Eur. Phys. J. C (2022) 82:275
https://doi.org/10.1140/epjc/s10052-022-10226-y

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Experimental Physics

Learning new physics from an imperfect machine

Raffaele Tito D'Agnolo[1], Gaia Grosso[2,3,a], Maurizio Pierini[2], Andrea Wulzer[3,4], Marco Zanetti[3]

[1] Institut de Physique Théorique, Université Paris Saclay, CEA, 91191 Gif-sur-Yvette, France
[2] Experimental Physics Department, CERN, Geneva, Switzerland
[3] Dipartimento di Fisica e Astronomia, Universitá di Padova and INFN, Sezione di Padova, via Marzolo 8, 35131 Padova, Italy
[4] Institut de Théorie des Phénomenes Physiques, EPFL, Lausanne, Switzerland

Received: 12 December 2021 / Accepted: 18 March 2022 / Published online: 30 March 2022

Abstract We show how to deal with uncertainties on the Standard Model predictions in an agnostic new physics search strategy that exploits artificial neural networks. Our approach builds directly on the specific Maximum Likelihood ratio treatment of uncertainties as nuisance parameters for hypothesis testing that is routinely employed in high-energy physics. After presenting the conceptual foundations of our method, we first illustrate all aspects of its implementation and extensively study its performances on a toy one-dimensional problem. We then show how to implement it in a multivariate setup by studying the impact of two typical sources of experimental uncertainties in two-body final states at the LHC.

Contents

1 Introduction . . . 1
1.1 Overview of the methodology . . . 2
1.2 Structure of the paper . . . 4
2 Foundations . . . 4
2.1 Hypothesis testing . . . 4
2.2 The central-value reference hypothesis . . . 5
2.3 Learning the effect of nuisance parameters . . . 5
2.4 Maximum likelihood from minimal loss . . . 7
2.5 Asymptotic formulae . . . 8
2.6 New physics in auxiliary measurements or in control regions . . . 9
3 Step-by-step implementation . . . 10
3.1 Model selection . . . 11
3.2 Learning nuisances . . . 13
3.3 Computing the test statistic . . . 14
3.4 Validation . . . 16
3.5 Sensitivity to new physics . . . 18
4 Two-body final state . . . 20
4.1 Model selection . . . 22
4.2 Learning nuisances and validation . . . 24
4.3 The τ-like scenario . . . 26
4.4 Sensitivity to new physics . . . 28
5 Conclusions and outlook . . . 31
Appendix A: Model-independent strategies . . . 34
References . . . 35

[a] e-mail: gaia.grosso@cern.ch (corresponding author)

1 Introduction

Experimental results in the last several decades consolidated our knowledge of fundamental physics as described by "standard" theoretical models such as the Standard Model (SM) of particle physics or the ΛCDM model of cosmology. On the other hand we lack understanding of the microscopic origin of several ingredients of these models, such as the Dark Matter and Dark Energy densities in ΛCDM, the electroweak scale and the Yukawa couplings structure in the SM. These considerations, as well as the theoretical incompleteness of our current theory of gravity, guarantee the existence of new fundamental laws waiting to be discovered, but do not sharply outline a path towards their actual experimental discovery.

One can take the incompleteness of the standard models as guidance to formulate putative "new physics" models or scenarios that complete the standard models in one or several aspects. Then one can organize the exploration of new fundamental laws as the search for the experimental manifestations of such models. We call these searches "model-dependent" as they target the signal expected in one specific model and have poor or no sensitivity to unexpected signals. The problem with this strategy is that each new physics model only offers one possible solution to the problems of the standard models. Even searching for all of them experimentally, we are not guaranteed to achieve a discovery, as the actual solution might be one that we have not yet hypothesized. This

possibility should be taken seriously also in light of the lack of discovery so far in the vast program of model-dependent searches carried out at past and ongoing experiments.

The development of "model-independent" strategies to search for new physics emerges in this context as a priority of fundamental physics. We dub model-independent those strategies that aim at assessing the compatibility of data with the predictions of a Reference theoretical Model, to be interpreted as one of the "standard" models previously discussed, rather than at probing the signatures of a specific alternative model, as in traditional model-dependent searches. It should be noted on the one hand that testing one Reference hypothesis with no assumption on the set of allowed alternative hypotheses is an ill-defined statistical concept. On the other hand, it is often trivial in practice to tell the level of compatibility of the Reference Model with the data of an experiment whose outcome consists of a single or a few measurements. The statistical distribution of the measurements is known and can be compared with the one predicted by the Reference Model. Combining a limited number of measurements does not spoil the sensitivity even if the departure from the Reference Model is present in one single measurement. However the problem becomes practically and conceptually non-trivial in modern fundamental physics experiments where the data are extremely rich and the number of possible measurements is essentially infinite. In model-dependent strategies one restricts the set of measurements to those where the specific new physics model is expected to contribute significantly, and/or one exploits the correlation between the outcome of different measurements predicted by the new physics model. Obviously this is not an option in the model-independent case.

