Uncovering latent jet substructure

Barry M. Dillon[*] and Darius A. Faroughy[†]
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

Jernej F. Kamenik[‡]
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia and Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia

(Received 16 April 2019; published 3 September 2019)

We apply techniques from Bayesian generative statistical modeling to uncover hidden features in jet substructure observables that discriminate between different *a priori* unknown underlying short distance physical processes in multijet events. In particular, we use a mixed membership model known as *latent Dirichlet allocation* to build a data-driven *unsupervised* top-quark tagger and $t\bar{t}$ event classifier. We compare our proposal to existing traditional and machine learning approaches to top-jet tagging. Finally, employing a toy vector-scalar boson model as a benchmark, we demonstrate the potential for discovering new physics signatures in multijet events in a model independent and unsupervised way.

DOI: 10.1103/PhysRevD.100.056002

I. INTRODUCTION

The use of jet substructure techniques in studying large area jets has played an important role in identifying hadronic decays of Higgs and electroweak gauge bosons in runs 1 and 2 of the LHC [1–4]. These techniques have also been used efficiently to tag jets arising from top quarks [5–15]. In the last few years, machine learning (ML) tools have extended the application of jet substructure in tagging jets at the LHC [16–32] through the use of neural networks (NNs) to process and "learn" from vast amounts of training data. Since these approaches rely on theoretical predictions for pure signal and background training data sets [typically through Monte Carlo (MC) generators], they (a) are exposed to MC mismodeling of realistic events as reconstructed from real data and detectors; (b) require exact model knowledge of both expected signal and backgrounds. This limits their use in searches for *a priori* unknown new phenomena in LHC jet events.

There have been recent advances in unsupervised or semisupervised ML techniques, based on NNs designed to be able to separate signal and background events in mixed samples, and could therefore be run directly on experimental data without the need for pure MC training samples, see e.g., Refs. [33–38] and [39–43]. They rely on categorizing and comparing datasets with different expected signal and background admixtures or identifying anomalous events inside large datasets. While these approaches ameliorate the model dependence of fully supervised ML, they are still potentially susceptible to correlated systematics (i.e., detector) effects and/or subject to large look-elsewhere effects. In addition, they generally work best when applied on very large datasets. Consequently their performance may suffer when looking for effects in tails of distributions.

In this article, we outline a new technique to classify jets and events *in situ* within a single mixed event sample, using tools developed in a branch of ML called generative statistical modeling [44,45]. Developed primarily to identify emergent themes in collections of documents, these models infer the hidden (or latent) structure of a document corpus using posterior Bayesian inference based on word and theme co-occurence [46]. Translated into the language jet physics, one assumes that observable jet substructure histograms (*words*) in events (*documents*) are generated by drawing from latent distributions (*themes*) of varying proportions. This allows us to construct so-called statistical mixed membership models of jet substructure.[1] Furthermore assuming that each event is a mixture of only few latent distributions and that within each of these only few histogram bins have high co-occurrence, such models can be solved using techniques of latent Dirichlet allocation

[*]barry.dillon@ijs.si
[†]darius.faroughy@ijs.si
[‡]jernej.kamenik@cern.ch

[1]Similar techniques have been used recently in a semisupervised way to reconstruct "pure" quark and gluon jet observable distributions from mixed event samples [47,48].

(LDA) [49]. Finally, with a trained model at hand, one can define robust parametric jet and event classifiers by inferring on the latent distribution proportions in tested events.

In the following we first present the main ingredients of our proposal in more detail. Then we discuss two proof of principle implementations based on benchmark examples: an unsupervised top-quark jet tagger and $t\bar{t}$ event classifier, as well as an unsupervised new physics (NP) search strategy able to identify boosted neutral scalar bosons decaying to pairs of W's (previously studied in Refs. [37,38,50]). We compare them to existing conventional and ML approaches and also outline possible further improvements and future directions.

II. GENERATIVE BAYESIAN MODELS OF JET SUBSTRUCTURE

We start by considering the formation of a jet stemming from an initial hard seed, as a sequential combination of QCD showering (followed by fragmentation and hadronization) and possibly massive particle decays. Next we assume that some relevant information on this intertwined sequence of processes can be recovered by looking at the clustering history of a jet-clustering algorithm. This is in fact the basis for many conventional taggers of massive jets [1,8,13].

