

Topic modeling in new physics detection

Alexandre Alves^{1,*}, Eduardo da Silva Almeida^{2,†}, and Douglas Roberto de Matos Pimentel^{1,‡}

¹*Departamento de Física, Universidade Federal de São Paulo, UNIFESP, 09972-270, Diadema-SP, Brazil*

²*Departamento de Física do Estado Sólido, Universidade Federal da Bahia, UFBA, 40170-115, Salvador-BA, Brazil*



(Received 26 January 2026; accepted 25 September 2026; published 14 October 2026)

In this work, we apply topic modeling to detect new physics in proton-proton collisions at the LHC in an unsupervised way. We investigate three new physics scenarios where fully leptonic $t\bar{t} \rightarrow b\bar{b}\ell^+\ell^-\nu_\ell\bar{\nu}_\ell$ is the main source of background without relying on jet substructure variables. We demonstrate that the algorithm remains effective even in this low-particle multiplicity framework, complementing jet tagging studies, where it is typically employed. Moreover, we demonstrate that the performance of topic modeling is competitive or even better than well-known semisupervised outlier detectors, such as isolation forest and variational autoencoders, and unsupervised algorithms like Gaussian mixture models and k-means clustering with moderate and high-background pollution in almost all new physics scenarios considered.

DOI: [10.1103/f6dm-g4kw](https://doi.org/10.1103/f6dm-g4kw)

I. INTRODUCTION

With the discovery of the Higgs boson in 2012 at the LHC, the Standard Model (SM) of particle physics filled its last major experimental gap. However, several fundamental questions remain unanswered within this framework, such as the need for dark matter to explain the cohesion of galactic structures, dark energy driving the accelerated expansion of the Universe, and the matter-antimatter asymmetry observed in the cosmos [1], among others.

Over the past decades, numerous theoretical frameworks have been proposed to extend the Standard Model and address these open questions. As a result, LHC searches have largely adopted a top-down strategy, in which analyses are designed to test specific beyond Standard Model (BSM) approaches by generating simulated signal and background events. While this approach has proven effective for well-motivated scenarios, it inherently narrows the exploration space and may overlook unexpected phenomena.

Although many analyses present their outcomes in a nominally “model-independent” form, the event selection and background estimation procedures are still guided by model-specific assumptions. This limitation highlights the

importance of developing truly model-independent techniques that can discover new patterns directly from experimental data [1].

In response to this challenge, the particle physics community has increasingly turned to model-independent search strategies capable of detecting potential signals without assuming a particular new physics model. In this context, semisupervised and unsupervised machine learning (ML) techniques have proven to be particularly promising and have received growing attention in recent years [1,2]. These methods establish a broader analytical context: rather than optimizing for a single theoretical signature, they systematically learn the underlying structures of the kinematic data, providing a comprehensive and less biased sensitivity to unexpected phenomena.

As a matter of fact, we can try to bypass the model dependency on prior knowledge about signals and backgrounds, which is required in the supervised framework, using some kind of unsupervised or semisupervised algorithm. In novelty detection [3], we simulate background events and train an algorithm to recognize them through some measure of, let us say, backgroundness—in autoencoders, for example, this is just the reconstruction error as the event is compressed to a latent representation and reconstructed back to its original representation. Signals will present larger reconstruction errors compared to backgrounds, denouncing them as novel events. Note that novelty detection is an inductive semisupervised learning technique that tacitly assumes a background hypothesis, so it is partially model dependent assuming prior physical knowledge. Moreover, it is sensitive to simulation biases and choices like parton distribution functions,

*Contact author: aalves@unifesp.br

†Contact author: almeidae@ufba.br

‡Contact author: douglas.pimentel@unifesp.br

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

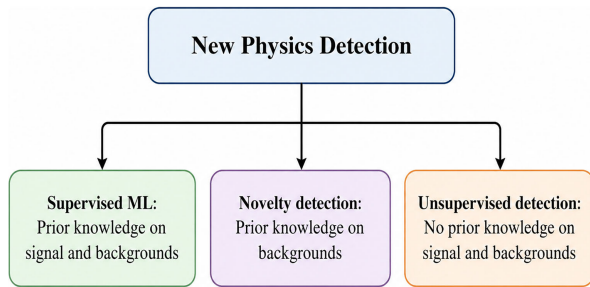


FIG. 1. Machine learning frameworks to separate signals from backgrounds. They distinguish among themselves by the degree of physics prior knowledge, the learning model and simulation settings required in training, and applying the algorithms to real data. In the case of unsupervised detection, model dependency manifest itself as choice of hyperparameters and collision simulations settings in order to evaluate its performance in-sample. Out-of-sample evaluation is intended to be performed in data.

factorization/renormalization scales, high-order corrections, hadronization, and detector modeling.¹ Unsupervised detection, a transductive learning example, does not depend on physics model assumptions other than a final-state topology where we seek new physics and can be applied directly to data, although it also depends on the ML algorithms and protocols, especially on choice of representation and hyperparameters. In Fig. 1, we categorize the three main ML frameworks in searching for signals in colliders in terms of physics prior knowledge assumed for training.

One unsupervised technique is topic modeling (TM), originally developed in the field of natural language processing (NLP) to extract latent topics from a collection of texts without the need for labeled data [5]. Although originally motivated by NLP problems, Metodiev and Thaler [6] observed a conceptual correspondence between the generative process of text documents and the formation of hadronic jets in high-energy collisions: each collision event can be viewed as a “document,” and the observables of the resulting jets as “words.” Under this analogy, the distributions of quarks and gluons produced in collisions can be interpreted as latent topics. The authors demonstrated a proof of concept using the DEMIX model [7] to extract quark and gluon fractions in mixed Z + jet and dijet samples. This approach has since been explored in several related works [8–12].

In the studies by Dillon *et al.* [13–16], the latent Dirichlet allocation (LDA) model was employed to identify signal and background components based on patterns learned from jet substructure in proton-proton collisions. Considering QCD dijet events as background, two physical scenarios were investigated in [13–16]: one involving $t\bar{t}$ production as a signal, and another simulating

¹A way out of this model dependency in semisupervised ML is training the model with real data as shown in [4].

a hypothetical BSM process with a heavy charged vector boson and a scalar field.

In these works, a detailed reconstruction of the jet history was required to compute suitable observables for training the algorithm. However, this dependency on jet-substructure information limits the applicability of such methods to more general searches. At the same time, the field of topic modeling encompasses a broad family of algorithms developed to address different types of data, document sizes, and analytical goals. Despite this diversity, no systematic comparison has yet been performed among distinct topic-modeling approaches applied to the same physical scenario.

In our work, we present, for the first time, an approach to apply topic modeling algorithms for detecting new physics signals in LHC collisions without relying on jet substructure variables and scenarios with low particle multiplicities. Our methodology explores only high-level final-state particle observables, significantly simplifying data preparation. Furthermore, we compare the performance of three distinct topic modeling algorithms [LDA [17], Biterm (BTM) [18], and ProLDA [19]] and discuss their respective strengths and limitations within the context of particle physics applications.

We focus our studies on three new physics scenarios that have totally leptonic $t\bar{t}$ backgrounds in common: (a) a heavy new Higgs boson decaying to SM Higgs bosons, hh , and also in $t\bar{t}$, (b) a nonresonant hh production with non-SM trilinear coupling, and (c) effective operators affecting top-quark pair production. We will show that the resonant production of Higgs and top-quark pairs can be efficiently identified even in scenarios with high-background pollution, while nonresonant hh and effective operators present more challenging scenarios.

We also compare topic modeling with other four algorithms: variational autoencoders (VAE) [20] and isolation forest (IForest) [21], both in the unsupervised and semisupervised settings, and k-means clustering (K-Means) [22] and Gaussian mixture models (GMM) [22], in the unsupervised setting, in the same physics problems. We conducted two experiments: (1) using VAE, IForest, K-Means, and GMM as unsupervised algorithms, that is, without previous training on backgrounds and (2) VAE and IForest as novelty detectors, as they are usually applied in particle physics. Our goal is to investigate topic modeling as a label-free, transductive alternative that does not require a separately specified signal or background training sample. This reduces one important source of model dependence, although the analysis still relies on prior choices concerning the event topology, observables, data representation, and number of latent components.

The manuscript is organized in this way. In Sec. II we describe the ML models, including topic modeling and baselines algorithms; new physics models targeted by the

algorithms are presented in Sec. III; details of simulations can be found in Sec. IV; in Sec. V we describe the training procedure of the algorithms; Sec. VI presents the results of the work, while Sec. VII is reserved for conclusions.

II. MACHINE LEARNING MODELS

A. Topic modeling

Topic modeling refers to a class of unsupervised methods whose goal is to identify latent distributions (commonly called topics) present in a collection of documents. These techniques are widely used in text-mining tasks such as theme and sentiment detection in social media, information retrieval, and recommendation systems [23,24].

A key aspect of topic modeling is that no single algorithm is universally optimal: long highly structured documents, short noisy texts, and sparse messages each require different modeling strategies. Since the introduction of latent semantic indexing in the 1990s [25], many algorithms have been developed, each tailored to specific data characteristics [5].

In particle physics, topic-based approaches have been explored at the jet level [6,13], where the distribution of constituents inside a jet can be interpreted as a “document,” enabling unsupervised separation of signal and background in LHC events. In this work, we extend this idea to events with final-state particles comprising jets, charged leptons, and recovered neutrino solutions in processes with leptonic $t\bar{t}$ as the main background source, without relying on the jet substructure. We investigate whether latent structures extracted directly from basic observables such as invariant masses can provide physically meaningful discrimination power, even under severe background contamination.

The models approached here belong to the class of probabilistic generative models, where documents are assumed to be generated through latent variables, typically discrete topics (as in LDA and BTM), or, in more recent neural approaches such as ProdLDA, continuous latent representations. All models operate under the bag-of-words assumption, treating word occurrences as exchangeable within each document, which allows ignoring word order while modeling word counts. This exchangeability is formally justified by de Finetti’s theorem [26], which states that any infinite sequence of exchangeable observations can be represented as a mixture model involving latent variables. In topic modeling, this means that word tokens in a document can be interpreted as conditionally independent given hidden structures such as topic proportions or latent vectors.

Once the generative process is defined, the objective is to characterize the posterior distribution, which gives the probability of the latent variable, z , conditioned on the observed data,

$$p(z|w) = \frac{p(w, z)}{p(w)}. \quad (1)$$

The core difficulty arises because the denominator, known as *evidence* $p(w)$, which is obtained by marginalizing all hidden variables, is generally analytically intractable to compute. As a result, training these models requires approximating the posterior through numerical or variational inference methods rather than computing it exactly.

1. LDA model

Latent Dirichlet allocation is a generative probabilistic model in which each document in a corpus is represented as a finite mixture over latent topics, denoted by the topic-proportion vector $\vec{\theta}$, and each topic is represented as a probability distribution over the vocabulary, denoted by $\vec{\phi}_k$ [17]. For each word in position n in a document, a latent topic variable z_{dn} is selected from the document-specific mixture $\vec{\theta}_d$, and then the observed word w_{dn} is drawn from the corresponding topic-word distribution $\vec{\phi}_{z_{dn}}$. The joint distribution of all hidden and observed variables for a document with N_d words is

$$p(\vec{\theta}_d, \vec{z}_d, \vec{w}_d | \vec{\alpha}, \vec{\phi}) = p(\vec{\theta}_d | \vec{\alpha}) \prod_{n=1}^{N_d} p(z_{dn} | \vec{\theta}_d) p(w_{dn} | z_{dn}, \vec{\phi}), \quad (2)$$

where $p(\vec{\theta}_d | \vec{\alpha})$ is the distribution of topics in a given document, which follows a Dirichlet distribution. The parameter $\vec{\alpha}$ controls the mixture of topics: high values of $\vec{\alpha}$ correspond to a document containing a broad mix of topics, while low values of $\vec{\alpha}$ result in a document concentrated on a few topics, and it is a hyperparameter of our model. $p(z_{dn} | \vec{\theta}_d)$ is the probability of observing a topic z_{dn} given the topic distribution $\vec{\theta}_d$. Thus, the likelihood of observing a topic z_{dn} depends on how the topics are distributed according to $p(\vec{\theta}_d | \vec{\alpha})$. On the other hand, $p(w_{dn} | z_{dn}, \vec{\phi})$ is the probability of observing a word w_{dn} given a specific topic z_{dn} . The topic-word distributions $\vec{\phi} = \{\phi_k\}_{k=1}^K$ are drawn independently from a Dirichlet prior $\vec{\phi}_k \sim \text{Dir}(\vec{\beta})$. The hyperparameter $\vec{\beta}$ controls how concentrated each topic is over the vocabulary. Since the choice of each topic for each word and each word itself are treated independently, the joint probability density function is the product of all these probabilities. Considering all documents in the corpus, we get

$$p(\vec{w}, \vec{\theta}, \vec{z}, \vec{\phi} | \vec{\alpha}, \vec{\beta}) = \left[\prod_{k=1}^K p(\vec{\phi}_k | \vec{\beta}) \right] \prod_{d=1}^D p(\vec{\theta}_d, \vec{z}_d, \vec{w}_d | \vec{\alpha}, \vec{\phi}). \quad (3)$$

The variables $\vec{\phi}$, $\vec{\theta}$, and $\vec{z}$ are latent and unobserved. By integrating over $\vec{\phi}$, $\vec{\theta}$, and summing over $\vec{z}$, we obtain the marginal distribution, which represents the probability of generating a document, given by

$$p(\vec{w}|\vec{\alpha}, \vec{\beta}) = \int \int \left[\prod_{k=1}^K p(\vec{\phi}_k|\vec{\beta}) \right] \left[\prod_{d=1}^D p(\vec{\theta}_d|\vec{\alpha}) \right] \times \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\vec{\theta}_d) p(w_{dn}|z_{dn}, \vec{\phi}) d\vec{\theta}_d d\vec{\phi}. \quad (4)$$

Integrating the result of the sum over all possible topic assignments is analytically intractable. For this reason, LDA is trained using variational inference [17] or collapsed Gibbs sampling [27].

