloud computing resources were provided through a Microsoft Azure for Research award. The work of E.M.M. and J.T. is supported by the DOE under grant contract numbers DE-SC-00012567 and DE-SC-00015476. The work of B.N. is supported by the DOE under contract DE-AC02-05CH11231.

References

[1] J. Cogan, M. Kagan, E. Strauss and A. Schwarztman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, *JHEP* **02** (2015) 118 [arXiv:1407.5675] [INSPIRE].

[2] L.G. Almeida, M. Backović, M. Cliche, S.J. Lee and M. Perelstein, *Playing Tag with ANN: Boosted Top Identification with Pattern Recognition*, *JHEP* **07** (2015) 086 [arXiv:1501.05968] [INSPIRE].

[3] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, *Jet-images — deep learning edition*, *JHEP* **07** (2016) 069 [arXiv:1511.05190] [INSPIRE].

[4] P. Baldi, K. Bauer, C. Eng, P. Sadowski and D. Whiteson, *Jet Substructure Classification in High-Energy Physics with Deep Neural Networks*, *Phys. Rev.* **D 93** (2016) 094034 [arXiv:1603.09349] [INSPIRE].

[5] J. Barnard, E.N. Dawe, M.J. Dolan and N. Rajcic, *Parton Shower Uncertainties in Jet Substructure Analyses with Deep Neural Networks*, *Phys. Rev.* **D 95** (2017) 014018 [arXiv:1609.00607] [INSPIRE].

[6] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning Top Taggers or The End of QCD?*, *JHEP* **05** (2017) 006 [arXiv:1701.08784] [INSPIRE].

[7] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, *JHEP* **01** (2017) 110 [arXiv:1612.01551] [INSPIRE].

[8] L. de Oliveira, M. Paganini and B. Nachman, *Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis*, arXiv:1701.05927 [INSPIRE].

[9] P.T. Komiske, E.M. Metodiev, B. Nachman and M.D. Schwartz, *Pileup Mitigation with Machine Learning (PUMML)*, arXiv:1707.08600 [INSPIRE].

[10] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics*, arXiv:1702.00748 [INSPIRE].

[11] ATLAS collaboration, *Performance of Top Quark and W Boson Tagging in Run 2 with ATLAS*, ATLAS-CONF-2017-064 (2017).

[12] ATLAS collaboration, *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*, ATL-PHYS-PUB-2017-013 (2017).

[13] ATLAS collaboration, *Identification of Hadronically-Decaying W Bosons and Top Quarks Using High-Level Features as Input to Boosted Decision Trees and Deep Neural Networks in ATLAS at* $\sqrt{s}$ *= 13 TeV*, ATL-PHYS-PUB-2017-004 (2017).

[14] CMS collaboration, *Heavy flavor identification at CMS with deep neural networks*, CMS-DP-2017-005 (2017).

[15] CMS collaboration, *CMS Phase 1 heavy flavour identification performance and developments*, CMS-DP-2017-013 (2017).

[16] ATLAS collaboration, *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*, ATL-PHYS-PUB-2017-003 (2017).

[17] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned Top Tagging using Lorentz Invariance and Nothing Else*, arXiv:1707.08966 [INSPIRE].

[18] J. Pearkes, W. Fedorko, A. Lister and C. Gay, *Jet Constituents for Deep Neural Network Based Top Quark Tagging*, arXiv:1704.02124 [INSPIRE].

[19] K. Datta and A. Larkoski, *How Much Information is in a Jet?*, *JHEP* **06** (2017) 073 [arXiv:1704.08249] [INSPIRE].

[20] ATLAS collaboration, *Quark versus Gluon Jet Tagging Using Jet Images with the ATLAS Detector*, ATL-PHYS-PUB-2017-017 (2017).

[21] CMS collaboration, *New Developments for Jet Substructure Reconstruction in CMS*, CMS-DP-2017-027 (2017).

[22] ATLAS collaboration, *Performance of b-Jet Identification in the ATLAS Experiment*, 2016 *JINST* **11** P04008 [arXiv:1512.01094] [INSPIRE].

[23] CMS collaboration, *Identification of b-quark jets with the CMS experiment*, 2013 *JINST* **8** P04013 [arXiv:1211.4462] [INSPIRE].

[24] ATLAS collaboration, *Light-quark and gluon jet discrimination in pp collisions at* $\sqrt{s} = 7$ TeV *with the ATLAS detector*, *Eur. Phys. J.* **C 74** (2014) 3023 [arXiv:1405.6583] [INSPIRE].