We consider here the model-independent method that we proposed and developed in Refs. [1,2] for data analysis at particle colliders such as the Large Hadron Collider (LHC). In this case the data $\mathcal{D} = \{x_1, \ldots, x_{\mathcal{N}_\mathcal{D}}\}$ consist of $\mathcal{N}_\mathcal{D}$ independent and identically-distributed measurements of a vector of features x. The physical knowledge of the Reference Model (the SM) can be used to produce a synthetic set of Reference data $\mathcal{R} = \{x_1, \ldots, x_{\mathrm{N}_\mathcal{R}}\}$, whose elements follow the probability distribution of x in the Reference hypothesis "R". In general, $\mathcal{R}$ could be a weighted event sample. The Reference Model can also predict the total number of events N(R) expected in the experiment, around which the number of observations $\mathcal{N}_\mathcal{D}$ is Poisson-distributed. Model-independent search strategies aim at exploiting these elements for a test of compatibility between the hypothesis R and the data.[1] In order to be useful, the test should be capable to detect "generic" departures of the data distribution from the Reference expectation. Moreover it should target "small" departures in the distribution. The significance of the discrepancy can be large, but the signal can be sizable (i.e., given by a number of events that is large, relative to the Reference model expectation) only in a small (low-probability) region of the features space, or its significance emerge from correlated small differences in a large region. This is because previous experiments and theoretical considerations generically exclude the viability of new physics models that produce a radical deformation of the LHC data distribution, which are furthermore easier to detect.

As said, the Reference sample $\mathcal{R}$ consists of synthetic instances of the variable x that follow the distribution predicted by the Reference Model. It plays conceptually the same role as the background dataset in regular model-dependent searches and it can be obtained either by a first-principle Monte Carlo simulation based on the fundamental physical laws of the Reference Model, or with data-driven methods. In the latter case, one could extrapolate the background from data measured in a control region, using transfer functions that are extracted from Monte Carlo simulations. In both cases, $\mathcal{R}$ results from a knowledge of the Reference Model that is unavoidably imperfect. Therefore it provides only an approximate representation of the data distribution in the Reference (or background) hypothesis. Uncertainties emerge from all the ingredients of the simulations such as the value of the Reference Model input parameters, of the parton distribution functions and of the detector response, as well as from the finite accuracy of the underlying theoretical calculations. The impact of all these uncertainties must be assessed and included if needed in any LHC analysis. In this paper we define a strategy to deal with them in our framework for model-independent new physics searches.[2]

1.1 Overview of the methodology

In this work we develop a full treatment of systematic uncertainties within a model-independent search. Our treatment follows closely the canonical high-energy physics profile likelihood approach, reviewed in Ref. [13]. Each source of imperfection in the knowledge of the Reference Model is associated with a nuisance parameter ν. Its (true) value is unknown but statistically constrained by an "auxiliary" dataset $\mathcal{A}$, which produces a ν-dependent multiplicative term in the likelihood, $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})$. The Reference Model prediction for the distribution of the variable x depends on the nui-

[1] A concise overview of the fast-growing literature on model-independent LHC searches and a categorization of the different approaches is reported in Appendix A. In particular the origin and the role played by the Reference data set $\mathcal{R}$ in each type of approach is discussed there.

[2] Previous attempts to include systematic uncertainties in Machine Learning applications [3–12] mainly focused on modeling nuisance parameters in more traditional setups, e.g., on classifiers for tagging.

sance parameters, which we collect in a vector $\boldsymbol{\nu}$. The Reference Model is thus interpreted as a composite (parameter-dependent) statistical hypothesis $\mathrm{R}_{\boldsymbol{\nu}}$, to be identified with the null hypothesis H_0 of the statistical test. The alternative hypothesis H_1 is defined as a local (in the features space) rescaling of the Reference distribution by the exponential of a neural network function $f(x;\mathbf{w})$. The H_1 hypothesis is clearly also a composite one. We denote it as $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$, where $\mathbf{w}$ represents the trainable parameters of the neural network. Our strategy consists of performing a hypothesis test, based on the Maximum Likelihood log-ratio test statistic [14–16], between the $\mathrm{R}_{\boldsymbol{\nu}}$ and $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ hypotheses. Namely our test statistic t (see Eq. (8)) is twice the logarithm of the ratio between the likelihood of $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ given the data (times the auxiliary likelihood $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})$), maximized over $\mathbf{w}$ and $\boldsymbol{\nu}$, and the likelihood of $\mathrm{R}_{\boldsymbol{\nu}}$ (times $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})$) maximized over $\boldsymbol{\nu}$.

The concept is literally the same as in Refs. [1,2], with the difference that the Reference hypothesis is now composite rather than simple (i.e., $\boldsymbol{\nu}$-independent) and the H_1 hypothesis also depends on the nuisances and not only on the neural network parameters $\mathbf{w}$. As in Refs. [1,2], the choice of a neural network model for H_1 is motivated by the quest for an unbiased flexible approximant that can adapt itself to generic departures of the data from the Reference distribution, in order to maximize the sensitivity of the hypothesis test to generic new physics.