Within this very simplified picture of jet formation and observation we can draw interesting parallels to so-called mixed membership models describing generation of documents in the context of text analysis [49], or genotypes in population studies [51]. In particular, we assume that the observable distribution bins in a clustering profile are populated by drawing from a few latent distributions—*themes*—corresponding to different contributing physical processes. The likelihood of populating a certain distribution bin o, given a theme t can then be described by a multinomial distribution $p(o|t,\beta)$ [a multicategory generalization of the binomial distribution, where the number of categories is given by the number of bins in the distribution and is parametrized by a set of parameters β]. In addition, we assume that the likelihood of a given theme contributing to any given event (and thus jet) $p(t|\omega)$ is also described by some multinomial distribution (parametrized by variables ω), where the number of categories now corresponds to the number of themes. The ω's themselves are drawn from a probability distribution $p(\omega|\alpha)$, reflecting the theme proportions in the dataset and parametrized by the *hyperparameter* α. In this picture the themes (β) as well as theme proportions (ω) are hidden variables reflecting the thematic structure of the studied event sample. With a given model, the probability that a certain event or jet distribution bin is populated can be written as a compact expression in terms of the latent variables. For example, the likelihood of generating a jet represented in terms of observables $j = (o_1, o_2, \ldots, o_n)$ is just

$$p(j|\alpha,\beta) = \int_\omega p(\omega|\alpha) \prod_{o\in j} \left(\sum_t p(t|\omega) p(o|t,\beta) \right) d\omega. \quad (1)$$

Statistical models defined in this way are generative in that given the latent variables (themes and theme proportions) the best model will be the one that best reproduces a set of jets or events, i.e., has the best generative power. Therefore, the task of finding the latent variables from a set of training events is specifically to invert the above expression and use the set of events to find the best fit for the latent variables. This can in fact be done using posterior Bayesian inference, i.e.,

$$p(a|x) \propto p(x|a) * p(a), \quad (2)$$

where $p(x|a)$ is the likelihood of observing x given a latent variable a, while $p(a)$ and $p(a|x)$ are the prior and posterior distributions of the latent variable itself. The main insight here is that $p(\omega|\alpha)$ in Eq. (1) is a conjugate prior to the multinomial likelihood $p(t|\omega)$ and thus forms the multicategory generalization of the beta distribution—the Dirichlet distribution. The model is thus called LDA and can be solved approximately (*trained*) in an iterative manner using variational inference [45,49] or Gibbs sampling [52].

The generative model defined by Eq. (1) does not include the conditional probabilities $p(o_i|o_{i-1})$ describing the ordering present in the (binary) clustering tree of the jet (or correspondingly in a Markov chain Monte Carlo jet generator). Therefore, the jet observables at each clustering step are assumed to be "conditionally independent," [49] i.e., they only depend conditionally on the same latent distributions (β, ω) of the model. This is reminiscent with the *bag-of-words* assumption widely used in probabilistic text modeling where the semantic structure relating different words in the vocabulary is completely neglected in the generative process for documents. While this simplifying assumption, of neglecting the clustering order information in jets, forbids us to use the probabilistic model (1) as a reliable jet or event generator,[2] it still comes in useful for jet or event classification tasks. As we show below, the LDA generative model is flexible enough to capture hidden features in the jet-clustering history, in particular, features produced by the decay chains of massive resonances.

Formally, a trained LDA model consists of the latent variables inferred from the training data and the probabilistic generative model used in constructing Eq. (1). In order to classify jets or events, we can perform statistical inference on the test sample. Once the LDA model is trained, the theme proportions $[\omega_t(j)]$ present in each new jet j (or event) can be estimated by maximizing the likelihood function for j while keeping the theme

[2]In the same way most generative text models cannot be used as reliable document generators.

distributions (β) fixed. As a result, each jet is described by a mixture of themes with corresponding weights $\omega_t(j)$ that can be directly used for classification. Since the extracted mixtures satisfy $\sum_t \omega_t(j) = 1$ and here we are focusing on only two themes (i.e., $t = 0, 1$) it suffices to choose just one of the weights to describe the jet. In this case, we define a simple classifier $h(j) = \omega_1(j)$ based on the proportion of one of the themes in the jet (or event).

Alternatively, one can directly use the latent themes $p(o|t_{1,2})$ discovered by the LDA algorithm and compute the likelihood ratio $\mathcal{L}(j) = p(j|t_1)/p(j|t_2)$ for every new jet j (or event) in a test sample, and use it as the classifier. While the likelihood ratio is known to be the optimal classifier given exact knowledge of pure distributions [53], it has been shown recently, that it remains optimal even for mixed distributions of *a priori* unknown but different mixture proportions [34]. Thus, $\mathcal{L}(j)$ is an optimal LDA classifier in the limit that the extracted themes correspond to pure distributions and the LDA model reduces to a simple mixture model. In general however, this will not be the case and we have checked explicitly that the inference and $\mathcal{L}(j)$ based classifiers based on LDA perform comparably. In the remainder of the paper we only present results based on the inference classifier $h(j)$.

