Exploring unsupervised top tagging using Bayesian inference

Ezequiel Alvarez[1*], Manuel Szewc[2†], Alejandro Szynkman[3‡], Santiago Tanco[3§],
Tatiana Tarutina[4¶]

[1]*International Center for Advanced Studies (ICAS) and CONICET, UNSAM,
Campus Miguelete, 25 de Mayo y Francia, CP1650, San Martín, Buenos Aires, Argentina*

[2]*Department of Physics, University of Cincinnati,
Cincinnati, Ohio 45221,USA*

[3]*IFLP, CONICET - Dpto. de Física, Universidad Nacional de La Plata,
C.C. 67, 1900 La Plata, Argentina*

[4]*IFLP, CONICET,
Diagonal 113 e/ 63 y 64, 1900 La Plata, Argentina*

Abstract

Recognizing hadronically decaying top-quark jets in a sample of jets, or even its total fraction in the sample, is an important step in many LHC searches for Standard Model and Beyond Standard Model physics as well. Although there exists outstanding top-tagger algorithms, their construction and their expected performance rely on Montecarlo simulations, which may induce potential biases. For these reasons we develop two simple unsupervised top-tagger algorithms based on performing Bayesian inference on a mixture model. In one of them we use as the observed variable a new geometrically-based observable $\tilde{A}_3$, and in the other we consider the more traditional τ_3/τ_2 N-subjettiness ratio, which yields a better performance. As expected, we find that the unsupervised tagger performance is below existing supervised taggers, reaching expected Area Under Curve AUC $\sim 0.80 - 0.81$ and accuracies of about $69\% - 75\%$ in a full range of sample purity. However, these performances are more robust to possible biases in the Montecarlo that their supervised counterparts. Our findings are a step towards exploring and considering simpler and unbiased taggers.

*sequi@unsam.edu.ar
†szewcml@ucmail.uc.edu
‡szynkman@fisica.unlp.edu.ar
§santiago.tanco@fisica.unlp.edu.ar
¶tarutina@fisica.unlp.edu.ar

1 Introduction

After an outstanding successful first half of the LHC lifespan [1, 2], we are heading for a second half in which one of the most important features is data gathering. Consequently, one of the main concerns for the community is how to take best and full advantage of a unique dataset of protons colliding in a controlled environment. Among the most relevant studies for this stage are those including heavy particles such as the top quark, the Higgs boson and the heavy gauge bosons Z and W, since they are not only important by themselves, but also because there are theoretical arguments to expect that these are the Standard Model (SM) particles with larger couplings to New Physics (NP). Improving identification techniques for these particles in a variety of scenarios at the LHC is a crucial tool for the search of NP. Along this work we study innovative techniques for identifying boosted top quarks within a framework of Bayesian unsupervised Machine Learning.

The top quark was discovered at Tevatron in 1995 [3,4], and since then it has become an object of exhaustive analysis [5]. With the start of the LHC, which can be seen as a top-factory, the study of the top quark entered into a new stage [6–11] with unique challenges. In particular, because of the high energies achieved at the LHC, highly boosted top quarks with transverse momentum above $p_T \gtrsim 300$ GeV are copiously produced. In this kinematic regime, the top decay products become collimated and the top jet is reconstructed as a fat jet by the standard jet clustering algorithms. Consequently, the techniques for identifying top quarks change drastically with respect to the resolved regime, relying on the analysis of the jet substructure. The study of top jet substructure for top quark identification has developed into an area of collider physics coined as *top-tagging*, which has reached many milestones [12].

The main challenge within this area is to distinguish hadronically decaying top quarks collimated as fat jets, which we shall call *top jets*, and jets that are induced by quarks or gluons, which we will name hereafter as *QCD jets*. This is of course a very difficult task, since it implies –among others– understanding and/or modelling hadronization, splittings and showering processes which occur at an energy level in which QCD is not totally understood. Traditional top taggers combine standard clustering algorithms, jet grooming techniques, declustering, mass drop, and soft drop

among others, in order to search for kinematic features in the jet [13–17]. With the emergence of high performance Machine Learning algorithms and associated boost in available hardware, even more powerful techniques have been developed [18], usually relying on Neural Networks trained to distinguish top from QCD jets using as a training set labelled Montecarlo generated samples. In the last few years, these techniques have improved considerably by defining and using specific objects within jets, such as for instance the Energy Flow Polynomials [19]. The state of the art in these techniques reach an oustanding tagging performance which is measured as an area under the ROC curve, AUC $\approx$ 0.988, and an accuracy of about 94% [20, 21] using the standard Top Quark Tagging Reference Dataset [22]. However because these are supervised techniques, the tagger performance is contingent on the specific Montecarlo simulations considered for labelled training and benchmarking at a given working point within the ROC curve. Any potential bias in the Montecarlo would lead to a bias in the use of the algorithm, and thus in any conclusions drawn from it; see for instance recent discussion in [23, 24]. Because of this, it is worth exploring other methods, such as unsupervised top-tagging techniques, even though they may be below the performances announced within supervised frameworks. An unsupervised approach, such as Latent Dirichlet Allocation (LDA) [25], could be run directly over experimental data, bypassing potential simulation biases, but generally giving a less efficient tagging performance.

In this manuscript we focus on exploring other unsupervised alternatives, by means of Bayesian inference over a mixture model [26] to simultaneously learn about top and QCD jet features, estimate the mixture fractions in a given sample, and create a top-tagging algorithm. Bayesian techniques are based in designing a Probability Density Function (PDF) from which one assumes that the data is sampled, and then, using the measured data, a probability distribution for the parameters of said PDF is inferred. The advantage comes from using robust knowledge to construct the PDF and leaving the unknowns as random variables whose distribution is to be inferred. A mixture model considers the particular case where each jet belongs to a given class, which in this manuscript are either top or QCD jets. Mixture models have previously been applied in High Energy Physics for samples containing signal and background [27–29]. In designing the PDF that can distinguish top jets from QCD jets, one should use robust knowledge. Along this manuscript we exploit the fact that the top has a well known hadronic decay to three particles, $t \to Wb \to jjb$. We work within two different frameworks, one of which appeals to a simple geometrical representation of the process whereas the other employs N-subjettiness [30] observables. Implementing these techniques in such a way that the robust knowledge is used to determine the observable and the shape of its PDF, while the unknown parameters left to be inferred from the data, is part of the *art et métiers* of a Bayesian inference framework, and it is what is pursued in the next Sections.