[25] CMS collaboration, *Performance of quark/gluon discrimination in 8 TeV pp data*, CMS-PAS-JME-13-002 (2013).

[26] CMS collaboration, *Performance of quark/gluon discrimination in 13 TeV data*, CMS-DP-2016-070 (2016).

[27] ATLAS collaboration, *Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV*, Eur. Phys. J. **C 76** (2016) 154 [arXiv:1510.05821] [INSPIRE].

[28] CMS collaboration, *Identification techniques for highly boosted W bosons that decay into hadrons*, JHEP **12** (2014) 017 [arXiv:1410.4227] [INSPIRE].

[29] ATLAS collaboration, *Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s}$ = 8 TeV with the ATLAS detector*, JHEP **06** (2016) 093 [arXiv:1603.03127] [INSPIRE].

[30] CMS collaboration, *Boosted Top Jet Tagging at CMS*, CMS-PAS-JME-13-007 (2014).

[31] G. Louppe, M. Kagan and K. Cranmer, *Learning to Pivot with Adversarial Networks*, arXiv:1611.01046 [INSPIRE].

[32] G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott, *Classification with asymmetric label noise: consistency and maximal denoising*, Electron. J. Stat. **10** (2016) 2780.

[33] J. Hernández-González, I. Inza, and J.A. Lozano, *Weak supervision and other non-standard classification problems: a taxonomy*, *Pattern Recogn. Lett.* **69** (2016) 49.

[34] L.M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly Supervised Classification in High Energy Physics*, JHEP **05** (2017) 145 [arXiv:1702.00414] [INSPIRE].

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

[36] G. Patrini, R. Nock, P. Rivera and T. Caetano, *(Almost) no label no cry*, in *Advances in Neural Information Processing Systems*, (2014), pp. 190–198.

[37] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer et al., *Systematics of quark/gluon tagging*, JHEP **07** (2017) 091 [arXiv:1704.03878] [INSPIRE].

[38] T. Cohen, M. Freytsis and B. Ostdiek, *(Machine) Learning to Do More with Less*, arXiv:1706.09451 [INSPIRE].

[39] J. Neyman and E.S. Pearson, *On the problem of the most efficient tests of statistical hypotheses*, in *Breakthroughs in statistics*, Springer (1992), pp. 73–108.

[40] N. Natarajan, I.S. Dhillon, P.K. Ravikumar and A. Tewari, *Learning with noisy labels*, in *Advances in neural information processing systems 26 (NIPS 2013)*, (2013), pp. 1196–1204.

[41] C. Scott, *A rate of convergence for mixture proportion estimation, with application to learning from noisy labels*, in Proceedings of the Eighteenth International Conference on *Artificial Intelligence and Statistics*, (2015), pp. 838–846.

[42] K. Cranmer, J. Pavez and G. Louppe, *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*, arXiv:1506.02169 [INSPIRE].

[43] H.P. Nilles and K.H. Streng, *quark-gluon Separation in Three Jet Events*, Phys. Rev. **D 23** (1981) 1944 [INSPIRE].

[44] L.M. Jones, *Tests for Determining the Parton Ancestor of a Hadron Jet*, Phys. Rev. **D 39** (1989) 2550 [INSPIRE].

[45] Z. Fodor, *How to See the Differences Between Quark and Gluon Jets*, Phys. Rev. **D 41** (1990) 1726 [INSPIRE].

[46] L. Jones, *Towards a systematic jet classification*, Phys. Rev. **D 42** (1990) 811 [INSPIRE].

[47] L. Lönnblad, C. Peterson and T. Rognvaldsson, *Using neural networks to identify jets*, *Nucl. Phys.* **B 349** (1991) 675 [INSPIRE].

[48] J. Pumplin, *How to tell quark jets from gluon jets*, *Phys. Rev.* **D 44** (1991) 2025 [INSPIRE].

[49] J. Gallicchio and M.D. Schwartz, *Quark and Gluon Tagging at the LHC*, *Phys. Rev. Lett.* **107** (2011) 172001 [arXiv:1106.3076] [INSPIRE].

[50] J. Gallicchio and M.D. Schwartz, *Quark and Gluon Jet Substructure*, *JHEP* **04** (2013) 090 [arXiv:1211.7038] [INSPIRE].

[51] A.J. Larkoski, J. Thaler and W.J. Waalewijn, *Gaining (Mutual) Information about Quark/Gluon Discrimination*, *JHEP* **11** (2014) 129 [arXiv:1408.3122] [INSPIRE].