III. UNSUPERVISED TOP TAGGER

Our first proof of principle example is a tagger discriminating between boosted hadronically decaying top quarks and QCD jets. Working with a single mixed ($t\bar{t}$ and QCD) multijet event sample we first need to construct the relevant jet substructure observable histograms (o). We do this by clustering the jets in an event using the Cambridge-Aachen (CA) [54,55] algorithm with a large radius R. We then proceed to uncluster the jets by reversing each step in the clustering, iteratively separating each (sub)jet into two objects $j_0 \to j_1 j_2$. Ordering the subjets by their invariant mass $m_{j_1} > m_{j_2}$ (and following the standard approach of Refs. [1,3]), we define the relevant clustering observables at each clustering step as

$$o_{j_0} = \left\{ m_{j_0}, \frac{m_{j_1}}{m_{j_0}}, \frac{m_{j_2}}{m_{j_1}}, \frac{\min(p_{T,1}^2, p_{T,2}^2)}{m_{j_0}^2} \Delta R_{1,2}^2 \right\}, \quad (3)$$

where $p_{T,i}$ is the transverse momentum of a given object j_i and $\Delta R_{1,2}^2 = (\phi_1 - \phi_2)^2 + (\eta_1 - \eta_2)^2$ is the so-called planar distance between j_1 and j_2 (ϕ_i and η_i being the azimuthal angle and pseudorapidity of j_i, respectively). The declustering step is then iteratively repeated on both $j_{1,2}$. The procedure is terminated once $m_{j_0} < m_{\min}$, where $m_{\min}$ is an algorithm parameter, which we choose to lie below the lowest massive resonance state of interest. In the case of the top tagger, we fix $m_{\min} = 30$ GeV $\ll m_W$, but have checked that lowering this threshold by a factor of a few does not significantly affect the results. The output of such a procedure is a (typically a rather sparse) four-dimensional histogram of o_j which can be defined either *per jet* or even *per event*. After mapping individual histogram bins into *words*, we feed individual jets or events as *documents* into an LDA implementation using the software package Gensim [56,57], fixing the number of themes to two ($\omega_{0,1}$). Further technical details of the required binning and mapping of data onto (one-dimensional) text vocabularies compatible with Gensim, as well as a detailed analysis of the convergence of the algorithm when applied on sparse jet substructure data will be presented elsewhere [58]. Here we only focus on the consistency and stability of the resulting trained models. For this purpose we use the k-folding method with $k = 10$. This involves splitting the training data into k different mutually exclusive blocks and then running the training k times on event samples built from $k - 1$ blocks, with the combination changing on each training run. The performance of the tagger is tested on events or jets from the remaining block.

In order to evaluate the performance of the tagger and compare it to existing methods, we construct a receiver operating characteristic (ROC) curve for our tagger. This is the only step where one needs to rely on access to pure samples (either MC generated or pre-tagged in some other way using observables orthogonal to o_j). In particular, we construct the ROC curve by performing the classification on such pure samples while continuously varying the threshold of the theme proportion defining the classifier $h(j)$. This is done for all k sets of results and we calculate the median mistag rate (ε_b) for each signal efficiency (ε_s), as well as the mean absolute deviation of the mistag rate to evaluate the stability and consistency of the tagger.

Our training samples for the QCD dijet background and the (hadronic) $t\bar{t}$ signal both consist of $\sim$84,000 13 TeV pp collision events, where the final state particles are clustered into $R = 1.5$ CA jets with p_T in the range [350, 450] GeV. The samples are generated using aMC@NLO 2.6.1 [59] interfaced with Pythia 8.2 [60] for showering and hadronization, while jet clustering is performed using FastJet 3.2.0 [61]. Note that no grooming is performed on the jets. We have also checked explicitly that applying jet (sub)cluster energy smearing consistent with the parametric fast detector simulation of ATLAS implemented in Delphes 3.4.1 [62] has no significant effect on our results.

We train the top tagger on four test cases: supervised, and unsupervised mixed samples with $S/B = 1, 1/9, 1/99$. In the supervised case we collapse the pure samples into single documents such that they are processed by the algorithm in a single block, essentially providing the labeling of the data required in supervised algorithms. For the different S/B ratios each jet or event is represented by a single document. However, we inform the tagger to search for certain S/B ratios by setting the hyperparameters of the Dirichlet distribution accordingly, i.e., $\alpha = [0.5, 0.5]$, [0.9, 0.1], and [0.99, 0.01]. Note that these may not be the

optimal choices, but they are based on the intuition from the values of S/B and give a useful parametrization to demonstrate the performance of the algorithm. We also stress that $\mathcal{O}(1)$ variations in α have only a small effect on the performance of the algorithm provided that the hierarchy in the elements of α approximately reflect the S/B ratio, and that the elements are smaller than one. More details on the dependence of the algorithm on these hyperparameters, and how to determine their optimal values without prior knowledge of the S/B ratios, will be presented elsewhere [58].