This work is divided as follows. In the next Section we discuss Bayesian inference techniques on mixture models for tagging at the LHC. In Sec. 3 we perform inference on a mixture model in which the observed variable is the third coefficient of the Fourier transform of the distribution of tracks around the center of the jet. In Sec. 4 we infer on a mixture model which takes as observed variable the well-known N-subjettiness ratio τ_3/τ_2. In Sec. 5 we discuss the obtained results and we summarize the conclusions of the article.

2   Mixture models for top-tagging at the LHC

In this Section we provide a brief summary of mixture models oriented to top-tagging. For each jet, we define a tagging observable x whose distribution we propose as a mixture model, in which

the likelihood function for a collection X of N independent measurements of this observable can be expressed as

$$p(X) = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \sum_{k=\{0,1\}} \pi_k \, p(x_i|k) \, , \tag{1}$$

where π_k is the probability of sampling a jet from jet class k and $p(x|k)$ the class-dependent probability functions for the tagging observable x. We use $k = 0\,(1)$ to denote QCD (top) jets. The mixture model is defined by providing appropriate probability densities $p(x|k)$ which need to capture the relevant physics. These probability densities are typically parametric functions dependent on a set of a few latent parameters ζ_k. In typical frequentist fashion, one can then find the set of parameters $\zeta = \{(\pi_k, \zeta_k), k = 0, 1\}$ that most accurately describe the observed data distribution by maximizing the Likelihood introduced in Eq. 1. In this work, we consider a Bayesian framework where the parameters are promoted to random variables and it is possible to obtain their posterior probability distribution $p(\zeta|X)$ by first proposing a prior probability $p(\zeta)$ and using Bayes' theorem,

$$p(\zeta|X) \propto p(X|\zeta)p(\zeta) \, , \tag{2}$$

where $p(X|\zeta)$ is given by Eq. (1). The posterior distribution can be estimated via approximate Bayesian inference techniques. Alternatively, there are techniques to directly obtain the set of model parameters that maximize the posterior probability, i.e. the maximum a posteriori (MAP). Along this work we use the latter option through Stochastic Variational Inference (SVI) [31], via the `Pyro` package [32] in `Python3`.

In order to construct a top tagger from the Bayesian inference process described above, one can get the MAP using SVI, and compute the assignment probability

$$p(\text{QCD}|x, \zeta^{\text{MAP}}) = \frac{\pi_0^{\text{MAP}} p(x|\zeta_0^{\text{MAP}})}{\sum_{k=\{0,1\}} \pi_k^{\text{MAP}} p(x|\zeta_k^{\text{MAP}})} \, , \tag{3}$$

with $p(\text{top}|x, \zeta^{\text{MAP}}) = 1 - p(\text{QCD}|x, \zeta^{\text{MAP}})$. A given jet with an observed value of x is then tagged as a QCD jet if its assignment probability is higher than a selected threshold, $p(\text{QCD}|x, \zeta^{\text{MAP}}) > t$ with $0 < t < 1$.

In the following Sections we explore two choices for the tagging observable x and its associated proposed mixture model. In Sec. 3 we introduce a new variable based on the 3-pronged nature of top jets and defined on the Fourier space. In Sec. 4 we use the ratio of N-subjettiness variables τ_3/τ_2, which shows a good discriminating power for tops.

Although we design an unsupervised algorithm that aims to be applicable directly to data, we make use of simulations to benchmark our proposal. To compare with other approaches more directly, we use the standard Top Quark Tagging Reference Dataset [22], which consists of 1.2M training events, 400k testing events and 400k validation events generated via `Pythia8` [33] at c.m. energy of 14 TeV and with the default `Delphes` [34] simulation of the ATLAS detector response. Fat jets are reconstructed with the anti-k_T algorithm at $R = 0.8$ and are required to have $p_T \in (550, 650)$ GeV and $|\eta| < 2$. We work with the train dataset, first applying a selection cut in the invariant mass of the jets m_j, such that $145 \text{ GeV} < m_j < 205 \text{ GeV}$.

Using this dataset, we always consider three different metrics to evaluate the learned model: the learned fraction to be compared against the true fraction, the accuracy when tagging at the fixed probabilistic threshold $t = 0.5$ and the AUC. The last two are standard in the top-tagging

community. However, we emphasize that our method is not meant to maximize these supervised metrics through a smart choice of observables and algorithms. We only need our choice of observables and density modelling to perform well enough so that overcoming the biases of simulations is a clear benefit without a large cost in tagging performance.

3 Mixture model in the Fourier framework

We want to work with observables that take advantage of the well understood physical differences between QCD and top jets. In particular, since a top jet can be interpreted as a superposition of three subjets, we first introduce a geometrical variable that captures this feature. In the (θ, φ) space of the detector, each jet consists of a collection of coordinate pairs for each constituent track

$$(\text{Jet})_i = \{(\theta^{(j)}, \varphi^{(j)})_i\,,\ j = 1, \ldots, N_i\}\,. \tag{4}$$

The constituent coordinates are all shifted in such a way that the center of momentum of the jet coincides with $(\theta, \varphi) = (0, 0)$. We define the angle ϖ with respect to the θ positive axis, such that $\varpi \in (-\pi, \pi)$, and heretofore we reduce the jets in our dataset to a collection of N_i angles $\{\varpi_1, \varpi_2, \ldots, \varpi_{N_i}\}$, each corresponding to a particular track. Observe that we do not use the distance of the tracks to the center of momentum of the jet. We now take 20 equally sized bins in this variable, which define angular slices of the 2-dimensional plane (θ, φ), and count the number of tracks in each bin, resulting in an array of N_bins integer numbers,

$$(a_0, a_1, \ldots, a_{N_\text{bins}-1})\,, \tag{5}$$

for each event. The aforementioned reduction aims to make explicit an arbitrary degree of freedom for the (a_i) histogram given by a rotation of all the jet constituents by the same angle around the center. We eliminate this arbitrariness by evaluating a discrete Fourier transform (DFT),

$$A_l = \sum_{m=0}^{N_\text{bins}-1} a'_m\, e^{-2\pi i \frac{ml}{N_\text{bins}}}\,, \quad l = 1, ..., 10\,, \tag{6}$$

where $a'_i = a_i - \bar{a}$, and $\bar{a}$ is the mean of (a_i). Therefore, the baseline A_0 frequency is removed. Finally we normalize the $|A_l|$ values, by defining

$$\tilde{A}_l = \frac{|A_l|}{\sum_i |A_i|}\,. \tag{7}$$

As we expect the top jets to show a higher density of tracks around the centers of the three subjets, the DFT of the angular variable should show a tendency towards higher values in the $l = 3$ frequency of the top events with respect to the QCD background.