2. ProdLDA model

A big challenge in topic modeling is that any modification in the topic structure typically requires designing a new inference algorithm. Autoencoder variational inference for topic models (AVITM) was developed to generalize inference in topic models [19]. ProdLDA arises as a concrete application of this framework.

While LDA assumes that words in a document are generated from a mixture of topics,

$$p(w_{dn} = v|\vec{\theta}_d, \vec{\phi}) = \sum_{k=1}^K \theta_{dk} \phi_{k,v}, \quad (5)$$

where the mixture is performed directly in the space of probability vectors, ProdLDA departs from this formulation by performing the combination of topics in the space of natural parameters of the multinomial distribution. Each topic is parametrized by an unconstrained vector $\vec{\beta}_k \in \mathbb{R}^V$, with $\vec{\phi}_k = \text{softmax}(\vec{\beta}_k)$. For a given document, the natural parameters are linearly combined as $\sum_k \theta_{dk} \vec{\beta}_k$, and the word likelihood is defined as

$$p(w_{dn} = v|\vec{\theta}_d, \vec{\phi}) = \frac{\prod_{k=1}^K \phi_{kv}^{\theta_{dk}}}{\sum_{v'=1}^V \prod_{k=1}^K \phi_{kv'}^{\theta_{dk}}}. \quad (6)$$

This formulation corresponds to a Product of Experts model and typically yields sharper and more distinct topics.

Unlike LDA, ProdLDA does not place a Dirichlet prior over $\vec{\theta}_d$. Instead, AVITM introduces a continuous latent variable $\vec{z}_d \sim N(\vec{\mu}_d, \vec{\Sigma}_d)$ and defines $\vec{\theta}_d = \text{softmax}(\vec{z}_d)$. The marginalized likelihood of the corpus is obtained by

$$p(\vec{w}|\vec{\phi}) = \prod_d p(\vec{w}_d|\vec{\phi}), \quad (7)$$

where each document likelihood is given by

$$p(\vec{w}_d|\vec{\phi}) = \int N(\vec{z}_d|\vec{\mu}_d, \vec{\Sigma}_d) \prod_{n=1}^{N_d} p(w_{dn}|\vec{\theta}_d, \vec{\phi}) d\vec{z}_d, \quad (8)$$

which is again intractable, now due to the nonlinear softmax transformation and the Product of Experts likelihood inside the integral. Therefore, ProdLDA is trained using AVITM. A neural encoder $q_\phi(\vec{z}_d|\vec{w}_d)$ approximates the posterior, while the decoder implements the Product of Experts likelihood. Training maximizes the Evidence Lower Bound (ELBO):

$$\mathcal{L} = \mathbb{E}_{q_\phi(\vec{z}_d|\vec{w}_d)} [\log p(\vec{w}_d|\vec{\theta}_d, \vec{\phi})] - \text{KL}(q_\phi(\vec{z}_d|\vec{w}_d) || p(\vec{z}_d)), \quad (9)$$

where KL is the Kullback-Liebler divergence [28].

3. Biterm model

The biterm topic model is a topic modeling method developed for discovering latent topics in a corpus of short texts, like tweets, SMS, and so on. This is important because traditional models, like LDA, need co-occurrence of words at the document level to capture patterns to define latent topics, which suffer from data sparsity in short texts. The BTM captures corpus-level word co-occurrence by modeling unordered word pairs, called biterm [18]. Each biterm represents two words that co-occur within a short context, and in the BTM generative process, every biterm is drawn from a particular topic within a mixture of topics.

Similarly to LDA, each topic z_n has an associated word distribution $\vec{\phi}_{z_n} \sim \text{Dir}(\vec{\beta})$, where $\vec{\phi}_{z_n}$ is a multinomial distribution over the vocabulary. However, now the corpus has a global topic mixture $\vec{\theta} \sim \text{Dir}(\vec{\alpha})$. Given a biterm $b_n = (w_{n1}, w_{n2})$, the probability that it is generated by the model is obtained by marginalizing over the latent topic assignment z_n :

$$p(b_n|\vec{\theta}, \vec{\phi}) = \sum_{z_n=1}^K p(z_n|\vec{\theta}) p(b_n|z_n, \vec{\phi}) = \sum_{z_n=1}^K \theta_{z_n} \phi_{z_n, w_{n1}} \phi_{z_n, w_{n2}}. \quad (10)$$

For all biterms in the corpus we have

$$p(\vec{B}|\vec{\theta}, \vec{\phi}) = \prod_{n=1}^{N_B} \left(\sum_{z_n=1}^K \theta_{z_n} \phi_{z_n, w_{n1}} \phi_{z_n, w_{n2}} \right). \quad (11)$$

The likelihood of the full set of biterms $\vec{B}$ is obtained by integrating out the Dirichlet priors:

$$p(\vec{B}|\vec{\alpha}, \vec{\beta}) = \int p(\vec{\theta}|\vec{\alpha}) \prod_{k=1}^K p(\vec{\phi}_k|\vec{\beta}) \times \prod_{n=1}^{N_B} \left(\sum_{z_n=1}^K \theta_{z_n} \phi_{z_n, w_{n1}} \phi_{z_n, w_{n2}} \right) d\vec{\phi} d\vec{\theta}. \quad (12)$$

Just like in LDA model, this integral is intractable, so BTM is trained using approximate inference, usually collapsed Gibbs sampling, where both $\vec{\theta}$ and $\vec{\phi}$ are analytically integrated out and topics for biterms are sampled directly from conditional distributions. Training therefore proceeds by iteratively sampling latent topics and estimating expected sufficient statistics, bypassing the need to compute the true evidence.

Since each biterm is associated with a topic, BTM can capture semantic relationships between words even when they appear in very short or isolated documents. On the other hand, modeling only global biterms does not retain information about which words appear together within each document. This limits its use when word order or local co-occurrence is important; as a consequence, the BTM is less interpretable at the document level than LDA, which provides the topic distribution per document. However, in practice, a biterm might be better than LDA even for long texts due to semantic relationships.

We now proceed to describe baseline classification algorithms other than topic modeling ones.

B. Baseline models

Topic modeling methods can be employed in a transductive setting, allowing them to be applied directly to real collider data. Similarly to clustering approaches such as K-Means and Gaussian mixture models, it does not require any prior training on simulated background samples although a tuning phase can be performed on simulated data to select appropriate hyperparameters. Novelty detectors, like VAE, IForest, and many others, however, need to be pretrained on background simulations in order to flag outliers. Yet, they could be used as unsupervised algorithms too. In this case, we should not expect them to perform well in regimes where there is not a clear majority class. This is the regime where $S \approx B$. Of course, this usage is not where they shine. On the other hand, semisupervised algorithms are noticeably sensitive to physics simulation and machine learning protocols and biases [29–32] so using them as unsupervised algorithms could weaken that model dependency.

In this section, we describe the models that we will compare topic modeling against. The selected baselines comprise both fully unsupervised methods (K-Means and GMM) and novelty detection methods (VAE and IForest). Because topic modeling belongs to the class of algorithms that can be applied directly to data without prior background training, we also compare it against VAE and

IForest operating in a fully unsupervised mode, where they are not usually employed.

1. Variational autoencoders

Reconstruction-based anomaly detection represents a specific class of unsupervised methods that identify anomalous data through the reconstruction capabilities of deep generative neural networks. These models reconstruct the input data and compare it to the original inputs, identifying anomalies based on the magnitude of the reconstruction error.

A prominent example used in this work is the variational autoencoder [20], a probabilistic generative model comprising an encoder and a decoder to capture complex data distributions. The encoder, denoted as $q_\phi(z|x)$, transforms the input kinematic data x into a latent representation z , characterized by a Gaussian distribution whose mean and standard deviation are parametrized by $\mu_\phi(x)$ and $\sigma_\phi(x)$. To allow gradient-based optimization through this stochastic sampling process, the reparametrization trick is employed. The decoder, $p_\theta(\hat{x}|z)$, then utilizes the sampled latent representation to reconstruct the input data into $\hat{x}$.

The model is trained by minimizing the negative ELBO, which consists of a reconstruction term, measuring the negative log-likelihood of the generated data, and a Kullback-Leibler divergence term that penalizes deviations from a predefined standard normal prior, $p(z) = \mathcal{N}(0, I)$. In our new physics searches, the VAE is trained over a dataset assumed to be background dominated. During inference, background events yield low anomaly scores (low reconstruction error), whereas rare signal events map into poorly generalized regions of the latent space, exhibiting high reconstruction errors and allowing for effective signal-to-background discrimination.

2. Isolation forest

Isolation forest [21] is an ensemble-based outlier detection method that explicitly isolates anomalies rather than profiling normal background instances. The algorithm constructs a forest of random binary trees by recursively partitioning the dataset, randomly selecting an attribute and a split value, until instances become isolated or a stopping criterion is reached. The fundamental premise of IForest is that anomalies are both “few and different” making them significantly more susceptible to isolation than normal points. Consequently, anomalous instances are isolated closer to the root of the trees, requiring fewer partitions. The anomaly score of a given event is derived from its average path length across the forest: events with shorter expected path lengths receive higher anomaly scores, indicating a greater likelihood of being anomalous.

In the context of our new physics searches, the IForest assumes that signal events constitute a small, distinct minority compared to the Standard Model background. It is worth noting, however, that if the expected signal

fraction is unexpectedly high, the underlying assumption that anomalies are “few” breaks down. When anomalous events become abundant, they form dense clusters that require more partitions to be isolated, causing signal events to hide each other (the masking effect). Conversely, this increased density of anomalies distorts the partition space, often causing normal background instances to be isolated too early (the swamping effect). These effects can reduce the effectiveness of the isolation mechanism and potentially degrade anomaly-detection performance.

3. K-Means

K-Means belongs to the class of centroid-based clustering algorithms. The method partitions a dataset of N observations into a predefined number of K clusters, operating by iteratively minimizing a distortion measure defined by the sum of squared Euclidean distances between each data point and its assigned cluster center [22].

Mathematically, the algorithm introduces a set of K prototype vectors, μ_k , representing the cluster centers, and a set of binary indicator variables, $r_{nk} \in 0, 1$, where $r_{nk} = 1$ if data point x_n is assigned to cluster k , and 0 otherwise. The objective is to minimize the cost function

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2. \quad (13)$$

The optimization alternates between assigning observations to the nearest cluster center and updating the cluster centers according to the current assignments, until convergence.

For the analyses considered here, K-Means provides a natural fully unsupervised baseline, operating under conditions similar to those of topic modeling. The algorithm requires only the number of clusters, fixed to $K = 2$, motivated by the presence of signal and background components in the data, without assuming class imbalance or a predefined notion of normality. Nevertheless, its hard-assignment nature imposes rigid cluster boundaries that may be unsuitable for kinematic distributions exhibiting significant overlap in feature space due to similarities in the underlying final-state kinematics.

4. Gaussian mixture model

Gaussian mixture models generalize traditional clustering methods by assigning posterior probabilities rather than deterministic labels. In this framework, the dataset is modeled as a linear superposition of K multivariate Gaussian components [22]. Mathematically, the probability distribution of a given data point x is defined as

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \quad (14)$$

where each component k is characterized by its mean vector μ_k , its covariance matrix Σ_k , and a mixing coefficient π_k . The mixing coefficients represent the prior probability of each component and must satisfy $\sum_{k=1}^K \pi_k = 1$.

A fundamental advantage of GMM is the use of soft assignments, managed through discrete latent variables. For any observed data point x_n , the model calculates the posterior probability, or responsibility $\gamma(z_{nk})$, that the k th Gaussian component generated that specific point. Derived via Bayes’ theorem, this responsibility is computed as

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}. \quad (15)$$

The parameters (μ_k, Σ_k, π_k) are determined through the expectation-maximization algorithm, which iteratively estimates cluster memberships and updates the mixture components until convergence.

By explicitly modeling covariance structures and replacing deterministic assignments with probabilistic responsibilities, GMM can naturally accommodate overlapping populations and nonspherical geometries. In the context of this work, it provides a fully unsupervised baseline operating under conditions similar to those of topic modeling and K-Means, while offering a more flexible representation of kinematic distributions in which signal and background events may exhibit substantial overlap.

In the following section, we provide details of the physics models, simulations of the collision events, and the representation of the data.

III. NEW PHYSICS MODELS

We investigate the performance of topic models in separating signal events from fully leptonic $t\bar{t}$ backgrounds in an unsupervised way. Three interesting new physics scenarios might arise in this context: a heavy neutral scalar that can decay into (1) SM Higgses, (2) top quarks, and (3) mass dimension-six effective operators affecting the top-quark production. We now present models that fit these phenomenological scenarios.

A. Heavy scalar

Consider a new real scalar singlet under the SM group, S . This model is also known as the xSM [33], and exhibits first-order electroweak phase transitions that can be detected, in principle, via gravitational waves in the next generation of satellite experiments like LISA [34]. The interaction Lagrangian of interest for us is

$$V(H, S) = -\mu^2 H H^\dagger + \lambda (H H^\dagger)^2 + \frac{a_1}{2} (H H^\dagger) S + \frac{a_2}{2} (H H^\dagger) S^2 + \frac{b_2}{2} S^2 + \frac{b_3}{3} S^3 + \frac{b_4}{4} S^4 \quad (16)$$

where $S = v_s + s$, and $H^T = (G^+, (v_{EW} + h_{SM} + iG_0)/\sqrt{2})$, and v_s , and v_{EW} represent the new scalar and the SM Higgs vacuum expectation value with $v_{EW} = 246$ GeV, respectively. All the parameters of this Lagrangian are real. The physical scalars of the model can be obtained after electroweak symmetry breaking by the rotation [33]

$$h = \cos \theta h_{SM} + \sin \theta s, \quad H = -\sin \theta h_{SM} + \cos \theta s. \quad (17)$$

The new heavy scalar, H , now interacts with all the SM spectrum and, in particular, with the SM Higgs bosons and top quarks. Within xSM, the couplings of H with hh and $t\bar{t}$, and the mass of the new scalar, as well, can be considered as independent parameters.