In Fig. 1 (upper panel) we plot the ROC curves for our top-jet taggers, where separate documents are represented by individual jets, and compare these to various supervised taggers in the literature [8,22,23]. We see that the taggers perform well and with relatively small variance, with the supervised tagger performing the best. An interesting observation is that at high background rejection rates $[1/\epsilon_b \gg \mathcal{O}\ (\text{few})]$ the taggers trained on smaller S/B perform slightly better than the tagger trained on the $S/B = 1$ sample, although the differences are comparable to the estimated uncertainties. This is essentially because the algorithm is designed to discern features in the jet substructure, which are subsequently used to tag jets and events. In the supervised and $S/B = 1$ case the algorithm discovers features in top jets both near $m_{j_0} \sim m_t$ and $m_{j_0} \sim m_W$ (see the right plot in Fig. 2), while in the lower S/B cases the algorithm is only able to identify $m_{j_0} \sim m_t$ as relevant. On the other hand, lower m_{j_0} regions generically feature more prominently in QCD jets (see left plot in Fig. 2). Thus, while a very accurate determination of the features near $m_{j_0} \sim m_W$ in the supervised case helps the performance of the tagging algorithm, the worse resolution in the unsupervised $S/B = 1$ case leads to worse tagging performance compared to lower S/B examples. We see that the performance of the unsupervised taggers is comparable to the original Johns Hopkins (JH) top tagger [8], although it falls short in comparison to the others. We note that the observables we use mostly match those used in the JH top tagger, hence the similar performance is indeed encouraging.

In Fig. 1 (lower panel) we plot the ROC curves for our $t\bar{t}$ event classifiers, where a single document now contains all jets within the selected p_T region in an event, and again compare these to the top-jet taggers in the literature. To make the comparison with other taggers fair, we rescale those results by defining an event tagging efficiency (ϵ_e) in terms of the jet tagging efficiency (ϵ_j) and the fraction of

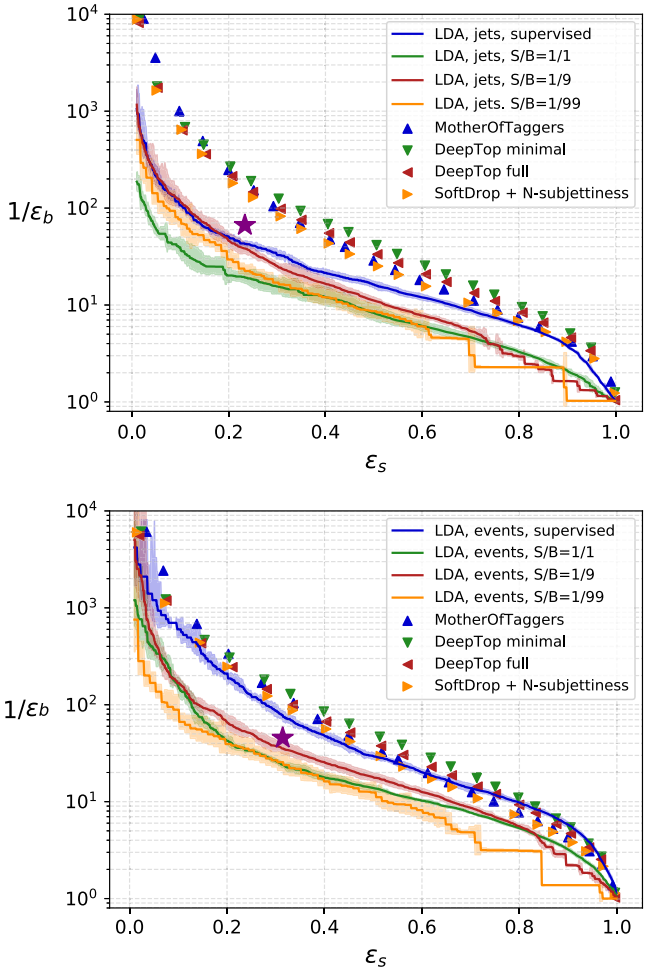

FIG. 1. (Upper plot) ROC curves for the LDA top-jet taggers compared to the DeepTop tagger [22,23] (colored triangles) for events with fat jets satisfying $p_T \in [350, 450]$ GeV. The purple star represents the default JH top tagger [8] reference point. (Lower plot) ROC curves for the $t\bar{t}$ LDA event classifiers compared to the classifiers from the DeepTop (colored triangles) and the JH top tagger (purple star). In both plots the shaded bands represent the mean-average deviation extracted from the k-folding procedure. See text for details.

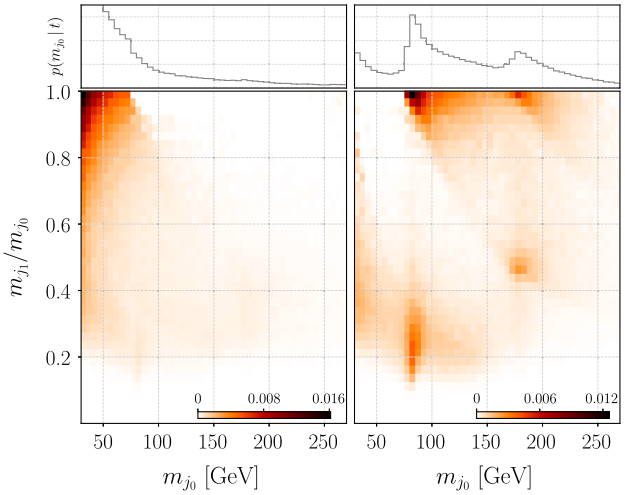

FIG. 2. 2D projected probability distributions (in the plane of m_{j_0} and m_{j_1}/m_{j_0}) of the two latent themes discovered in mixed $(S/B = 1)$ QCD and $t\bar{t}$ event samples with fat jets satisfying $p_T \in [350, 450]$ GeV.