To illustrate this new set of variables we build a toy model, described in Appendix A, which also has two classes. These two classes reflect the expected differences between QCD and top jets, with the former generating random points in 2d space while the latter having a clear three-pronged structure. We show in Fig. 1 two typical examples of this pipeline for a toy model with a typical toy QCD (top) jet shown in the upper (lower) row. In the left column we show the constituents in the (θ, φ) plane, called (x, y) for the toy model, with angular bins drawn in gray. In the middle column we have the histogram for the angular variable ϖ that represents the (a_i) values, and the
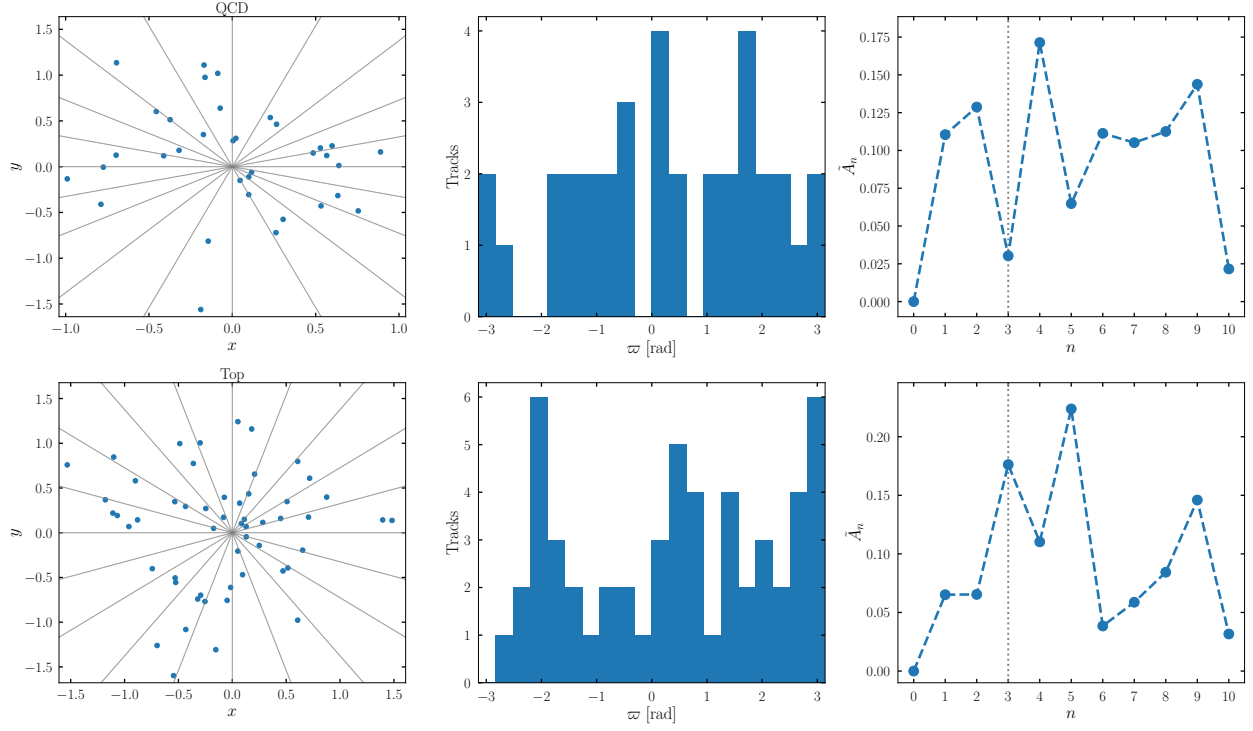
Figure 1: Left column: jet constituents in a bidimensional space of coordinates (x, y), with angular slices shown by gray lines. Center column: histogram of constituents per angular slice. Right column: normalized values of the DFT results $\tilde{A}_n$, with $\tilde{A}_3$ highlighted by the dotted line. Top (bottom) row corresponds to a QCD (top) jet. Constituents were drawn from a toy model described in Appendix A.

corresponding $\tilde{A}_l$ distribution is shown in the right column. Fig. 1 shows that the resulting $\tilde{A}_n$, in particular $\tilde{A}_3$, may highlight the three-pronged structure of top jets in comparison to QCD jets even if it is not clearly discernible in the original (x, y) space.

Having discussed the Fourier representation within the framework of a toy model, we now return to the Bayesian inference over $\tilde{A}_3$ with MC simulated data. The $p(\tilde{A}_3|k)$ modeling for Bayesian inference is done as follows. Since by definition $\tilde{A}_3$ is bound between 0 and 1 we propose, as a working approximation, a probability density function corresponding to a Beta distribution for each class,

$$B(\tilde{A}_3; \alpha_k, \beta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \tilde{A}_3^{\alpha_k-1}(1 - \tilde{A}_3)^{\beta_k-1} \ . \tag{8}$$

Therefore, $\tilde{A}_3$ has the following total probability density given by the mixture model

$$\tilde{A}_3 \sim \pi_0 \, B(\alpha_0, \beta_0) + (1 - \pi_0) \, B(\alpha_1, \beta_1) \,, \tag{9}$$

where the mixing fraction π_0 is the probability of sampling a QCD jet, and consequently $(1 - \pi_0)$ corresponds to top jets. By following the general procedure described in Sec. 2 we can infer the MAP by first providing a prior. We use uniform priors for all latent variables,