The first goal of the present paper is to construct a practical algorithm that computes the Maximum Likelihood log-ratio test statistic as defined above, including the effect of nuisance parameters. The basic idea is to normalize the $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ and $\mathrm{R}_{\boldsymbol{\nu}}$ likelihoods to the likelihood of the "central-value" Reference hypothesis $\mathrm{R}_{\mathbf{0}}$, namely the one where the nuisance parameters are set to their central value ($\boldsymbol{\nu}=0$) that maximizes the observed auxiliary likelihood. In this way we divide the calculation of the test statistic t in the evaluation of two separate terms. One of them merely consists of the likelihood log-ratio between the nuisance-dependent $\mathrm{R}_{\boldsymbol{\nu}}$ likelihood maximized over $\boldsymbol{\nu}$, and the likelihood of the central-value $\mathrm{R}_{\mathbf{0}}$ hypothesis. Maximizing the background-only likelihood as a function of the nuisance parameters is a necessary step of any LHC analysis. It serves in the first place to quantify the pull of the best-fit values of the nuisances, that maximize the complete likelihood (including the likelihood of the data of interest and of the auxiliary data, $\mathcal{A}$), relative to their central value estimates and uncertainties as obtained from the auxiliary likelihood alone. Therefore the determination of the first term in t does not pose any novel challenge, and could be in principle performed with the standard strategy of employing a binned approximation of the likelihood after modeling the dependence of the cross section in each bin on the nuisances. For the specific applications studied in this paper we have found more effective and more easy to employ an un-binned likelihood reconstructed by neural networks [17–23].

The other term required for the determination of the test statistic t involves the neural network and requires the maximization over the neural network parameters $\mathbf{w}$ (and over $\boldsymbol{\nu}$). It will be obtained by neural network training (with simultaneous minimization over $\boldsymbol{\nu}$), with a strategy that is a relatively straightforward generalization of the one we already employed [1,2] in the absence of nuisance parameters. As in Refs. [1,2], the training data are the observed dataset $\mathcal{D}$ and the Reference dataset $\mathcal{R}$. The Reference data are supposed to represent the distribution in the central-value hypothesis $\mathrm{R}_{\mathbf{0}}$, therefore they are obtained fixing each nuisance parameter to its central value. They do not contain any information on the variability of the Reference distribution due to the nuisances, which is taken into account by the first term of the test statistic. This avoids employing in the training Reference samples with multiple values of the nuisance parameters. The algorithm is thus not more computationally expensive than the one in the absence of nuisances.

Like any other frequentist hypothesis test, the practical feasibility of our strategy is linked to the validity of asymptotic formulae for the distribution of the test statistic t in the null hypothesis $\mathrm{R}_{\boldsymbol{\nu}}$, $P(t|H_0) = P(t|\mathrm{R}_{\boldsymbol{\nu}})$. In particular the asymptotic formulae are needed to ensure the independence of $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ on the nuisance parameters $\boldsymbol{\nu}$ [13,24]. The Wilks–Wald Theorem [15,16] predicts a χ^2 distribution for t in the asymptotic (infinite sample) limit, but it gives no quantitative information on how "large" the dataset should be, in order for $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ to be similar to a χ^2. Furthermore there is obviously no universal lower threshold on the data statistics after which the asymptotic result starts applying. The threshold depends on the problem and, crucially, on the complexity of the statistical model that is being considered. For instance if a simple one-parameter linear model was used for the numerator hypothesis instead of a neural network, a statistics of a few data events might suffice to reach the asymptotic limit accurately. Larger and larger datasets will be needed if the expressivity of the model is increased using neural networks of increasing complexity. One can of course also adopt the opposite viewpoint, which is more convenient in our case where the statistics of the data is fixed, and consider the upper threshold for the model complexity below which the asymptotic limit is reached and the distribution of t starts following the χ^2 distribution.

We need the asymptotic formula to hold in order to eliminate or mitigate the dependence of $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ on $\boldsymbol{\nu}$. On the other hand, we would like our model to be as complex and expressive as possible in order to be sensitive to the largest possible variety of putative new physics effects. Therefore the optimal complexity for the neural network model is right at the threshold of loosing the χ^2 compatibility. In Ref. [2] we already advocated this χ^2 compatibility criterion for the selection of the neural network model, with the motivation that the t distribution not following the asymptotic formula signals that

t is sensitive to low-statistics regions of the dataset, a fact which in turn can be interpreted as "overfitting" in our context. This heuristic motivation remains, but it is accompanied by the stronger technical argument associated with the feasibility of the hypothesis test including nuisance parameters.

1.2 Structure of the paper

The rest of the paper is organized as follows. In Sect. 2 we describe the statistical foundations of our method. Namely we show how to turn the mathematical definition of the Maximum Likelihood ratio test statistic into a practical algorithm for its evaluation along the lines described above. The implementation of the algorithm in all its aspects, including the selection of the neural network hyperparameters by the χ^2 compatibility criterion, is described in Sect. 3 for an illustrative univariate problem. In that section we will obtain a first validation of our method by studying how it reacts to toy datasets generated with values of the nuisance parameters that are different from the central values employed for the Reference training set. We will see that the term in t coming from the neural network is typically large, its distribution over the toys shifts to the right and gets strongly distorted with respect to the distribution one obtains when the toy data are instead generated with central-value nuisances. The other term in t, associated with the $\mathrm{R}_{\boldsymbol{\nu}}/\mathrm{R}_{\mathbf{0}}$ likelihood ratio as previously described, engineers a non-trivial cancellation on the total value of t for each individual toy. A χ^2 distribution is eventually recovered for the total t distribution, compatibly with the Wilks–Wald Theorem, regardless of the value of $\boldsymbol{\nu}$ used in the generation of the toy data. Similar tests are performed in Sect. 4 in a slightly more realistic problem with five features (kinematical variables) that represent a dataset that one might encounter in the study of the production of two particles at the LHC. Two common sources of uncertainties are included, and their impact on the sensitivity of our strategy to benchmark putative signals is quantified. We report our Conclusions in Sect. 5. Appendix A provides an overview of model-independent strategies in connection and comparison with ours.

