Eur. Phys. J. C (2022) 82:993
https://doi.org/10.1140/epjc/s10052-022-10944-3

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Theoretical Physics

A method for approximating optimal statistical significances with machine-learned likelihoods

Ernesto Arganda[1,2,a], **Xabier Marcano**[1,3,b], **Víctor Martín Lozano**[4,5,c], **Anibal D. Medina**[2,d], **Andres D. Perez**[2,e], **Manuel Szewc**[6,7,f], **Alejandro Szynkman**[2,g]

[1] Instituto de Física Teórica UAM-CSIC, C/ Nicolás Cabrera 13-15, Campus de Cantoblanco, 28049 Madrid, Spain
[2] IFLP, CONICET-Dpto. de Física, Universidad Nacional de La Plata, C.C. 67, 1900 La Plata, Argentina
[3] Departamento de Física Teórica, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain
[4] Departament de Física Teòrica and IFIC, Universitat de València-CSIC, 46100 Burjassot, Spain
[5] Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany
[6] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[7] International Center for Advanced Studies (ICAS) and ICIFI, UNSAM, Campus Miguelete, 25 de Mayo y Francia, CP1650 San Martín, Buenos Aires, Argentina

Received: 18 May 2022 / Accepted: 23 October 2022 / Published online: 5 November 2022

Abstract Machine-learning techniques have become fundamental in high-energy physics and, for new physics searches, it is crucial to know their performance in terms of experimental sensitivity, understood as the statistical significance of the signal-plus-background hypothesis over the background-only one. We present here a simple method that combines the power of current machine-learning techniques to face high-dimensional data with the likelihood-based inference tests used in traditional analyses, which allows us to estimate the sensitivity for both discovery and exclusion limits through a single parameter of interest, the signal strength. Based on supervised learning techniques, it can perform well also with high-dimensional data, when traditional techniques cannot. We apply the method to a toy model first, so we can explore its potential, and then to a LHC study of new physics particles in dijet final states. Considering as the optimal statistical significance the one we would obtain if the true generative functions were known, we show that our method provides a better approximation than the usual naive counting experimental results.

[a] e-mail: ernesto.arganda@csic.es
[b] e-mail: xabier.marcano@uam.es
[c] e-mail: victor.lozano@desy.de
[d] e-mail: anibal.medina@fisica.unlp.edu.ar
[e] e-mail: andres.perez@iflp.unlp.edu.ar
[f] e-mail: manuel.szewc@ijs.si (corresponding author)
[g] e-mail: szynkman@fisica.unlp.edu.ar

Contents

1 Introduction

Machine learning (ML) techniques have become basic tools for data analysis in recent years, and particle physics is no exception. In fact, ML algorithms are playing a fundamental role in collider physics (for seminal papers see, for instance [1–3] and for recent reviews see [4–11]) and are already used practically as a standard tool in the experimental LHC searches carried out by the ATLAS and CMS collaborations (see, for instance, [12–20]). These ML methods can also be applied at the ensemble level of data [21–26] and it has been demonstrated that, under the assumption of independent and identically distributed events, one can construct the optimal multi-event classifier from a single-event classifier and, moreover, that these multi-event classifiers give rise to opti-

mal single-event classifiers [27,28], see also Refs. [29,30] for some examples applied in the HEP field.

Once we have a trained classifier, the fundamental question from the point of view of the search for beyond Standard Model (BSM) physics is how to quantify its performance in terms of experimental sensitivities. The most extended use of such a classifier considers a cut or working point (WP) of its output, which defines a region that ideally favors the signal over the background, and then it computes the sensitivities following standard techniques [31] taking into account only those events passing this cut. Nevertheless, this is nothing but a refined procedure of defining a signal region where to perform a search, and one might still wonder whether it is possible to directly connect the ML classifiers with the standard statistical tests, using its output in full glory and without the need of defining a working point. Experimental collaborations such as ATLAS and CMS do have a method to incorporate the full output distribution to a larger extent, see e.g. Ref. [32].[1] They treat the classifier output simply as a better variable to bin and perform a Binned Likelihood fit on. Although powerful, this treatment can be unsatisfactory as it washes over the probabilistic interpretation of the trained classifier. This is evident in the incorporation of systematic uncertainties which are propagated to the classifier output as it would be for any other high-level observable, without any re-training. This strategy means that the learned classifier is not necessarily a monotonous function of the Likelihood Ratio and perhaps the obtained significance is a sub-optimal approximation of the achievable significance. An analysis strategy that incorporates the full probabilistic structure of the classifier would perhaps be a more natural fit for incorporating systematic uncertainties and would be able to guarantee a better approximation to the full Likelihood ratio.

Some work in this direction has been recently done, see e.g. [33–47]. In particular, Ref. [33] established that training a classifier and then "calibrating" it to learn its distribution under the relevant hypotheses guarantees a proper estimation of the Likelihood Ratio and thus of the optimal significance of an analysis. In the absence of systematic uncertainties, when the distribution of the classifier output is approximated by a binned Likelihood the resulting statistical model coincides with the previously detailed strategy employed by the ATLAS and CMS collaborations.

In this paper we propose a simplification of Ref. [33] that can be used for any ensemble of events, combining the current ML-technique power to deal with high-dimensional data with the likelihood-based inference tests used in traditional analyses to discriminate between signal-plus-background and background-only hypotheses. Our method allows us to obtain the expected sensitivity when using these ML algorithms, both for discovery and for exclusion limits.

In order to assess the potential of our method, we will first consider a toy example in which we generate random data samples from multivariate Gaussian distributions. The motivation to do this is that we can compare the output of our method to the optimal classifier, which we can build since we know the actual generative functions. As we will see, when facing low-dimensional problems, our method gives close-to-optimal results[2] and performs similarly to those obtained by following a standard binned Poisson log-likelihood approach. On the other hand, while binning a multidimensional space becomes intractable, we will show that our method is still easy to apply when the dimensionality of the problem increases and, moreover, that it leads to results that are closer to the optimal classifier than those obtained by fixing a working point. Finally, and as a more practical example, we will apply our method to a realistic problem of searching for heavy W' bosons at the LHC, where we show that the statistical power of the analysis benefits greatly from the implementation of the method.

The paper is organized as follows: in Sect. 2 we detail the method proposed to enhance statistical tests through ML classifiers, discussing in Sect. 2.1 the differences between our method and previous ones proposed in the literature; Sect. 3 is dedicated to the application of this method to two examples, a toy model consisting of two Gaussians in varying dimensions and a realistic example extracted from the LHC Olympics datasets [48]; finally, Sect. 4 is left to summarize our results and discuss future improvements.

2 Method

Our method combines the power of current ML techniques, see e.g. Ref. [49] for a very pedagogical introduction, to deal with high-dimensional data with the likelihood-based inference tests used in traditional analyses to discard different hypotheses [31]. It is aimed as a different way to incorporate ML techniques to supervised searches for different Standard Model and BSM processes.

Suppose we have a set of N independent measurements, each of which consisting of an arbitrarily high-dimensional set of observables x.[3] We are interested in modelling the likelihood $\mathcal{L}$ of the data as a function of a background process b, a signal process s and a signal strength parameter μ, which defines the hypothesis we are testing for: a background-only hypothesis corresponds to $\mu = 0$ while the background-

[1] We thank Pietro Vischia and Sergio Sánchez Cruz for pointing the existing literature to us.

[2] We mostly focus on discovery sensitivities, although we provide all the relevant formulae for computing the exclusion limits.

[3] In the context of collider physics this could refer to kinematical variables such as p_T, η, invariant mass, etc.

plus-signal hypothesis corresponds to $\mu = 1$. This likelihood function is nothing more than the probability of obtaining a given dataset conditioned on the aforementioned information and parameters:

$$\mathcal{L}(\mu, s, b) = p(N, \{x_i, i = 1, \ldots, N\}|\mu, s, b). \tag{1}$$

A choice of likelihood function is a choice of a specific statistical model of the data. Following Ref. [50] , we define the statistical model of N independent measurements using the extended Likelihood

$$\mathcal{L}(\mu, s, b) = \text{Poiss}\big(N|\mu S + B\big) \prod_{i=1}^{N} p(x_i|\mu, s, b), \tag{2}$$

where S (B) is the expected total signal (background) yield, Poiss stands for a Poisson probability mass function $\text{Poiss}(n|\lambda) = \frac{e^{-\lambda}\lambda^n}{n!}$ and $p(x|\mu, s, b)$ is the probability density for a single measurement x. Looking at this equation, there is an interplay between local and global information. The global Poisson term reflects the ensemble factor while $p(x)$ encodes the event-by-event information. The latter is the one that could be enhanced by ML analyses.

We can model the probability density as a mixture of signal and background densities

$$p(x|\mu, s, b) = \frac{B}{\mu S + B}\, p_b(x) + \frac{\mu S}{\mu S + B}\, p_s(x), \tag{3}$$

where $p_s(x) = p(x|s)$ and $p_b(x) = p(x|b)$ are, respectively, the signal and background probability densities for a single measurement x, and $\frac{\mu S}{\mu S+B}$ and $\frac{B}{\mu S+B}$ are the probabilities of an event being sampled from said probability densities.