$$P(\zeta) = \begin{cases} 1, \ \zeta \in \mathcal{D} \\ 0, \ \text{otherwise} \end{cases} \tag{10}$$
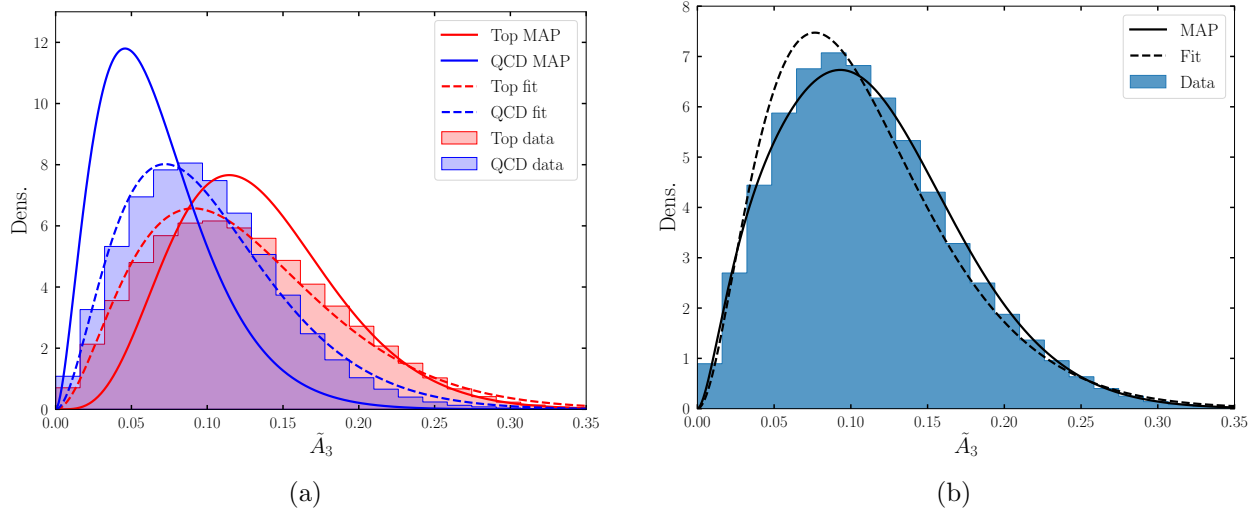
Figure 2: Left: Distribution of the variable $\tilde{A}_3$ for the QCD (top) jets in blue (red). Fitted Beta distributions for each individual class is shown in dashed lines. Our inferred MAP is shown in solid lines. Right: data without labels and weighted sum of MAP and fitted beta distributions.

where $\mathcal{D}$ is a rectangular region, such that $\zeta_i \in (\zeta_i^{\min}, \zeta_i^{\max})$ for each of the model parameters. We choose $\mathcal{D}$ to be large enough to have the maximum a posteriori (MAP) away from the region boundaries.

As detailed in Sec. 2, we estimate the MAP through Stochastic Variational Inference (SVI) by using the `Pyro` package [32] in `Python3`. In Fig. 2a we plot the data distribution with true QCD (top) labels in blue (red), using a true QCD ratio of $p = 0.5$. We perform a supervised fit of a Beta distribution over each class, with the resulting distributions shown in dashed lines. Solid lines represent the distributions given by MAP estimates, and the inferred QCD probability at the MAP is $\pi_0^{\mathrm{MAP}} = 0.32$. In Fig. 2b the total data distribution is shown, where the fitted and MAP distributions are taken as the weighted sum of the corresponding Beta distributions, as in Eq. (9). Note that the MAP distribution correctly captures the data without labels, at the cost of estimating a lower π_0^{MAP} value. Also, for low $\tilde{A}_3$ values the top distribution given by the MAP is underestimated, which is compensated by overestimating the QCD distribution. On the other hand, the QCD distribution is underestimated at large $\tilde{A}_3$, where the top distribution is correctly reproduced. Our model succeeds at capturing some features of the class distributions and it correctly fits the data without labels. Nonetheless, given the significant overlap between both classes, it tends to infer distributions that are distant from the truth data resulting in poor tagging performance, with a top tagger based on these inferred distributions yielding an AUC of 0.62 and an accuracy of 0.55 for a sample with $p = 0.5$. This is because the mixture model is appropriately learning the total density without the need to separate explicitly into the true distributions of QCD and top jets. This is similar to the results reported in Ref. [28] where it was shown that learning the top and QCD defining features is enough for unsupervised density estimation.

A clear improvement for the Bayesian Fourier approach in this Section would be to go beyond the Beta distribution to better model the expected shape for $\tilde{A}_3$ and propose the corresponding PDF. However, even though the Fourier simplification is attractive, the approach itself has an intrinsic limitation as a tagger due to the little discrimination power one can assess from the Montecarlo data in Fig. 2a. In any case, even without a good tagger, having a better model for the

$\tilde{A}_3$ distributions could provide the total fraction in the sample, as well as relevant information for the physics involved.

In the next Section we develop a similar analysis as in this Section, but using the N-subjettiness ratio τ_3/τ_2, which shows less overlap between classes and thus it is in principle better to match density estimation to classification performance.

4 Mixture model for the N-subjettiness ratio τ_3/τ_2

A variable introduced in the literature to study jet substructure, the N-subjettiness [30, 35] for a given identified jet, is obtained from N defined candidate subjets by computing

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min\{(\Delta R_{1,k})^\beta, (\Delta R_{2,k})^\beta, \dots, (\Delta R_{N,k})^\beta\}\,, \tag{11}$$

where the sum is done over all jet constituents, $p_{T,k}$ is their transverse momenta, $\Delta R_{i,k}$ is the distance in the (η, φ) plane between the k-th constituent and the i-th candidate subjet, β is an angular weighting exponent, and d_0 is a normalization factor,

$$d_0 = \sum_k p_{T,k} R_0^\beta\,, \tag{12}$$

where R_0 is the original jet clustering radius. Although N subjets should in principle be chosen by minimizing τ_N, we follow Ref. [30,35] and use the N subjets obtained by applying the exclusive k_T clustering algorithm to the jet via `FastJet` [36]. This approximation yields a small fraction of events for which $\tau_3/\tau_2 > 1$, which is not an obstacle for our algorithm. The variable τ_N quantifies the degree to which a jet can be regarded as N collimated subjets. In particular, for 3-pronged objects such as top jets, the ratio τ_3/τ_2 has shown to be an effective tagging variable using simple one-dimensional cuts [35]. To take advantage of this discriminating power, we implement the Bayesian mixture model to construct an unsupervised top tagger.

We propose a model in which τ_3/τ_2 is given by a mixture of Gamma distributions,

$$\frac{\tau_3}{\tau_2} \sim \pi_0\,\Gamma(\alpha_0, \beta_0) + (1 - \pi_0)\,\Gamma(\alpha_1, \beta_1)\,, \tag{13}$$

where

$$\Gamma(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\,, \tag{14}$$

where $x = \tau_3/\tau_2$, and the shape and rate parameters α_k and β_k are given for the QCD (top) class ($k = 0\,(1)$), and π_0 is the probability of sampling a QCD event. Given our mixture model we can evaluate the likelihood function of the sample data, and then construct a posterior distribution over the distribution parameters $\boldsymbol{\zeta} = (\pi_0, \alpha_0, \beta_0, \alpha_1, \beta_1)$ in an analogous manner to the one described in Sec. 2. The prior distribution we propose to be uniform in each of the inferred variables, and we infer the MAP of these parameters with SVI via the `Pyro` package; exactly as in the previous Section.