2 Foundations

2.1 Hypothesis testing

As explained in Sect. 1, our method consists of a hypothesis test between a null hypothesis $H_0 = \mathrm{R}_{\boldsymbol{\nu}}$ and an alternative $H_1 = \mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$. We now characterize the two hypotheses in turn, starting from the null $\mathrm{R}_{\boldsymbol{\nu}}$ Reference (i.e., the SM) hypothesis. The data collected in the region of interest for the analysis are denoted as $\mathcal{D} = \{x_1, \ldots, x_{\mathcal{N}_\mathcal{D}}\}$ and consist of $\mathcal{N}_\mathcal{D}$ instances of a multi-dimensional variable x. For instance, the region of interest for the analysis could be defined as the subset of the entire experimental dataset where a given experimental signature (e.g., two high-p_T muons reconstructed within a certain detector acceptance) has been observed. The features x would then consist of the reconstructed momenta of these particles. The region of interest might be further restricted by selection cuts that define the region X of the phase space ($x \in X$) to which the particle momenta belong. Each instance of x in $\mathcal{D}$ is thrown with a probability distribution that we denote as $P(x\,|\mathrm{R}_{\boldsymbol{\nu}})$ in the Reference hypothesis $\mathrm{R}_{\boldsymbol{\nu}}$. The total number of instances of x, $\mathcal{N}_\mathcal{D}$, is Poisson-distributed with a mean $\mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}})$ that equals the total cross section in the region X times the integrated luminosity. The likelihood of the $\mathrm{R}_{\boldsymbol{\nu}}$ hypothesis, given the observation of the dataset $\mathcal{D}$, is thus provided by the extended likelihood

$$\begin{aligned}\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D}) &= \frac{\mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}})^{\mathcal{N}_\mathcal{D}}}{\mathcal{N}_\mathcal{D}!} e^{-\mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}})} \prod_{x\in\mathcal{D}} P(x|\mathrm{R}_{\boldsymbol{\nu}}) \\ &= \frac{e^{-\mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}})}}{\mathcal{N}_\mathcal{D}!} \prod_{x\in\mathcal{D}} n(x|\mathrm{R}_{\boldsymbol{\nu}}). \end{aligned} \tag{1}$$

In the previous equation we defined for shortness

$$n(x|\mathrm{R}_{\boldsymbol{\nu}}) = \mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}})\, P(x|\mathrm{R}_{\boldsymbol{\nu}}). \tag{2}$$

We will denote $n(x|H)$, in different hypotheses H, the "distribution" of the variable x.

The Reference hypothesis distribution for x depends on a set of nuisance parameters $\boldsymbol{\nu}$. They model all the imperfections in the knowledge of the Reference Model, ranging from theoretical uncertainties like those in the determination of the parton distribution functions, to the calibration of the detector response. The nuisance parameters are (often, see below) statistically constrained by "auxiliary" measurements performed using data sets independent of $\mathcal{D}$, that we collectively denote as $\mathcal{A}$. The $\mathrm{R}_{\boldsymbol{\nu}}$ hypothesis provides a $\boldsymbol{\nu}$-dependent prediction also for the statistical distribution of the auxiliary measurements. The total likelihood of $\mathrm{R}_{\boldsymbol{\nu}}$, given the observation of both the data of interest and of the auxiliary data, thus reads

$$\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D},\mathcal{A}) = \mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D}) \cdot \mathcal{L}(\boldsymbol{\nu}|\mathcal{A}), \tag{3}$$

where we denoted, for brevity, $\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{A})$ as $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})$.

We now turn to the alternative hypothesis $H_1 = \mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$. This hypothesis should include potential departures in the distribution of the variable x from the Reference (i.e., SM) expectation. As anticipated in Sect. 1, we parametrize these departures as a local rescaling of the Reference distribution by the exponential of a single-output neural network. Following the approach of Refs. [1,2] we postulate

$$n(x|\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}) = e^{f(x;\mathbf{w})} n(x|\mathrm{R}_{\boldsymbol{\nu}}), \tag{4}$$

where f is the neural network and $\mathbf{w}$ denotes its trainable parameters. The neural network architecture and hyper-

parameters are problem-dependent. The general criteria for their optimization are discussed in Sect. 2.5 and illustrated in Sects. 3.1 and 4.1 in greater detail.

We further postulate that new physics is absent in the auxiliary data. Namely that the distribution of the auxiliary data in the $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ hypothesis is the same one as in hypothesis $\mathrm{R}_{\boldsymbol{\nu}}$

$$\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{A}) = \mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{A}) = \mathcal{L}(\boldsymbol{\nu}|\mathcal{A}). \tag{5}$$

Therefore the total likelihood of $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ is

$$\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D},\mathcal{A}) = \mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D}) \cdot \mathcal{L}(\boldsymbol{\nu}|\mathcal{A}), \tag{6}$$

where $\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D})$ is the extended likelihood

$$\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D}) = \frac{e^{-\mathrm{N}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}})}}{\mathcal{N}_{\mathcal{D}}!} \prod_{x\in\mathcal{D}} n(x|\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}), \tag{7}$$

with $n(x|\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}})$ as in Eq. (4). The total number of expected events $\mathrm{N}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}})$ is the integral of $n(x|\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}})$ over the features space. A discussion of the implications of postulating the absence of new physics in the auxiliary data as in Eq. (5), and of related aspects, is postponed to Sect. 2.6.