Having defined a statistical model, we can follow Ref. [31] and define the relevant test statistic $\tilde{t}_\mu$:

$$\tilde{t}_\mu = \begin{cases} -2\,\text{Ln}\,\dfrac{\mathcal{L}(\mu, s, b)}{\mathcal{L}(\hat{\mu}, s, b)} & \text{if } \hat{\mu} \geq 0, \\ -2\,\text{Ln}\,\dfrac{\mathcal{L}(\mu, s, b)}{\mathcal{L}(0, s, b)} & \text{if } \hat{\mu} < 0, \end{cases} \tag{4}$$

where $\hat{\mu}$ is the parameter that maximizes the likelihood $\mathcal{L}(\mu, s, b)$. By differentiating Eq. (2) with respect to μ and finding its zeroes one can show that $\hat{\mu}$ is such that

$$\sum_{i=1}^{N} \frac{p_s(x_i)}{\hat{\mu} S\, p_s(x_i) + B\, p_b(x_i)} = 1. \tag{5}$$

Notice that we recover the traditional counting experiment result $\hat{\mu} = \frac{N-B}{S}$ when the x offers no discrimination power between b and s, which implies $p_s = p_b$.

With the test statistic $\tilde{t}_\mu$, we can study the expected discovery potential and the expected upper limits of the analysis. The discovery potential corresponds to studying the background-only hypothesis $\mu = 0$, where the test statistic $q_0 \equiv \tilde{t}_0$ takes the form

$$q_0 = \begin{cases} -2\,\text{Ln}\,\dfrac{\mathcal{L}(0, s, b)}{\mathcal{L}(\hat{\mu}, s, b)} & \text{if } \hat{\mu} \geq 0, \\ 0 & \text{if } \hat{\mu} < 0, \end{cases} \tag{6}$$

and plugging Eq. (2) explicitly in

$$q_0 = \begin{cases} -2\hat{\mu} S + 2 \sum_{i=1}^{N} \text{Ln}\left(1 + \dfrac{\hat{\mu} S}{B} \dfrac{p_s(x_i)}{p_b(x_i)}\right) & \text{if } \hat{\mu} \geq 0, \\ 0 & \text{if } \hat{\mu} < 0. \end{cases} \tag{7}$$

In general $p_{s,b}(x)$ are not known and are usually approximated by discrete binned distributions. For D bins, one obtains in each bin d the expected number of background events B_d, the expected number of signal events S_d and the measured number of events N_d, so Eq. (2) turns to [51]

$$\mathcal{L}(\mu, s, b) = \prod_{d=1}^{D} \text{Poiss}\big(N_d|\mu S_d + B_d\big). \tag{8}$$

This binned log-likelihood approximation is very effective but runs into trouble when the dimensionality of the data grows, as the finite statistics renders the density estimation unreliable. For this reason, we propose a different way of dealing with the high-dimensional dataset. We train a classifier to distinguish between the signal and background hypotheses with a balanced large dataset,[4] obtaining a classification score $o(x)$ that maximizes the binary cross-entropy (BCE) and thus approaches

$$o(x) = \frac{p_s(x)}{p_s(x) + p_b(x)}, \tag{9}$$

as the classifier approaches its optimal performance, see e.g. the Machine Learning Chapter in Ref. [52]. This means that the classifier learns the per-instance likelihood ratio $\frac{p_s(x)}{p_b(x)}$, precisely the information needed in Eq. (7). We can then reduce the dimensionality by dealing with $o(x)$ instead of x, using

$$p_s(x) \to \tilde{p}_s(o(x)), \quad \text{and} \quad p_b(x) \to \tilde{p}_b(o(x)), \tag{10}$$

where $\tilde{p}_{s,b}(o(x))$ are the distributions of $o(x)$ for signal and background, obtained by evaluating the classifier on a set of pure signal or background events, respectively. Notice that this allows us to approximate both signal and background distributions individually, although only the ratio will be relevant for estimating the expected sensitivities. Since these distributions are one-dimensional, they can be easily binned and incorporated into Eq. (2). Therefore the test statistic of Eq. (7) becomes

$$q_0 = \begin{cases} -2\hat{\mu} S + 2 \sum_{i=1}^{N} \text{Ln}\left(1 + \dfrac{\hat{\mu} S}{B} \dfrac{\tilde{p}_s(o(x_i))}{\tilde{p}_b(o(x_i))}\right) & \text{if } \hat{\mu} \geq 0, \\ 0 & \text{if } \hat{\mu} < 0, \end{cases} \tag{11}$$

[4] Notice that this does not aim to reflect the measured set of N events, as at this point, we are interested in estimating only the densities.

and the condition on $\hat{\mu}$ from Eq. (5)

$$\sum_{i=1}^{N} \frac{\tilde{p}_s(o(x_i))}{\hat{\mu} S\, \tilde{p}_s(o(x_i)) + B\, \tilde{p}_b(o(x_i))} = 1. \tag{12}$$

We shall name the resulting statistical model Machine-Learned Likelihood (ML Likelihood). In this sense, we are treating the algorithm as a dimensionality-reduction technique where we learn the appropriate one-dimensional manifolds that best discriminates between signal and background. This is different from the usual way of incorporating these algorithms to experimental analyses. We are neither assuming a working point and counting events selected by the algorithm in this working point nor interpolating the Likelihood as in Ref. [53]. Our method has a more concise goal which is to take advantage of the full information of the data in a supervised analysis by replacing the cut and count procedure for the likelihood-ratio information.

This method is a simplification of the one detailed in Ref. [33] for likelihood-free inference, where we do not construct an unbinned likelihood ratio but use instead the "calibrated" estimated likelihoods obtained by applying density estimation techniques to the learned output function for each process. Although we are also using machine learning to reduce the dimensionality of the problem, we are taking an intermediate step where we only aim to approximate individual likelihoods and not to replace the likelihood-based test statistics with a learned, likelihood-free generalized log-likelihood ratio. This is evidenced by the fact that we exclude the signal strength from the training step, being instead a parameter to maximize in the manner detailed in Ref. [33]. The simplification is possible because we are dealing with additive signal whose probability distribution does not depend on the signal strength. The parameterization of the Likelihood implemented here is in some sense analogous to the use of parameterized Likelihood Ratios for Effective Field Theory searches, see e.g. Ref. [38]. Our parameterization is even simpler but, as we show in Sect. 3, still very useful to increase the statistical power of a given analysis.

The test statistic in Eq. (11) is estimated through a finite dataset of N events and thus has a probability distribution conditioned on the true unknown signal strength μ'. For a given hypothesis described by the μ' value, we can estimate numerically the q_0 distribution. With this distribution, one can estimate the median expected discovery significance med $[Z_0|\mu']$ by considering the median of the test statistic

$$\text{med}\,[Z_0|\mu'] = \sqrt{\text{med}\,[q_0|\mu']}. \tag{13}$$

In particular, in our results we will report the discovery significance of the signal-plus-background hypothesis med $[Z_0|1]$, where the significance encodes how likely is to the background-only hypothesis to explain data that follows the signal-plus-background hypothesis. A higher significance will thus imply that the background-only hypothesis can be excluded in favor of the signal-plus-background hypothesis with a larger confidence.

Notice that we do not introduce Asimov datasets here to provide an asymptotic estimation of the significance. This is because the introduction of $p(x)$ renders the definition of an Asimov dataset more complicated. We do instead a full numerical estimation where we generate a set of datasets generated under the signal-plus-background hypothesis and compute for each of them the test statistic q_0. Since our method is relatively simple, a numerical estimation of the q_0 distribution is a feasible task. Indeed, this is an advantage of the one-dimensional representation of the data.

Nevertheless, since we will be interested in comparing our method to other standard techniques, we also introduce here the median discovery significance estimate for the binned likelihoods in Eq. (8) through the use of Asimov datasets, given by the well known formula [31]:

$$\text{med}\,[Z_0^{\text{binned}}|1] = \left[2 \sum_{d=1}^{D} \left((S_d + B_d)\, \text{Ln}\left(1 + \frac{S_d}{B_d}\right) - S_d \right)\right]^{1/2}, \tag{14}$$

where again B_d and S_d are the expected number of background and signal events in bin d.

For a realistic problem, the trained classifier is usually suboptimal and the learned observable o is an approximation of the log-likelihood ratio which may miss relevant information, thus reducing the power of the considered tests. Since we are not considering a specific working point but instead taking advantage of the full information retained in o, the degree of classification power of a classifier is captured by global metrics such as the Area-Under-Curve (AUC). The AUC is the integral of the Receiver Operating Characteristic (ROC) curve $\epsilon_{s/b}(\text{WP})$, where $\epsilon_{s/b}(\text{WP})$ are the fraction of correctly classified signal/background events as a function of the WP, with a higher AUC signaling a higher overall performance. The closer the AUC is to its largest possible value (which usually is below 1), the better o captures the full distributions' information. Because of this, a higher AUC correlates with a larger significance, with an upper limit set by the optimal classifier. We emphasize that our method works regardless of whether the classifier is optimal or not, with optimality providing an upper limit of the test performance (which holds regardless of the method considered). The usefulness of this method is that it approximates the true likelihoods better than a binned log-likelihood analysis and thus provides a larger significance.