To evaluate and interpret the learned model we plot the MAP mixture components and the corresponding data distribution in Fig. 3 for a dataset with true QCD ratio $p = 0.5$. We compare the learned model both to the data and to the supervised fits to labeled data, where each Gamma
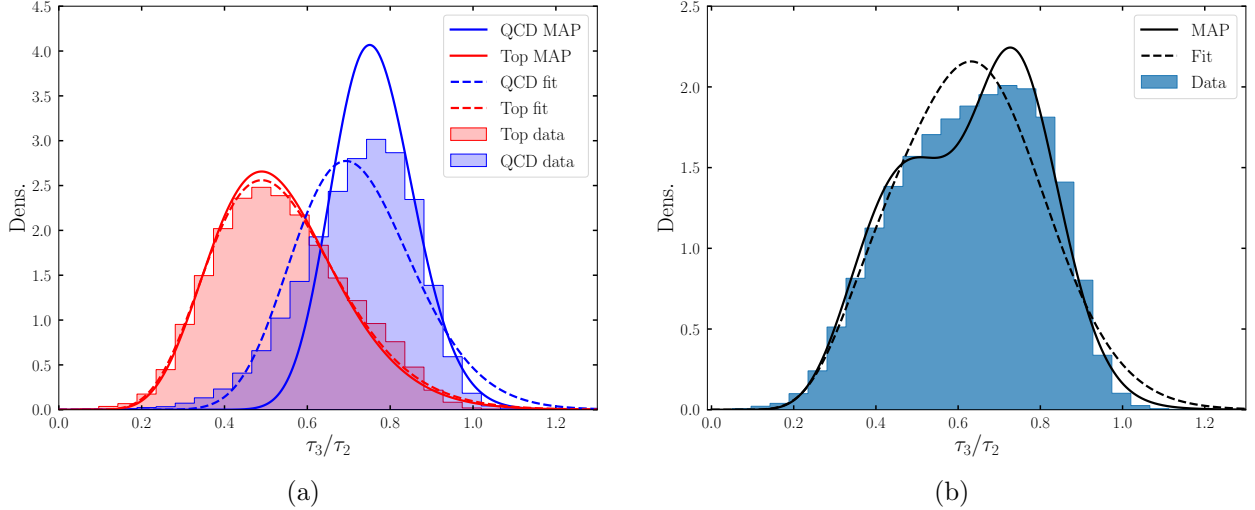
Figure 3: Distribution of the τ_3/τ_2 data with $p = 0.5$, where fitted and MAP Gamma distributions are plotted in dashed and solid lines respectively. The inferred value for the fraction in the sample is $\pi_0^{\text{MAP}} = 0.42$. Left plot shows separate QCD and top distributions, while the right plot corresponds to the total distribution. Within the allowed freedom the MAP distribution describes slightly better the data distribution, as it is what really captures the algorithm.

distribution is fitted separately to its corresponding process. In Fig. 3a we observe that both the MAP (solid) distribution and the fitted (dashed) distributions are closer to truth level for top jets (red) than for QCD jets (blue). However, the fitted and learned QCD approximations are not identical. The MAP better captures the maximum position while the fitted distribution better reproduces the mean by sacrificing an accurate maximum in favor of accurate tails. This is caused by the differing objectives of the two approximations. The fit aims to capture the individual labelled process accurately, while the MAP aims to capture the unlabelled data distribution. In Fig. 3b we observe that the MAP captures much better the multimodality of the full data, while the fitted distributions provide a worse fit. The inadequacy of both procedures highlights the fact that for QCD the τ_3/τ_2 distribution cannot be approximated by a Gamma distribution. However, the MAP objective ensures that the MAP estimates capture the difference between QCD and top in order to explain the data more accurately, as in Sec. 3.

As detailed in Sec. 2, after obtaining the MAP parameters we can construct a jet classifier by calculating the assignment probabilities of the class $\mathcal{C}_k$, with $k = 0\,(1)$ representing QCD (top) events,

$$P(\mathcal{C}_k|\tau_3/\tau_2, \boldsymbol{\zeta}^{\text{MAP}}) = \frac{\pi_k^{\text{MAP}}\,\Gamma(\tau_3/\tau_2; \alpha_k^{\text{MAP}}, \beta_k^{\text{MAP}})}{\sum_{i=0,1}\pi_i^{\text{MAP}}\,\Gamma(\tau_3/\tau_2; \alpha_i^{\text{MAP}}, \beta_i^{\text{MAP}})}\,. \tag{15}$$

The classifier assigns a jet to the top class if $P(\text{top}|\tau_3/\tau_2, \boldsymbol{\zeta}^{\text{MAP}}) > t$, where t is a chosen probability threshold. We evaluate our tagger over the test dataset of [22] with a cut in the invariant mass of the jets, $145 \text{ GeV} < m_j < 205 \text{ GeV}$, in order to compute the usual classifier performance metrics. In Fig. 4a we show the assignment probabilities for each class given by the MAP estimates as a function of τ_3/τ_2. By varying t in the interval $(0, 1)$ we can obtain the true and false positive rates (TPR and FPR respectively) and construct the ROC curve, which is plotted in Fig. 4b for the classifiers corresponding to the MAP and fitted (optimal) parameters.