events in our pure samples with one (f_1) and two (f_2) jets passing the selection cuts,[3] $\epsilon_e = (2\epsilon_j - \epsilon_j^2)f_2 + \epsilon_j f_1$. This means in practice that tagging an event as $t\bar{t}$ requires at least one jet in the event to be tagged as a top jet. The ROC curves do not change significantly under this rescaling, instead the points move along a trajectory towards higher efficiencies approximately equal to that of the ROC curve for jet tagging. We see again that the classifier performs very well in all cases, performing as well as the JH top tagger even for low S/B.

We observe that the LDA algorithm performs relatively better when characterizing and tagging events than jets, mainly due to the larger amount of substructure (*words*) in each document. With more data per document it is easier for the algorithm to identify co-occurrences between the different features shared by jets in the same event. For this reason it is also easier for the trained model to infer the correct thematic structure from events, than from jets.

The themes discovered by the unsupervised training algorithm contain valuable information about the substructure of the events or jets. In Fig. 2 we plot the substructure probability distributions of the two themes discovered by the top-jet tagger (with $S/B = 1$) projected onto the plane of m_{j_0} and m_{j_1}/m_{j_0}. We observe that while the distribution on the left-hand side plot (the "QCD" theme) is fairly unremarkable (mostly monotonic and smooth) and peaks towards ($m_j \to 0$, $m_{j_1}/m_{j_0} \to 1$), the theme on the right-hand side plot (the "$t\bar{t}$" theme) clearly exhibits a heavily weighted feature at both $m_{j_0} \sim m_t$ and $m_{j_0} \sim m_W$, even identifying the W subjet arising from the decay of the top quark within the jet resulting in a mass drop of $m_{j_1}/m_{j_0} \sim m_W/m_t \simeq 0.45$. On the other hand, the broad $m_{j_1}/m_{j_0} \sim 0.2 \gtrsim 0$ feature at $m_{j_0} \sim m_W$ is expected due to the fact that the mass drop is defined with the heaviest daughter subjet in the numerator thus skewing the m_{j_1}/m_{j_0} distribution away from zero.

IV. UNSUPERVISED NP SEARCH

As a second example, we consider a NP model [63,64] containing a heavy W' boson plus a heavy scalar ϕ. Signal events thus consist of resonant W' production (at $m_{W'} = 3$ TeV), followed by $W' \to W\phi$ decays (where we choose $m_\phi = 400$ GeV $\ll m_{W'}$ such that both the W and the ϕ coming from W' decays are boosted). Finally, the scalar further decays as $\phi \to W^+W^-$. Using the same event generation, jet-clustering/declustering procedure, observable basis o_j, and the same LDA tagging algorithm as before, we apply our procedure to the all-hadronic final state of this NP process in a region dominated by QCD background. The same model has been previously studied using the unsupervised ML approach called *classification*