For completeness, we also provide the relevant steps to derive upper limits on μ. In this case, we need to consider the test statistic [31]:

$$\tilde{q}_\mu = \begin{cases} 0 & \text{if } \hat{\mu} > \mu, \\ -2\,\text{Ln}\,\frac{\mathcal{L}(\mu,s,b)}{\mathcal{L}(\hat{\mu},s,b)} & \text{if } 0 \le \hat{\mu} \le \mu, \\ -2\,\text{Ln}\,\frac{\mathcal{L}(\mu,s,b)}{\mathcal{L}(0,s,b)} & \text{if } \hat{\mu} < 0, \end{cases} \tag{15}$$

and to look at its median expected significance when the true hypothesis is assumed to be the background-only one:

$$\text{med}\,[Z_\mu|0] = \sqrt{\text{med}\,[\tilde{q}_\mu|0]}, \tag{16}$$

where we estimate the $\tilde{q}_\mu$ distribution by generating a set of datasets with background-only events. Then, to set upper limits to a certain level, we select the lowest μ which achieves the required median expected significance.

Finally, we would like to mention that we have neglected the different systematic uncertainties that arise when performing any measurement. It is important to notice that one should also include in Eq. (2) a set of nuisance parameters θ to capture these systematic uncertainties, so S, B, p_s and p_b will be functions of these parameters. Extending our method with systematics could be relatively straightforward, with problems potentially arising when obtaining the ML Likelihood. The reason for this issue is that we need to compute ML Likelihood by training a ML algorithm, so dealing with these systematic errors requires some ingenuity. A possibility is to extend our training dataset from x to (x, θ) with θ sampled from a prior distribution $p(\theta|x')$ from any additional measurements x', in line with the treatment detailed in Ref. [33,54]. For the sake of simplicity, we will not include them in this analysis and leave them for future works.

2.1 Similar approaches in the literature

There have already been several approaches to marry ML classifiers and statistical tests, see e.g. [33–47]. To our knowledge, the most similar methods to the current proposal can be found in Refs. [35,37,38,42,44,45], with the latter two appearing during the completion of this work. Although we share several aspects, there are enough differences that warrant this proposal.

In Refs. [35,37,42,45], the authors propose a method to detect deviations from a reference dataset (namely, the SM). They parameterize the alternative hypothesis in terms of a learnable function, a Neural Network in Refs. [35,37,42] and a non-parametric kernel in Ref. [45], which is trained to quantify discrepancies between the data and the reference model. The function then provides the log-likelihood ratio between the two hypotheses, which can then be used for hypothesis testing to discard the reference hypothesis. This is a very powerful tool for anomaly detection, but is not exactly what we are proposing in this work. We consider a supervised search where the two hypotheses are well defined and we are reducing the hypothesis test to a single parameter of interest, the signal strength μ. Our method is thus simpler and easier to implement because we are not asking the algorithm to learn the dataset but only to learn the discriminator between two different processes in a high-dimensional space. The trade-off is a lack of flexibility and, at least in this form, an impossibility to perform a model-agnostic search.

On the other hand, Ref. [44] states a similar goal to what we present in this work: to obtain the significance of a supervised search that incorporates ML classifiers. They also extend the method to unsupervised searches, which is something our method is currently not designed to do. However, we note that the proposed statistical model and thus the questions that the statistical tests can answer are different. While our formulation in terms of statistical mixture models is an enhancement of traditional analyses, their likelihood proposal intends to differentiate between different types of ensembles. The authors obtain the output distribution for the background-only and for the signal-plus-background cases, while we obtain the background-only and signal-only distributions and introduce them as part of the mixture model. The question we aim to answer is thus different: while Ref. [44] aims to discriminate between different types of ensembles, we intend to discriminate between different possible compositions of a single measured ensemble of data. The statistical test obtained reflects this difference. In our case, the test statistic is a simple extension of the usual methodology while in their case it is a different test. An additional advantage of our method is that we are reducing the problem to a single parameter of interest which we can then study.

The search for optimal sensitivity through parameterization in this work relates it to Ref. [38]. There, the parameterization occurs at the Likelihood ratio level and is specific to the Effective Field Theory scenario, see also Refs. [55,56] for a detailed explanation of the method. Parameterizing $\frac{p_s}{p_b}$ explicitly in terms of the parameters of interest, the Wilson coefficients, the learning task changes from learning the Likelihood Ratio to learning a set of functions which combined with the parameters of interest yield the Likelihood Ratio. The authors of Ref. [38] show that this parameterization provides optimal discriminating power for a wide range of possible values of the parameters of interest without specific retraining. Our method is different in the sense that it tackles a different physics scenario, additive resonant physics as opposed to non-resonant EFT, and thus considers a different parameterization. The additive resonant physics scenario implies that we learn a mixture model where the parameter of interest is the signal strength μ, which we do by estimating the individual likelihoods in the learned one-dimensional embedding space. This is in contrast to Ref. [38] where the unbinned Likelihood ratio is learned by parameterizing it in terms of the known dependence on the parameters of interest at parton level.

3 Applications

To explore the Machine-Learned Likelihood approach that connects ML classifiers with the standard statistical tests, in this section we compare the expected discovery significance Z_0 estimated following the method described in the previous section against the usual and naive counting experiment result. First we consider a toy model, where the data is generated from Gaussian variables. In this simple example, we can explore the potential and robustness of our ML Likelihood approach by comparing its performance against the optimal log-likelihood ratio statistical test obtained using the true underlying probability density functions (pdfs). Moreover, we also compare it with the calculation of $S/\sqrt{B}$ considering a subset of events obtained by applying different cuts, i.e. defining a working point, with the same classifier used in our estimation. Finally we study a more realistic situation, where the true generative functions are unknown, by considering a search for new BSM particles in a dijet final state at the LHC.

In both examples we train our per-event classifiers using `XGBoost` [57], an optimized gradient boosting library that provides a parallel tree boosting. Maximum depth was set to 5, the number of estimators up to 500, and `binary:logistic` as objective to perform a logistic regression for binary classification. The evaluation metric for validation data is `logloss`, and early stopping was established after 50 rounds to avoid overtraining. We have checked that modifying slightly the `XGBoost` parameters does not change significantly our results. Furthermore, other ML algorithms suitable for the classification problem can also be used as long as they give good performance, for example deep neural networks. The use of Boosting algorithms for High Energy Physics is certainly not new, but its implementation for Likelihood estimation is not so common. This is probably due to the fact that its basis algorithm, Decision Trees, has been known to introduce non-smooth regions in the Likelihood Ratio estimator due to its very nature [33]. However, the Boosting strategy circumvents that problem by the recursive application of Decision Trees. See Ref. [58] for a similar example of the power of Boosting Decision Trees for Likelihood Ratio estimation. In each scenario involving toy models 1M events per class where generated, while in the BSM analysis 100k signal and 100k background events were used. The training procedure was performed with half of the dataset available.

3.1 Toy example with multivariate Gaussian distributions

We begin with our first example involving events generated by multivariate Gaussian variables, $\mathcal{N}_{dim}(\boldsymbol{m}, \boldsymbol{\Sigma})$, in several scenarios by increasing the dimensionality $dim = 1, \ldots, 10$ of the problem. For each dim we consider two multivariate Gaussian distributions with their covariance matrices fixed to the $dim \times dim$ identity matrix, $\boldsymbol{\Sigma} = \mathbb{I}_{dim \times dim}$, i.e. with no correlation between them, but with different means of $\boldsymbol{m} = +0.3\ \mathbb{1}_{dim}$ for the signal and $\boldsymbol{m} = -0.3\ \mathbb{1}_{dim}$ for the background, with $\mathbb{1}_{dim}$ the size dim vector of ones.