From Fig. 4b, we observe that our classifier is very close to the optimal performance even if

(a)   (b)
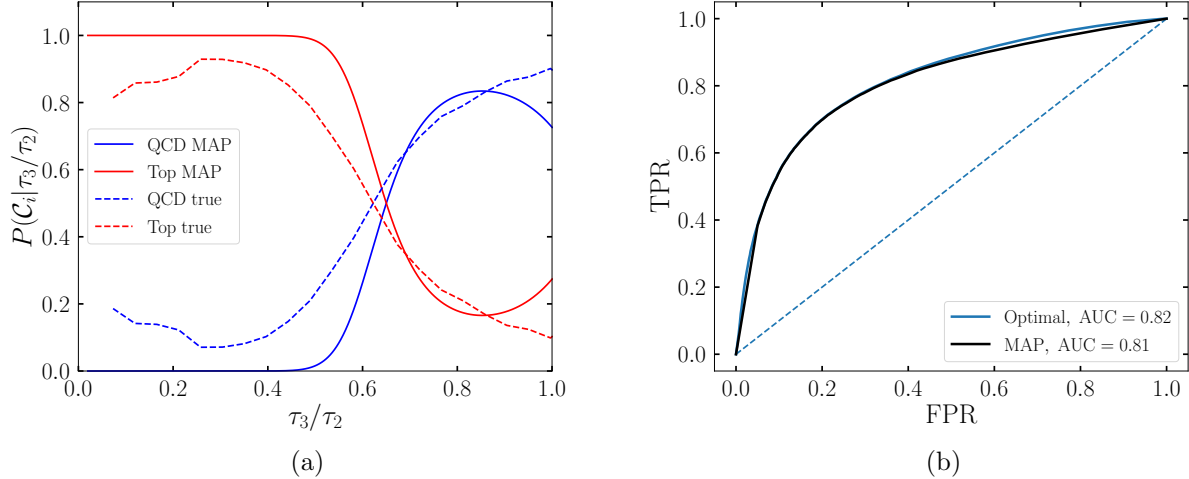
Figure 4: Left: Assignment probabilities for each class as a function of τ_3/τ_2, given the MAP estimates (solid lines), or given truth data (dashed). Right: ROC curves for the jet classifiers constructed with fitted parameters (blue) and MAP values (black).

the underlying distributions do not match perfectly the true distributions. This reinforces the observation that the mixture model is learning the appropriate differences between top and QCD jets. This is also seen in Fig. 4a where the learned probabilities for each class have the appropriate behavior close to the $p(C_i) = 0.5$ in the intermediate τ_3/τ_2 region while the model is overconfident assigning very small $p(C_0)$ for small τ_3/τ_2 and underconfident with $p(C_0)$ too small for large τ_3/τ_2.

As a final test, we benchmark our model for different true fractions of QCD p. To compare all cases, we compute the AUC and the accuracy for each set of learned parameters over the same balanced sample of QCD and top jets unseen during inference. The results are shown in Table 1. We observe that the model performs consistently in terms of accuracy and AUC for all fractions, while the differences between p and π_0^{MAP} suggest that the model does not need to learn the exact distributions to do so. The difference between true and learned distributions is larger for larger p due to the QCD distribution being the most distinct from a true Gamma distribution. In Fig. 5 we plot the true QCD fraction in the sample versus the corresponding MAP for its inferred value, which yields a satisfactory agreement of about 20% for the worst cases of large QCD fraction.

p	π_0^{MAP}	Acc	AUC
0.1	0.16	0.69	0.80
0.3	0.30	0.74	0.81
0.5	0.42	0.75	0.81
0.7	0.54	0.75	0.81
0.9	0.71	0.74	0.81

Table 1: Inferred π_0^{MAP} values and tagging performance metrics for different true mixing ratios (p) of QCD events.
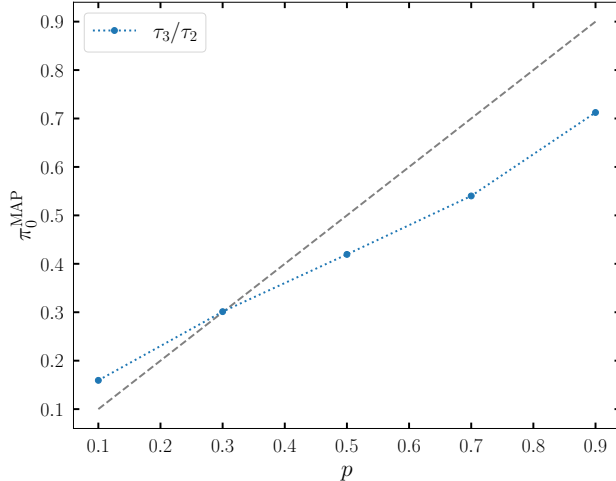
Figure 5: Learned π_0^{MAP} values as a function of true QCD mixture fraction p used for inference. The agreement is slightly better for low QCD fractions, since the top distribution is better captured by the model.

5   Discussion and outlook

Along this manuscript we have explored Bayesian inference techniques through two different mixture models to unsupervisedly recognize top quark jets from QCD jets. In one model (Sec. 3), we have used as the observed variable the normalized third coefficient $\tilde{A}_3$ (Eq. 7) of the Fourier transform of the angular distributions of the tracks around the center of the jet, and proposed for its PDF a different Beta distribution for each of the classes (top and QCD). For the other model (Sec. 4), we have used as the observed variable the N-subjettiness ratio τ_3/τ_2, and proposed for its PDF a different Gamma distribution for each of the classes. The inference procedure has been carried out using Stochastic Variational Inference which has the virtue of being fast and reliable, at the price of obtaining the Maximum A Posteriori (MAP) instead of the full posterior.

We find that the results for the Fourier framework, which has the advantage of a pure geometrical interpretation based on the 3-pronged nature of top jets, are subpar, the main reasons being that the $\tilde{A}_3$ variable has poor discriminating power and that the modelling of its PDF through a Beta distribution is not entirely satisfactory. An alternative framework that keeps the purely geometrical interpretation would be to consider a mixture model in which the tracks of the jet are sampled either from three overlapping Normal distributions or from one Normal distribution. Such an analysis seems appealing and it could have very challenging and promising approaches.

On the other hand, using the N-subjettiness ratio τ_3/τ_2 variable yields considerably better results. The PDF for the top quark class is captured well through the propose Gamma distribution. Although the PDF for τ_3/τ_2 in the QCD class is not well reproduced, still the tagger is satisfactory with an AUC $\sim 0.80-0.81$ and an accuracy of about $0.69-0.75$ for samples with top-quark fraction ranging from 10% to 90%. Moreover, the model also infers the total fraction of top-quark in the sample with very good agreement to the true value (see Fig. 5).

We find that the expected performance of the studied unsupervised taggers is below the expected performance of supervised taggers. However, the former is more robust than the latter to potential mismatches between Montecarlo simulations and real data inside the jets, since its foundations are

related to robust knowledge. Such as for instance that the top-quark has three decay products whereas other information is inferred directly from the data itself. This kind of unsupervised taggers are more simple and robust at the price of reducing the expected performance. Another non-negligible advantage of this unsupervised tagger is that the used parameters to determine the tagging are extracted from the sample itself. That is, they are not only extracted from real data, but also they are not imported from another kinematic region.