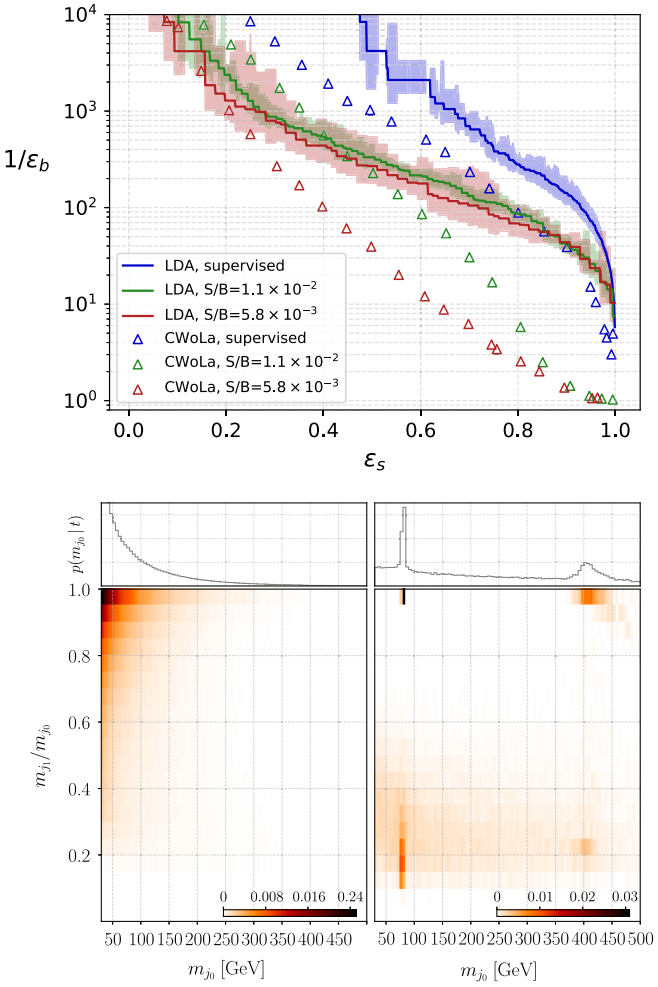

FIG. 3. (Upper plot) ROC curves comparing the performance of the LDA event classifier to CWoLa [38]. (Lower plot) 2D projected probability distributions (in the plane of m_{j_0} and m_{j_1}/m_{j_0}) of the two latent themes discovered in mixed ($S/B = 1.1 \times 10^{-2}$) QCD and W' event samples with invariant mass $2730 \le m_{jj} \le 3189$ GeV with fat jets satisfying $p_T > 400$ GeV.

without labels (CWoLa) [37,38]. It is based on mixed sample classification using phase space regions with vastly different S/B ratios processed by deep NNs. In order to quantitatively compare our results to CWoLa, our signal and background event samples mirror directly those in Ref. [38]. In particular, we consider just the signal region, $2730 \le m_{jj} \le 3189$ GeV, and cut jets with p_T below 400 GeV. The 30 GeV cut on the subjet invariant mass is also applied, just as in the top tagger case. After the selection cuts we work with ~60,000 events in both the signal and background samples. We train three different taggers; a supervised tagger, and two taggers with $S/B = 1.1 \times 10^{-2}$ and 5.8×10^{-3}. The Dirichlet hyperparameters α are chosen in the same way as in the previous section, i.e., $\alpha = [0.5, 0.5]$, [0.989, 0.011], and [0.942, 0.058].

[3]We have checked that the fractions of events with zero or more than two jets passing the selection cuts are negligible.

To evaluate the robustness of the taggers we again employ the k-folding procedure with $k = 10$.

In the upper plot of Fig. 3 we show the ROC curves for our taggers and compare the results to those from CWoLa [38]. We see that in most of the parameter space the LDA-based tagger outperforms the CWoLa tagger, most notably at high signal efficiencies.

In the lower plot of Fig. 3 we also show the probability distributions of the discovered themes in the plane of m_{j_0} and m_{j_1}/m_{j_0} for the LDA model trained on event samples with $S/B = 1.1 \times 10^{-2}$. Features in the subjet mass at $m_{j_0} \sim m_W$ and at $m_{j_0} \sim m_\phi$ are clearly discernible in one of the themes (the "ϕW" theme), as well as mass drops related to the decays of the heavy scalar and the W bosons.

V. CONCLUSIONS

We have demonstrated a new unsupervised ML technique for disentangling signal and background events in mixed samples by identifying features in jet substructure observables that differentiate between the two. To do so we have mapped jet substructure distributions onto a LDA model, a generative probabilistic model (*mixed membership model*) widely used in Bayesian statistics approaches to unsupervised ML. Assuming that the kinematic observable distributions within jets or events are sampled from a fixed set of (latent) themes, LDA can learn the thematic structure that most likely generated the observed data (the later being either in the form of reconstructed real LHC events or unlabeled MC-generated samples). Furthermore, we have shown that the learned structure from a two-theme LDA model can be used to build unsupervised jet taggers or event classifiers that efficiently discriminate between signal and background in previously unseen data.

As a first example we have trained a two-theme LDA model on MC-generated event samples consisting of different mixtures of $pp \to t\bar{t}$ and QCD dijet events. Our results show that the top-jet taggers and $t\bar{t}$ event classifiers constructed from the discovered themes have a very good discrimination power when applied to previously unseen pure samples, even if trained on data with S/B ratios as low as 1%. Our results are in some cases comparable even with fully supervised taggers in the literature. In addition we have explored the viability of LDA discovering NP phenomena in multijet events. Using a benchmark NP (vector W', scalar ϕ) model we have studied $pp \to W' \to \phi W \to WWW$ with hadronically decaying W bosons and a (boosted) new scalar ϕ with mass $m_\phi \ll m_{W'}$. The resulting LDA event classifiers from training samples with S/B as low as a few per-mille, when applied to pure samples, produce excellent signal efficiencies and QCD rejection rates that can outperform other existing approaches.

Besides being a fully unsupervised ML technique, one advantage of performing LDA on jet-clustering history observables, is the possibility of interpreting the thematic structure discovered by the model from the data. In both examples presented here, the features in the probability distributions over the kinematical observables of the two uncovered themes match to a high degree the expected features of the underlying hard processes—hadronic decays of top quarks (or $\phi \to W^+W^-$) and the QCD background, respectively, allowing for an intuitive and physical understanding of the high tagging performance as demonstrated by the ROC curves.

