Bayesian probabilistic modeling for four-top production at the LHC

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

Barry M. Dillon[†]
Institut für Theoretische Physik, Universität Heidelberg, 69120 Heidelberg, Germany

Darius A. Faroughy[‡]
Physik-Institut, Universität Zürich, CH-8057, Switzerland

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

Federico Lamagna[‖]
Centro Atómico Bariloche, Instituto Balseiro and CONICET, Bariloche CP8400, Argentina

Manuel Szewc[¶]
International Center for Advanced Studies (ICAS) and CONICET, UNSAM, Campus Miguelete, 25 de Mayo y Francia, CP1650, San Martin, Buenos Aires, Argentina and Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

(Received 18 July 2021; revised 29 September 2021; accepted 17 April 2022; published 5 May 2022)

Monte Carlo (MC) generators are crucial for analyzing data in particle collider experiments. However, often even a small mismatch between the MC simulations and the measurements can undermine the interpretation of the results. This is particularly important in the context of LHC searches for rare physics processes within and beyond the standard model (SM). One of the ultimate rare processes in the SM currently being explored at the LHC, $pp \to t\bar{t}t\bar{t}$ with its large multidimensional phase-space is an ideal testing ground to explore new ways to reduce the impact of potential MC mismodeling on experimental results. We propose a novel statistical method capable of disentangling the 4-top signal from the dominant backgrounds in the same-sign dilepton channel, while simultaneously correcting for possible MC imperfections in modeling of the most relevant discriminating observables—the jet multiplicity distributions. A Bayesian mixture of multinomials is used to model the light-jet and b-jet multiplicities under the assumption of their conditional independence. The signal and background distributions generated from a deliberately mistuned MC simulator are used as model priors. The posterior distributions, as well as the signal and background fractions, are then learned from the data using Bayesian inference. We demonstrate that our method can mitigate the effects of large MC mismodelings in the context of a realistic $t\bar{t}t\bar{t}$ search, leading to corrected posterior distributions that better approximate the underlying truth-level spectra.

DOI: 10.1103/PhysRevD.105.092001

[*]sequi@unsam.edu.ar
[†]dillon@thphys.uni-heidelberg.de
[‡]faroughy@physik.uzh.ch
[§]jernej.kamenik@cern.ch
[‖]federico.lamagna@cab.cnea.gov.ar
[¶]mszewc@unsam.edu.ar

I. INTRODUCTION

In recent years, the large abundance of LHC data on one hand, and the absence of clear new physics (NP) signals in theory driven analyses of this data on the other, have motivated the development of novel, more data driven approaches to LHC data analysis and NP searches. In particular, the advent of unsupervised and weakly supervised machine learning (ML) techniques has allowed for the development of broad model independent NP search and characterization strategies [1]. Simultaneously, there

have been important efforts to reduce reliance of LHC measurements on Monte Carlo (MC) simulations of hadronic processes [2–6].

The simultaneous production of four top quarks represents an important NP benchmark (see, e.g., Refs. [7–18]), but also an interesting point of coalescence for several of these developments [19]. One of the main issues in studying this final state is its tiny cross section (12 fb) compared to its main backgrounds (∼600 fb), which is compounded by the challenges to correctly model the complex final states through MC simulations. To address these issues, we have previously studied the two lepton same sign channel (2LSS $\pm\pm$) [20] which in the SM may contain signal and background events up to the same order of magnitude and furthermore exhibits somewhat reduced complexity of the (multijet) final state, compared to the single lepton channel [21,22]. In the 2LSS $++$ channel $t\bar{t}W^+$ production represents the main and most challenging background for the 4-top signal.[1] Recent experimental analyses in this channel [23,24] have highlighted difficulties in reliably modeling the signal and background kinematics using state of the art MC tools. This in turn hinders the sensitivity of this important signature to possible NP effects in four-top production.

Using the experimental challenge described above as an example and motivation, in the present paper we describe a novel Bayesian statistical framework to disentangle in-situ signal and background distributions of categorical data. Our method can be used to simultaneously identify and correct potential (MC) mismodeling of discrete distributions as well as extract signal and background admixtures in the data close to their truth values.

The paper is organized as follows. In Sec. II we introduce our statistical model of multinomial mixtures with Bayesian inference and demonstrate its use on a toy example. We apply the model to jet multiplicity distributions in the 2LSS $++$ channel of 4-top production at the LHC in Sec. III and show how it can be used to identify and correct MC mismodeling and extract signal and background fractions. Section IV is devoted to a detailed study of the assumptions and consistency checks of the model when applied to realistic datasets. Finally, we summarize our findings in Sec. V.

II. CATEGORICAL MIXTURE MODEL FOR FOUR-TOPS

Anticipating the application to 4-top production, in the following we represent an event generation process by a pair of random variables (N_j, N_b) indicating the number of clustered light-jets and b-jets, respectively. Our starting point is that a collection of such events can be described using a likelihood with a joint probability density $p(j, b)$ where j (b) are the observed number of light-jets (b-jets) in an event. The most general discrete model for this likelihood is the multinomial distribution[2] with $d_j \times d_b - 1$ parameters, where $d_{j,b}$ are the number of possible light-jets and b-jets to be expected in an event. However, our goal is to disentangle the contributions to this joint likelihood arising from four-top events and $t\bar{t}W$ events. To do so we introduce two mixture components, one for $t\bar{t}W$ and one for four-top. If we simply describe each mixture with a multinomial distribution $p(j, b|z)$ with $z \in [0, 1]$ representing the mixture label, we would have a mixture model with $2 \times (d_j \times d_b - 1) + 1$ parameters. Since each event is independent and consists of just a single draw from this distribution, each mixture can describe all possible combinations of N_j and N_b values in the data and therefore all correlations by itself. The model would thus overparametrize the data making the inclusion of mixtures redundant.[3]

Therefore the key insight is to instead write down a mixture model in terms of $p(j|z)$ and $p(b|z)$, such that the correlations between N_j and N_b in the dataset are parameterized by the class label alone. The number of parameters in this model is $2 \times (d_j + d_b - 2) + 1$. To be explicit, we optimize the model to parameterize the correlations between N_j and N_b in terms of a discrete variable Z, and interpret this as a class label for four-top and $t\bar{t}W$ events. We are making the simplifying assumption that N_j and N_b are *conditionally independent variables*, that all correlations between them in the dataset arise only from assignments to the two classes. Conditional independence is of course an approximation. In particular, in a realistic measurement setting, N_j and N_b are not strictly conditionally independent due to mistagging or other reconstruction imperfections. The degree to which the method succeeds is limited by this approximation. Conversely, a failure of the method to converge to a consistent description of the measured distributions would be a clear sign that the assumptions of the statistical model are not respected by the dataset. We return to this important caveat and discuss its mitigation in Sec. IV.[4] However, as we will show, in the case at hand, the method exhibits good convergence indicating that conditional independence holds sufficiently well in practice.