In brief, the present work explores a class of unsupervised taggers that apply Bayesian inference on a mixture model. Our findings aim at reflecting on potential improvements in unsupervised taggers for their possible use in real analyses. In the future, one possible path is to look for more appropriate distributions or better suited observables. For example, the energy of each constituent track could be incorporated more explicitly in the $\tilde{A}_n$. Another possible path is to increase the power of the mixture model to capture arbitrary distributions. A Normalizing Flow-inspired architecture could be developed provided the only multimodality comes from the mixture of the processes, which is the case for the observables considered in this work.

Acknowledgement

E.A., T.T., S.T. and A.S. thank CONICET and ANPCyT (under projects PICT 2017-2751 and PICT 2018-03682), and M.S. acknowledges support in part by the DOE grant DE-SC0011784 and NSF OAC-2103889. We all thank the *Muchachos* for the great effort poured during the development of this article.
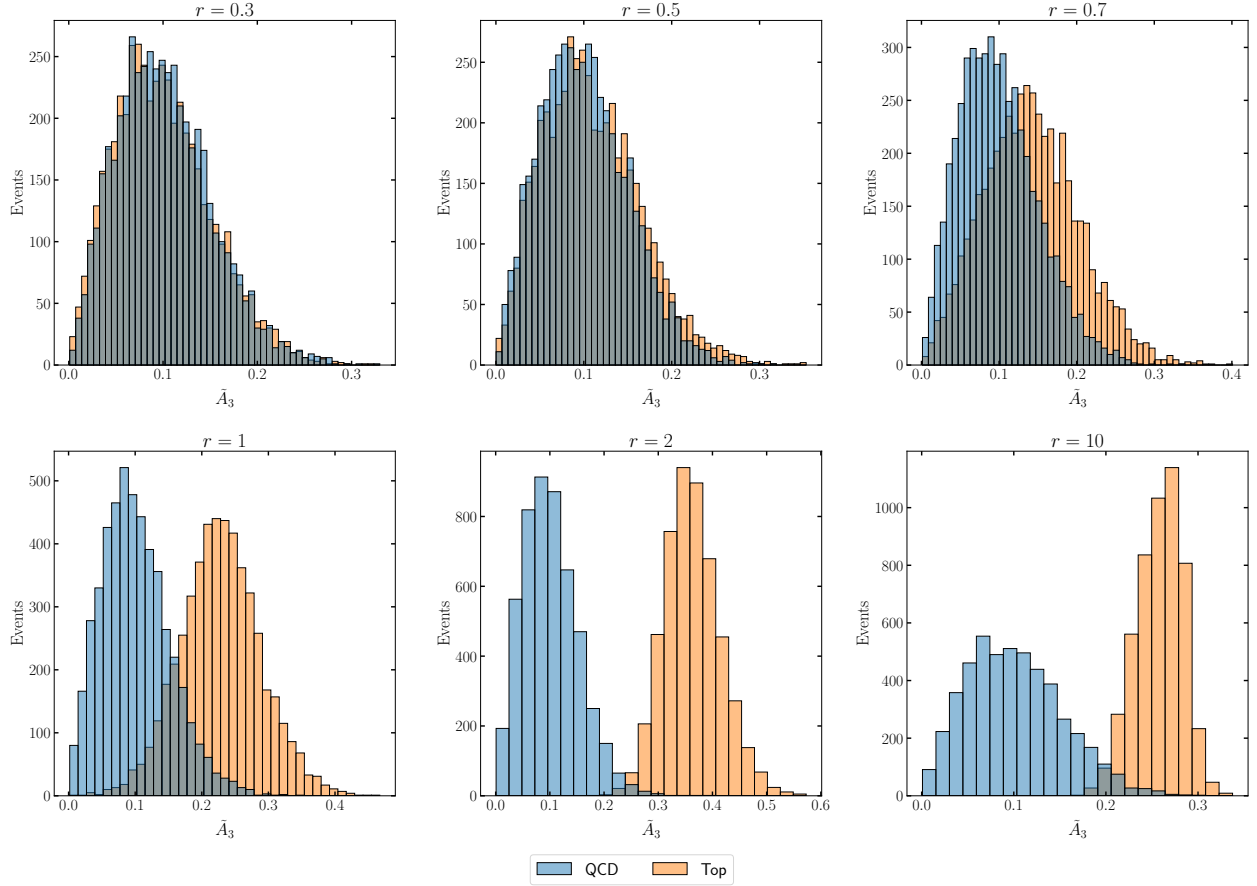
Figure 6: We show the obtained $\tilde{A}_3$ distributions for the toy QCD and top jets for different choices of the position r (see the text). As r increases, the differing physics between the two processes is picked up by the $\tilde{A}_3$ distributions.

A   Toy Model

In order to test the discriminating power of the variable $\tilde{A}_3$ constructed in Sec. 3, we first build a toy model that tries to capture the basic features of the top and QCD jets. For each event in the generated data, we draw a QCD event with probability p, or a top event with probability $1-p$. We represent each event by N_i points, *i.e.* tracks, in the 2-dimensional plane (x, y). The number of tracks N_i is drawn for each event from a Poisson distribution with a given expected value N_{tr}. Tracks corresponding to a QCD jet are drawn from a 2-dimensional isotropic Gaussian distribution centered at the origin,

$$(x, y)_{\mathrm{QCD}} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\Sigma}_0 = \mathrm{diag}(\sigma_0, \sigma_0), \tag{16}$$

while the substructure of $t \to bjj$ events is simulated as a mixture of three isotropic Gaussian distributions centered at positions $\mathbf{r}_i$, $i = 1, 2, 3$, each corresponding to a jet in the top decay,

$$(x, y)_{\mathrm{top}} \sim p_1 \mathcal{N}(\mathbf{r}_1, \boldsymbol{\Sigma}_1) + p_2 \mathcal{N}(\mathbf{r}_2, \boldsymbol{\Sigma}_2) + p_3 \mathcal{N}(\mathbf{r}_3, \boldsymbol{\Sigma}_3), \quad \boldsymbol{\Sigma}_i = \mathrm{diag}(\sigma_i, \sigma_i), \tag{17}$$

where p_i is the probability of a particular track to correspond to one of the three jets, so $\sum_{i=1}^3 p_i = 1$. For definiteness, in what follows we fix some of the model parameters at