The analysis presented here is a first exploration of what can be achieved when applying probabilistic mixed membership models to high-energy collider data. For example, with the addition of more jet substructure observables the discriminating power of the LDA classifiers could be further optimized and increased. Furthermore, relaxing the fixed number of themes of the LDA model applied to mixed event samples could allow us to classify multiple backgrounds together with the signal. In future work we will also detail how these techniques can be employed as part of a broad search strategy for new phenomena in multijet invariant mass spectra with the aim of performing unsupervised data-driven searches for NP at high p_T.

ACKNOWLEDGMENTS

We thank Jasna Urbančič, Erik Novak and Klemen Kenda for initial involvement in the project as well as Jack Collins for generously providing the $W' - \phi$ NP model implementation for use in aMC@NLO. We also thank César A. Ojeda and Bryan Zaldivar for useful discussions. D. A. F. is supported by the Young Researchers Programme of the Slovenian Research Agency under Grant No. 37468. J. F. K. and B. M. D. acknowledge the financial support from the Slovenian Research Agency (research core funding No. P1-0035 and No. J1-8137).

[1] 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**, 242001 (2008).

[2] J. M. Butterworth, B. E. Cox, and J. R. Forshaw, WW scattering at the CERN LHC, Phys. Rev. D **65**, 096014 (2002).

[3] J. M. Butterworth, J. R. Ellis, and A. R. Raklev, Reconstructing sparticle mass spectra using hadronic decays, J. High Energy Phys. 05 (2007) 033.
[4] Y. Cui, Z. Han, and M. D. Schwartz, W-jet tagging: Optimizing the identification of boosted hadronically-decaying W bosons, Phys. Rev. D **83**, 074023 (2011).
[5] W. Skiba and D. Tucker-Smith, Using jet mass to discover vector quarks at the LHC, Phys. Rev. D **75**, 115010 (2007).
[6] B. Holdom, t-prime at the LHC: The physics of discovery, J. High Energy Phys. 03 (2007) 063.
[7] M. Gerbush, T. J. Khoo, D. J. Phalen, A. Pierce, and D. Tucker-Smith, Color-octet scalars at the CERN LHC, Phys. Rev. D **77**, 095003 (2008).
[8] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, Top tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks, Phys. Rev. Lett. **101**, 142001 (2008).
[9] L. G. Almeida, S. J. Lee, G. Perez, I. Sung, and J. Virzi, Top Jets at the LHC, Phys. Rev. D **79**, 074012 (2009).
[10] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung, and J. Virzi, Substructure of high-pT Jets at the LHC, Phys. Rev. D **79**, 074017 (2009).
[11] L. G. Almeida, S. J. Lee, G. Perez, G. Sterman, and I. Sung, Template overlap method for massive jets, Phys. Rev. D **82**, 054034 (2010).
[12] M. Backović and J. Juknevich, TemplateTagger v1.0.0: A template matching tool for jet substructure, Comput. Phys. Commun. **185**, 1322 (2014).
[13] T. Plehn, G. P. Salam, and M. Spannowsky, Fat Jets for a Light Higgs, Phys. Rev. Lett. **104**, 111801 (2010).
[14] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, Stop reconstruction with tagged tops, J. High Energy Phys. 10 (2010) 078.
[15] D. E. Soper and M. Spannowsky, Finding top quarks with shower deconstruction, Phys. Rev. D **87**, 054012 (2013).
[16] A. J. Larkoski, I. Moult, and B. Nachman, Jet substructure at the large hadron collider: A review of recent advances in theory and machine learning, arXiv:1709.04464.
[17] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, Jet-images: Computer vision inspired techniques for jet tagging, J. High Energy Phys. 02 (2015) 118.
[18] L. G. Almeida, M. Backovic, M. Cliche, S. J. Lee, and M. Perelstein, Playing tag with ANN: Boosted top identification with pattern recognition, J. High Energy Phys. 07 (2015) 086.
[19] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, Jet-images—deep learning edition, J. High Energy Phys. 07 (2016) 069.
[20] 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**, 094034 (2016).
[21] 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**, 014018 (2017).
[22] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, Deep-learning top taggers or the end of QCD, J. High Energy Phys. 05 (2017) 006.
[23] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, Deep-learned top tagging with a Lorentz layer, SciPost Phys. **5**, 028 (2018).
[24] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, Deep learning in color: Towards automated quark/gluon jet discrimination, J. High Energy Phys. 01 (2017) 110.
[25] G. Louppe, K. Cho, C. Becot, and K. Cranmer, QCD-aware recursive neural networks for jet physics, J. High Energy Phys. 01 (2019) 057.
[26] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, Jet constituents for deep neural network based top quark tagging, arXiv:1704.02124.
[27] K. Datta and A. Larkoski, How much information is in a jet, J. High Energy Phys. 06 (2017) 073.
[28] K. Datta and A. J. Larkoski, Novel jet observables from machine learning, J. High Energy Phys. 03 (2018) 086.
[29] K. Fraser and M. D. Schwartz, Jet charge and machine learning, J. High Energy Phys. 10 (2018) 093.
[30] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, JUNIPR: A framework for unsupervised machine learning in particle physics, Eur. Phys. J. C **79**, 102 (2019).
[31] S. Macaluso and D. Shih, Pulling out all the tops with computer vision and deep learning, J. High Energy Phys. 10 (2018) 121.
[32] K. Datta, A. Larkoski, and B. Nachman, Automating the construction of jet observables with machine learning, arXiv:1902.07180.
[33] L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman, Weakly supervised classification in high energy physics, J. High Energy Phys. 05 (2017) 145.
[34] E. M. Metodiev, B. Nachman, and J. Thaler, Classification without labels: Learning from mixed samples in high energy physics, J. High Energy Phys. 10 (2017) 174.
[35] 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**, 011502 (2018).
[36] T. Cohen, M. Freytsis, and B. Ostdiek, (Machine) learning to do more with less, J. High Energy Phys. 02 (2018) 034.
[37] J. H. Collins, K. Howe, and B. Nachman, Anomaly Detection for Resonant New Physics with Machine Learning, Phys. Rev. Lett. **121**, 241803 (2018).
[38] J. H. Collins, K. Howe, and B. Nachman, Extending the search for new resonances with machine learning, Phys. Rev. D **99**, 014038 (2019).
[39] J. A. Aguilar-Saavedra, J. H. Collins, and R. K. Mishra, A generic anti-QCD jet tagger, J. High Energy Phys. 11 (2017) 163.
[40] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, Novelty detection meets collider physics, arXiv:1807.10261.
[41] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or What? SciPost Phys. **6**, 030 (2019).
[42] M. Farina, Y. Nakai, and D. Shih, Searching for new physics with deep autoencoders, arXiv:1808.08992.
[43] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Variational autoencoders for new physics mining at the large hadron collider, J. High Energy Phys. 19 (2019) 36.
[44] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. **41**, 391 (1990).
[45] T. Hofmann, Probabilistic latent semantic analysis, in *UAI'99 Proceedings of the Fifteenth conference on*