[1]Our results and discussion would apply equally well to other non-negligible backgrounds such as $t\bar{t}h$ and $t\bar{t}Z$.

[2]Along this work we refer to multinomial distribution although in all cases it consists of a single drawing per event and therefore it is also a categorical distribution, which is a special case of the former.

[3]Note that this would not be the case if each event was generated by several draws from $p(j, b|z)$, since there would then be additional correlations between the multiple draws per event. This is the case in s.c. mixed membership models [25–27] used in jet substructure analyses where the mixtures describe correlations between the multiple draws per event.

[4]A systematic study of statistical models which go beyond strict conditional independence assumptions is in progress and will be presented elsewhere.

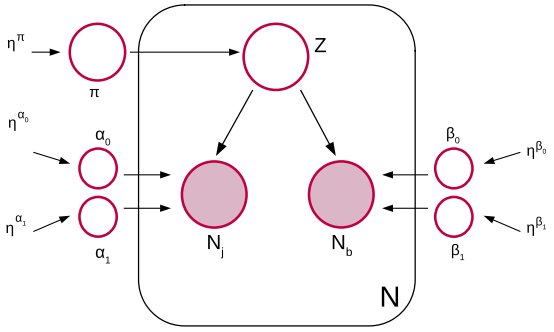


FIG. 1. Plate diagram of (Bayesian) 2-mixture model of multinomials for (N_b, N_j) N-event dataset. From the Dirichlet prior distributions (with hyperparameters η_i) the multinomial parameters $(\pi, \alpha_i$ and $\beta_i)$ are sampled, then N events are sampled through a latent variable Z that determines in turn from which multinomial the two observables in each event (N_j and N_b) are sampled.

Within the limitations described above, the generative process for the dataset proceeds as follows: for each event (n) a class label z_n is first drawn from a binomial probability distribution parametrized by $\pi \in [0, 1]$. Then j_n and b_n are sampled from separate multinomials corresponding to the drawn class and parametrized by $\alpha_{z,i}$ and $\beta_{z,k}$, respectively, where i and k run up to d_j and d_b, respectively. We assume that the whole dataset X, consisting of $n \in N$ pairs of measurements $x_n = (j_n, b_n)$ for the 2LSS++ selected events, is generated through this probabilistic model and we want to infer the values of its parameters, namely $\pi, \alpha_{0,j}, \beta_{0,i}, \alpha_{1,j}$ and $\beta_{1,i}$, which we collectively indicate as θ. Observe that the described model corresponds to a special case of a *mixture of multinomials* [28].

Adopting a Bayesian framework, we consider the model parameters (θ) to be random variables as well and we want to update our knowledge of these random variables after measuring X. However, it is more convenient in practice to consider explicitly also the latent variables Z which represent the class assignments of each event. Graphically, the probabilistic model can be represented through the plate diagram in Fig. 1 and leads to the posterior:

$$p(Z, \pi, \alpha, \beta|X) = \frac{p(X, Z, \pi, \alpha, \beta)}{p(X)}, \tag{1}$$

where the joint distribution $p(X, Z, \pi, \alpha, \beta)$ is given explicitly by

$$p(X, Z, \pi, \alpha, \beta) = \prod_{n=1}^{N} p(x_n|z_n, \alpha, \beta) p(z_n|\pi) \times p(\pi|\eta_\pi) \prod_{k=0}^{1} p(\alpha_k|\eta_{\alpha_k}) p(\beta_k|\eta_{\beta_k}).$$

Here $p(x_n|z_n, \alpha, \beta) = \alpha_{z_n j_n} \beta_{z_n b_n}$, $p(z_n|\pi) = \pi_{z_n}$ and $p(\pi|\eta^\pi)$, $p(\alpha_k|\eta^{\alpha_k})$ and $p(\beta_k|\eta^{\beta_k})$ are Dirichlet distributions with the corresponding η^i set of parameters.

The main idea in this expression is that given the dataset X, a probabilistic model that allows us to write down an expression for $p(X|\theta)$ and a reasonable prior $p(\theta)$, we can in principle determine the probability density function (pdf) for the parameters $p(\theta|X)$. This is a powerful result, since it gives us not only the fraction of signal to background and its uncertainty through $p(\pi|X)$ marginalizing over the other parameters, but it can also give us the N_j and N_b distributions of both individual classes. If the probabilistic model describes the data well and the prior is reasonable, then these should match within uncertainties the true underlying background and signal N_j and N_b distributions.

There are many known approaches to solving Eq. (1) using Bayesian inference; including mean-field techniques such as variational inference (VI) [28] and numerical Markov Chain Monte Carlo methods such as Gibbs Sampling (GS) [28]. Below we focus on the latter numerical approach which turns out to be preferred to the mean-field methods which approximate the posterior with a fully factorized model that neglects possible correlations between the inferred parameters. As we are interested in finding the correlations between N_j and N_b through class assignment, VI is challenged by definition to find the appropriate correlations.

The goal of the GS algorithm is to approximate the posterior through the use of a finite number of samples. These samples can then be used to obtain any desired expected values such as the mean of the relevant parameters $\mathbb{E}[\theta_i]$. To obtain samples from the posterior, each iteration samples an observation of each parameter θ_i from the marginal distribution conditioned on the remaining parameters $p(\theta_i|\theta_{\setminus i}, X)$. When implementing a Gibbs sampler to approximate Eq. (1), the conditional distributions can be obtained and sampled from efficiently, being either Dirichlet or multinomial distributions. Our algorithm implemented in python is available at GitHub [29].