The test statistic variable we aim at computing and employing for the hypothesis test is the Maximum Likelihood log ratio [13,14,24]

$$t(\mathcal{D},\mathcal{A}) = 2\,\log \frac{\max\limits_{\mathbf{w},\boldsymbol{\nu}} \left[\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D},\mathcal{A})\right]}{\max\limits_{\boldsymbol{\nu}} \left[\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D},\mathcal{A})\right]}. \tag{8}$$

Notice that this definition of the test statistic, and in turn its properties [15,16], assumes that the composite hypothesis in the denominator (H_0) is contained in the numerator hypothesis (H_1). This holds in our case since the neural network function in Eq. (4) is equal to zero when all its weights and biases $\mathbf{w}$ vanish. Therefore $(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}})|_{\mathbf{w}=0} = \mathrm{R}_{\boldsymbol{\nu}}$. Also notice that the test statistic variable t depends on all the data that are employed in the analysis. In particular it depends on the auxiliary data $\mathcal{A}$ as well as on the data of interest $\mathcal{D}$. We now address the problem of evaluating t, once the data are made available either from the actual experiment or artificially by generating toy datasets.

2.2 The central-value reference hypothesis

In order to proceed, we consider the special point in the space of nuisance parameters that corresponds to their central-value determination as obtained from the auxiliary data alone. If we call $\mathcal{A}_0$ the observed auxiliary dataset, namely the one that is observed in the actual experiment, the central values of the nuisance parameters are those maximizing the auxiliary likelihood function $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A}_0)$. It is always possible to choose the coordinates in the nuisance parameters space such that the central values of all the parameters sit at $\nu = 0$. So we have, by definition

$$\max_{\boldsymbol{\nu}} \left[\mathcal{L}(\boldsymbol{\nu}|\mathcal{A}_0)\right] = \mathcal{L}(\mathbf{0}|\mathcal{A}_0). \tag{9}$$

We stress again that $\mathcal{A}_0$ represents one single outcome of the auxiliary measurements (the one observed in the actual experiment), unlike $\mathcal{A}$ (and $\mathcal{D}$) that describe all the possible experimental outcomes. Therefore $\mathcal{A}_0$, and in turn the central value of the nuisance parameters that we have set to $\boldsymbol{\nu} = \mathbf{0}$, is not a statistical variable and therefore it will not fluctuate when we will generate toy experiments, unlike $\mathcal{A}$ and $\mathcal{D}$.

The central-value Reference hypothesis $\mathrm{R}_{\mathbf{0}}$ predicts a distribution for the variable x, $n(x|\mathrm{R}_{\mathbf{0}})$, that can be regarded as the "best guess" we can make for the actual SM distribution of x before analyzing the dataset of interest $\mathcal{D}$. Correspondingly, $\boldsymbol{\nu} = \mathbf{0}$ is the best prior guess for the value of the nuisances. The likelihood of $\mathrm{R}_{\mathbf{0}}$, given by

$$\begin{aligned}\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D},\mathcal{A}) &= \mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D}) \cdot \mathcal{L}(\mathbf{0}|\mathcal{A}) \\ &= \frac{e^{-\mathrm{N}(\mathrm{R}_{\mathbf{0}})}}{\mathcal{N}_{\mathcal{D}}!} \prod_{x\in\mathcal{D}} n(x|\mathrm{R}_{\mathbf{0}}) \cdot \mathcal{L}(\mathbf{0}|\mathcal{A}),\end{aligned} \tag{10}$$

is thus conveniently used to "normalize" the likelihoods at the numerator and denominator in Eq. (8). Namely we multiply and divide the argument of the log by $\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D},\mathcal{A})$ and we obtain

$$t(\mathcal{D},\mathcal{A}) = \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A}), \tag{11}$$

where τ involves the maximization over the neural network parameters $\mathbf{w}$ and over $\boldsymbol{\nu}$

$$\tau(\mathcal{D},\mathcal{A}) = 2 \max_{\mathbf{w},\boldsymbol{\nu}} \log \left[\frac{\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})} \cdot \frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right], \tag{12}$$

while the "correction" term Δ does not contain the neural network and involves exclusively the Reference hypothesis

$$\Delta(\mathcal{D},\mathcal{A}) = 2 \max_{\boldsymbol{\nu}} \log \left[\frac{\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})} \cdot \frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right]. \tag{13}$$

Both τ and Δ are positive-definite. Since they contribute with opposite sign, the test statistic t will emerge from a cancellation between these two terms. The cancellation is more and more severe the more the data happen to favor a value of $\boldsymbol{\nu}$ that is far from the central value. In Sect. 2.5 we will describe the nature and the origin of this cancellation in connection with the asymptotic formulae for the distribution of t. Below we outline our strategy for computing τ and Δ, starting from the latter term.

2.3 Learning the effect of nuisance parameters

The correction term Δ in Eq. (13) is the log-ratio between the likelihood of the Reference hypothesis evaluated with best-fit values of the nuisance parameters, and the one with central-value nuisance parameters. This object is of interest for any statistical analysis to be performed on the dataset $\mathcal{D}$, as it provides a first indication of the data compatibility with

the Reference hypothesis. In particular a sizable departure of the best-fit nuisance parameters from the central values should be monitored as an indication of a mis-modeling of the Reference hypothesis or possibly of a new physics effect.