$$p_1 = p_2 = p_3 = 1/3, \tag{18}$$

$$|\mathbf{r}_1| = |\mathbf{r}_2| = |\mathbf{r}_3| = r \,, \tag{19}$$

$$\mathrm{ang}(\mathbf{r}_i, \mathbf{r}_j) = 2\pi/3 \,, \ i, j = 1, 2, 3 \,, \tag{20}$$

$$\sigma_i = 0.5 \,, \ i = 0, 1, 2, 3 \,. \tag{21}$$

Therefore, with this parametrization r represents how separate are the top subjets from each other. We show in Fig. 6 the $\tilde{A}_3$ distribution for different r values, for both QCD and top classes. Note that the distributions become distinguishable from each other around $r \sim 0.7$, which corresponds to $r > \sigma$. On the other hand, as we increase r, the $\tilde{A}_3$ distribution of the top events does not reach values arbitrarily close to $\tilde{A}_3 = 1$. This is due to the fact that, as r increases, so do the $l = 6, 9$ frequencies of the DFT, and since they are normalized they constrain the maximum value of $\tilde{A}_3$ for any radius. This toy model captures the essential physics and we obtain Fig. 1 by setting $r = 0.5$.

References

[1] **ATLAS** Collaboration, G. Aad et al., *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B* **716** (2012) 1–29, [arXiv:1207.7214].

[2] **CMS** Collaboration, S. Chatrchyan et al., *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC, Phys. Lett. B* **716** (2012) 30–61, [arXiv:1207.7235].

[3] **D0** Collaboration, S. Abachi et al., *Observation of the top quark, Phys. Rev. Lett.* **74** (1995) 2632–2637, [hep-ex/9503003].

[4] **CDF** Collaboration, F. Abe et al., *Observation of top quark production in $\bar{p}p$ collisions, Phys. Rev. Lett.* **74** (1995) 2626–2631, [hep-ex/9503002].

[5] C. E. Gerber and C. Vellidis, *Review of Physics Results from the Tevatron: Top Quark Physics,* arXiv:1409.5038.

[6] **CMS** Collaboration, *Measurement of the jet mass distribution and top quark mass in hadronic decays of boosted top quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV,* arXiv:2211.01456.

[7] **ATLAS** Collaboration, G. Aad et al., *Measurement of the top-quark mass in $t\bar{t} + 1$-jet events collected with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV, JHEP* **11** (2019) 150, [arXiv:1905.02302].

[8] **CMS** Collaboration, A. Tumasyan et al., *Measurement of the inclusive $t\bar{t}$ production cross section in proton-proton collisions at $\sqrt{s} = 5.02$ TeV, JHEP* **04** (2022) 144, [arXiv:2112.09114].

[9] **ATLAS** Collaboration, *Measurement of the $t\bar{t}$ production cross-section in pp collisions at $\sqrt{s} = 5.02$ TeV with the ATLAS detector,* arXiv:2207.01354.

[10] **CMS** Collaboration, G. K. Krintiras, *Results on TOP physics from CMS,* in *International Workshop on Future Linear Collider*, 1, 2018. arXiv:1801.08398.

[11] R. Naranjo, *Top physics in ATLAS,* in *17th Lomonosov Conference on Elementary Particle Physics*, pp. 223–226, 2017. arXiv:1603.06203.

[12] 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** (2008) 142001, [arXiv:0806.0848].

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

[14] G. Kasieczka, T. Plehn, T. Schell, T. Strebler, and G. P. Salam, *Resonance Searches with an Updated Top Tagger*, JHEP **06** (2015) 203, [arXiv:1503.05921].

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

[16] **CMS** Collaboration, *Jet algorithms performance in 13 TeV data*, tech. rep., CERN, Geneva, 2017 [http://cds.cern.ch/record/2256875].

[17] G. Kasieczka, *Boosted Top Tagging Method Overview*, in *10th International Workshop on Top Quark Physics*, 1, 2018. arXiv:1801.04180.

[18] A. Butter et al., *The Machine Learning landscape of top taggers*, SciPost Phys. **7** (2019) 014, [arXiv:1902.09914].

[19] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, JHEP **04** (2018) 013, [arXiv:1712.07124].

[20] H. Qu, C. Li, and S. Qian, *Particle Transformer for Jet Tagging*, arXiv:2202.03772.

[21] B. Bhattacherjee, C. Bose, A. Chakraborty, and R. Sengupta, *Boosted top tagging and its interpretation using Shapley values*, arXiv:2212.11606.

[22] G. Kasieczka, T. Plehn, J. Thompson, and M. Russel, *Top quark tagging reference dataset*, Mar., 2019 [https://doi.org/10.5281/zenodo.2603256].

[23] D. Yallup and W. Handley, *Hunting for bumps in the margins*, arXiv:2211.10391.

[24] A. Butter, B. M. Dillon, T. Plehn, and L. Vogel, *Performance versus Resilience in Modern Quark-Gluon Tagging*, arXiv:2212.10493.

[25] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, *Uncovering latent jet substructure*, Phys. Rev. D **100** (2019), no. 5 056002, [arXiv:1904.04200].

[26] C. M. Bishop, *Pattern recognition and machine learning*. Springer, New York, NY, 2006.

[27] E. Alvarez, M. Spannowsky, and M. Szewc, *Unsupervised Quark/Gluon Jet Tagging With Poissonian Mixture Models*, Front. Artif. Intell. **5** (2022) 852970, [arXiv:2112.11352].

[28] E. Alvarez, B. M. Dillon, D. A. Faroughy, J. F. Kamenik, F. Lamagna, and M. Szewc, *Bayesian probabilistic modeling for four-top production at the LHC*, Phys. Rev. D **105** (2022), no. 9 092001, [arXiv:2107.00668].

[29] E. Alvarez, *Bayesian inference to study a signal with two or more decaying particles in a non-resonant background*, arXiv:2210.07358.

[30] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, JHEP **03** (2011) 015, [arXiv:1011.2268].

[31] M. D. Hoffman, D. M. Blei, C. Wang, and J. W. Paisley, *Stochastic variational inference*, *J. Mach. Learn. Res.* **14** (2013) 1303–1347.

[32] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, *Pyro: Deep universal probabilistic programming*, *J. Mach. Learn. Res.* **20** (2019) 28:1–28:6.

[33] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, arXiv:2203.11601.

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

[35] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, *JHEP* **02** (2012) 093, [arXiv:1108.2701].

[36] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet user manual*, *The European Physical Journal C* **72** (mar, 2012).