In practice, subsequently drawn samples are highly correlated. To mitigate this we drop the first M samples, which constitute what is called the burn-in phase, and then apply a "thinning" procedure which consists in only keeping every lth sample. We also implement different chains, or walkers, initialized at different randomly chosen starting points. We estimate sufficient M and l values by computing the integrated autocorrelation time τ as defined in Ref. [30] and adapting its implementation in EMCEE [31] accounting for the fact that we do not have an ensemble sampler. We find that with 30 walkers and 1000 saved iterations per walker after thinning with $l = 100$ we have τ's in the range $\tau \in [1, 2.5]$. We consider a burn-in phase of $M = 1000$ after which we save the aforementioned 1000

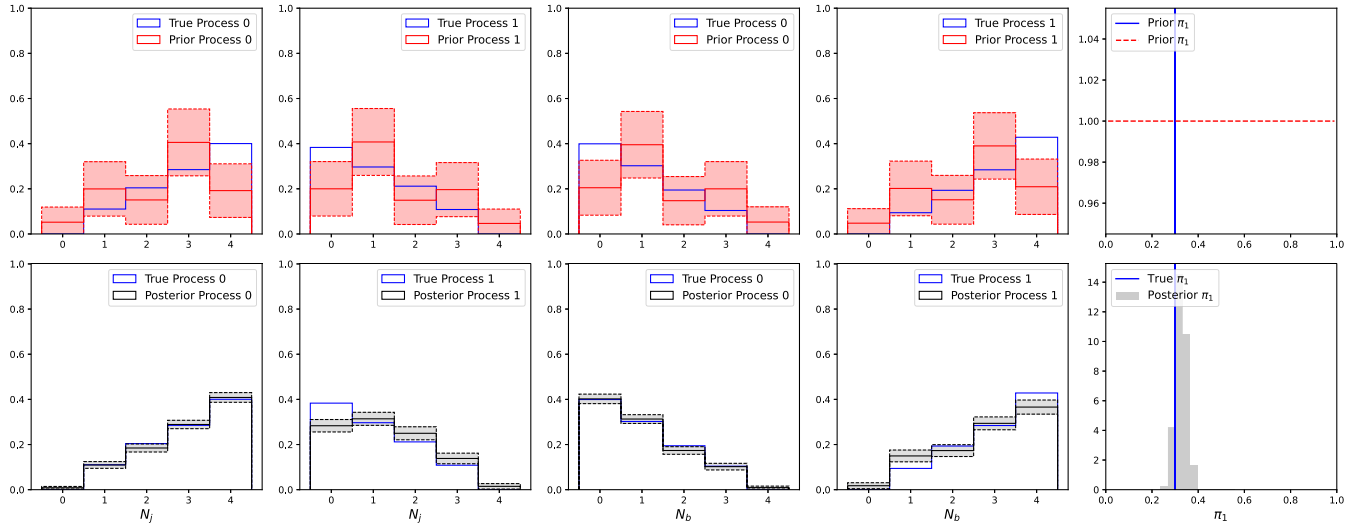


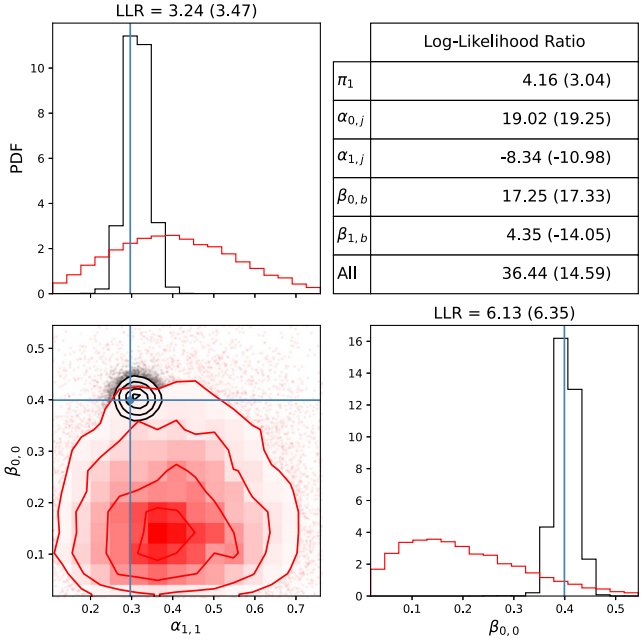
FIG. 2. N_j, N_b, and π_1 distributions: true values (blue), priors (red), and posterior (black) for the toy model. Shaded regions in first four plots indicate the 1σ uncertainty region. Comparing the posteriors to the priors one can appreciate the improvement in estimating the true distributions departing from incorrect and uncertain priors using Bayesian inference on the data.

samples with thinning.[5] Once we have an accurate approximation of $p(Z, \pi, \alpha, \beta|X)$, we can marginalize over the class assignments by neglecting the sampled values Z.

A. A simple toy example

To demonstrate the efficiency of this approach, as well as the limitations due to the approximations we make, we will first apply it to inference in a very simple toy model. We take a sample of "events," each with just two features. The sample is comprised of two types of events, which for the sake of analogy we call background and signal. These signal and background events are sampled from sets of overlapping distributions in the feature-space. The features for each event are sampled independently, therefore in this simple toy example these two features are completely uncorrelated from each other. We consider the case in which the prior distributions for these features are not too far from the truth. In contrast, we consider a uniform prior distribution for the π parameter giving the fraction of signal and background in the sample. This indicates no prior knowledge on how much background and signal we can expect in the dataset and is the most conservative assumption we can make in this regard. We show the prior distributions as well as the true values of the parameters in the upper row of Fig. 2.

After numerically solving the Bayes inference problem using GS, we compare the class-0 and class-1 inferred distributions for N_j and N_b to the truth-level background and signal distributions in X. A good summary to assess the success of the algorithm is the corner-plot which visualizes the distribution through marginalizing to either two or one parameter dimensions and the true values. An excerpt is shown in Fig. 3. In each panel we show the corresponding



Log-Likelihood Ratio	
π_1	4.16 (3.04)
$\alpha_{0,j}$	19.02 (19.25)
$\alpha_{1,j}$	-8.34 (-10.98)
$\beta_{0,b}$	17.25 (17.33)
$\beta_{1,b}$	4.35 (-14.05)
All	36.44 (14.59)

FIG. 3. Excerpt from the corner-plot for the toy model. Red indicates the prior distribution, black the posterior distribution obtained through GS and blue is the true value. We see how the posterior distribution captures the correlation between N_j and N_b. The titles of each 1D histogram contain the log-likelihood ratio between the posterior and the prior using either GS or VI for the posterior estimation, with the latter shown in parentheses. The table contains the sum of log-likelihood ratios per parameter block, again considering the posteriors obtained through GS and through VI. We see that VI is a bad approximation to GS, failing to improve on the prior for several parameter blocks.

[5]The GS algorithm and techniques described above are well known in other disciplines, in particular computer sciences, however they have to our knowledge not been applied before in the context of (high energy) physics.