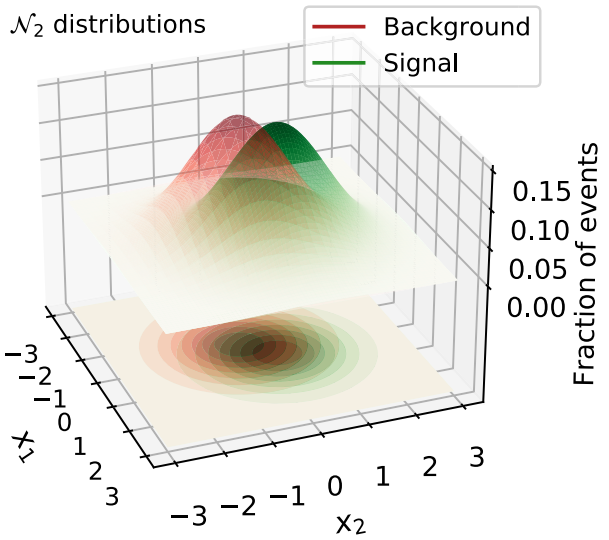


Fig. 1 Multivariate Gaussian random variables, $dim = 2$ example. $\mathcal{N}_2(+0.3\ \mathbb{1}_2, \mathbb{I}_{2\times 2})(x)$ for signal, $\mathcal{N}_2(-0.3\ \mathbb{1}_2, \mathbb{I}_{2\times 2})(x)$ for background

Dimension 2 case

To ease visualization, we consider first $dim = 2$, as shown in Fig. 1 with the signal in green and background in red. For concreteness, let us consider a fixed expected number of background events, $\langle B \rangle = 50$k, and a free number of signals events, $\langle S \rangle$, that we vary to evaluate the performance of the ML Likelihood method. On the left panel of Fig. 2 we show an example of how an ensemble with $S = 500$ events would look like. We stress again that this is just a toy model in abstract space (x_1, x_2). In a real life experiment, such as a collider analysis, this could correspond for example to the transverse momentum and pseudorapidity of a jet, $x_1 = p_T$ and $x_2 = \eta$. Moreover, the expected signal-to-background ratio would be set by the relative cross-sections and the total amount of events by the effective luminosity.

In order to apply the ML Likelihood method introduced in the previous section, we need to estimate the likelihood ratio $p_s(x)/p_b(x)$, which we obtain by training a supervised per-event classifier, `XGBoost`. At this stage we are only interested in obtaining a classifier to distinguish between signal and background, therefore to train and test the algorithm we employ all the events in our dataset, i.e. a large and balanced sample. As usual, we label signal events with a 1 and background events with a 0. A histogram of the resulting classification score, $o(x)$, can be seen on the right panel of Fig. 2 for two new independent datasets of pure signal (green) and pure

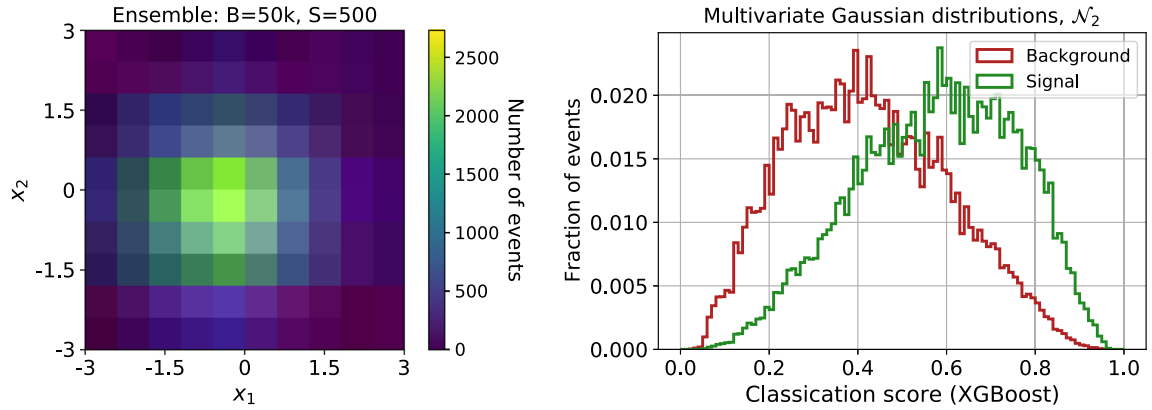


Fig. 2 Left panel: histogram of an hypothetical experiment with $B = 50$k and $S = 500$ events divided into 10×10 bins. Signal and background distributions can be seen in Fig. 1. Right panel: classification score, $o(x)$, for a binary classifier using the `XGBoost` algorithm

background (red) events. Then, we use these classification output distributions to estimate the per-instance likelihood ratio

$$\frac{p_s(x)}{p_b(x)} \to \frac{\tilde{p}_s(o(x))}{\tilde{p}_b(o(x))}. \tag{17}$$

This is a key step of the method, where we approximate the signal and background *dim*-dimensional pdfs, $p_{s,b}(x)$, by binning the 1-dimensional ML classification score $o(x)$, and thus taking advantage of its dimensionality-reduction power.

Now we can focus on the test statistic of Eq. (11). Following the procedure of Sect. 2, first we need to compute $\hat{\mu}$, the signal strength value that maximizes the likelihood. To do this we construct 10k ensembles mixing background and signal events such that the number of events per class are taken from Poisson distributions with means $\langle B \rangle = 50$k and a value of $\langle S \rangle$. For each ensemble, solving Eq. (12) we obtain numerically a value of $\hat{\mu}$ that we finally use to calculate the test statistic q_0.

The median expected discovery significance for the ML Likelihood method, estimated as the median of the test statistic, is shown in Fig. 3 as a red curve. Notice that Z_0 is depicted as a function of $S/\sqrt{B}$, therefore the identity relation (black dotted curve) represents the naive significance approximation of Eq. (14) considering a single bin and $S \ll B$. We can see that the curve is above this naive estimate even for low significances, $Z_0 < 1$.

On the other hand, since we are dealing with a low dimensional problem, we can also employ a binned Poisson log-likelihood approximation, Eq. (8), and its median discovery significance estimated by introducing the Asimov dataset, Eq. (14). Then, we calculate numerically B_d and S_d constructing ensembles of 10×10 bins with $\langle B \rangle = 50$k and different values of $\langle S \rangle$ from a 1M events database of per class. The resulting median significance is also shown in Fig. 3 as a light blue curve.

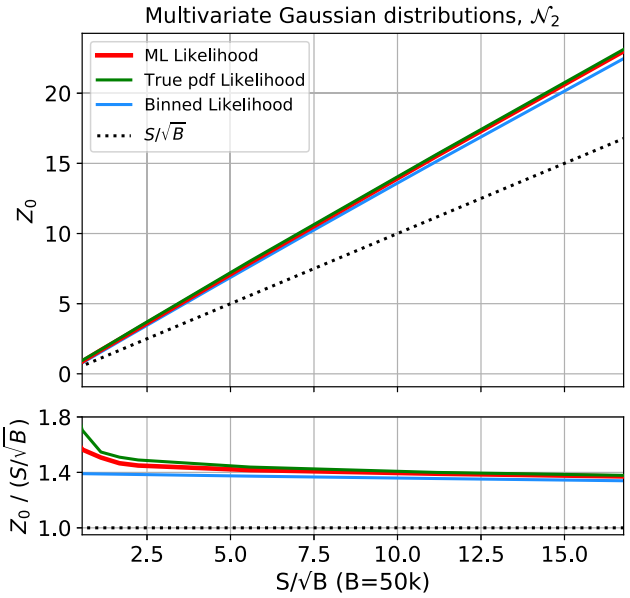


Fig. 3 Discovery significance calculated with various methods for the example in Fig. 2 for fixed background, $\langle B \rangle = 50$k, and different signal strengths $\langle S \rangle$. Red lines show the results implementing the ML Likelihood method with `XGBoost` used to estimate the probability density for single events and green lines when using the true multivariate distributions. The black dotted line represents the result of the usual counting method, $S/\sqrt{B}$, in the entire range of interest (only one bin), and the light blue curve the result of a binned counting experiment

Finally, we also present in Fig. 3 as a green curve the significance using the true probability density functions, i.e. $p_{s,b}(x) = \mathcal{N}_2(\pm \boldsymbol{m}, \boldsymbol{\Sigma})(x)$ for our example. Since we do not use a classifier to approximate $p_{s,b}(x)$, we can consider it as an optimal scenario and, thus, the green curve represents an upper limit for Z_0. For this simple example, we see that the results obtained using the ML Likelihood approach are very close to the optimal scenario and slightly outperforms the binned Poisson method.

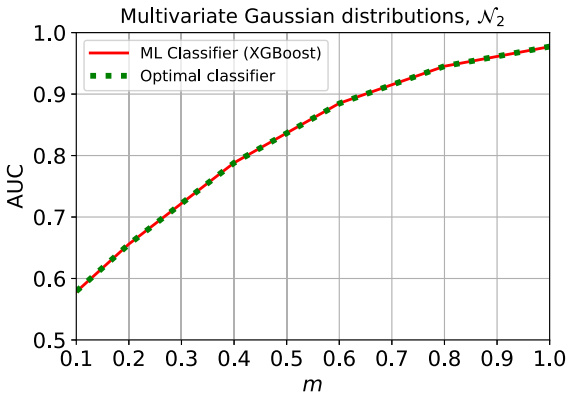
Before moving to a higher dimensional scenario, let us explore the new method in more detail. First of all, it is important to highlight that the classification score is one-dimensional by construction regardless of the dimensionality (the number of components) of our data and therefore can be easily binned. This is not the case for the binned Poisson log-likelihood approximation, where the number of bins needed to estimate the density increases with the number of components and can eventually be problematic. In this $dim = 2$ example each feature range on the left panel of Fig. 2 is divided in 10 bins, therefore we end up with $10^{dim=2} = 100$ bins. However, as the complexity, i.e. the dimension, of the problem grows, the problem of efficiently binning every feature becomes exponentially more challenging, and in practice intractable for a finite number of events. On the other hand, with the ML Likelihood method we do not need to bin the original features, but the one-dimensional classification score, which was divided into 100 bins to approximate $\tilde{p}_{s,b}(o(x))$ as can be seen on the right panel of Fig. 2. The introduction of binning can reduce the obtained significance for the ML Likelihood as the Likelihood ratio $\tilde{p}_s(o(x))/\tilde{p}_b(o(x))$ is approximated to an averaged version, just as it does when going from the True Likelihood to the Binned Likelihood. However, we emphasize again that because we are binning in one dimension, optimal choices of binning can be easily explored.