The mixture of the scalars also changes the SM Higgs self-interactions. In this case, the trilinear self-coupling of the SM Higgs and the coupling Hhh are no longer independent. However, it is possible to show that there exist blind spots in $H \rightarrow hh$ [35], that is, the coupling vanishes for certain combinations of the Lagrangian parameters, but with sizable trilinear SM Higgs couplings. On the other hand, for a relatively narrow heavy Higgs, the resonant production can be the relevant contribution after selection cuts. We thus can find scenarios with dominant resonant or nonresonant hh production.

Concerning a new heavy scalar, we will consider, effectively, the following interactions in our studies:

$$\mathcal{L}_{\text{int}} \supset \kappa \lambda_{SM} h^3 + \lambda_{Hhh} H h^2 + g_H H t\bar{t}, \quad (18)$$

where κ is a coupling modifier, and the SM corresponds to $\kappa = +1$.

All these interactions can give rise to final states with two bottom jets, two charged leptons, and neutrinos in pp collisions. In the case of the new scalar production and decay to hh or $t\bar{t}$, we should expect hard yields compared to backgrounds. The new scalar interactions are of SM type, and only the mass of the heavy scalar can change the kinematic distributions of bottom jets, leptons, and neutrinos. In the case of the nonresonant double Higgs production, the interplay between the trilinear Higgs diagram and the top-quark box diagram determines the kinematic distributions of the final-state particles, that is, κ is the parameter of control. We will assume that resonant and nonresonant hh and $H \rightarrow t\bar{t}$ will not occur simultaneously. It is important to note that many other interesting models containing new scalars could be a target for this method, for example, the model of Ref. [36] that explains various excesses observed in recent data.

B. Effective operators

We also study model-independent modifications of the gluon triple gauge coupling and four-fermion interactions, using as illustrative examples the following dimension-six operators in the Warsaw basis [37]:

$$\mathcal{O}_{\tilde{G}} = f^{ABC} \tilde{G}^{A\mu}_{\nu} G^{B\rho}_{\nu} G^{C\mu}_{\rho}, \quad (19)$$

$$\mathcal{O}_{qq}^{(1)} = (\bar{q}_L \gamma_{\mu} q_L)(\bar{t}_L \gamma^{\mu} t_L), \quad (20)$$

where f^{ABC} are the structure constants of $SU(3)_C$, $G^{A\mu}_{\nu}$ denotes the gluon field strength, $\tilde{G}^{A\mu}_{\nu}$ its dual, q_L (t_L) the left-handed light-quark (top quark) doublet, and γ^{μ} the gamma matrices. Note that the gluon operator is CP odd. While $\mathcal{O}_{\tilde{G}}$ modifies the trilinear gluon coupling in $gg \rightarrow t\bar{t}$, $\mathcal{O}_{qq}^{(1)}$ represents a contact interaction involving top-quark production in $q\bar{q}$ collisions. At order $1/\Lambda^2$ they contribute via interference with the SM diagrams. We set $\Lambda = 1$ TeV and consider only one operator at a time in our simulations.

IV. SIMULATIONS AND DESCRIPTION OF THE VARIABLES

All simulations of signals and $t\bar{t}$ backgrounds in $pp \rightarrow b\bar{b}\ell^+\ell^-\nu_{\ell}\bar{\nu}_{\ell}$, $\ell = e, \mu$ were performed using MadGraph5 [38] at leading order QCD. We chose the NNPDF2.3QED [39] for the parton distribution functions. Hadronization and showering of quarks and gluons were performed using Pythia8 [40], while FastJet [41] was employed for jet clustering. Detector effects were estimated with Delphes3 [42] using default setting cards.

The following basic selections are required for signals and backgrounds containing two b -tagged jets, two opposite-sign leptons (electrons or muons), and missing energy

$$p_T(b) > 30 \text{ GeV}, \quad |\eta_b| < 2.5, \quad (21)$$

$$p_T(\ell) > 20 \text{ GeV}, \quad |\eta_{\ell}| < 2.5, \quad (22)$$

$$\cancel{E}_T > 20 \text{ GeV}. \quad (23)$$

The variables that we employ to populate the histograms for each event use combinations of two, three, and four visible particles: the two bottom jets and the two charged leptons, plus reconstructed neutrinos using the Higgsness and Topness method [43]. In this method, the neutrinos' momenta must minimize mass constraints that characterize double Higgs or top-pair production. For example, the invariant mass of the two charged leptons and the two reconstructed neutrinos should reconstruct a Higgs boson, while a bottom-jet plus a charged lepton and a reconstructed neutrino compound a top-quark mass. For more details, we refer the reader to the original paper [43]. The quality of the neutrino momentum solution was investigated in Ref. [44] and found to be useful for reconstructing kinematic variables of interest, such as the mass of a heavy-Higgs boson decaying into two SM Higgs bosons, as well as other kinematic variables.

In this work, we demonstrate that these solutions are also beneficial for the unsupervised learning of events with jets, charged leptons, and neutrinos. Discarding the neutrino

solutions, however, would leave us with too few particles to create a useful event representation for topic modeling to work properly. Yet, compared to the hundreds of particles available in jets, our data representation solution for this low-particle-multiplicity scenario was successful, as we will show.

Taking into account the four visible particles and the two reconstructed neutrinos, we can construct, for each event, all possible subsets of final-state momenta containing k particles, with $k = 2, \dots, 6$. For a given value of k , the number of such subsets is given by the binomial coefficient $C_{6,k} = \binom{6}{k}$. This results in a total of $\sum_{k=2}^6 C_{6,k} = 57$ different variables that will define the size of the document (event), where each entry is drawn from the discrete vocabulary (all possible bins) defined by the full binning of the observables. Most variables that we use for the representation require more than one input 4-momenta, so we discard combinations with only one particle, $k = 1$. Note that for jet topics, the size of the documents tends to be much larger, with hundreds of words (bins), as a jet might be composed of hundreds of particles. Another important difference is that, in our case, the number of particles in the final state is fixed, while for jets, the number of components varies.

Each event, ev , is a set of six final-state momenta: $ev = [p_{b_1}, p_{b_2}, p_{\ell_1}, p_{\ell_2}, p_{\nu_1}, p_{\nu_2}]$ that can be combined into subsets of momenta $\{p\}$ using 2- (up to 6-) momenta to form a variable.

The kinematic variables used for the search of data representation are:

- (i) Invariant mass: $M(\{p\}) = (\sum_{p_i \in \{p\}} p_i^\mu)^2$.
- (ii) Transverse momentum: $P_T(\{p\}) = P_T(\sum_{p_i \in \{p\}} p_i^\mu)$.
- (iii) Transverse momentum fraction: $z(\{p\}) = \frac{\min(P_T(\{p\}))}{\sum_{p_i \in \{p\}} P_T(p_i)}$.
- (iv) $\cos \theta^*(\{p\}) = \arctan(\frac{\Delta \eta(p_i, p_j \in \{p\})}{2})$.
- (v) $k_T(\{p\}) = \min(p_T(p_i), p_T(p_j), p_i, p_j \in \{p\}) \Delta R(\{p\})$, where the difference in the $\eta \times \phi$ plane is given by $\Delta R(\{p\}) = \Delta R(p_i, p_j \in \{p\}) = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$, in this case, p_i or p_j can be itself the sum of a subset of the events' momenta.
- (vi) $\cos \theta_{WW}(\{p\}) = \frac{p_i \cdot p_j}{\|p_i\| \|p_j\|}$, $p_i, p_j \in \{p\}$, where the 3-momenta are calculated in the rest frame of the $\ell^+, \ell^-, \nu_\ell, \bar{\nu}_\ell$ system. This is inspired by events containing a Higgs boson decaying to WW .
- (vii) $RIV_t(\{p\}) = \frac{\mathcal{M}}{M_{\text{vis}}(\{p\})}$, where the invisible mass, $\mathcal{M}$, is defined as $\sqrt{p_T(\nu_{t_1})^2 + p_T(\nu_{t_2})^2 + M^2(\nu_{t_1}, \nu_{t_2})}$. Here, ν_{t_i} represents a reconstructed neutrino from the computation of Topness ordered in p_T . The visible mass, $M_{\text{vis}} = M(\{p\})$. This variable is shown to improve the prospects of observation of hh [44] at the LHC.

Now, for each subset $\{p\} \in ev$ we can calculate a 7-D vector $[M, p_T, z, \cos \theta^*, k_T, \cos \theta_{WW}, RIV_t]$. To form the document that represents the event, we stack all these vectors into fixed-size $57 \times N$, $1 \leq N \leq 7$ arrays. Since each choice of N directly affects the dimensionality of the binning space and, consequently, the size of the vocabulary, it may introduce either additional discriminative information or unnecessary noise. In practice, not all variables are equally relevant for discrimination, so we searched for the subset of N variables that delivered the best discriminating performance, as we will discuss in the next sections. We display, in Fig. 2, the variables used to represent the events plus the $\Delta R(\{p\})$ one, which composes the k_T variable for the case of a 1 TeV Higgs, $\Gamma_H = 10$ GeV, decaying to hh and the $t\bar{t}$ background.

It is important to note that the distributions shown in Fig. 2 take into account all the events of the signal and background samples. The distributions of individual events are much sparser once they are populated by 57 values only. In this respect, the binning of the distributions plays an important role.

An exploratory data analysis was therefore performed to define a binning scheme that spans the full physical range of each observable while preserving the relevant kinematic information and avoiding excessive sparsity at the event level. This discretization enables each event to be represented as a bag-of-words, where each populated bin corresponds to a word, allowing the application of probabilistic topic modeling techniques.

For the observables M , p_T , and k_T , which extend up to approximately 1 TeV, a uniform bin width of 20 GeV was adopted. Angular variables ($\cos \theta^*$ and $\cos \theta_{WW}$) were binned with a width of 0.1 within the interval $[-1, 1]$. The variable z , defined in the range $[0, 0.5]$, was discretized using bins of width 0.02, while for RIV_t , which extends up to values of order 5, a bin width of 0.2 was used.

After binning, each event is represented by the set of bins populated by its final-state observables. This discrete representation is then used as input to the topic modeling algorithms, treating each event as a document composed of a small number of words drawn from the global vocabulary. Table I summarizes the relationship between topic modeling and particle physics features.

V. TRAINING PROCEDURE

Unsupervised algorithms are never exposed to true labels, eliminating the risk of label leakage. For each physical scenario and signal-to-background fraction, the models are trained and evaluated directly on the complete mixed corpus of simulated events, typically containing $\mathcal{O}(10^4)$ events. This approach is consistent with data-driven, model-independent search protocols in high-energy physics, such as some of the unsupervised strategies benchmarked in the LHC Olympics 2020 challenge [1], where underlying structures are inferred directly from

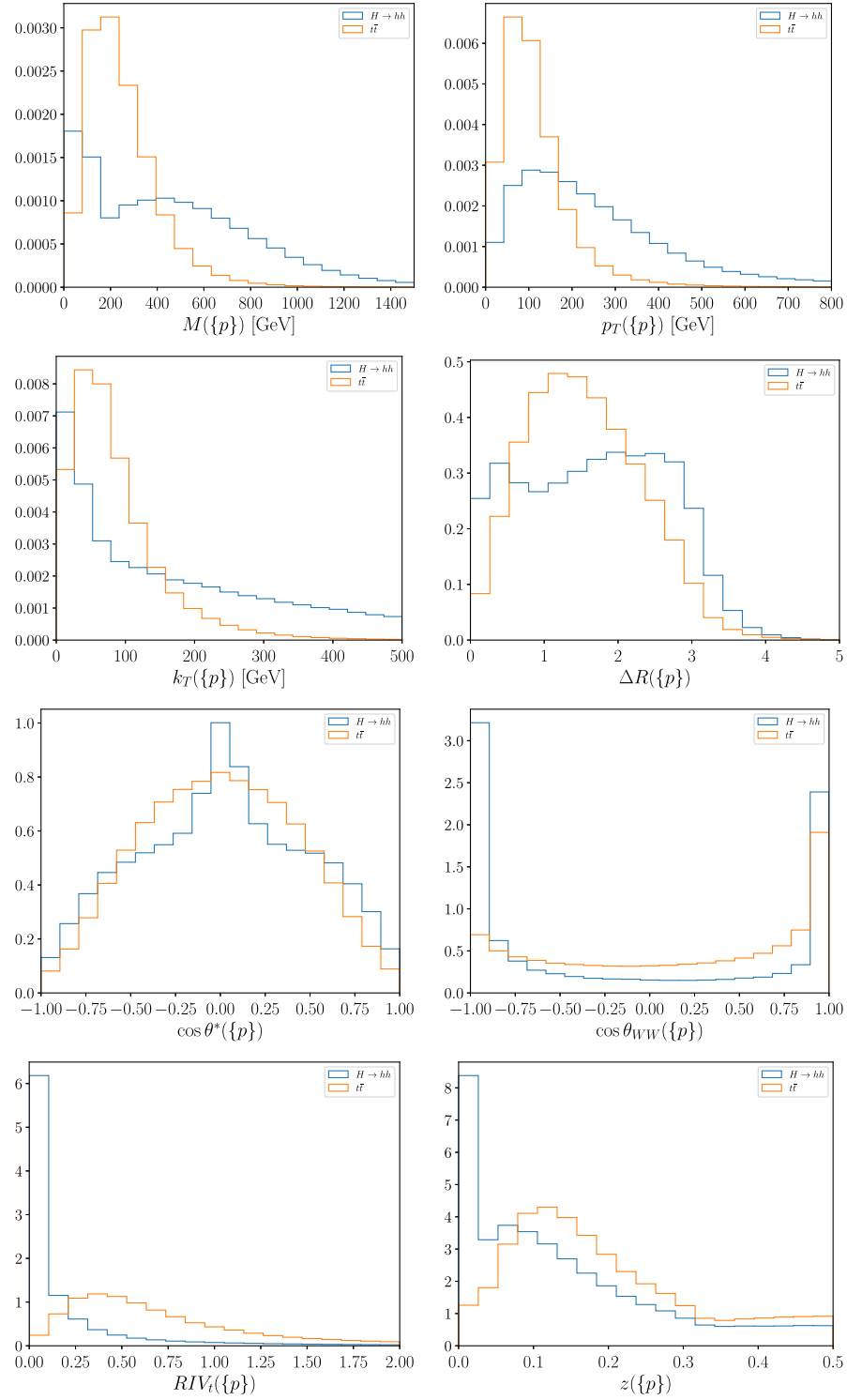


FIG. 2. The normalized distributions of the kinematic variables used for the search for the best representation of the data. The signal events correspond to $H \rightarrow hh$. The ΔR variable is used in the computation of k_T but not included explicitly in the representation of the data.

unlabeled data to disentangle potential signal events from the background. This allows us to quantify the inherent physical separability of the populations without any supervised bias. On the other hand, only in-sample testing is conducted in our

study. Out-of-sample inference with real data or independent Monte Carlo samples created with possibly different settings would be desirable but it is beyond the scope of this work. The transferability of learned topics/clusters to independent

TABLE I. Correspondence between topic modeling concepts and the representation of final-state particle collision events. The dimensionality N depends on the selected subset of kinematic variables.