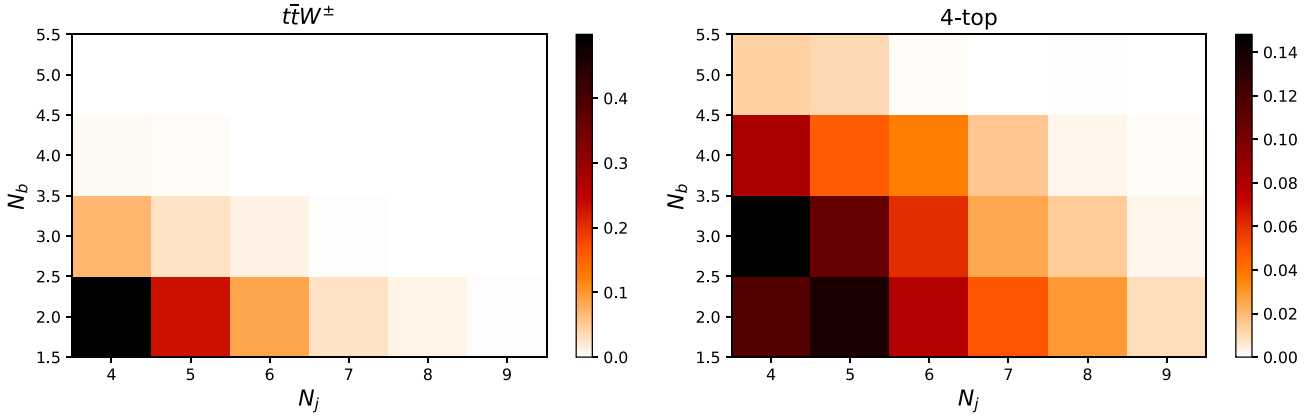


FIG. 4. N_j and N_b two dimensional distributions for $t\bar{t}W^{\pm}$ and 4-top.

prior distribution (red), posterior distribution (black) and the true values (blue). Quantitatively, one can also compare the level of improvement between the prior and the posterior by computing their log-likelihood ratio (LLR) with respect to the true value for each parameter. We display these numbers above the diagonal panels of the corner plot, and we see a robust improvement in most of them. To compute the LLR of the posterior and prior of the complete model one would in principle need to evaluate the joint density distributions of all pairs of parameters (off-diagonal elements in the corner-plot) which is beyond the scope of this work. Instead, as a rough approximation, neglecting the correlations between the parameters, we obtain a global LLR as a sum of the individual parameter LLRs, LLR ≈ 36. We display this global sum as well as partial sums grouping different parameters together in Fig. 3. We also include the partial and global sums of LLR obtained when approximating the posterior through VI. We observe that although VI captures the maximum of the posterior accurately, it consistently underestimates the variance of the distribution yielding a too narrow approximation to the GS obtained posterior. This is reflected in a lower improvement over the prior (LLR ≈ 15).

Finally, in the bottom row of Fig. 2 we group together the one-dimensional marginalized posterior distributions for each parameter to obtain the N_j, N_b distributions of the signal and background, as well as for the π parameter, i.e., the fraction of signal in the sample. In the plot the true value of the parameters is shown in solid blue. Notice that the posterior exhibits good convergence to the true values as well as a considerable reduction of the uncertainty, when compared to the prior, which emulates the imperfect MC. We find an improvement in both the N_j and N_b distributions for each process as expected from Fig. 3. It is also interesting to notice in Fig. 2 how from a complete ignorance of the signal and background fractions in the sample, the algorithm recovers a pdf for π in good agreement with its true value.

III. APPLICATION TO FOUR-TOP MEASUREMENTS

In the 2LSS$++$ channel, the final state is usually characterized by at least $2\ell^{+}$, at least $2b$-tagged jets, and at least 4 light jets. Additional cuts on missing transverse energy and transverse momentum may be invoked to enhance the signal fraction in the sample. The exact details of the event selection are however not important for the purposes of this work. From the decay products at matrix-element level of the signal, one expects *a priori* that the N_j and N_b distributions to be skewed toward higher values when compared to the background process, thus providing enough separation for disentangling them using statistical inference.

In our setup we have simulated 4-top and $t\bar{t}W^{\pm}$ events using MADGRAPH [32], PYTHIA [33], and DELPHES [34] to account for matrix level calculations and showering, hadronization and detector simulation, respectively. We selected $N = 500$ events, roughly equivalent to $\mathcal{L} = 800$ fb^{-1}, in the 2LSS$++$ channel with 70% background and 30% signal (we also tested for other signal fractions and obtained similar results). Using this data we created a dataset X, represented by N pairs (j_n, b_n), $n = 1, \ldots, N$, to serve as our benchmark truth-level sample. The resulting two dimensional distributions are shown in Fig. 4.

We observe from Fig. 4 that the N_j and N_b distributions do not appear to be strictly (conditionally) independent. This is evidenced by the fact that different rows (columns) show different bin hierarchies depending on the column (row) they are conditioned on. These effects arise from experimental systematics such as imperfect b-tagging and different (b-)jet acceptances, as well as from statistical fluctuations due to finite sample sizes involved: the sample size of the Monte Carlo simulation and the sample size of the expected events at to the collider luminosity considered. Category bins with very small event yields are particularly affected by these later effects. In Sec. IV we study in detail how well our model approximates the true data

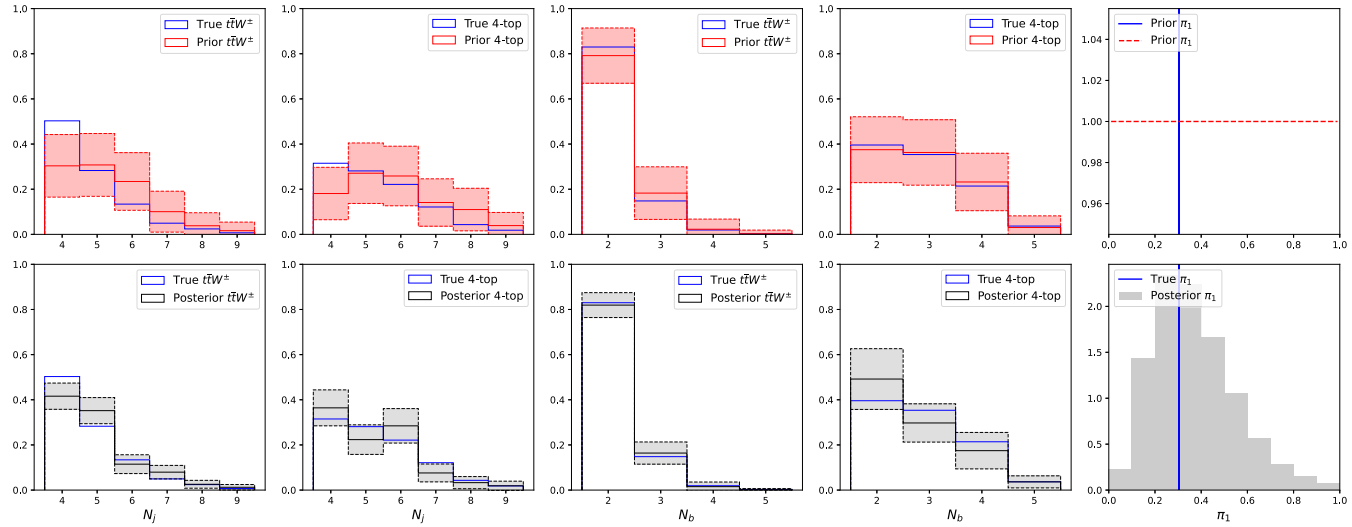


FIG. 5. N_j, N_b, and π_1 distributions: true values (blue), priors (red) and posterior (black). Shaded regions in first four plots indicate the 1σ uncertainty region. Comparing the posteriors to the priors one can appreciate the improvement in estimating the true distributions departing from incorrect and uncertain priors using Bayesian inference on the data.

distributions even when conditional independence is not exact. We find that for the foreseen (HL)LHC luminosities our model is statistically indistinguishable from the data while retaining classification power to infer the 4-top and $t\bar{t}W^{\pm}$ distributions.