Second, we want to explore how the new method behaves when changing the performance of the classifier itself. In order to do that, we still consider the same two $dim = 2$ multivariate Gaussian distributions, but we vary their means, m. Notice that increasing values of m imply larger separation between signal and background, and therefore the classifier performs better. This is seen on the left panel of Fig. 4, where we display the classification power of the ML algorithm, measured by its AUC. As expected, from the left and right panels of Fig. 4 we can see that larger values of AUC imply higher significances.

We also observe that the difference between the ML Likelihood significance and the True Likelihood significance increases with the AUC, even when the classifier is approximately optimal as seen in the left panel of Fig. 4. This is an effect of the suboptimal choice of binning. There is thus an additional approximation involved beyond the dimensionality reduction mentioned in Sect. 2 that can reduce the significance of the ML Likelihood method compared to the True Likelihood significance. Recall that, if the classifier is suboptimal, the ML Likelihood significance will be inevitably lower than the true significance as the dimensionality reduction causes information loss. Additionally, the ML Likelihood significance can be further reduced if the $o(x)$ binning is suboptimal in the sense of capturing the $p_s(o(x))/p_b(o(x))$ behavior appropriately. This is amenable by exploring optimal choices of one dimensional binning. In Fig. 4 we include additional binning choices and observe a clear dependence on the ML Likelihood significance. From this we can be certain that the classifier is indeed optimal and thus any loss in significance can be attributed to the choice of binning. We also include the Significance obtained by using the classifier output to estimate directly the unbinned Likelihood Ratio $p_s(x)/p_b(x)$ through Eq. (9) that is needed to estimate the test statistic detailed in Eq. (7). We observe how the unbinned Likelihood Ratio performance lies between the 100 bins performance and the 1000 bins performance. If the unbinned Ratio was perfectly estimated, its performance should be identical to the true Likelihood performance. However, finite statistics of the training sample lead to imperfect estimation and, specially, to numerical uncertainties. The latter is one of the main motivations behind the introduction of calibration in Ref. [33]. We thus show how our method, although binning dependent, yields a comparable performance to the unbinned method but with increased stability and robustness against numerical effects on the Likelihood Ratio estimation.

On the remainder of this work we consider the intermediate choice of 100 bins as it is a good compromise between approximating the optimal results and increasing computing costs. In realistic applications, the optimal AUC is not known a priori and thus one should explore binning choices to achieve maximum significance for a given AUC. A possible avenue is suggested by previous experimental analyses [12–20], where the binning is decided in terms of a number of bins and a maximum statistical uncertainty per bin. An alternative is to implement other density estimation techniques such as Kernel Density Estimation or Normalizing Flows.

Finally, a practical comment about the implementation is in order. Notice that we treat each value of $\langle S \rangle$ as a different hypothesis, thus we construct independent ensembles and compute $\hat{\mu}$ for each scenario. This can be a somewhat tedious task and could be the source of numerical errors in the significance estimation, so one possible simplification is to repeat the significance estimation procedure fixing $\mu = 1$. The change in strategy between learning the parameter of interest and keeping it fixed amounts to change from a composite hypothesis test to a simple hypothesis test where the two hypothesis are fixed. In Fig. 5 we compare the results when $\hat{\mu}$ is calculated numerically from data satisfying Eq. (12), and when we fix $\mu = 1$, using both the true probability densities and the ML Likelihood. We see that the relative differences are small. This has a two-fold importance: it shows that the numerical procedure for $\hat{\mu}$ is correct, and also that the simple and composite hypotheses yield consistent significances, i.e. that the simple hypothesis with no learned parameters of interest and the composite hypothesis where we estimate $\hat{\mu}$ have the same discriminating power. The latter is extremely useful to assess optimality, as the Neyman–Pearson Lemma that ensures optimal power for the Likelihood Ratio test is valid for simple hypothe-



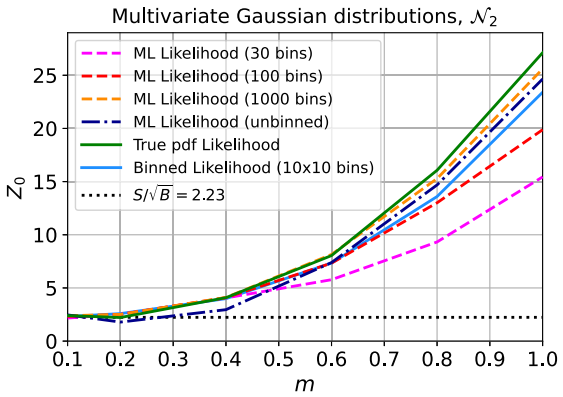


Fig. 4 Left panel: AUC obtained with XGBoost (red curve) and the optimal classifier calculated with Eq. (9) (green dotted curve) for the multivariate Gaussian example of Fig. 2, but with increasing signal-background separation m. Right panel: discovery significance calculated with various methods for the same example. Here we fixed $\langle B \rangle = 50\text{k}$ and $\langle S \rangle = 500$. Color coding is the same as in Fig. 3, but we also include the results for the ML Likelihood method with different binning choices of the classification score and the significance obtained using the estimated unbinned Likelihood Ratio

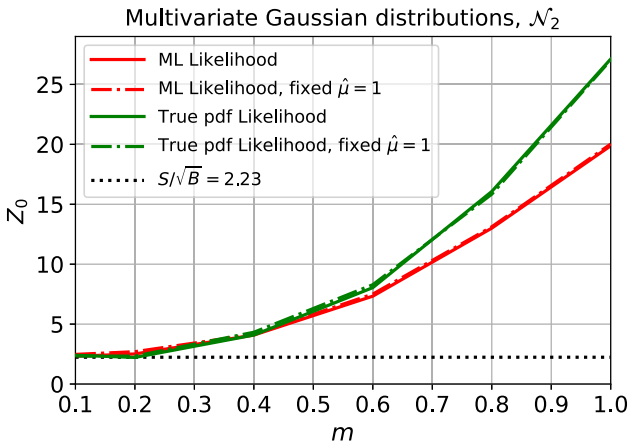


Fig. 5 Discovery significance calculated with various methods for the same multivariate Gaussian example used in Fig. 4, increasing signal-background separation m, and fixed $\langle B \rangle = 50\text{k}$ and $\langle S \rangle = 500$. We compare the results when computing $\hat{\mu}$ numerically from data (solid) with those obtained after the approximation $\hat{\mu} = 1$ (dot-dashed). Both ML Likelihood results use a classification score divided into 100 bins

ses. Notice moreover that this is true when using either the true Likelihoods or our ML-estimated Likelihoods, showing that our method yields an optimal test statistic that can be obtained using the one-dimensional learned output. Indeed, we have checked that the simple and composite hypotheses are consistent for all the examples considered in this work both for the ML Likelihood and the true Likelihoods when available, although for brevity we only report the cases with estimated $\hat{\mu}$.

High-dimensional cases

We repeat the procedure considering higher dimensional data, $\mathcal{N}_{dim}(\boldsymbol{m}, \boldsymbol{\Sigma})$, with $dim = 1, \ldots, 10$, $\boldsymbol{\Sigma} = \mathbb{I}_{dim \times dim}$, and $\boldsymbol{m} = +0.3\ \mathbb{1}_{dim}$ for the signal and $\boldsymbol{m} = -0.3\ \mathbb{1}_{dim}$ for the background. Notice that we treat each value of dim as an independent hypothesis. The results are shown in Fig. 6. On the left panel we present the AUC of the classifier and, as expected, it increases with the data dimensionality, since we introduce more information with each extra component making it increasingly easier to distinguish signal from background. We can also see that the classifier found is approximately optimal. On the right panel we show the discovery significances obtained with several procedures for a fixed value of $\langle B \rangle = 50\text{k}$ and $\langle S \rangle = 500$.