Topic modeling	Particle physics
Word	Bin in an N -dimensional kinematic histogram
Vocabulary	Set of all possible N -D bins
Document	57 bins inside an event, one from each final-state combination ($C_{6,k}$)
Topic	Latent bin distribution (signal/background)
Corpus	Set of all collision events

datasets is thus an important direction for future work. It is also important to note that even though a training phase is absent, a tuning phase is necessary to boost the performance of the algorithms.

For topic modeling, the optimal kinematic representation of the physical events, however, depends on the specific underlying physics process. Therefore, for each distinct physical scenario evaluated in this work (e.g., $H \rightarrow hh$, $H \rightarrow t\bar{t}$, and effective operators), we perform an independent combinatorial scan across multiple combinations of the binned final-state observables. In this procedure, different sets of variables are grouped to construct the “words” and “documents” fed into the topic models.

To prevent any selection bias tailored to highly imbalanced regimes, this combinatorial scan is exclusively performed in the balanced scenario (background fraction $B/(S+B) = 0.5$). The discriminative power of each combination is evaluated extrinsically using the area under the receiver operating characteristic (ROC) curve (AUC). Once the optimal joint representation is identified for the balanced case, such as the combination of invariant mass, transverse momentum (p_T), and k_T which emerged as the most informative for the 1 TeV $H \rightarrow hh$ process, this specific set of variables is rigidly fixed and systematically applied to all other background fractions within that same physical scenario.

Biterm topic modeling was performed using the `bitermplus` [45] package, latent Dirichlet allocation using `gensim` [46], and ProDLDA was implemented from scratch using the PyTorch framework. For each model, a set of algorithm-specific hyperparameters was allowed to vary during training:

- (i) Biterm topic modeling: α and β ;
- (ii) Latent Dirichlet allocation: α , β , number of passes, and number of iterations;
- (iii) ProDLDA: dropout rate, learning rate, and hidden layer size.

In both LDA and BTM, Dirichlet hyperparameters α and β can be defined either as scalar (symmetric) or vector-valued (asymmetric) parameters. In this work, we adopt symmetric priors, where α and β are single scalar values that are applied uniformly across all topics and vocabulary entries, respectively. This choice reduces the number of free parameters and reflects the absence of prior preference for

specific topics or observables, which is well suited for exploratory and physics-driven applications.

For LDA, two additional training parameters are left free: the parameter `passes` controls the number of full passes over the corpus during model training, while `iterations` sets the maximum number of internal inference steps used to estimate the document-topic distributions.

For the ProDLDA model, we tune three main hyperparameters: the hidden layer size, the learning rate, and the dropout rate. The size of the hidden layer controls the expressive capacity of the encoder network, determining how complex the latent document representations can be. The learning rate governs the step size of the gradient-based optimization and directly affects convergence speed and training stability. The dropout rate acts as a regularization mechanism, randomly deactivating neurons during training to mitigate overfitting and improve generalization, particularly in regimes with limited or noisy data.

For all topic modeling algorithms, the number of latent topics was strictly fixed to $K = 2$, reflecting the Standard Model background and the targeted new physics signal. To determine the optimal configuration for the remaining algorithm-specific hyperparameters, we conduct an automated hyperparameter search.

For the BTM and LDA models, the symmetric Dirichlet priors α and β were allowed to vary continuously within the range of $[0.01, 5.0]$. For the ProDLDA network, the optimization grid was defined by sampling the learning rate from a logarithmic distribution ($[10^{-4}, 10^{-2}]$), the dropout rate from a uniform distribution ($[0.1, 0.5]$), and exploring discrete variations in the hidden layer size. During this automated tuning phase, the quality of the models was assessed intrinsically through perplexity minimization over the corpus. Ultimately, the AUC was calculated as the extrinsic in-sample evaluation metric to quantify the final signal-background separation achieved by the optimized configurations.

Finally, for comparative purposes, the following baseline methods were considered: VAE, IForest, K-Means, and GMM. VAE and IForest were implemented using `PyOD` [47], while K-Means and GMM were implemented using `scikit-learn` [48]. Unlike topic models, which require discretized bag-of-words representations, these baselines

were trained directly on continuous inputs. The input feature space consisted of nine-dimensional continuous vectors composed of kinematic variables, ensuring they operate at maximum resolution without arbitrary binning artifacts. These observables are: the invariant masses of $b\bar{b}$, $\ell^+\ell^+\nu\bar{\nu}$, and $b\bar{b}\ell^+\ell^+\nu\bar{\nu}$ which capture resonances in the SM Higgs decay to $b\bar{b}$, SM Higgs decay to W^+W^- , and the heavy-Higgs resonance, respectively, plus the $\ell^+\ell^-$ mass; the transverse momentum and distance in η - ϕ plane of $b\bar{b}$, and $\ell^+\ell^-$ pairs; and the RIV_i variable.

Due to the nature of these baseline models, they were trained in two different settings: semisupervised (VAE and IForest) and unsupervised (VAE, IForest, K-Means, and GMM). In the semisupervised setting, the VAE and isolation forest models were trained using a pure background dataset containing 25,000 $t\bar{t}$ events. Once trained, these models were applied to the background-signal datasets used in the analysis described in Sec. VI. In the transductive unsupervised setting, each model was fitted directly to the unlabeled mixed sample whose latent structure was to be inferred. No signal or background labels were provided during model fitting or hyperparameter optimization. Monte Carlo labels were used only after inference to quantify the correspondence between the learned clusters or topics and the known physical components. Consequently, the reported AUC values characterize in-sample separability for the transductive decomposition and should not be interpreted as conventional held-out test performance or as a direct measurement of out-of-sample generalization.

The isolation forest algorithm was configured with an ensemble of 100 isolation trees, utilizing automated subsampling constrained to 256 instances per tree (`max_samples`) to mitigate masking effects, and a predefined contamination fraction of 0.1 (`contamination`). The variational autoencoder was implemented as a fully connected deep neural network. The encoder was structured with two hidden layers (32 and 16 neurons), progressively bottlenecking the data into a three-dimensional latent space. The decoder followed a symmetric architecture. All hidden layers utilized rectified linear unit activation functions, while the output layer employed a sigmoid function. The network was trained for 30 epochs (`epoch_num`) with a `batch_size` of 32, using the Adam optimizer with a learning rate (`lr`) of 0.001 and a weight decay of 10^{-5} . Furthermore, the K-Means algorithm was configured to partition the data into two clusters (`n_clusters`), corresponding to the expected signal and background distributions, with the algorithm initialized ten times (`n_init`) using different centroid seeds to ensure robust convergence. The GMM was similarly configured to estimate the underlying data density using two components (`n_components`). Both clustering models employed a fixed random state to guarantee reproducibility. Similarly to the topic models, their discriminative performance was ultimately evaluated using the AUC metric.

VI. RESULTS

This section presents the quantitative and qualitative results obtained from the proposed approach. As primary performance indicators, we display the background rejection versus signal efficiency curves, and summarize the models overall performance using the AUC value. For the topic modeling algorithms, the raw classification score used to construct these metrics is defined as the event-level probability of belonging to a specific latent topic, mathematically represented as $p(z_k|d)$.

It is important to emphasize that topic modeling is a purely unsupervised clustering technique; consequently, the extracted latent topics (e.g., Topic 0 and Topic 1) are permutation-invariant and do not inherently carry “signal” or “background” labels. To evaluate the fundamental discriminative capacity of the algorithms to separate distinct kinematic populations, we calculate the ROC AUC using the topic probabilities and report the absolute separation power defined as $\max(\text{AUC}, 1 - \text{AUC})$. This geometric inversion ensures that the metric reflects the true physical distinguishability of the background and signal distributions in the latent space, regardless of the arbitrary algorithmic indexing of the extracted topics. The same reasoning applies to the baseline models K-Means and GMM.

For all collision processes, each algorithm is evaluated under various contrasting conditions: from a balanced situation with $B/(S+B) = 50\%$ background pollution up to an unbalanced case with 99.5% background events. In general, the unbalanced case is more likely, but it is interesting to investigate the performance of the algorithms in unexpected scenarios once we do not know for sure the exact composition of signal and backgrounds in the experimental data sample [13,14]. Larger signal-to-background ratios might occur when harder selections are imposed on the events or in specific regions of a BSM parameter space.

To ensure the robustness of our results against statistical fluctuations in the dataset composition, we employed a repeated random subsampling approach. For each physics scenario and background-to-signal fraction, ten independent datasets were generated by randomly sampling events from the full simulated data. The performance metrics reported correspond to the average across these ten independent samples, with the variance represented by the uncertainty bands. We first compare the three topic-modeling approaches described in Sec. II A and subsequently contrast the best-performing model with the baseline models introduced in Sec. II B.

A. Comparing topic modeling algorithms

We begin by evaluating the performance of the three topic-modeling approaches described in Sec. II A: LDA, ProLDA, and Biterm. In Fig. 3, we see that all models display strong discriminative power between signal and

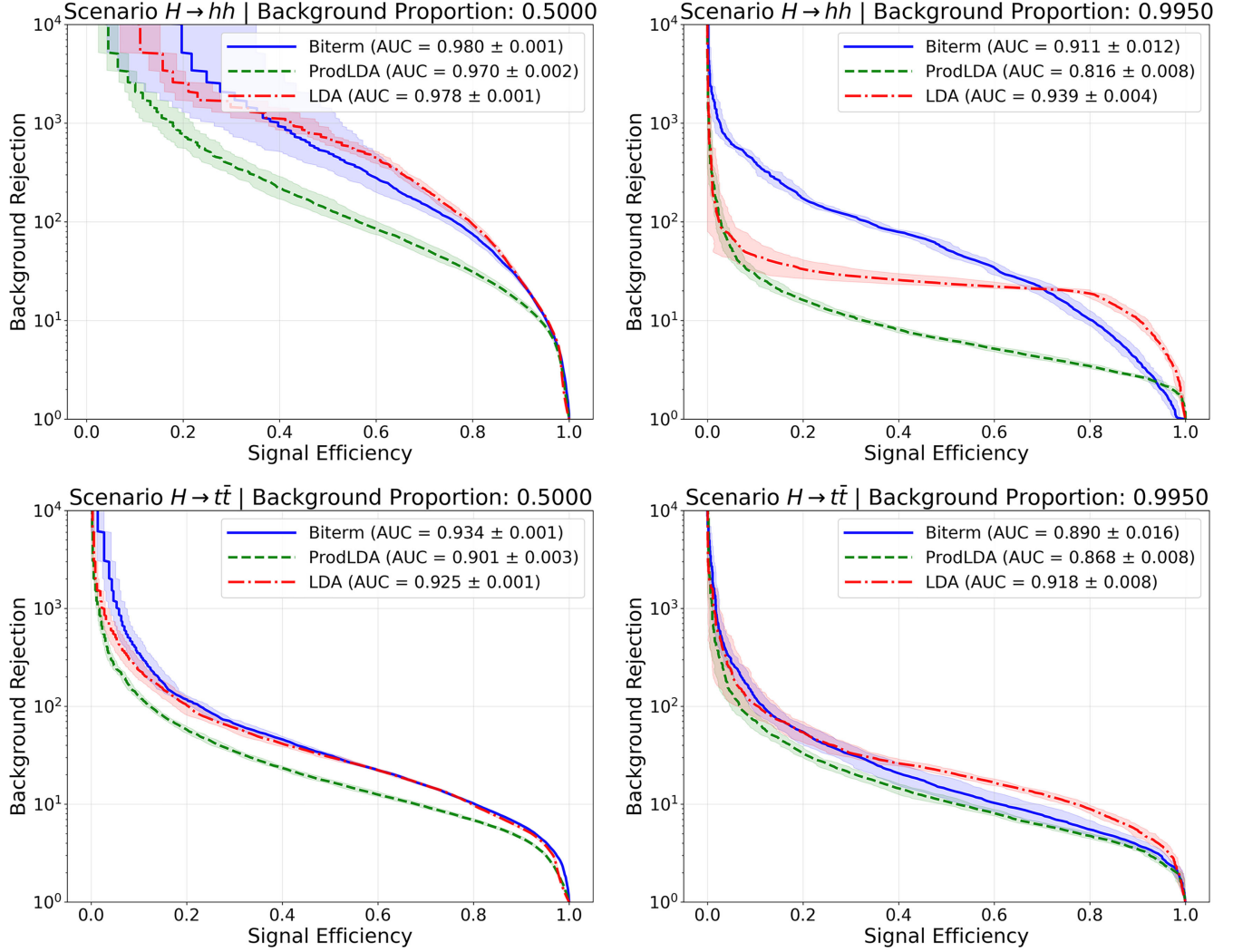


FIG. 3. Background rejection versus signal efficiency curves for topic modeling models: LDA, Biterm, and ProdLDA in $H \rightarrow hh$ (top panels) and $H \rightarrow t\bar{t}$ (bottom panels) processes. The left (right) column shows the case of 50% (99.5%) of background pollution scenario. The legends show the AUC values, which are used to discriminate the ability of each model in disentangling the signal from the background. In this figure and all the subsequent ones, the shaded bands represent the result of a ten-fold random subsampling.

background, with AUC values above 0.9 in 50% of background pollution cases, and above 0.8 in the scenario with 99.5% of background contamination, both in the $H \rightarrow hh$ (upper plots) and in the $H \rightarrow t\bar{t}$ (lower plots) processes. We observe similar behavior in other physics scenarios.