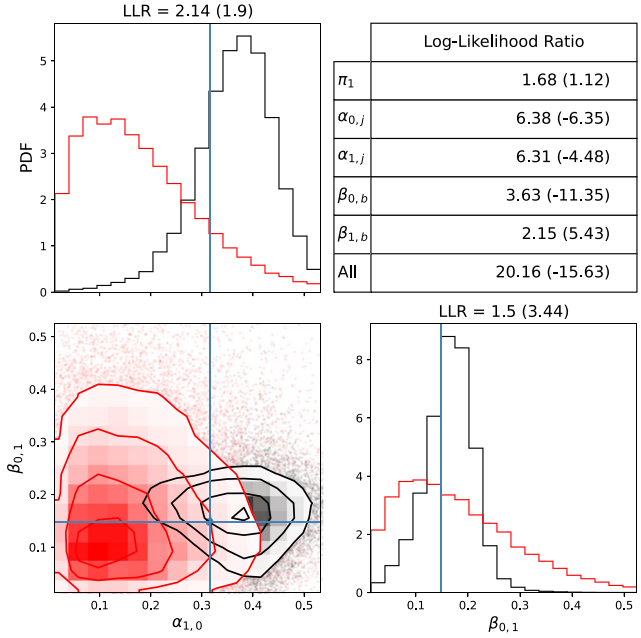
On the other hand, regarding the potential MC mismodeling, we would like to emphasize that our model is aimed to work directly on data and thus address this very kind of problem. That is, we care that our model recovers the true underlying distribution with imperfect (i.e., MC based) priors. In this context we use MC simulations as stand-in mock data for actual (mixed) distribution measurements and apply our model to this mock data with imperfect knowledge encoded in the priors. In order to emulate an imperfect MC prior we skewed the corresponding N_j and N_b distributions from X to higher values and incorporated this into our model through the prior hyperparameters. In general, we can write the hyperparameters η of a V-dimensional Dirichlet distribution of a random variable θ as $\eta_v = \Sigma \cdot p_v$, for $v = 1, \ldots, V$. Here p is a multinomial probability distribution and Σ is a normalization factor. The role of p_v and Σ can be understood by looking at the mean and variance of θ_v:

$$\mathbb{E}[\theta_v] = p_v$$
$$\mathrm{Var}[\theta_v] = \frac{p_v(1-p_v)}{\Sigma+1}. \tag{2}$$

From these equations, we see that p_v represents the expected value of θ_v while Σ controls the confidence we have on that expectation. We fixed the p_v values of the priors for α and β in their respective Dirichlets to the normalized N_j and N_b populations given by the imperfect MC predictions. To reflect our confidence in this estimate, in this example we chose $\Sigma = 10$ for each Dirichlet. See Fig. 5 upper row, where we plot the central values and 1σ ranges for the prior distributions for α and β. In an actual experimental analysis, Σ could be chosen such that the priors cover all reasonable ranges of the modeled observables. As an extreme example, for the prior on the π parameter, giving the fraction of signal and background in the sample, we take a uniform distribution, indicating no prior knowledge on how much background and signal we can expect in the dataset.

As we do for the toy model, we study the posterior distribution obtained using GS through the corner-plot, with its LLR partial and global and sums, and through histograms that condense the class-0 and class-1 N_j and N_b probability distributions and the π probability distribution. We show an excerpt of the corner-plot in Fig. 6. The global sum of the LLRs is ≈20, reflecting an improvement over the prior. In comparison, the VI estimated posterior does not show an improvement over the prior. This is due to the narrow width of the approximation which excludes the true values of the parameters to a higher level than the more accurate GS obtained posterior estimation.

In Fig. 5 we show the results for N_j and N_b distributions of the signal and background, as well as for the π parameter, i.e., the fraction of signal in the sample. As in the toy model case, the posterior exhibits good convergence to the true values as well as a considerable reduction of the uncertainty when compared to the prior which emulates the imperfect MC. However, in this case the improvement is different for each feature. The N_j distribution shows a larger improvement, as expected from Fig. 6, while the N_b distribution is harder to reconstruct due to the much larger fraction of events populating the first bin. Similar results are obtained



	Log-Likelihood Ratio
π_1	1.68 (1.12)
$\alpha_{0,j}$	6.38 (-6.35)
$\alpha_{1,j}$	6.31 (-4.48)
$\beta_{0,b}$	3.63 (-11.35)
$\beta_{1,b}$	2.15 (5.43)
All	20.16 (-15.63)

FIG. 6. Excerpt from the corner-plot. Red indicates the prior distribution, black the posterior distribution obtained through GS and blue is the true value. We see how the posterior distribution captures the correlation between N_j and N_b. The titles of each 1D histogram contain the log-likelihood ratio between the posterior and the prior using either GS or VI for the posterior estimation, with the latter shown in parentheses. The table contains the sum of log-likelihood ratios per parameter block, again considering the posteriors obtained through GS and through VI. We see that VI is a bad approximation to GS, failing to improve on the prior for several parameter blocks.

for other cases which differ in signal-to-background ratio and number of events. It is also interesting to notice in Fig. 5 how again from a complete ignorance of the signal and background fractions in the 2LSS++ sample the algorithm recovers a pdf for π in good agreement with its true value. We also checked that this agreement holds for other truth values of π, and that the matching only worsens as the value of π approaches the boundaries of [0, 1].

In summary, we find that the algorithm successfully infers the N_j and N_b distributions as well as the signal/background fractions. Notably, the best inference occurs for the N_j distribution, which is usually the hardest to predict correctly through MC simulations based on perturbative QCD calculations matched to parton shower algorithms.