In order to introduce our strategy for the evaluation of Δ, it is convenient to first recall the standard approach, employed in most LHC analyses, based on a binned Poisson likelihood approximation of $\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})$. In this approach, the dataset gets binned and the observed counting in each bin is compared with the corresponding $\boldsymbol{\nu}$-dependent cross section prediction. The predictions are obtained by computing each cross section for multiple values of the nuisance parameters and interpolating with a polynomial (or with the exponential of a polynomial, to enforce cross section positivity) around the central value $\boldsymbol{\nu} = 0$. A simple polynomial is sufficient to model the dependence of the cross section on the nuisances if their effect is small. The polynomial interpolation produces analytic expressions for the cross sections as a function of $\boldsymbol{\nu}$, which are fed into the Poisson likelihood. Clearly, if the analytic dependence of the cross section on one or more nuisance parameters is known then the polynomial approximation is not needed and the exact form can be used. The maximization over $\boldsymbol{\nu}$ in Eq. (13) is then performed with standard computer packages.

In principle we could proceed to the evaluation of Δ exactly as described above. However we found it simpler and more effective to employ an un-binned $\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})$ likelihood, obtained by reconstructing the ratio between the $n(x|\mathrm{R}_{\boldsymbol{\nu}})$ and $n(x|\mathrm{R}_{\mathbf{0}})$ distributions locally in the feature space. This is achieved by a rather straightforward adaptation of likelihood-reconstruction techniques based on neural networks developed in the literature [17–23]. In particular, our implementation (briefly summarized below) closely follows Refs. [21–23] to which we refer the reader for a more in-depth exposition. As for the regular binned approach, the basic idea is to employ a polynomial approximation for the dependence of the distribution on the nuisances. The polynomial coefficients, functions of the input x, are expressed as suitably trained neural networks. For instance, in the case of a single nuisance parameter ν we would write

$$r(x;\boldsymbol{\nu}) \equiv \frac{n(x|\mathrm{R}_{\boldsymbol{\nu}})}{n(x|\mathrm{R}_{\mathbf{0}})} = \exp\left[\nu\,\delta_1(x) + \frac{1}{2}\nu^2\,\delta_2(x) + \cdots\right], \tag{14}$$

with the Taylor series expansion in the exponent truncated at some finite order. Clearly the truncation is justified only if the effect of the nuisance is a relatively small correction to the central-value distribution. More precisely, nuisance effects must be small when ν is in a "plausibility" range around 0, compatibly with the shape of the auxiliary likelihood $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})$. For instance, if the auxiliary likelihood is Gaussian with standard deviation σ_ν, we should worry about the validity of the approximation in Eq. (14) only for ν within few times $\pm\sigma_\nu$. Larger values are not relevant for the maximization in Eq. (13) because they are suppressed by $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})$. Notice that in Eq. (14) we might have opted for a polynomial approximation of the ratio r rather than of its logarithm. However the latter choice guarantees the positivity of r even when the numerical minimization algorithm is led to explore regions where ν is large. Furthermore working with $\log r(x;\boldsymbol{\nu})$ is more convenient for our purposes, as we will readily see. The polynomial expansion in Eq. (14) can be straightforwardly generalized to deal with several nuisance parameters, including if needed mixed quadratic terms to capture the correlated effects of two different parameters.

Approximations $\widehat{\delta}(x)$ of the $\delta(x)$ coefficient functions are obtained as follows. Consider a continuous-output classifier $c(x;\boldsymbol{\nu}) \in (0,1)$ defined as

$$c(x;\boldsymbol{\nu}) \equiv \frac{1}{1+\widehat{r}(x;\boldsymbol{\nu})}, \tag{15}$$

where $\widehat{r}$ has the same dependence on the nuisance parameter as the true distribution ratio r. For instance in the case of a single nuisance parameter, and truncating Eq. (14) at the quadratic order, we have

$$\widehat{r}(x;\boldsymbol{\nu}) = \exp\left[\nu\,\widehat{\delta}_1(x) + \frac{1}{2}\nu^2\,\widehat{\delta}_2(x)\right], \tag{16}$$

where $\widehat{\delta}_{1,2}(x)$ represents two suitably trained single-output neural network models.[3]

The training is performed on a set of data samples $\mathrm{S}_0(\boldsymbol{\nu}_i)$ that follow the distribution of x in the $\mathrm{R}_{\boldsymbol{\nu}}$ hypothesis at different points $\boldsymbol{\nu} = \boldsymbol{\nu}_i \neq \mathbf{0}$ in the nuisance parameters space. Two distinct $\boldsymbol{\nu}_i$ points are sufficient to learn the two coefficient functions associated to a single nuisance parameter at the quadratic order. Employing more points is possible and typically convenient for the accuracy of the coefficient functions reconstruction. Data samples produced in the central-value Reference hypothesis $\boldsymbol{\nu} = \mathbf{0}$ are also employed, one for each $\mathrm{S}_0(\boldsymbol{\nu}_i)$ sample. These central-value Reference samples are denoted as $\mathrm{S}_1(\boldsymbol{\nu}_i)$, in spite of the fact that they all follow the $\mathrm{R}_{\mathbf{0}}$ hypothesis. Each event "e" in the samples has a weight w_e, normalized such that the sum of the weights in each sample equals the total number of expected events in the corresponding hypothesis (i.e., $\mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}_i})$ for $\mathrm{S}_0(\boldsymbol{\nu}_i)$ and $\mathrm{N}(\mathrm{R}_{\mathbf{0}})$ for $\mathrm{S}_1(\boldsymbol{\nu}_i)$). The loss function is

$$L[\widehat{\delta}(\cdot)] = \sum_{\boldsymbol{\nu}_i}\left\{\sum_{\mathrm{e}\in\mathrm{S}_0(\boldsymbol{\nu}_i)} w_\mathrm{e}[c(x_\mathrm{e};\boldsymbol{\nu}_i)]^2\right.$$