In the 99.5% case, Biterm shows superior mean background-rejection capability across a broad range of signal efficiencies compared to LDA and ProdLDA. LDA has the best overall performance with $\text{AUC} \sim 0.94$ and a narrow band of uncertainties, but the average curve of the Biterm has better performance for signal efficiencies up to ~ 0.4 in both processes.

BTM is very effective in capturing rare global co-occurrence patterns but this global feature, operating at the corpus level, also makes it more prone to variations of the dataset. Especially in comparison to LDA and

ProdLDA, which function at document level, BTM performance varies more strongly due to corpus of documents (the dataset events) changes; see Fig. 3. We perceive this behavior in the tenfold random subsampling band of the $H \rightarrow hh$ process. Nevertheless, the variation in BTM AUC is comparable to the other algorithms because the band widens for very low signal and background efficiencies where the contribution to AUC is small. In other physics scenarios, the band of variation of Biterm is much narrower, like in $H \rightarrow t\bar{t}$, as we are going to see in the next sections.

Considering its computation cost-performance benefit ratio, algorithmic simplicity, and because it is designed for short documents, that is, events with reduced number of particles, we adopt Biterm as the topic modeling method for all subsequent analyses. Now we proceed to compare Biterm against other algorithms in all the collision processes investigated in this work.

B. Heavy Higgs to SM Higgses

First, we investigate the detection of heavy-Higgs bosons to $H \rightarrow hh$. This process exhibits distinctive distributions, which makes it a suitable benchmark for verifying whether the topic-modeling approach can detect the underlying patterns that distinguish signal from background, and how its behavior changes as the background becomes dominant.

Investigating the performance of Biterm against other types of algorithms, we display, in Fig. 4, the background rejection versus signal efficiency curves in extreme cases, balanced (left panel), and unbalanced (middle panel) scenarios. The top row exhibits the Biterm performance versus models used in unsupervised way, while the bottom plots contrast the topic modeling against the semisupervised models.

Let us consider the unsupervised case first. In the 50% background scenario, Biterm, GMM, and K-Means deliver the best overall performances, with $\text{AUC} \gtrsim 0.97$. VAE and IForest, however, perform very poorly. This behavior is not surprising once VAE and IForest are being employed as unsupervised classifiers in a dataset with no majority class.

In the 99.5% background regime, Biterm shows a decrease in performance. GMM still retains a strong capability for signal/background identification, while K-Means presents a similar reduction in AUC compared to Biterm. However, its background rejection curve reveals

a very poor discriminative power for signal efficiencies ranging from 0 to 0.6. The behavior of VAE and IForest in this extreme scenario is particularly noteworthy, since their performances remain essentially unchanged regardless of whether they operate in unsupervised or semisupervised mode, maintaining high AUC values. This indicates that training with a background-only dataset has a very limited impact on these models in extremely rare signal scenarios, so they can be used as true unsupervised models.

In the case of Biterm versus semisupervised algorithms, VAE and IForest as true novelty detectors, algorithms perform similarly for $B/(S+B) = 50\%$, with $\text{AUC} \sim 0.98$, despite Biterm presenting higher-background rejection for fixed signal acceptance up to ~ 0.8 . With 99.5% background pollution, however, Biterm lags behind the other two, with AUC decreasing to ~ 0.91 .

Inspired by these findings, we investigated the discriminative power of the algorithms, in terms of AUC, as a function of the background contamination. In realistic experimental scenarios, one expects a high level of background pollution, although the true signal-to-background ratio is generally unknown prior to the observation. Motivated by this uncertainty, we vary the background fraction from 50% up to 99.5% and compute the AUC for the Biterm and baseline models. The results are presented in the right panels of Fig. 4. In the unsupervised setting,

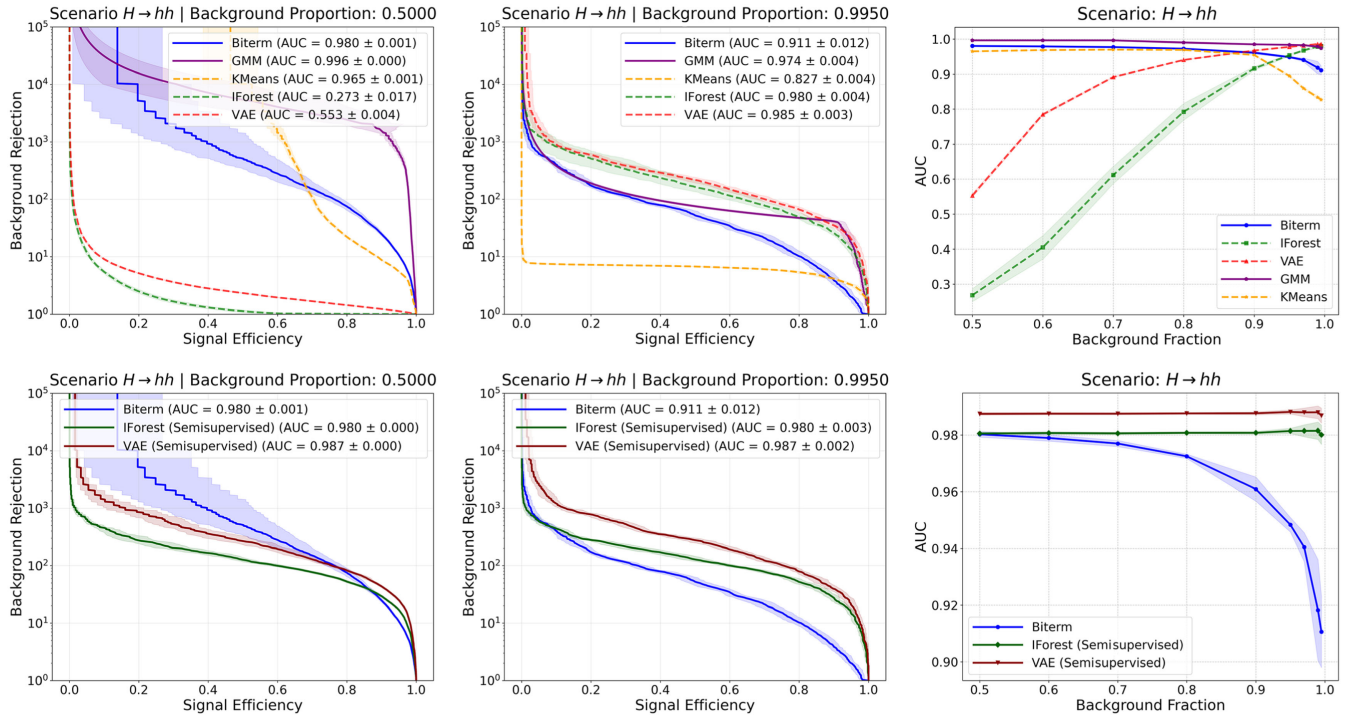


FIG. 4. Background rejection versus signal efficiency curves for BTM and alternative algorithms in $H \rightarrow hh$ process. The first row depicts comparisons among the unsupervised models, while the second row compares BTM against the semisupervised models. The left (middle) column shows the case of 50% (99.5%) of background pollution. The legends show the AUC values, which are used to discriminate the ability of each model in disentangling the signal from the background. The right panel depicts AUC as a function of $B/(S+B)$.

topic modeling and clustering approaches outperform VAE and IForest up to background contamination of approximately $\sim 90\%$ and $\sim 95\%$, respectively. Moreover, unlike IForest and VAE, Biterm, GMM, and K-Means remain remarkably robust under variations of $B/(S+B)$, with Biterm exhibiting only a mild decrease in performance. In particular, for relatively low background contamination, Biterm, GMM, and K-Means achieve excellent discriminative performance, whereas the outlier detection methods, especially IForest, struggle to correctly separate signal and background events. On the other hand, when VAE and IForest are trained to detect background events, they perform better than BTM for all background pollution scenarios as we see in the bottom right panel of Fig. 4.

This behavior can be understood from the intrinsic nature of the algorithms. IForest and VAE are fundamentally outlier detection methods. Therefore, in the regime where signal and background are present in comparable proportions, the notion of an outlier becomes ill defined, since the algorithms effectively learn a nearly balanced mixture of the two classes with no clear dominant component. In contrast, topic modeling approaches learn latent structures through the identification of co-occurrence patterns, making them substantially less sensitive to variations in $B/(S+B)$. The gradual decrease in the AUC of the Biterm model as the background contamination increases can be mainly attributed to the limited signal statistics available in highly imbalanced samples. Since the underlying signal and background event distributions remain unchanged, the observed degradation does not indicate a fundamental loss of separability between the two classes. Instead, the scarcity of signal events affects the stability of the topic estimation process, leading to noisier latent representations and a moderate reduction in discriminative

performance. More interestingly, though, is the behavior of IForest and VAE when used as unsupervised models. In this case, they perform as well as in the semisupervised setting when signals are rare in the sample.

The repeated random subsampling analysis indicates that these conclusions are not an artifact of cherry-picking a favorable test set. This interpretation is further supported by the progressively wider shaded bands observed at high-background fractions, which reflect the larger statistical fluctuations across the resampled datasets in low-signal regimes.

As a final remark, Biterm uses only three final-state observables $\{M, p_T, k_T\}$, but it consistently extracts latent patterns that separate signal from background without relying on jet-substructure information. These results indicate that topic modeling provides a robust unsupervised alternative for signal-background separation and can complement other traditional unsupervised and semisupervised approaches.

1. Invariant mass reconstruction

Although background-rejection curves quantify the statistical discrimination achieved by each model, it is essential to verify whether this separation corresponds to physically meaningful structures. To illustrate this capability, Fig. 5 presents the distribution for the kinematic combination corresponding to the heavy-Higgs invariant mass in $H \rightarrow hh$, $M(b\bar{b}\ell^-\ell^+\nu\bar{\nu})$, where neutrinos momenta are reconstructed from the Higgsness variable method [43].

As shown in Fig. 5, the distribution reconstructed by the Biterm model exhibits a clear agreement with the true physical ones. In the balanced 50% contamination scenario (left panel), the model precisely recovers both the signal and background shapes. More importantly, even in the

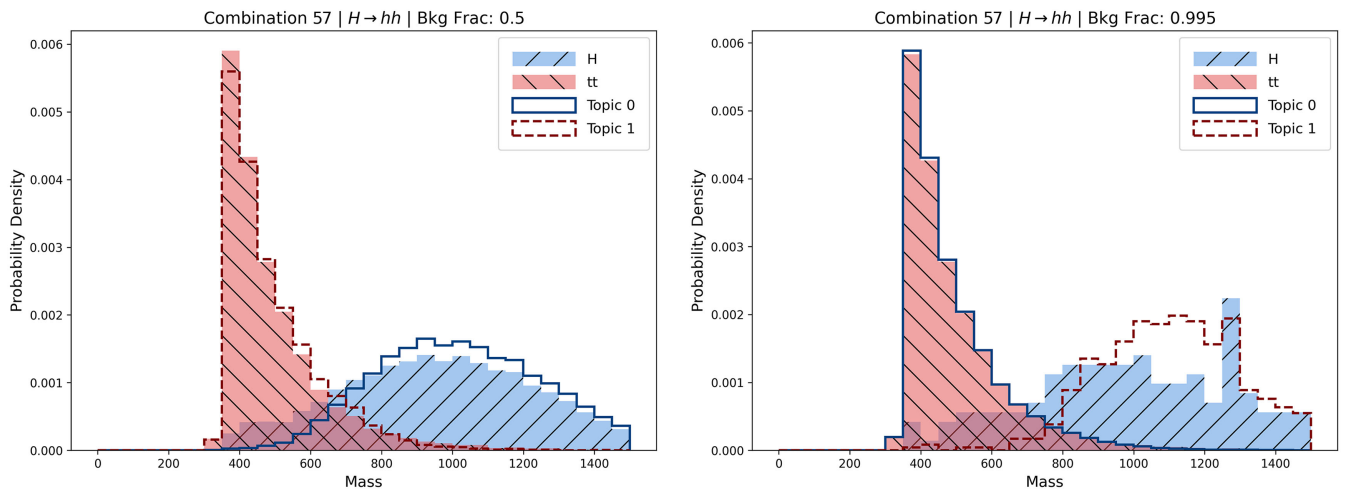


FIG. 5. The $b\bar{b}\ell^+\ell^-\nu\bar{\nu}$ invariant mass distribution for the $H \rightarrow hh$ process. Filled histograms correspond to the true signal and background distributions, while the step histograms represent the reconstructed distributions based on the Biterm model classification. The left panel illustrates the balanced scenario with 50% background contamination, whereas the right panel displays the highly imbalanced 99.5% background scenario.

highly polluted 99.5% background regime (right panel), this combination yields a reconstructed distribution that closely matches the true mass peak at approximately 1 TeV. Although signal identification naturally becomes more challenging in this extreme regime, the reconstructed tail reproduces the expected physical behavior, reinforcing the physical consistency of the separation learned by the topic modeling approach.

For the sake of completeness, an extended set of kinematic distributions for the $H \rightarrow hh$ process is provided in Appendix.

C. Heavy Higgs to $t\bar{t}$

The second process approached is the heavy Higgs decaying into a pair of a quark and an antiquark top, $H \rightarrow t\bar{t}$. Similarly to the $H \rightarrow hh$ case, the set of variables $\{M, p_T, k_T\}$ provides the best discriminative power between signal and background.

As shown in Fig. 6, in this case, BTM outperforms the alternative models, both unsupervised and semisupervised, in both background pollution scenarios with AUC values exceeding 0.93 and 0.89 in the 50% and 99.5% cases, respectively. The background-rejection curve for 99.5% pollution shows that all models are very competitive in rejecting background over a wide range of signal efficiency,

but BTM achieves the best performance for almost all signal efficiencies.

This result reinforces the capacity of topic modeling models to separate background from signal in resonant cases, even in extreme scenarios of high-background pollution. It should be emphasized that $H \rightarrow t\bar{t}$ is naturally harder to disentangle from the top-quark pair background than $H \rightarrow hh$ as reflected in the smaller AUC compared to the resonant SM Higgs pair production. The drop in the BTM performance from background fractions from 80% up to 99.5%, seen at the rightmost panel of Fig. 6, is due to a diminished signal statistics compared to the background which presents much more events.