IV. TESTING MODEL VALIDITY

Our method hinges on the validity of the underlying statistical (generative) model. Thus it is imperative to understand how well our model that assumes conditional independence, approximates the true data distributions even when their conditional independence is not exact. To quantify the agreement between the data and our model we consider the mutual information (MI) $I(N_j, N_b)$ between N_j and N_b,

$$I(N_j, N_b) = D_{\mathrm{KL}}(p(j,b)||p(j)p(b)) = \sum_{j=4}^{9}\sum_{b=2}^{5} p(j,b)\mathrm{Ln}\frac{p(j,b)}{p(j)p(b)}. \quad (3)$$

The MI encodes how much information is lost by approximating the full distribution with the product of the two marginal distributions. We can also condition the MI on the class label and obtain the MI for each process $I(N_j, N_b|z)$. By combining the per process MI, we build the conditional MI $I(N_j, N_b|Z) = \sum_z p(z) I(N_j, N_b|z)$ which encodes our exact model hypothesis: the data follows a probability distribution which can be written as a combination of two processes, each of which presents a factorized probability distribution. We should note that $I(N_j, N_b|z)$ and $I(N_j, N_b|Z)$ depend explicitly on the availability of labeled data and thus are not computable purely from measured distributions. However, because we expect the simulations to be qualitatively reasonable approximations to measurements, studying the validity of the modeling hypothesis using MC simulations is justified.

Using our finite 4-top and $t\bar{t}W^{\pm}$ dataset, we can estimate the relevant probability distributions and obtain finite sample estimations of the relevant MIs. In the large statistics limit, the estimator follows compact asymptotic distributions [35]. However, we are dealing with finite event samples where some category bins are scarcely populated. Thus, in order to quantify the compatibility of our model with the data, we do a series of pseudoexperiments according to the following procedure:

(1) We take the expected event rates obtained from the MADGRAPH+ PYTHIA+ DELPHES pipeline and their uncertainties to generate 2500 pseudodatasets. For each pseudodataset, we sample the expected event rate for each bin according to a Gaussian centered in the MC central value and with the appropriate uncertainty. Then, we sample the observed events for that bin through a Poisson distribution.

(2) For each of these pseudodatasets, we compute the two-dimensional probability distribution and the marginals for each process and for the full dataset. With these, we obtain the estimators of all four relevant MIs $\hat{I}(N_j, N_b|z)$, with $z = t\bar{t}W^{\pm}, 4-\mathrm{top}$, $\hat{I}(N_j, N_b)$, and $\hat{I}(N_j, N_b|Z)$.

(3) We use these estimators to study the validity of approximating the joint probability distribution with a certain modeling hypothesis. To this end we construct the probability distribution of the estimator by generating another batch of 2500 pseudodatasets. This time, each pseudodataset is generated using the relevant approximation: for $I(N_j, N_b|z)$, we generate the pseudodatasets with $p(j|z)p(b|z)$; for $I(N_j, N_b)$,

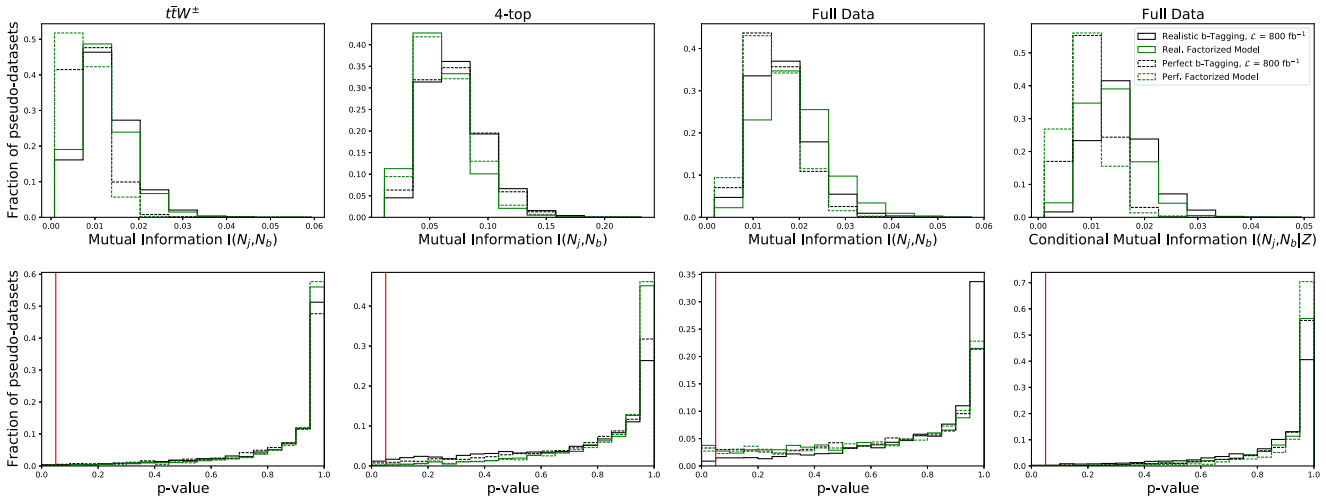


FIG. 7. Top row: we show in solid (dashed) black lines the MI between N_j and N_b with realistic (perfect) b-tagging. In solid (dashed) green we show the MI distribution for the expected event rates which respect conditional independence with realistic (perfect) b-tagging. Bottom row: we show with the same color and line conventions the p-value of the null hypothesis distribution of each estimator. We show in red the $p = 0.05$ conventional exclusion value. We can see that for the considered luminosity, N_j and N_b cannot be ruled out to be conditionally mutually independent. See text for details.

we generate the pseudodatasets with $p(j)p(b)$; and for $I(N_j, N_b|Z)$, we generate the pseudodatasets with $\sum_z p(z)p(j|z)p(b|z)$. The hypothesis that the obtained estimators $\hat{I}$ are sampled according to the model is the null hypothesis H_0.

(4) Having obtained the probability distribution of each estimator conditioned on its null hypothesis H_0 using these additional pseudodatasets, we compute the one-sided p-value for the "measured" estimator which allows us to discard the null hypothesis with a certain confidence level.[6] The p-value can be computed as

$$\text{p-value} = \int_{\hat{I}}^{\infty} p(I|H_0)dI$$

where I can be any of the four metrics considered and H_0 its associated null hypothesis. In the large statistics limit, this one-sided test asymptotically converges to the compact formulae considered in Ref. [35].

We show the results of this procedure in Fig. 7 for four types of pseudodatasets. In solid black line we show the pseudo-dataset generated with the expected events as obtained from the MADGRAPH+ PYTHIA+ DELPHES pipeline. In dashed black line we consider the event rates we obtain when considering perfect b-tagging in DELPHES. We do this to verify whether the introduction of imperfect b-tagging, and the resulting correlations between the number of light- and b-jets, spoil conditional independence. In green solid and dashed lines we modify the sampled expected event rates to ensure conditional independence for realistic and perfect b-tagging. These two pseudodatasets thus agree with our modeling hypothesis and provide a self-consistency check. One should note that the Poisson sampling with relatively small event rates induces a slight violation of conditional independence as it is done in a bin by bin basis.