As before, the green line is computed using the true pdf distributions, and therefore provides an upper limit for the performance of obtaining Z_0. The results of the ML Likelihood method, which can be easily computed also for higher dimensions, are shown in red. By comparing the two lines, we see that both have the same tendency, although the difference between the two increases for higher dim. This can again be attributed to a suboptimal choice of $o(x)$ binning. The increase of dimensions produces the same behavior as the increase in separation for a fixed dimension (shown in Fig. 4): the two Gaussians are more distinguishable and the optimal AUC increases. Since the classifier is able to capture the increase in AUC, the $o(x)$ distribution gets more concentrated on the boundaries and thus the binning is not able to capture the likelihood ratio granularity as efficiently as for lower AUCs. Although it lies beyond the scope of this work, a ML Likelihood implementation in a real analysis where we do not know the true pdf can explore different binning choices. Even if suboptimal, ML Likelihood still provides a good estimate of the significance for relatively high dimensional problems.

Nevertheless, the main difference of increasing the dimensionality of the problem is the challenge it implies for com-

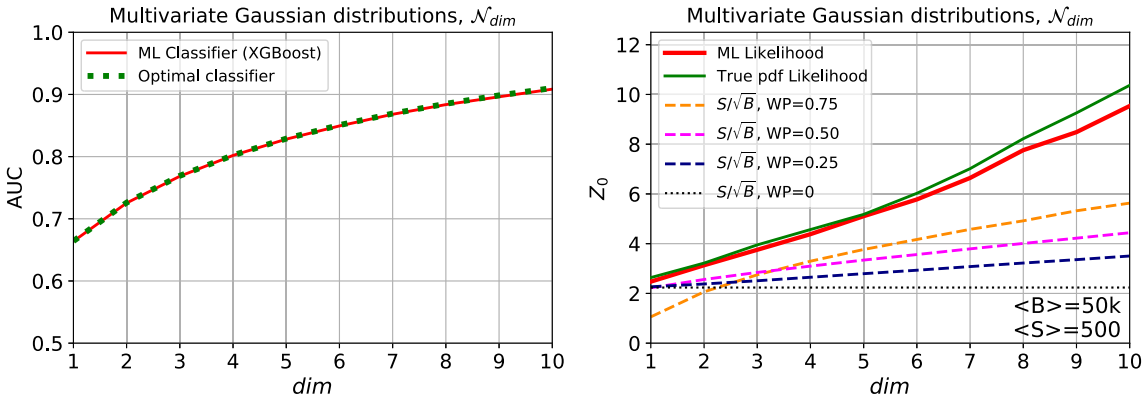


Fig. 6 Left panel: AUC for the binary classifier using `XGBoost` (red curve) and the optimal classifier calculated with Eq. (9) (green dotted curve) as a function of the data dimension, dim, for multivariate Gaussian variables. Right panel: significance calculated with various methods as a function of dim. For every case, the background and signal strengths were fixed, $\langle B \rangle = 50$k and $\langle S \rangle = 500$. Solid lines show the results implementing the method described in this work with: `XGBoost` used to estimate the probability density for single events (red), the true multivariate distributions (green). The dashed curves represent the result of the usual counting method (only one bin, $S/\sqrt{B}$), but for a subsample of the original data found with `XGBoost` assuming several working points, WP $= 0.75, 0.5, 0.25$ to obtain signal enriched regions. The black dashed line also represents the result of the usual counting procedure, but considering the entire dataset (equivalent to WP $= 0$)

puting the binned Poisson likelihood, as we did before for $dim = 2$. For example, if every component range is divided in 10 we get 10^{dim} bins, which rapidly becomes intractable for a finite amount of statistics (we recall that each of our ensemble has $\sim$ 50k events). A commonly followed procedure to face this kind of situations is to used a ML algorithm as a classifier and to define a lower cut or working point in its output $o(x)$ to define signal enriched regions and to calculate $Z_0 = S/\sqrt{B}$ on the resulting subset. We do this using the already trained `XGBoost` classifier and defining several WP, including the particular case of WP=0 equivalent to applying no cut. The results of this method are shown as dashed lines in the right panel of Fig. 6. For this example, we see that increasing the value of the WP does help improving the significances, and for values of WP=0.75 they approach to our results from the ML likelihood. Notice that this is an interesting result, since the ML Likelihood method was able to perform better, i.e. closer to the optimal green line, without the need of defining and optimizing a WP. This is due to the fact that it makes use of the complete output of the classifier, thus including as much information as possible. In addition to including all WPs, this method has the advantage that it does not need to lose events by defining a more exclusive signal region. For relatively low total number of events, this is an important enhancement.

3.2 Realistic application: a W' study at the LHC

In this subsection we will focus on a collider physics example taken from the LHC Olympics [48] challenge. The database is comprised by dijets events from two different sources: SM quantum chromodynamics (QCD) processes (background), and the production of a BSM new resonance W' with mass $m_{W'} = 3.5$ TeV. This new particle decays to two new particles X and Y with masses $m_X = 500$ GeV and $m_Y = 100$ GeV, which in turn both decay promptly to a pair of quarks producing two large-radius jets with two-prong substructure (signal). The selected events have a reconstructed dijet mass within [3.3, 3.7] TeV.

Four features are considered to characterize the process and used to test the ML Likelihood method (thus we are dealing with a $dim = 4$ problem): the invariant mass of the lighter jet (m_{j1}), the mass difference of the leading two jets (Δm_j), and the N-subjettiness ratios of the leading two jets ($\tau_{21,1}$ and $\tau_{21,2}$) [59,60]. The latter parameters quantify if a jet is described by one or two subjets, indicating a two-prong substructure for smaller values. In Fig. 7 we show the distributions of these parameters for the signal and background database.

As in the multivariate Gaussian example, we train `XGBoost` with the same hyper parameters to obtain a per-event binary classifier. We obtain AUC $= 0.96$, meaning that the algorithm can distinguish between signal and background very efficiently, as can be seen on the left panel of Fig. 8 where the classification score, $o(x)$, is shown divided into 100 bins. Then, we estimate $p_{s,b}(x)$ by the discrete binned distribution of the ML output: $\tilde{p}_{s,b}(o(x))$. We would like to highlight again that we are binning a one-dimensional distribution, while in the usual binned Poisson approximation a four-dimensional space (the number of features that describes the process) would have to be binned.

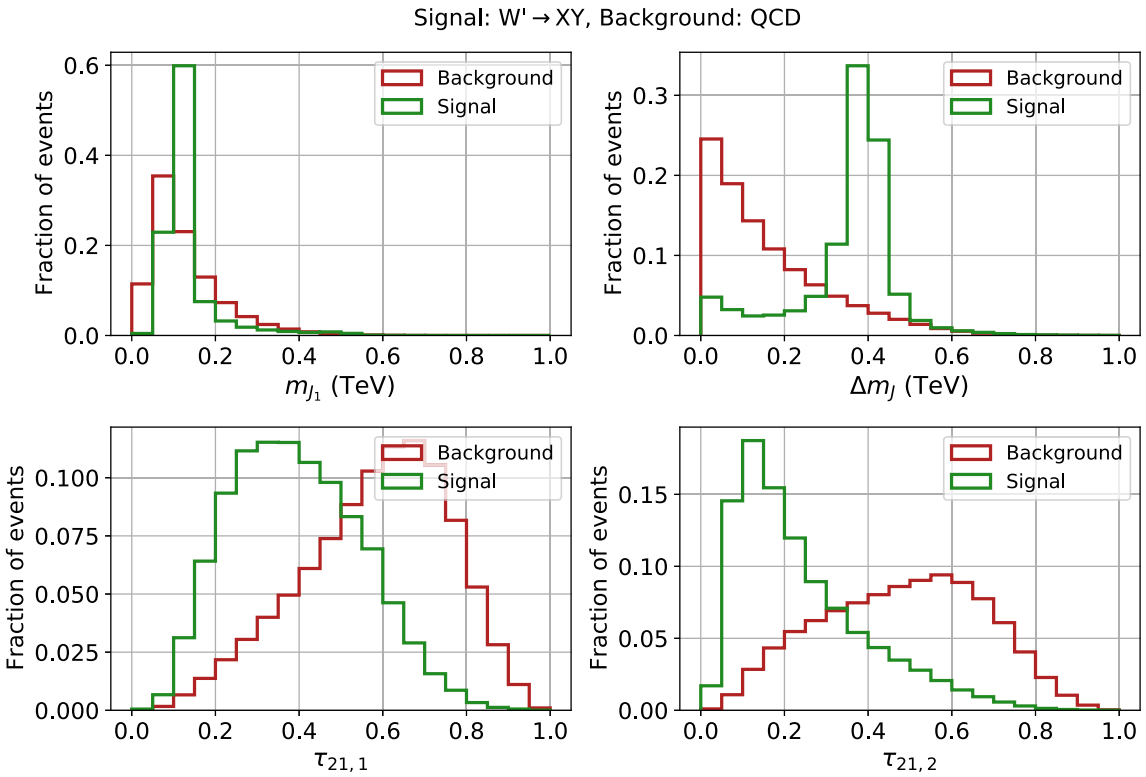


Fig. 7 Distribution of the four features that characterize the dijet events. Signal corresponds to a new resonance $W' \to XY$ both decaying to two large-radius jets with two-prong substructure, and background to QCD dijet processes. The signal distributions on the top panels are centered at $m_Y = 100$ GeV (left) and $m_X - m_Y = 400$ GeV (right), while on the bottom panels the lower values τ_{21} signal distributions indicate a two-prong substructure

To illustrate the method we fixed the expected number of background events $\langle B \rangle = 50$k and vary the expected signal events $\langle S \rangle$ within the range $[10, 300]$, since the signal-to-background ratio value would be determined by a particular model. We obtain numerically $\hat{\mu}$ satisfying Eq. (12) by constructing 10k ensembles mixing B and S events taken from Poisson distributions with means $\langle B \rangle = 50$k and different fixed $\langle S \rangle$. We obtain the ML Likelihood test statistic of Eq. (11) for each value of $\langle S \rangle$, and finally estimate the median expected discovery significance as the median of the test statistic. The results are shown on the right panel of Fig. 8 as a red curve.