Table II summarizes the results for the two heavy-Higgs scenarios approached, decaying into two SM Higgses and in a $t\bar{t}$ pair.

D. Nonresonant double Higgs production

Now we focus on the nonresonant production of a double Higgs, where three possible scenarios are investigated: the SM coupling $\kappa = +1$ and two BSM scenarios that are also not experimentally excluded, $\kappa = -1$ and $\kappa = +3$. The set of optimum BTM variables for $\kappa = +1$ is $\{p_T, k_T, z, \cos\theta_{WW}\}$, for $\kappa = +3$ it is $\{p_T, z, \cos\theta_{WW}\}$, and for $\kappa = -1$ it is $\{p_T, k_T\}$.

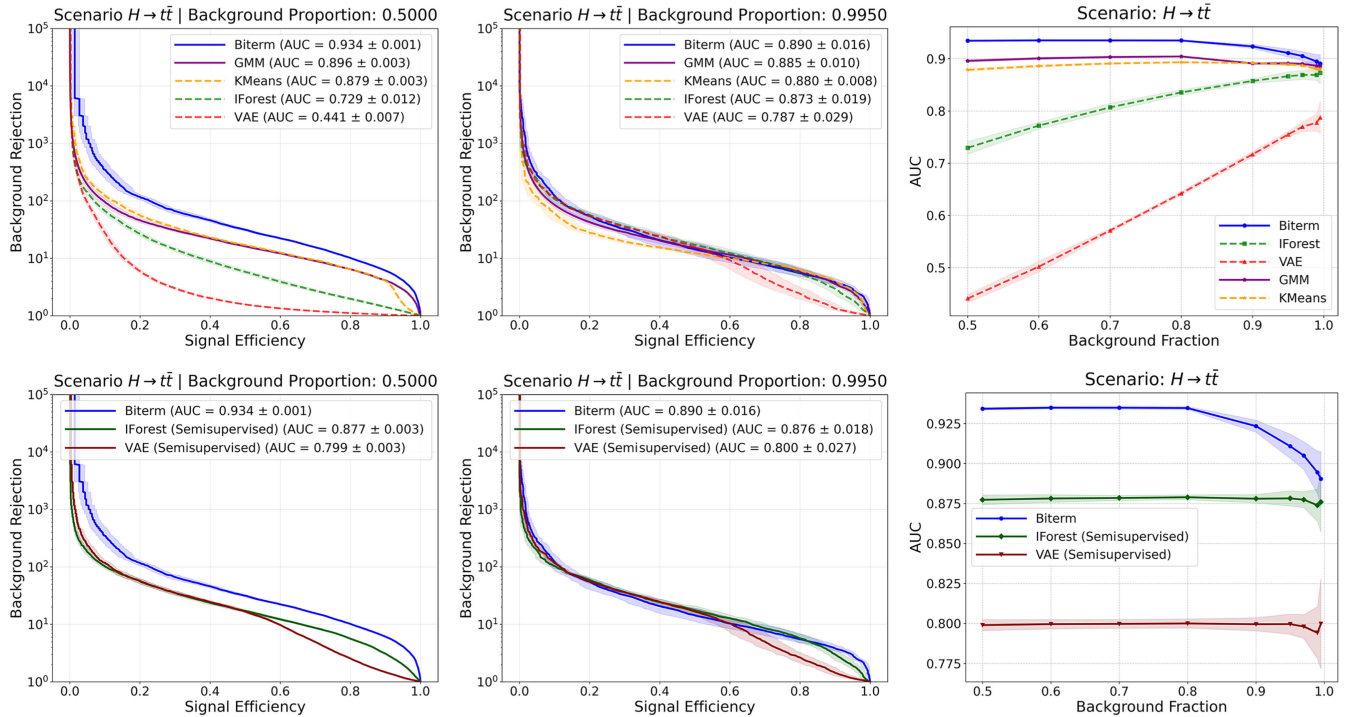


FIG. 6. Background rejection versus signal efficiency curves for BTM and alternative algorithms in the $H \rightarrow t\bar{t}$ process. The first row depicts comparisons among the unsupervised models, while the second row compares BTM against the semisupervised models. The left (middle) column shows the case of 50% (99.5%) of background pollution. The legends show the AUC values, which are used to discriminate the ability of each model in disentangling the signal from the background. The right panel depicts AUC as a function of $B/(S+B)$.

TABLE II. AUC values summarizing the ROC curves for each model and heavy-Higgs production scenarios. Background contamination levels of 50% and 99.5% were considered.

Model	$H \rightarrow hh$		$H \rightarrow t\bar{t}$	
	50% bkg	99.5% bkg	50% bkg	99.5% bkg
BTM	0.980 ± 0.001	0.911 ± 0.012	0.934 ± 0.001	0.890 ± 0.016
IForest (Semisupervised)	0.980 ± 0.000	0.980 ± 0.003	0.877 ± 0.003	0.876 ± 0.018
VAE (Semisupervised)	0.987 ± 0.000	0.987 ± 0.002	0.799 ± 0.003	0.800 ± 0.027
IForest (Unsupervised)	0.273 ± 0.017	0.980 ± 0.004	0.729 ± 0.012	0.873 ± 0.019
VAE (Unsupervised)	0.553 ± 0.004	0.985 ± 0.003	0.441 ± 0.007	0.787 ± 0.029
GMM	0.996 ± 0.000	0.974 ± 0.004	0.896 ± 0.003	0.885 ± 0.010
K-Means	0.965 ± 0.001	0.827 ± 0.004	0.879 ± 0.003	0.880 ± 0.008

Figure 7 shows the SM case ($\kappa = +1$). As in the heavy-Higgs cases, the topic modeling model exhibits a similar power of discrimination independent of the background pollution, with $AUC = 0.7$ in the balanced scenario and $AUC = 0.64$ for the extreme case. In the unsupervised case, GMM is the best model for the 50% background pollution, in terms of AUC, followed by BTM. In the 99.5% scenario, however, IForest and VAE win. On the other hand, as can be seen in the upper right panel of Fig. 7, BTM is the more stable algorithm across all the background

pollution scenarios. The comparison with IForest and VAE as semisupervised models show that they are consistently better than BTM for $\kappa = +1$, see bottom row of plots of Fig. 7.

Figures 8 and 9 correspond to the BSM $\kappa = -1$ and $\kappa = +3$ cases, respectively, which show similar behaviors to the case $\kappa = +1$ in all aspects. BTM have an AUC around 0.7 in all cases and show no impact on performance due to the degree of imbalance in the data. Table III summarizes the AUC results of these cases.

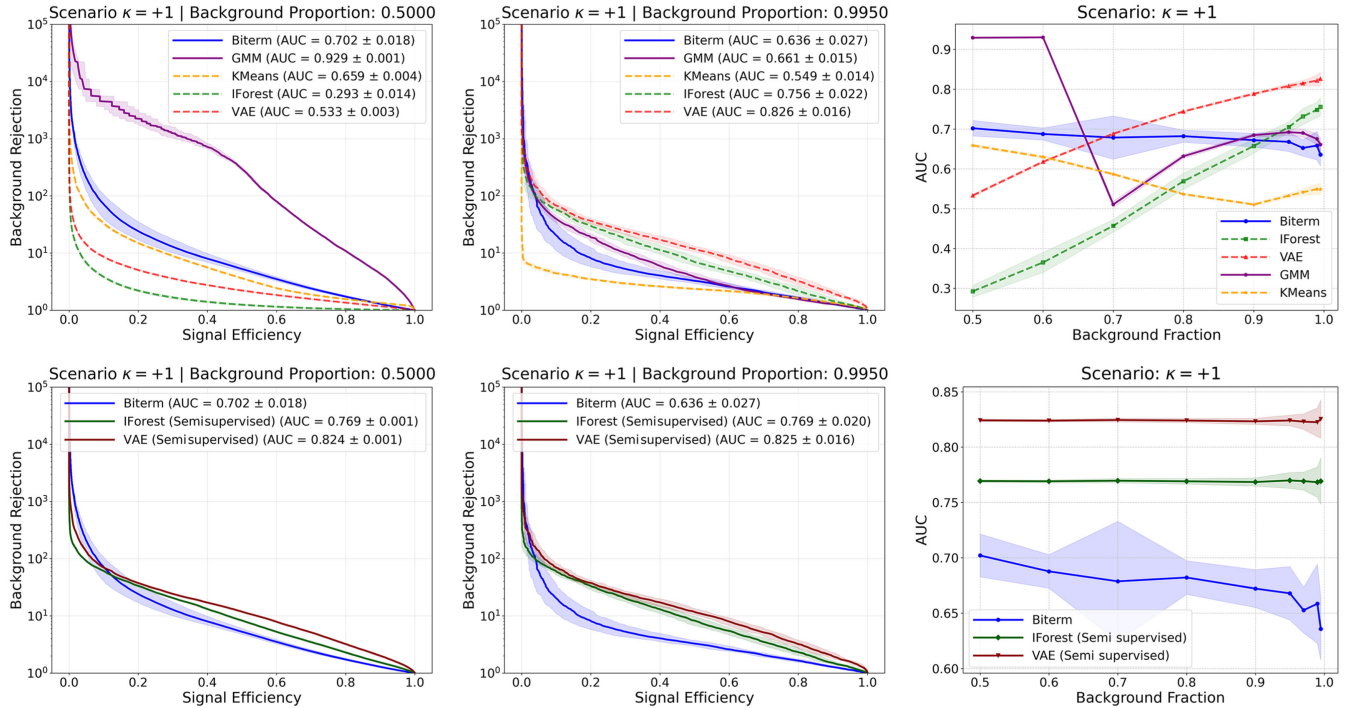


FIG. 7. Background rejection versus signal efficiency curves for BTM and alternative algorithms in nonresonant SM-like hh production. The first row depicts comparisons among the unsupervised models, while the second row compares BTM against the semisupervised models. The left (middle) column shows the case of 50% (99.5%) of background pollution. The legends show the AUC values, which are used to discriminate the ability of each model in disentangling the signal from the background. The right panel depicts AUC as a function of $B/(S+B)$.

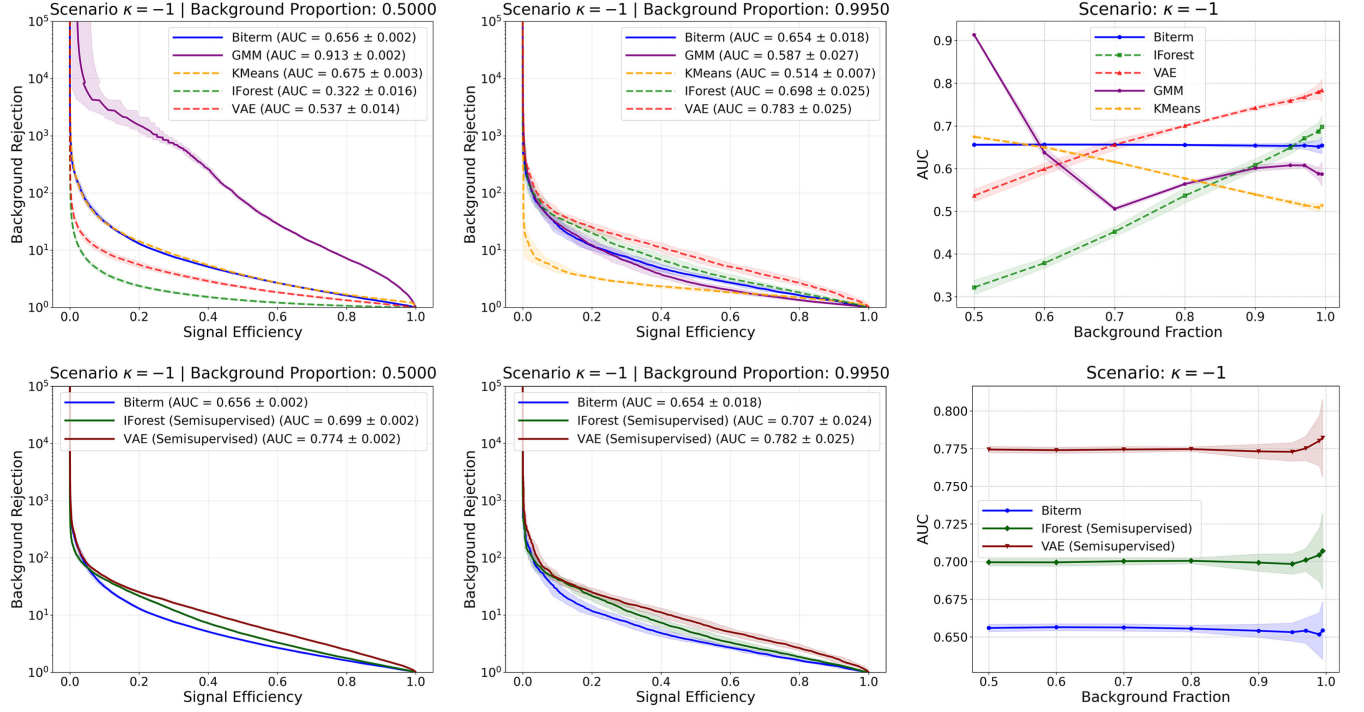


FIG. 8. Background rejection versus signal efficiency curves for BTM and alternative algorithms in nonresonant BSM hh production scenario with $\kappa = -1$. The first row depicts comparisons among the unsupervised models, while the second row compares BTM against the semisupervised models. The left (middle) column shows the case of 50% (99.5%) of background pollution. The legends show the AUC values, which are used to discriminate the ability of each model in disentangling the signal from the background. The right panel depicts AUC as a function of $B/(S+B)$.