In Fig. 7, we observe that for the considered luminosity $\mathcal{L} \simeq 800$ fb^{-1}, the data and our model are not statistically distinguishable from each other. This can be seen from the first, second, and fourth columns, where the null hypothesis coincides with the green curves. The p-value distributions in the first and second column imply that 4-top and $t\bar{t}W^{\pm}$ cannot be ruled out to have factorized (N_j, N_b) distributions while the fourth column implies the same for the full data and the model which assumes conditional independence. For the third column, both the data and the model are different from the null hypothesis that considers full independence between N_j and N_b. In that case, both the model and the data show slight disagreements with the null hypothesis although they remain compatible with it. We observe how the p-value distribution is more tilted toward the discarded region for the MI compared to the Conditional MI for the full data distribution, specially for perfect b-tagging. Conditional independence is thus a reasonable modeling hypothesis that yields qualitatively different behavior than assuming a single process with a factorized (N_j, N_b) distribution. Because conditional independence assumes that correlations between light- and

[6]Although not explicit, there is an assumed alternative hypothesis H_1: the saturated model. For a given pseudodataset of N events sampled from a multinomial distribution, its MI is nothing more than $2N$ times its saturated log-likelihood [36].

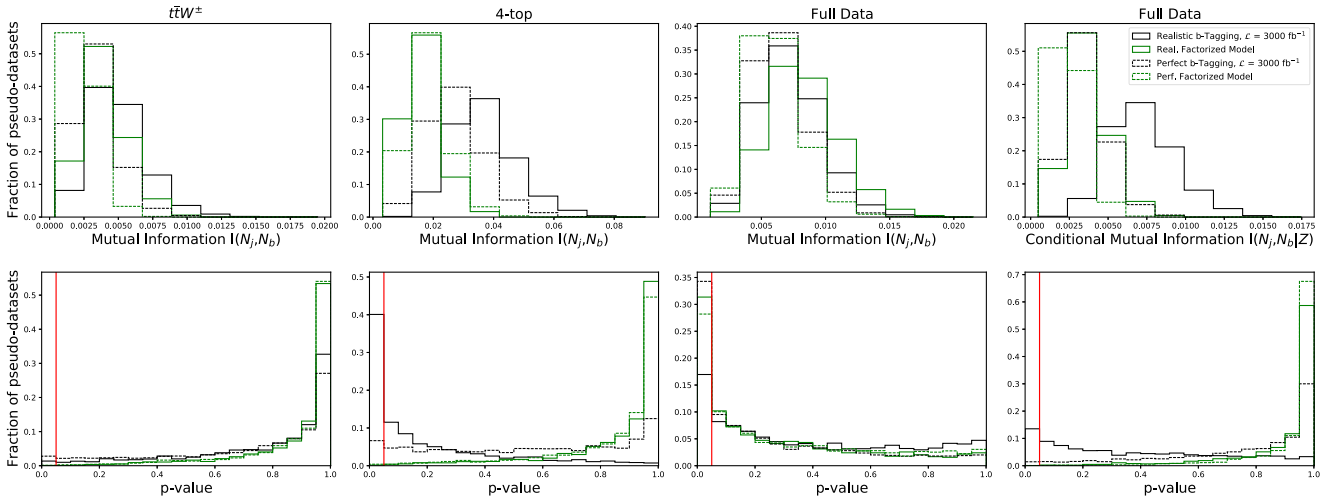


FIG. 8. Same as Fig. 7 but for projected high-luminosity expected event rates.

b-jets are induced by marginalizing over the labels, the model acquires classification power for the underlying processes (that we can match to 4-top and $t\bar{t}W^{\pm}$) by learning the induced correlations to achieve explanatory power over the full data distribution.

The different hypotheses become better distinguishable at larger luminosities. This is seen in Fig. 8 where we show the results for high-luminosity LHC projections with $\mathcal{L} = 3000$ fb^{-1}. We observe that 4-top exhibits larger deviations from independence than $t\bar{t}W^{\pm}$. In particular N_j and N_b independence can be ruled out for the 4-top distribution with realistic b-tagging. This in turn causes the full data distribution to be tilted toward lower p-values for the conditionally independent null hypothesis. The $t\bar{t}W^{\pm}$ does not exhibit the same behavior. We verify that the MI of both processes decreases considerably in the case of (near) perfect b-tagging. In particular, joint (black) 4-top distribution is much closer to its marginalized (green) counterpart which is also reflected in the full data conditional MI distribution This implies that imperfect b-tagging is indeed an important factor behind observed deviations from the conditional independence hypothesis although it is not the only one. Because we are considering a probabilistic model for the data, a feasible sophistication of this model that includes b-tagging efficiencies as a random variable could restore conditional independence while keeping the number of parameters under control. Such incorporation of the b-tagging efficiency would be the analogue to the introduction of an associated nuisance parameter in traditional statistical analyses. Another key feature at HL-LHC luminosity is that for all four pseudodatasets full independence between N_j and N_b can be ruled out, as evidenced by the third column. For perfect b-tagging, we can conclude that conditional independence is a valid approximation which yields learnable distributions with discriminatory power between processes. If imperfect b-tagging is taken into account in the generative model then conditional independence remains a valid modeling hypothesis with explanatory power for the full range of luminosities expected at the LHC.

V. CONCLUSIONS

In summary, we have proposed a new technique to extract signal and background features and fractions relevant for measurements of four-top production at the LHC using Bayesian inference on the N_j and N_b jet multiplicity distributions. It relies on the assumption of conditional (upon signal and background class) independence of the inferred distributions and harnesses the resulting correlations between N_j and N_b within each class. The algorithm is weakly supervised since, in addition to data (in the signal region), it only relies on imperfect *a priori* knowledge how the signal and background differ in their N_j and N_b distributions. Using these results we have proposed a novel approach to test or tune MC predictions in the signal region. Alternatively, it could allow to measure four-top production cross section and/or test for NP effects in a novel way that alleviates the dependence on MC simulations altogether, as also proposed in Ref. [19]. One could for instance tune the MC in the signal region using the class-0 (background) N_j and N_b distributions and then simulate the signal using the tuned MC to check whether its predicted fraction in the 2LSS $++$ sample agrees with the predictions in $p(\pi|X)$. Moreover, one can also check whether the MC signal N_j and N_b distributions match the $p(\alpha_{1,i}|X)$ and $p(\beta_{1,i}|X)$ inferred by the algorithm. Using these ideas one would be able effectively to compute acceptances with a MC tuned *in-situ* in the signal region, while simultaneously measure the four-top cross-section, or study potential NP contributions to the signal or the backgrounds.