We also use the already trained `XGBoost` classifier to calculate the significance by a traditional method. We consider the following working points to define a subsample within an enriched signal region, WP $= 0.75, 0.5, 0.25$. We count only the events that satisfy $o(x) >$WP and compute $S/\sqrt{B}$, shown as dashed lines on the right panel of Fig. 8. The special case WP $= 0$ shown as a black dotted line represent the naive counting significance $S/\sqrt{B}$ in the entire range, i.e. using all the events. Comparing the results it is clear that the ML Likelihood method exceeds the usual ones. For example a 5σ discovery significance would be obtained for $S \gtrsim 85$, while the number of expected signal events needs to be $\gtrsim 151, 200, 272, 1118$ for WP $= 0.75, 0.5, 0.25, 0$, respectively.

4 Discussion and outlook

In this paper we have developed a simple method, called Machine-Learned Likelihood (ML Likelihood), which can be used for any ensemble of events, that combines the current ML-technique power to deal with high-dimensional data, with the likelihood-based inference tests used in standard analyses to discriminate between different hypotheses in a minimal way, which makes it amenable for exploratory analyses without high computational costs. It allows to obtain the expected experimental sensitivity when using ML algorithms, both for discovery and exclusion limits, evidencing the utility of these sort of algorithms even for resonance searches where the parameter of interest is a signal strength.

Unlike other methods, the one proposed here makes use of all the output of the classifier, taking advantage of the entire ROC curve, and therefore its performance is better described by global quantities such as the AUC. Nevertheless, it is a

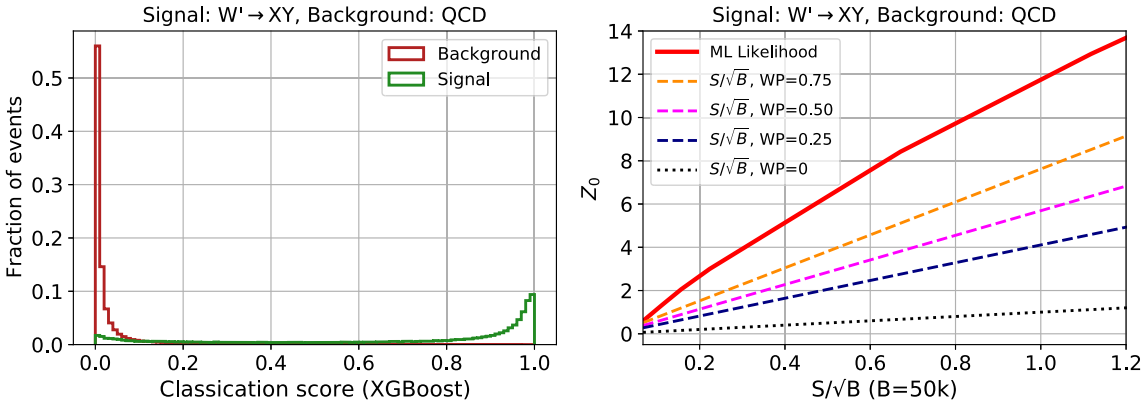


Fig. 8 Left panel: classification score for the BSM search of W' using a `XGBoost` binary classifier. Independent pure signal and pure background test samples are evaluated to estimate $p_{s,b}(x)$. Right panel: significance calculated with several methods as a function of the signal-to-background ratio for $\langle B\rangle = 50$k and $\langle S\rangle$ within [10, 300]. Same color coding as in Fig. 6. A 5σ discovery significance could be found for $S \gtrsim 85$ with our method, and for $S \gtrsim 151, 200, 272, 1118$ with a usual counting method on enriched signal regions obtained with the same classifier but WP $= 0.75, 0.5, 0.25, 0$, respectively

rather simple method, as it is based on a single classifier that is used to estimate the individual probability densities and then to evaluate the significance in the statistical test given the values of S and B for the considered ensemble, in the manner of the calibrated classifiers proposed in Ref. [33]. Notice that this is actually a key ingredient of the method, since the output of the classifier is one-dimensional by construction, and therefore the method has the advantage of always being easy to bin or even fit accurately, irrespectively of the actual dimensionality of the problem at hand.

In order to test the potential of the proposed method, we have applied it to two cases: a toy model with data generated from Gaussian variables, and a W' search in dijet final states at the LHC. The former was particularly relevant to understand the performance of the method, as we were able to compare the results obtained using the ML-estimated pdf against the true generative ones. We found that the new method leads to results that are close to the optimal case and, as expected, they remain so also for high-dimensional problems. Moreover, while its performance is similar to traditional binning analysis for low-dimensional problems, we saw that the ML Likelihood method is particularly effective for more complex problems, where traditional binning is no longer possible and standard ML analysis are used by defining a particular working point. This improvement has been found also in the more realistic analyses of W' searches at LHC, where the true generative functions are unknown. Again, we obtain higher significances with the ML Likelihood method.

To summarize, we proposed a simple method to estimate statistical significances when using ML classification algorithms. It has the main advantages of remaining simple and reliable also for high dimensional problems, and of making use of the full knowledge of the ML algorithm, without the need of relying on a given working point for the analysis. Yet, we have seen that it leads to excellent results, approaching the optimal ones computed with the true generative functions, and improving those obtained by traditional analysis techniques.

Finally, it should be mentioned that the main lacks of the ML Likelihood method are that it is not valid for anomaly detection nor can be applied to unsupervised analyses. Besides, we have not incorporated systematic uncertainties in the calculation of significance presented here. All these issues are relevant for any real analysis and will thus be addressed in future publications.

Acknowledgements The authors thank Martin de los Rios and Rosa María Sandá Seoane for useful discussions. This work is partially supported by the "Atracción de Talento" program (Modalidad 1) of the Comunidad de Madrid (Spain) under the grant number 2019-T1/TIC-14019 (EA), by the Spanish Research Agency (Agencia Estatal de Investigación) through the Grant IFT Centro de Excelencia Severo Ochoa No CEX2020-001007-S, funded by MCIN/AEI/10.13039/501100011033 (EA, XM), and by CONICET and ANPCyT under projects PICT 2017-2751, and PICT 2018-03682 (EA, AM, AP, AS). XM acknowledges partial financial support also by Grant PID2019-108892RB-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860881-HIDDeN. VML acknowledges the financial support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2121 "Quantum Universe" – 39083330 and grant María Zambrano UP2021-044 funded by Ministerio de Universidades and "European Union-NextGenerationEU/PRTR". MS acknowledges the financial support from the Slovenian Research Agency (grant No. J1-3013 and research core funding No. P1-0035).

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: This manuscript uses public datasets and there is no additional associated data.]

Funded by SCOAP[3]. SCOAP[3] supports the goals of the International Year of Basic Sciences for Sustainable Development.