[52] ATLAS collaboration, *Measurement of the charged-particle multiplicity inside jets from* $\sqrt{s} = 8$ *TeV pp collisions with the ATLAS detector*, *Eur. Phys. J.* **C 76** (2016) 322 [arXiv:1602.00988] [INSPIRE].

[53] J. Gallicchio and M.D. Schwartz, *Pure Samples of Quark and Gluon Jets at the LHC*, *JHEP* **10** (2011) 103 [arXiv:1104.1175] [INSPIRE].

[54] J.M. Butterworth, A.R. Davison, M. Rubin and G.P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys. Rev. Lett.* **100** (2008) 242001 [arXiv:0802.2470] [INSPIRE].

[55] S.D. Ellis, C.K. Vermilion and J.R. Walsh, *Techniques for improved heavy particle searches with jet substructure*, *Phys. Rev.* **D 80** (2009) 051501 [arXiv:0903.5081] [INSPIRE].

[56] S.D. Ellis, C.K. Vermilion and J.R. Walsh, *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*, *Phys. Rev.* **D 81** (2010) 094023 [arXiv:0912.0033] [INSPIRE].

[57] D. Krohn, J. Thaler and L.-T. Wang, *Jet Trimming*, *JHEP* **02** (2010) 084 [arXiv:0912.1342] [INSPIRE].

[58] M. Dasgupta, A. Fregoso, S. Marzani and G.P. Salam, *Towards an understanding of jet substructure*, *JHEP* **09** (2013) 029 [arXiv:1307.0007] [INSPIRE].

[59] A.J. Larkoski, S. Marzani, G. Soyez and J. Thaler, *Soft Drop*, *JHEP* **05** (2014) 146 [arXiv:1402.2657] [INSPIRE].

[60] C. Frye, A.J. Larkoski, M.D. Schwartz and K. Yan, *Precision physics with pile-up insensitive observables*, arXiv:1603.06375 [INSPIRE].

[61] C. Frye, A.J. Larkoski, M.D. Schwartz and K. Yan, *Factorization for groomed jet substructure beyond the next-to-leading logarithm*, *JHEP* **07** (2016) 064 [arXiv:1603.09338] [INSPIRE].

[62] C.F. Berger, T. Kucs and G.F. Sterman, *Event shape/energy flow correlations*, *Phys. Rev.* **D 68** (2003) 014012 [hep-ph/0303051] [INSPIRE].

[63] L.G. Almeida, S.J. Lee, G. Perez, G.F. Sterman, I. Sung and J. Virzi, *Substructure of high-*p_T *Jets at the LHC*, *Phys. Rev.* **D 79** (2009) 074017 [arXiv:0807.0234] [INSPIRE].

[64] S.D. Ellis, C.K. Vermilion, J.R. Walsh, A. Hornig and C. Lee, *Jet Shapes and Jet Algorithms in SCET*, *JHEP* **11** (2010) 101 [arXiv:1001.0014] [INSPIRE].

[65] D. Bertolini, T. Chan and J. Thaler, *Jet Observables Without Jet Algorithms*, *JHEP* **04** (2014) 013 [arXiv:1310.7584] [INSPIRE].

[66] A.J. Larkoski, D. Neill and J. Thaler, *Jet Shapes with the Broadening Axis*, *JHEP* **04** (2014) 017 [arXiv:1401.2158] [INSPIRE].

[67] G. Salam, *E_t^∞ Scheme*, unpublished.

[68] CMS collaboration, *Search for a Higgs boson in the decay channel $H \to ZZ^{(*)} \to q\bar{q}\ell^-\ell^+$ in pp collisions at $\sqrt{s} = 7$ TeV*, *JHEP* **04** (2012) 036 [arXiv:1202.1416] [INSPIRE].

[69] J.R. Andersen et al., *Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report*, in *9th Les Houches Workshop on Physics at TeV Colliders (PhysTeV 2015)*, Les Houches, France, 1–19 June 2015, (2016) [arXiv:1605.04692] [INSPIRE].

[70] T. Sjöstrand, S. Mrenna and P.Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852 [arXiv:0710.3820] [INSPIRE].

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

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

[73] F. Chollet, *Keras*, (2017) https://github.com/fchollet/keras.

[74] M. Abadi et al., *Tensorflow: a system for large-scale machine learning*, proceedings of the *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, volume 16, (2016), pp. 265–283.

[75] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: surpassing human-level performance on imagenet classification*, in proceedings of the *IEEE international conference on computer vision*, (2015), pp. 1026–1034.

[76] D. Kingma and J. Ba, *Adam: a method for stochastic optimization*, arXiv:1412.6980.