Uncertainty in artificial intelligence, Stockholm, Sweden (1999), pp. 289–296.

[46] K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell, Text classification from labeled and unlabeled documents using EM, *Machine Learning* (Kluwer Academic Publishers, Dordrecht, 2000), Vol. 39, pp. 103–134, https://doi.org/10.1023/A:1007692713085.

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

[48] P. T. Komiske, E. M. Metodiev, and J. Thaler, An operational definition of quark and gluon jets, J. High Energy Phys. 11 (2018) 059.

[49] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, Latent dirichlet allocation, J. Mach. Learn. Res. **3**, 2003 (2003).

[50] K. Agashe, J. H. Collins, P. Du, S. Hong, D. Kim, and R. K. Mishra, Detecting a boosted Diboson resonance, J. High Energy Phys. 11 (2018) 027.

[51] J. K. Pritchard, M. Stephens, and P. Donnelly, Inference of population structure using multilocus genotype data, Genetics **155**, 945 (2000).

[52] T. L. Griffiths and M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. U.S.A. **101**, 5228 (2004).

[53] J. Neyman, On the problem of the most efficient tests of statistical hypotheses, Phil. Trans. R. Soc. A **231**, 289 (1933).

[54] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, Better jet clustering algorithms, J. High Energy Phys. 08 (1997) 001.

[55] M. Wobisch and T. Wengler, Monte Carlo generators for HERA physics, in *Proceedings, Workshop, Hamburg, Germany, 1998–1999*, edited by A. T. Doyle, G. Grindhammer, G. Ingelman, and H. Jung (1998), pp. 270–279.

[56] R. Řehůřek and P. Sojka, Software framework for topic modelling with large corpora, in *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (University of Malta, Valletta, Malta, 2010), pp. 46–50.

[57] M. Hoffman, D. M. Blei, and F. Bach, Online learning for latent dirichlet allocation, Advances in Neural Information Processing Systems 856 (2010), Vol. **23**, pp. 856–864.

[58] B. Dillon, D. A. Faroughy, and J. F. Kamenik (to be published).

[59] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.

[60] T. Sjostrand, S. Mrenna, and P. Z. Skands, A brief introduction to PYTHIA 8.1, Comput. Phys. Commun. **178**, 852 (2008).

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

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

[63] K. Agashe, P. Du, S. Hong, and R. Sundrum, Flavor universal resonances and warped gravity, J. High Energy Phys. 01 (2017) 016.

[64] K. Agashe, J. H. Collins, P. Du, S. Hong, D. Kim, and R. K. Mishra, Dedicated strategies for triboson signals from cascade decays of vector resonances, Phys. Rev. D **99**, 075016 (2019).