[3] Alternatively, the $\widehat{\delta}_{1,2}(x)$ coefficient functions might be described by a single network with two outputs. The choice between the two options, as well as the choice of the neural networks hyper-parameters, obviously depends on the specific problem. Other models could also be considered for $\widehat{\delta}$ in alternative to neural networks.

$$+ \sum_{e \in S_1(\boldsymbol{\nu}_i)} w_e [1 - c(x_e; \boldsymbol{\nu}_i)]^2 \Bigg\} . \tag{17}$$

It is not difficult to show [21] that the $\widehat{\delta}$ networks trained with the loss in Eq. (17) converge to the corresponding coefficient function δ in the limit where the samples are large, provided of course the true distribution ratio is in the form of Eq. (14).

The basic strategy outlined above can be improved and refined in several aspects [22,23], whose detailed description falls however outside the scope of the present paper. For our purposes it is sufficient to know that the coefficient functions in Eq. (14) can be rather easily and accurately reconstructed. As such, the dependence on $\boldsymbol{\nu}$ of the distribution ratio $r(x; \boldsymbol{\nu})$ is known analytically at each point x of the features space. This solves our problem of evaluating the correction term Δ in Eq. (13), because Δ is

$$\Delta(\mathcal{D}, \mathcal{A}) = 2 \max_{\boldsymbol{\nu}} \Bigg\{ \sum_{x \in \mathcal{D}} \log[r(x; \boldsymbol{\nu})] - \mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}}) + \mathrm{N}(\mathrm{R}_{\mathbf{0}}) + \log \left[\frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})} \right] \Bigg\} . \tag{18}$$

Thanks to the fact that we adopted an exponential parametrization for r (14), the first term in the curly brackets is a polynomial in $\boldsymbol{\nu}$. The constant term of the polynomial vanishes. The higher degree terms are the sum over $x \in \mathcal{D}$ of the corresponding $\delta(x)$ coefficients, approximated with the reconstructed $\widehat{\delta}(x)$ that are provided by the trained neural networks. The second term, $\mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}})$, is proportional to the total cross section in the $\mathrm{R}_{\boldsymbol{\nu}}$ hypothesis. It can be approximated with a polynomial or with the exponential of the polynomial as in regular binned likelihood analyses. Finally, $\mathrm{N}(\mathrm{R}_{\mathbf{0}})$ is a constant and the log ratio between the $\boldsymbol{\nu}$ and the $\mathbf{0}$ auxiliary likelihoods is also known in an analytical form. Actually in most cases the auxiliary likelihood is Gaussian and $\log[\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})/\mathcal{L}(\mathbf{0}|\mathcal{A})]$ is merely a quadratic polynomial. In summary, all the terms in the curly brackets of Eq. (18) are known analytically. The maximization required to evaluate Δ is thus a trivial numerical operation for dedicated computer packages.

2.4 Maximum likelihood from minimal loss

We now turn to the evaluation of the τ term defined in Eq. (12). This term involves the $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ hypothesis, which foresees possible non-SM effects (i.e., departures from the Reference Model) in the distribution of x. Non-SM effects are parametrized by the neural network $f(x; \mathbf{w})$ as in Eq. (4). The calculation of τ involves the maximization over the neural network weights and biases, $\mathbf{w}$, and over the nuisance parameters $\boldsymbol{\nu}$. The maximization will be performed by running a training algorithm, treating both $\mathbf{w}$ and $\boldsymbol{\nu}$ as trainable parameters. The algorithm will exploit the knowledge of the δ coefficient functions that is provided by the $\widehat{\delta}$ neural networks as explained in the previous section. However the latter networks are pre-trained. Therefore their parameters are not trainable during the evaluation of τ, even if they do appear in the loss function as we will readily see.

In order to turn the evaluation of τ into a training problem, the first step is to combine Eq. (4) with the definition of r in Eq. (14), obtaining

$$n(x|\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}) = e^{f(x;\mathbf{w})} r(x; \boldsymbol{\nu})\, n(x|\mathrm{R}_{\mathbf{0}}). \tag{19}$$

We then rewrite τ in the form

$$\tau(\mathcal{D}, \mathcal{A}) = 2 \max_{\mathbf{w},\boldsymbol{\nu}} \Bigg\{ \sum_{x \in \mathcal{D}} \big[f(x; \mathbf{w}) + \log(r(x; \boldsymbol{\nu})) \big] - \mathrm{N}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}) + \mathrm{N}(\mathrm{R}_{\mathbf{0}}) + \log \left[\frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})} \right] \Bigg\} . \tag{20}$$