References

1. B.H. Denby, Neural networks and cellular automata in experimental high-energy physics. Comput. Phys. Commun. **49**, 429–448 (1988)
2. L. Lonnblad, C. Peterson, T. Rognvaldsson, Finding gluon jets with a neural trigger. Phys. Rev. Lett. **65**, 1321–1324 (1990)
3. P. Baldi, P. Sadowski, D. Whiteson, Searching for exotic particles in high-energy physics with deep learning. Nat. Commun. **5**, 4308 (2014). arXiv:1402.4735
4. A.J. Larkoski, I. Moult, B. Nachman, Jet substructure at the large hadron collider: a review of recent advances in theory and machine learning. Phys. Rep. **841**, 1–63 (2020). arXiv:1709.04464
5. D. Guest, K. Cranmer, D. Whiteson, Deep learning and its application to LHC physics. Annu. Rev. Nucl. Part. Sci. **68**, 161–181 (2018). arXiv:1806.11484
6. K. Albertsson et al., Machine learning in high energy physics community white paper. J. Phys. Conf. Ser. **1085**(2), 022008 (2018). arXiv:1807.02876
7. A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao, T. Wongjirad, Machine learning at the energy and intensity frontiers of particle physics. Nature **560**(7716), 41–48 (2018)
8. G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborová, Machine learning and the physical sciences. Rev. Mod. Phys. **91**(4), 045002 (2019). arXiv:1903.10563
9. D. Bourilkov, Machine and deep learning applications in particle physics. Int. J. Mod. Phys. A **34**(35), 1930019 (2020). arXiv:1912.08245
10. G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, D. Shih, Machine learning in the search for new fundamental physics. arXiv:2112.03769
11. M. Feickert, B. Nachman, A living review of machine learning for particle physics. arXiv:2102.02770
12. ATLAS Collaboration, Identification of hadronically-decaying W bosons and top quarks using high-level features as input to boosted decision trees and deep neural networks in ATLAS at $\sqrt{s} = 13$ TeV, ATL-PHYS-PUB-2017-004 (2017)
13. ATLAS Collaboration, Generalized numerical inversion: a neural network approach to jet calibration, ATL-PHYS-PUB-2018-013 (2018)
14. ATLAS Collaboration, Convolutional neural networks with event images for pileup mitigation with the ATLAS detector, ATL-PHYS-PUB-2019-028 (2019)
15. CMS Collaboration, A.M. Sirunyan et al., A deep neural network for simultaneous estimation of b jet energy and resolution. Comput. Softw. Big Sci. **4**(1), 10 (2020). arXiv:1912.06046
16. CMS Collaboration, A.M. Sirunyan et al., Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques. JINST **15**(06), P06005 (2020). arXiv:2004.08262
17. ATLAS Collaboration, Deep learning for pion identification and energy calibration with the ATLAS detector, ATL-PHYS-PUB-2020-018 (2020)
18. ATLAS Collaboration, Measurement of the properties of Higgs boson production at $\sqrt{s} = 13$ TeV in the $H \to \gamma\gamma$ channel using 139 fb^{-1} of pp collision data with the ATLAS experiment, ATLAS-CONF-2020-026 (2020)
19. CMS Collaboration, A.M. Sirunyan et al., Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC. JINST **16**(05), P05014 (2021). arXiv:2012.06888
20. CMS Collaboration, A.M. Sirunyan et al., Measurements of Higgs boson production cross sections and couplings in the diphoton decay channel at $\sqrt{s} = 13$ TeV. JHEP **07**, 027 (2021). arXiv:2103.06956
21. C.K. Khosa, V. Sanz, M. Soughton, Using machine learning to disentangle LHC signatures of Dark Matter candidates. SciPost Phys. **10**(6), 151 (2021). arXiv:1910.06058
22. A. Mullin, S. Nicholls, H. Pacey, M. Parker, M. White, S. Williams, Does SUSY have friends? A new approach for LHC event analysis. JHEP **02**, 160 (2021). arXiv:1912.10625
23. S. Chang, T.-K. Chen, C.-W. Chiang, Distinguishing W' signals at hadron colliders using neural networks. Phys. Rev. D **103**(3), 036016 (2021). arXiv:2007.14586
24. F. Flesher, K. Fraser, C. Hutchison, B. Ostdiek, M.D. Schwartz, Parameter inference from event ensembles and the top-quark mass. JHEP **09**, 058 (2021). arXiv:2011.04666
25. Y.S. Lai, D. Neill, M. Płoskoń, F. Ringer, Explainable machine learning of the underlying physics of high-energy particle collisions. arXiv:2012.06582
26. E. Arganda, A.D. Medina, A.D. Perez, A. Szynkman, Towards a method to anticipate dark matter signals with deep learning at the LHC. SciPost Phys. **12**, 063 (2022). arXiv:2105.12018
27. M. Anthony, P.L. Bartlett, *Neural Network Learning: Theoretical Foundations* (Cambridge University Press, Cambridge, 1999)
28. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org
29. E.M. Metodiev, B. Nachman, J. Thaler, Classification without labels: learning from mixed samples in high energy physics. JHEP **10**, 174 (2017). arXiv:1708.02949
30. B. Nachman, J. Thaler, Learning from many collider events at once. Phys. Rev. D **103**(11), 116013 (2021). arXiv:2101.07263
31. G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics. Eur. Phys. J. C **71**, 1554 (2011). arXiv:1007.1727 [Erratum: Eur. Phys. J. C 73, 2501 (2013)]
32. ATLAS Collaboration, G. Aad et al., Measurement of the t-channel single top-quark production cross section in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. Phys. Lett. B **717**, 330–350 (2012). arXiv:1205.3130
33. K. Cranmer, J. Pavez, G. Louppe, Approximating likelihood ratios with calibrated discriminative classifiers. arXiv:1506.02169
34. A. Elwood, D. Krücker, Direct optimisation of the discovery significance when training neural networks to search for new physics in particle colliders. arXiv:1806.00322
35. R.T. D'Agnolo, A. Wulzer, Learning new physics from a machine. Phys. Rev. D **99**(1), 015014 (2019). arXiv:1806.02350
36. B. Nachman, A guide for deploying Deep Learning in LHC searches: how to achieve optimality and account for uncertainty. SciPost Phys. **8**, 090 (2020). arXiv:1909.03081
37. R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Learning multivariate new physics. Eur. Phys. J. C **81**(1), 89 (2021). arXiv:1912.12155

38. S. Chen, A. Glioti, G. Panico, A. Wulzer, Parametrized classifiers for optimal EFT sensitivity. JHEP **05**, 247 (2021). arXiv:2007.10356
39. K.T. Matchev, P. Shyamsundar, J. Smolinsky, A quantum algorithm for model independent searches for new physics. arXiv:2003.02181
40. A.S. Cornell, W. Doorsamy, B. Fuks, G. Harmsen, L. Mason, Boosted decision trees in the era of new physics: a smuon analysis case study. arXiv:2109.11815
41. J.A. Aguilar-Saavedra, Anomaly detection from mass unspecific jet tagging. Eur. Phys. J. C **82**(2), 130 (2022). arXiv:2111.02647
42. R.T. d'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Learning new physics from an imperfect machine. arXiv:2111.13633
43. V. Mikuni, B. Nachman, D. Shih, Online-compatible unsupervised nonresonant anomaly detection. Phys. Rev. D **105**(5), 055006 (2022). arXiv:2111.06417
44. C.K. Khosa, V. Sanz, M. Soughton, A simple guide from Machine Learning outputs to statistical criteria. arXiv:2203.03669
45. M. Letizia, G. Losapio, M. Rando, G. Grosso, A. Wulzer, M. Pierini, M. Zanetti, L. Rosasco, Learning new physics efficiently with nonparametric methods. arXiv:2204.02317
46. T. Finke, M. Krämer, M. Lipp, A. Mück, Boosting mono-jet searches with model-agnostic machine learning. arXiv:2204.11889
47. F.F. Freitas, J.A. Gonçalves, A.P. Morais, R. Pasechnik, Phenomenology at the Large Hadron Collider with Deep Learning: the case of vector-like quarks decaying to light jets. arXiv:2204.12542
48. G. Kasieczka et al., The LHC Olympics 2020: a community challenge for anomaly detection in high energy physics. arXiv:2101.08320
49. C.M. Bishop, *Pattern Recognition and Machine Learning. Information Science and Statistics* (Springer, New York, 2006). Softcover published in 2016
50. G. Cowan, *Statistical Data Analysis* (Oxford Science Publications, Clarendon Press, Oxford, 1998)
51. K. Cranmer et al., Publishing statistical models: getting the most out of particle physics experiments. SciPost Phys. **12**, 037 (2022). arXiv:2109.04981
52. Particle Data Group Collaboration, R.L. Workman et al., Review of particle physics. PTEP **2022**, 083C01 (2022)
53. A. Coccaro, M. Pierini, L. Silvestrini, R. Torre, The DNNLikelihood: enhancing likelihood distribution with Deep Learning. Eur. Phys. J. C **80**(7), 664 (2020). arXiv:1911.03305
54. A. Ghosh, B. Nachman, D. Whiteson, Uncertainty-aware machine learning for high energy physics. Phys. Rev. D **104**(5), 056026 (2021). arXiv:2105.08742
55. J. Brehmer, K. Cranmer, G. Louppe, J. Pavez, A guide to constraining effective field theories with machine learning. Phys. Rev. D **98**(5), 052004 (2018). arXiv:1805.00020
56. J. Brehmer, K. Cranmer, G. Louppe, J. Pavez, Constraining effective field theories with machine learning. Phys. Rev. Lett. **121**(11), 111801 (2018). arXiv:1805.00013
57. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM (2016)
58. S. Chatterjee, S. Rohshap, R. Schöfbeck, D. Schwarz, Learning the EFT likelihood with tree boosting. arXiv:2205.12976
59. J. Thaler, K. Van Tilburg, Identifying boosted objects with N-subjettiness. JHEP **03**, 015 (2011). arXiv:1011.2268
60. J. Thaler, K. Van Tilburg, Maximizing boosted top identification by minimizing N-subjettiness. JHEP **02**, 093 (2012). arXiv:1108.2701