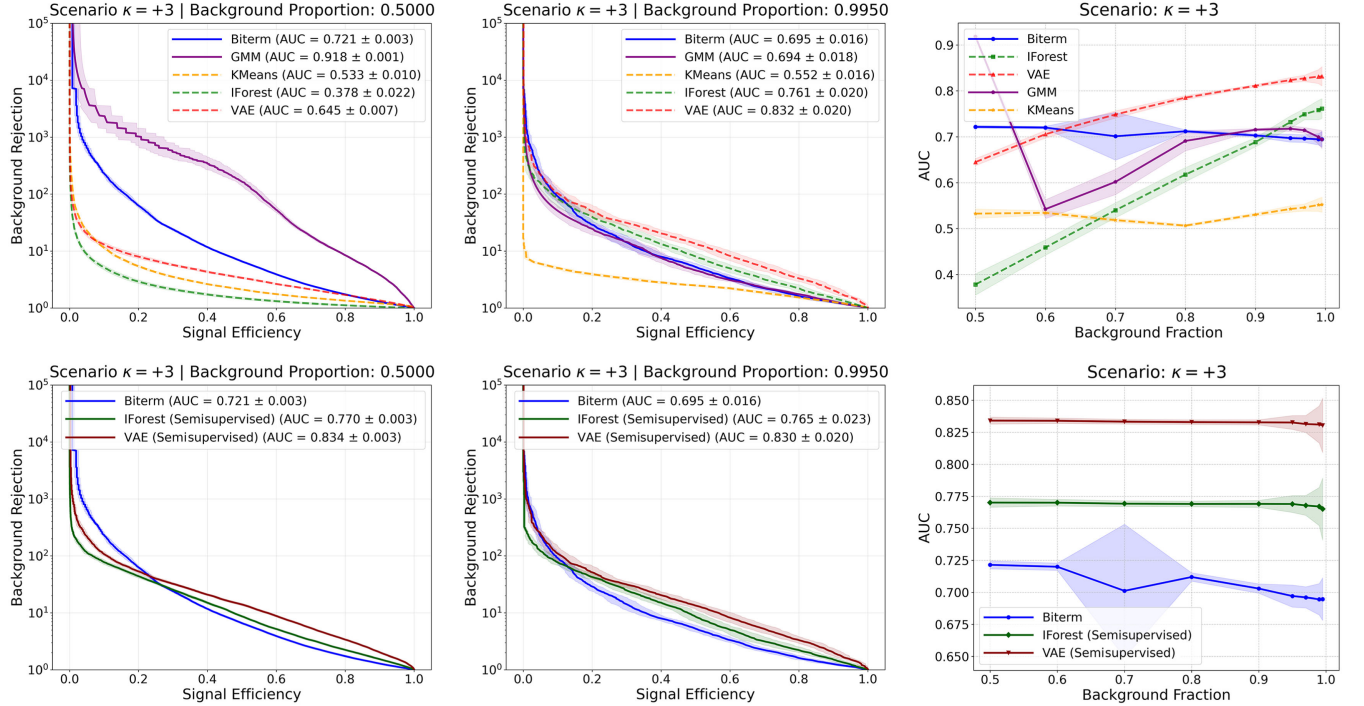


FIG. 9. Background rejection versus signal efficiency curves for BTM and alternative algorithms in nonresonant BSM hh production scenario with $\kappa = +3$. The first row depicts comparisons among the unsupervised models, while the second row compares BTM against the semisupervised models. The left (middle) column shows the case of 50% (99.5%) of background pollution. The legends show the AUC values, which are used to discriminate the ability of each model in disentangling the signal from the background. The right panel depicts AUC as a function of $B/(S+B)$.

TABLE III. AUC values summarizing the ROC curves for each model and coupling scenario. Background contamination levels of 50% and 99.5% were considered.

Model	$\kappa = +1$		$\kappa = -1$		$\kappa = +3$	
	50% bkg	99.5% bkg	50% bkg	99.5% bkg	50% bkg	99.5% bkg
BTM	0.702 ± 0.018	0.636 ± 0.027	0.656 ± 0.002	0.654 ± 0.018	0.721 ± 0.003	0.695 ± 0.016
IForest (Semisupervised)	0.769 ± 0.001	0.769 ± 0.020	0.699 ± 0.002	0.707 ± 0.024	0.770 ± 0.003	0.765 ± 0.023
VAE (Semisupervised)	0.824 ± 0.001	0.825 ± 0.016	0.774 ± 0.002	0.782 ± 0.025	0.834 ± 0.003	0.830 ± 0.020
IForest (Unsupervised)	0.293 ± 0.014	0.756 ± 0.022	0.322 ± 0.016	0.698 ± 0.025	0.378 ± 0.022	0.761 ± 0.020
VAE (Unsupervised)	0.533 ± 0.003	0.826 ± 0.016	0.537 ± 0.014	0.783 ± 0.025	0.645 ± 0.007	0.832 ± 0.020
GMM	0.929 ± 0.001	0.661 ± 0.015	0.913 ± 0.002	0.587 ± 0.027	0.918 ± 0.001	0.694 ± 0.018
Kmeans	0.659 ± 0.004	0.549 ± 0.014	0.675 ± 0.003	0.514 ± 0.007	0.533 ± 0.010	0.522 ± 0.016

E. Effective operators

The last case considered in this work involves the insertion of the effective operators $\mathcal{O}_{\tilde{G}}$ and $\mathcal{O}_{qq}^{(1)}$. The best set of variables found in the analysis is $\{M, p_T\}$ for $\mathcal{O}_{\tilde{G}}$ and $\{\cos\theta^*\}$ for $\mathcal{O}_{qq}^{(1)}$.

Figure 10 shows the results for $\mathcal{O}_{\tilde{G}}$. The topic modeling model returns an AUC of 0.6 in both balanced and unbalanced cases. With an AUC ~ 0.5 , alternative models are compatible with random guessing, indicating that there is no discriminative power of separation, irrespective the background pollution. Very similar behavior is found in the

case $\mathcal{O}_{qq}^{(1)}$, which is shown in Fig. 11. In both regimes of background pollution, BTM obtains AUC of around 0.6, while the alternative models perform as random classifiers. Table IV summarizes these results.

As we can see, effective operators arise as the more challenging case. Contrary to other physics models, for those effective operators, no algorithm presents any improvement from the 50% to the 99.5% background pollution situation, despite the BTM model being the only one that exhibits an ROC curve consistently above the random-guessing diagonal across the entire threshold

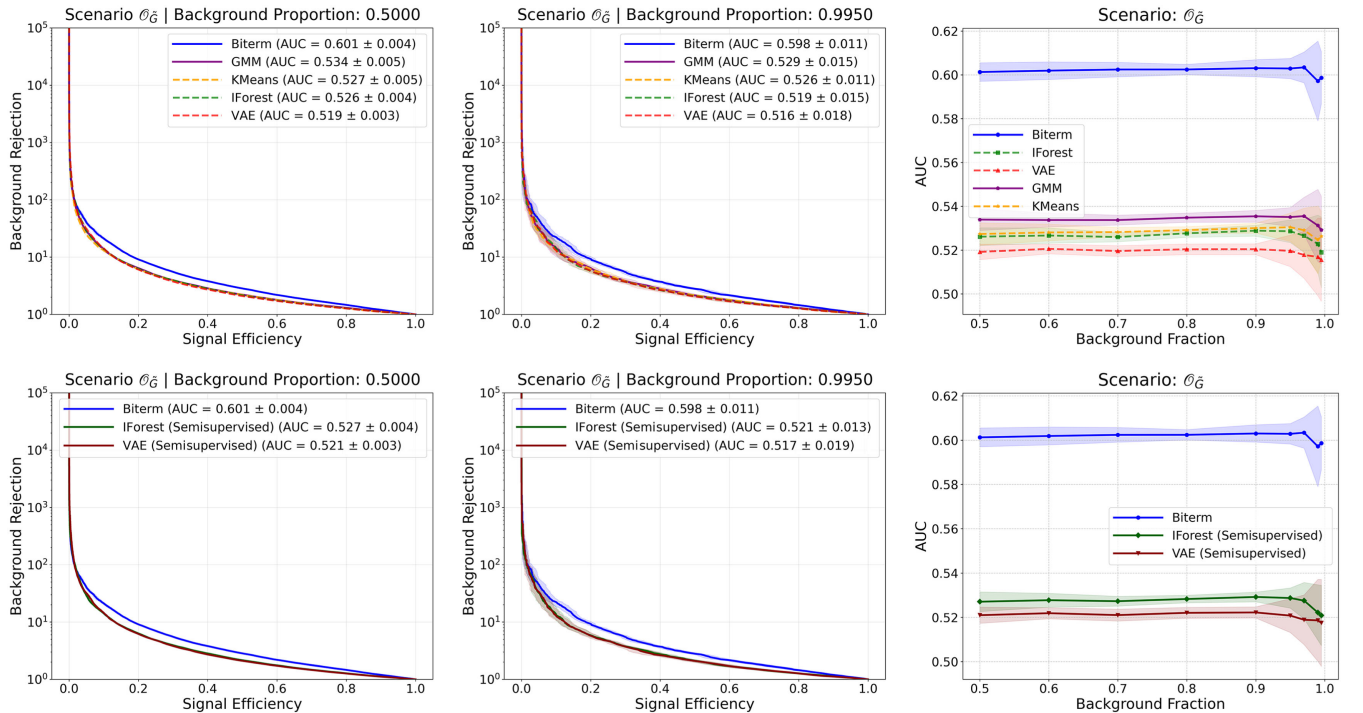


FIG. 10. Background rejection versus signal efficiency curves for BTM and alternative algorithms for $\mathcal{O}_{\tilde{G}}$ effective interaction. The first row depicts comparisons among the unsupervised models, while the second row compares BTM against the semisupervised models. The left (middle) column shows the case of 50% (99.5%) of background pollution. The legends show the AUC values, which are used to discriminate the ability of each model in disentangling the signal from the background. The right panel depicts AUC as a function of $B/(S+B)$.

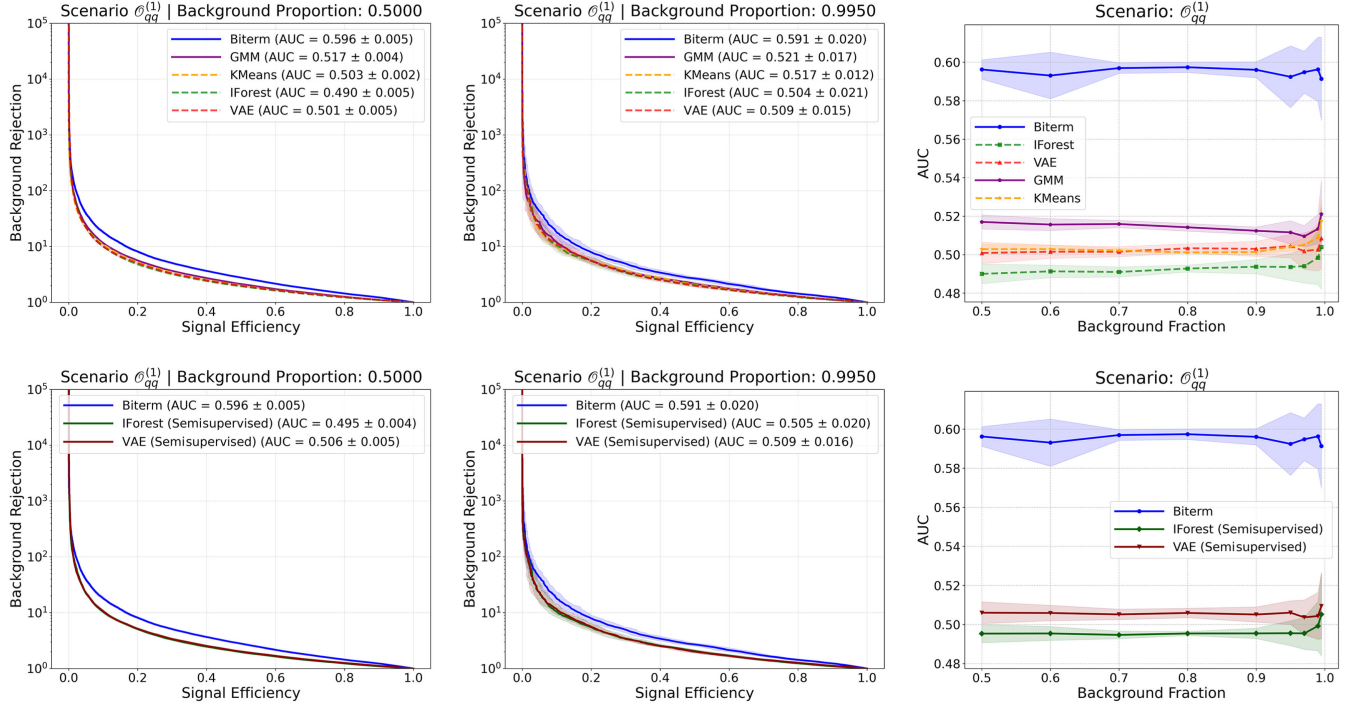


FIG. 11. Background rejection versus signal efficiency curves for BTM and alternative algorithms for $\mathcal{O}_{qq}^{(1)}$ effective interaction. The first row depicts comparisons among the unsupervised models, while the second row compares BTM against the semisupervised models. The left (middle) column shows the case of 50% (99.5%) of background pollution. The legends show the AUC values, which are used to discriminate the ability of each model in disentangling the signal from the background. The right panel depicts AUC as a function of $B/(S+B)$.

range. The observed discriminatory power of BTM is stable, although weak, and therefore unlikely to result from statistical fluctuations.

The difficulty in disentangling effective interactions from QCD in $t\bar{t}$ events is well known. The only observable effect is encoded in distortions of variables like transverse momentum and rapidity of the top quark induced by the interference between the SM and the new physics contribution [49]. Depending of the strength of effective coupling, the interference is weak and the distortion is barely noticeable.

VII. CONCLUSIONS

In this work, we investigated the potential of topic-modeling-based methods, in particular the Biterm topic model, for unsupervised signal detection in LHC collision events using leptons, b -jets, and reconstructed neutrino momenta exclusively, but without relying on jet substructure data. The approach was systematically tested across nonresonant double SM Higgs production and resonant hh and $t\bar{t}$ production channels through a new heavy-Higgs decay, as well as in scenarios described by effective

TABLE IV. AUC values summarizing the ROC curves for each model and effective operator scenarios. Background contamination levels of 50% and 99.5% were considered.