Certainly, our method as presented in Sec. II is general and applicable also to other particle physics scenarios

beside four-top production and potentially opens new venues of searches for NP at colliders. Certainly however, much further work is needed to implement these techniques into feasible experimental analyses.

ACKNOWLEDGMENTS

Acknowledgments J. F. K. acknowledges the financial support from the Slovenian Research Agency (Grant No. J1-3013 and research core funding No. P1-0035). B. D. M. acknowledges funding from BMBF. D. A. F. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under Grant agreement No. 833280 (FLAY), and by the Swiss National Science Foundation (SNF) under Contract No. 200021-175940. We thank the referee for his/her report, which has considerably improved the content of the article.

[1] G. Kasieczka *et al.*, The LHC Olympics 2020: A community challenge for anomaly detection in high energy physics, Rep. Prog. Phys. **84**, 124201 (2021).
[2] G. Kasieczka, B. Nachman, M. D. Schwartz, and D. Shih, ABCDisCo: Automating the ABCD method with machine learning, Phys. Rev. D **103**, 035021 (2021).
[3] A. Ghosh, B. Nachman, and D. Whiteson, Uncertainty aware learning for high energy physics, Phys. Rev. D **104**, 056026 (2021).
[4] K. Benkendorfer, L. L. Pottier, and B. Nachman, Simulation-assisted decorrelation for resonant anomaly detection, Phys. Rev. D **104**, 035003 (2021).
[5] S. Choi, J. Lim, and H. Oh, Data-driven estimation of background distribution through neural autoregressive flows, arXiv:2008.03636.
[6] F. Flesher, K. Fraser, C. Hutchison, B. Ostdiek, and M. D. Schwartz, Parameter inference from event ensembles and the top-quark mass, J. High Energy Phys. 09 (2021) 058.
[7] B. Lillie, J. Shu, and T. M. P. Tait, Top compositeness at the Tevatron and LHC, J. High Energy Phys. 04 (2008) 087.
[8] K. Kumar, T. M. P. Tait, and R. Vega-Morales, Manifestations of top compositeness at colliders, J. High Energy Phys. 05 (2009) 022.
[9] B. S. Acharya, P. Grajek, G. L. Kane, E. Kuflik, K. Suruliz, and L.-T. Wang, Identifying multi-top events from gluino decay at the LHC, arXiv:0901.3367.
[10] J. H. Kim, K. Kong, S. J. Lee, and G. Mohlabeng, Probing TeV scale top-philic resonances with boosted top-tagging at the high luminosity LHC, Phys. Rev. D **94**, 035023 (2016).
[11] D. Liu and R. Mahbubani, Probing top-antitop resonances with $t\bar{t}$ scattering at LHC14, J. High Energy Phys. 04 (2016) 116.
[12] J. A. Aguilar-Saavedra and J. Santiago, Four tops and the $t\bar{t}$ forward-backward asymmetry, Phys. Rev. D **85**, 034021 (2012).
[13] J. E. Camargo-Molina, A. Celis, and D. A. Faroughy, Anomalies in bottom from new physics in top, Phys. Lett. B **784**, 284 (2018).
[14] E. Alvarez, A. Juste, and R. M. S. Seoane, Four-top as probe of light top-philic new physics, J. High Energy Phys. 12 (2019) 080.
[15] L. Darmé, B. Fuks, and F. Maltoni, Top-philic heavy resonances in four-top final states and their EFT interpretation, J. High Energy Phys. 09 (2021) 143.
[16] S. Khatibi and H. Khanpour, Probing four-fermion operators in the triple top production at future hadron colliders, Nucl. Phys. **B967**, 115432 (2021).
[17] G. Banelli, E. Salvioni, J. Serra, T. Theil, and A. Weiler, The present and future of four top operators, J. High Energy Phys. 02 (2021) 043.
[18] Q.-H. Cao, J.-N. Fu, Y. Liu, X.-H. Wang, and R. Zhang, Probing top-philic new physics via four-top-quark production, Chin. Phys. C **45**, 093107 (2021).
[19] E. Alvarez, F. Lamagna, and M. Szewc, Topic model for four-top at the LHC, J. High Energy Phys. 01 (2020) 049; 20 (2020) 049.
[20] E. Alvarez, D. A. Faroughy, J. F. Kamenik, R. Morales, and A. Szynkman, Four tops for LHC, Nucl. Phys. **B915**, 19 (2017).
[21] G. Aad *et al.* (ATLAS Collaboration), Measurement of the $t\bar{t}t\bar{t}$ production cross section in pp collisions at $\sqrt{s} =$ 13 TeV with the ATLAS detector, J. High Energy Phys. 11 (2014) 118.
[22] A. M. Sirunyan *et al.* (CMS Collaboration), Search for the production of four top quarks in the single-lepton and opposite-sign dilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 11 (2019) 082.
[23] A. M. Sirunyan *et al.* (CMS Collaboration), Search for production of four top quarks in final states with same-sign or multiple leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV, Eur. Phys. J. C **80**, 75 (2020).
[24] G. Aad *et al.* (ATLAS Collaboration), Evidence for $t\bar{t}t\bar{t}$ production in the multilepton final state in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Eur. Phys. J. C **80**, 1085 (2020).
[25] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, Uncovering latent jet substructure, Phys. Rev. D **100**, 056002 (2019).
[26] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szewc, Learning the latent structure of collider events, J. High Energy Phys. 10 (2020) 206.
[27] B. M. Dillon, T. Plehn, C. Sauer, and P. Sorrenson, Better latent spaces for better autoencoders, SciPost Phys. **11**, 061 (2021).

[28] C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics (Springer, New York, NY, 2006) softcover published in 2016.

[29] Bayesian inference for four tops at the lhc, https://github.com/ManuelSzewc/bayes-4tops (2021).

[30] A. Sokal, Monte carlo methods in statistical mechanics: Foundations and new algorithms note to the reader (unpublished), https://www2.stat.duke.edu/~scs/Courses/Stat376/Papers/Sokal.ps.

[31] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC hammer, Publ. Astron. Soc. Pac. **125**, 306 (2013).

[32] 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.

[33] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015).

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

[35] B. Goebel, Z. Dawy, J. Hagenauer, and J. Mueller, *An Approximation to the Distribution of Finite Sample Size Mutual Information Estimates* (IEEE, Seoul, Korea (South), 2005), Vol. 2, pp. 1102–1106.

[36] S. Baker and R. D. Cousins, Clarification of the use of chi-square and likelihood functions in fits to histograms, Nucl. Instrum. Methods Phys. Res. **221**, 437 (1984).