Model	$\mathcal{O}_{\tilde{G}}$		$\mathcal{O}_{qq}^{(1)}$	
	50% bkg	99.5% bkg	50% bkg	99.5% bkg
BTM	0.601 ± 0.004	0.598 ± 0.011	0.596 ± 0.005	0.591 ± 0.020
IForest (Semisupervised)	0.527 ± 0.004	0.521 ± 0.013	0.495 ± 0.004	0.505 ± 0.020
VAE (Semisupervised)	0.521 ± 0.003	0.517 ± 0.019	0.506 ± 0.005	0.509 ± 0.016
IForest (Unsupervised)	0.526 ± 0.004	0.519 ± 0.015	0.490 ± 0.005	0.504 ± 0.021
VAE (Unsupervised)	0.519 ± 0.003	0.516 ± 0.018	0.501 ± 0.005	0.509 ± 0.015
GMM	0.534 ± 0.005	0.529 ± 0.015	0.517 ± 0.004	0.521 ± 0.017
Kmeans	0.527 ± 0.005	0.526 ± 0.011	0.503 ± 0.002	0.517 ± 0.012

operators, under both balanced and highly imbalanced background conditions.

Our results demonstrate that BTM provides a robust and stable discrimination capability across all scenarios, with performance largely insensitive to the degree of background contamination. In resonant heavy-Higgs searches, the method achieves AUC values above 0.9 and exhibits excellent background rejection at low-signal efficiencies, competing with, and in some cases, surpassing established unsupervised models such as GMM and K-Means, and semisupervised models like VAE and isolation forest. For nonresonant Higgs pair production, all models show moderate discrimination power, reflecting the intrinsic complexity of this process, while BTM remains consistently competitive and stable across different constant couplings.

The effective operator scenarios constitute the most challenging regime. In this case, BTM is the only method that maintains classification performance systematically above random guessing, indicating that it captures subtle latent structures encoded in the final-state observables, even when semisupervised anomaly-detection models fail to extract meaningful separation.

A key outcome of this study is that topic modeling can achieve competitive signal-background separation without relying on jet substructure observables, using only a compact set of final-state kinematic variables. This highlights the ability of topic models to uncover physically meaningful latent patterns in sparse event representations, offering a complementary strategy to traditional supervised, unsupervised, and anomaly-detection approaches.

Although the classification performance in some scenarios remains limited by the intrinsic overlap between signal and background, the stability of the topic-modeling approach under extreme class imbalance and its minimal

reliance on feature engineering makes it particularly attractive for exploratory searches and model-independent analyses.

Future studies may extend this framework to new sets of observables, approach other physical scenarios, and explore hybrid strategies that combine topic modeling with other techniques. Overall, our results indicate that topic-modeling-based methods constitute a promising and versatile tool for unsupervised new physics searches at the LHC.

ACKNOWLEDGMENTS

This study was financed in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), via Grant No. 307317/2021-8 (A. A.). A. A. also acknowledges support from the FAPESP (Grant No. 2021/01089-1) grant.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX: RECONSTRUCTED KINEMATIC DISTRIBUTIONS

In Fig. 12, we depict six out of seven kinematic distributions listed in Sec. IV to represent $H \rightarrow hh$ events for the Biterm algorithm in two background pollution scenarios: $B/(S+B) = 50\%$ (left panels) and 99.5% (right panels). In the case of low-background pollution, Biterm is highly efficient in reconstructing the distributions. For higher-background pollution, it is more efficient to reconstruct the invariant masses, including the one shown in Fig. 5, p_T , k_T , and $\cos(\theta^*)$. We then understand why BTM performs better in the representation $\{M, p_T, k_T\}$ in $H \rightarrow hh$ events.

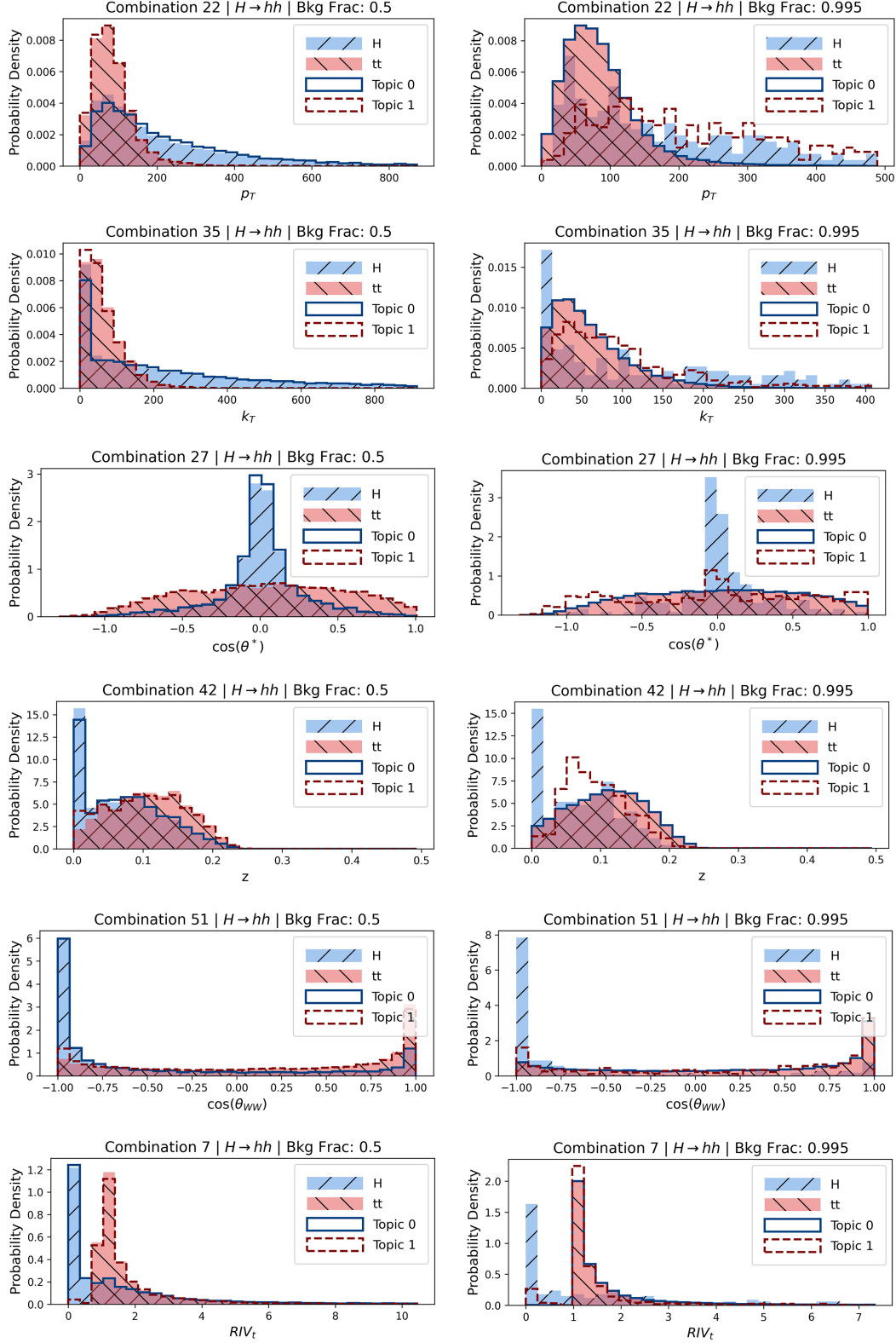


FIG. 12. Distributions of kinematic variables for the $H \rightarrow hh$ process. Filled histograms correspond to the true signal and background distributions, while the step histograms represent the reconstructed distributions based on the Biterm model classification. The left panel shows the scenario with 50% background contamination, whereas the right panel exhibits the extremely imbalanced case with 99.5% contamination. Each row corresponds to a specific variable for a randomly chosen combination.

- [1] G. Kasieczka, B. Nachman, D. Shih *et al.*, *Rep. Prog. Phys.* **84**, 124201 (2021).
- [2] J. Bardhan, T. Mandal, S. Mitra, C. Neeraj, and M. Patra, *Eur. Phys. J. Special Topics* **233**, 2559 (2024).
- [3] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, *Phys. Rev. D* **101**, 076015 (2020).
- [4] G. Aad *et al.* (ATLAS Collaboration), *Phys. Rev. Lett.* **132**, 081801 (2024).
- [5] R. Churchill and L. Singh, *ACM Comput. Surv.* **54**, 1 (2022).
- [6] E. M. Metodiev and J. Thaler, *Phys. Rev. Lett.* **120**, 241602 (2018).
- [7] G. Blanchard and C. Scott, in *AISTATS 2014*, Proceedings of Machine Learning Research, Vol. 33, edited by S. Kaski and J. Corander (PMLR, Reykjavik, Iceland, 2014), pp. 1–9.
- [8] P. T. Komiske, E. M. Metodiev, and J. Thaler, *J. High Energy Phys.* **11** (2018) 059.
- [9] J. Brewer, J. Thaler, and A. P. Turner, *Phys. Rev. C* **103**, L021901 (2021).
- [10] G. Aad, B. Abbott, D. C. Abbott *et al.*, *Phys. Rev. D* **100**, 052011 (2019).
- [11] E. Alvarez, F. Lamagna, and M. Szwec, *J. High Energy Phys.* **01** (2020) 049.
- [12] M. LeBlanc, B. Nachman, and C. Sauer, *J. High Energy Phys.* **02** (2023) 150.
- [13] B. M. Dillon, D. A. Faroughy, and J. F. Kamenik, *Phys. Rev. D* **100**, 056002 (2019).
- [14] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szwec, *J. High Energy Phys.* **10** (2020) 206.
- [15] B. M. Dillon, D. A. Faroughy, J. F. Kamenik, and M. Szwec, *Symmetry* **13**, 1167 (2021).
- [16] D. A. Faroughy, in *Proceedings of Science*, Vol. 390 (2021).
- [17] D. M. Blei, A. Y. Ng, and M. I. Jordan, *J. Mach. Learn. Res.* **3** (2003).
- [18] X. Yan, J. Guo, Y. Lan, and X. Cheng, in *Proceedings of the 22nd International Conference on World Wide Web* (2013).
- [19] A. Srivastava and C. Sutton, in *ICLR 2017—Conference Track Proceedings* (2017).
- [20] H. H. Nguyen, C. N. Nguyen, X. T. Dao, Q. T. Duong, D. P. T. Kim, and M.-T. Pham, [arXiv:2408.13561](https://arxiv.org/abs/2408.13561).
- [21] F. T. Liu, K. M. Ting, and Z.-H. Zhou, in *2008 Eighth IEEE International Conference on Data Mining* (IEEE, 2008), pp. 413–422.
- [22] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, Vol. 4 (Springer, New York, 2006).
- [23] R. Krestel, P. Fankhauser, and W. Nejdl, in *RecSys'09* (2009).
- [24] X. Yi and J. Allan, in *International Conference on Information and Knowledge Management, Proceedings* (2008).
- [25] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, *J. Am. Soc. Inf. Sci.* **41**, 391 (1990).
- [26] E. S. Hewitt and L. J. Savage, *Trans. Am. Math. Soc.* **80**, 470 (1955).
- [27] T. L. Griffiths and M. Steyvers, *Proc. Natl. Acad. Sci. U. S. A.* **101**, 5228 (2004).
- [28] I. Csiszar, *Ann. Probab.* **3**, 146 (1975).
- [29] F. A. de Souza, M. Barros, N. F. Castro, M. Crispim Romão, C. Neiva, and R. Pedro, *Phys. Rev. D* **113**, 035003 (2026).
- [30] E. Alvarez, M. Szwec, A. Szynekman, S. A. Tanco, and T. Tarutina, *SciPost Phys. Core* **6**, 046 (2023).
- [31] X. Ding, L. Zhao, and L. Akoglu, [arXiv:2206.07647](https://arxiv.org/abs/2206.07647).
- [32] J. Soenen, E. Van Wolputte, L. Perini, V. Vercruyssen, W. Meert, J. Davis, and H. Blockeel, The effect of hyperparameter tuning on the comparative evaluation of unsupervised anomaly detection methods, (2021-08-15).
- [33] S. Profumo, M. J. Ramsey-Musolf, and G. Shaughnessy, *J. High Energy Phys.* **08** (2007) 010.
- [34] P. Amaro-Seoane, H. Audley, S. Babak, J. Baker, E. Barausse *et al.*, [arXiv:1702.00786](https://arxiv.org/abs/1702.00786).
- [35] A. Alves, D. Gonçalves, T. Ghosh, H.-K. Guo, and K. Sinha, *Phys. Lett. B* **818**, 136377 (2021).
- [36] S. Banik, G. Coloretti, A. Crivellin, and B. Mellado, *J. High Energy Phys.* **01** (2025) 155.
- [37] B. Grzadkowski, M. Iskrzyński, M. Misiak, and J. Rosiek, *J. High Energy Phys.* **10** (2010) 085.
- [38] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *J. High Energy Phys.* **07** (2014) 079.
- [39] R. D. Ball, V. Bertone, S. Carrazza, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland, and J. Rojo (NNPDF Collaboration), *Nucl. Phys. B* **877**, 290 (2013).
- [40] T. Sjostrand, S. Mrenna, and P. Z. Skands, *Comput. Phys. Commun.* **178**, 852 (2008).
- [41] M. Cacciari, G. P. Salam, and G. Soyez, *Eur. Phys. J. C* **72**, 1896 (2012).
- [42] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaitre, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), *J. High Energy Phys.* **02** (2014) 057.
- [43] J. H. Kim, K. Kong, K. T. Matchev, and M. Park, *Phys. Rev. Lett.* **122**, 091801 (2019).
- [44] A. Alves, E. d. S. Almeida, and D. S. V. Gonçalves, *Eur. Phys. J. C* **86**, 128 (2026).
- [45] M. Terpilovskii and C. Kälén, maximtrp/bitermplus: bitermplus: v0.9.1, (2025).
- [46] R. Rehurek and P. Sojka, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA, Valletta, Malta, 2010).
- [47] Y. Zhao, Z. Nasrullah, and Z. Li, *J. Mach. Learn. Res.* **20**, 1 (2019), <http://jmlr.org/papers/v20/19-011.html>.
- [48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [49] N. Kidonakis and K. Şimşek, *Phys. Rev. D* **114**, 055019 (2026).