The first, third and fourth terms in the curly brackets are easily available. The first one depends on the neural network $f(x; \mathbf{w})$, as well as on the coefficient functions δ (approximated by the neural networks $\widehat{\delta}$) through $r(x; \boldsymbol{\nu})$ in Eq. (14). The second term is the total number of events in the $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ hypothesis, given by

$$\mathrm{N}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}) = \int_X dx\, n(x|\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}) = \int_X dx\, n(x|\mathrm{R}_{\mathbf{0}}) \cdot \exp\big[f(x; \mathbf{w}) + \log(r(x; \boldsymbol{\nu})) \big] . \tag{21}$$

Clearly $\mathrm{N}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}})$ is not easily available because $n(x|\mathrm{R}_{\mathbf{0}})$ is not known in closed form and even if it was, computing the integral as a function of $\mathbf{w}$ and $\boldsymbol{\nu}$ is numerically unfeasible.

Evaluating $\mathrm{N}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}})$ requires us to employ a Reference data set $\mathcal{R} = \{x_1, \ldots, x_{\mathrm{N}_\mathcal{R}}\}$. As described in Sect. 1, $\mathcal{R}$ consists of synthetic instances of the variable x that follow the Reference Model distribution. The $\mathcal{R}$ set plus the data $\mathcal{D}$ constitute the sample that we will employ for training the neural network $f(x; \mathbf{w})$. Notice that the $\mathcal{R}$ dataset follows, by construction, the central-value distribution $n(x|\mathrm{R}_{\mathbf{0}})$. It might result from a first-principle Monte Carlo simulation, or have data-driven origin. In both cases it might take the form of a weighted event sample. [4] We choose the normalization of the weights such that

$$\sum_{e \in \mathcal{R}} w_e = \mathrm{N}(\mathrm{R}_{\mathbf{0}}). \tag{22}$$

If the $\mathcal{R}$ sample is "unweighted", all the weights are equal, and equal to $w_e = \mathrm{N}(\mathrm{R}_{\mathbf{0}})/\mathrm{N}_\mathcal{R}$, with $\mathrm{N}_\mathcal{R}$ the Reference sample size. The Reference sample plays conceptually the same role as the central-value in regular model-dependent LHC searches. Its composition and origin is the same one of the samples $\mathrm{S}_1(\boldsymbol{\nu}_i)$ employed to learn the effect of nuisance parameters with the strategy outlined in the previous section.

[4] For instance, a data-driven background sample could be obtained from a MC-assisted reweighting of control region data as it is often done in SUSY searches.

With the normalization (22), the weighted sum of a function of x over the Reference sample approximates the integral of the function with integration measure $n(x|\mathrm{R}_{\boldsymbol{\nu}})dx$. Therefore

$$\mathrm{N}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}) \simeq \sum_{\mathrm{e}\in\mathcal{R}} w_{\mathrm{e}} \exp\left[f(x_{\mathrm{e}};\mathbf{w}) + \log(r(x_{\mathrm{e}};\boldsymbol{\nu}))\right], \tag{23}$$

where the accuracy of the approximation improves with (square root of) the size of the Reference sample. In what follows we are going to assume an infinitely abundant Reference sample and turn the approximate equality above into a strict equality. Clearly in so doing we are ignoring the uncertainties associated with finite statistics of $\mathcal{R}$. This is justified if $\mathrm{N}_{\mathcal{R}} \gg \mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}}) \sim \mathcal{N}_{\mathcal{D}}$, because in this case the statistical variability of τ is expectedly dominated by the statistical fluctuation of the data sample $\mathcal{D}$. All the results of the present paper are compatible with this expectation for $\mathrm{N}_{\mathcal{R}}$ a few times larger than $\mathcal{N}_{\mathcal{D}}$.

By combining Eqs. (20) and (23) (and (22)) and by factoring out a minus sign to turn the maximization into a minimization, we express

$$\tau(\mathcal{D},\mathcal{A}) = -2 \min_{\mathbf{w},\boldsymbol{\nu}} \left\{L\left[f(\cdot;\mathbf{w}),\ \boldsymbol{\nu};\ \widehat{\delta}(\cdot)\right]\right\}, \tag{24}$$

where L has the form of a loss function for a supervised training between the $\mathcal{D}$ and $\mathcal{R}$ samples

$$\begin{aligned} L\left[f(\cdot;\mathbf{w}),\ \boldsymbol{\nu};\ \widehat{\delta}(\cdot)\right] = &- \sum_{x\in\mathcal{D}} \left[f(x;\mathbf{w}) + \log(r(x;\boldsymbol{\nu}))\right] \\ &+ \sum_{\mathrm{e}\in\mathcal{R}} w_{\mathrm{e}} \left[e^{f(x_{\mathrm{e}};\mathbf{w})+\log(r(x_{\mathrm{e}};\boldsymbol{\nu}))} - 1\right] \\ &- \log\left[\frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right]. \end{aligned} \tag{25}$$

The loss depends on the neural network function $f(\cdot;\mathbf{w})$ and in particular on its trainable parameters $\mathbf{w}$. It also depends on the nuisance parameters $\boldsymbol{\nu}$, through the ratio r and through the auxiliary likelihood ratio term. The minimization over the nuisances is requested by Eq. (24), therefore the nuisances should be treated as trainable parameters on the same footing as the neural network parameters $\mathbf{w}$. This is relatively straightforward to implement in standard deep learning packages, provided the loss depends on $\boldsymbol{\nu}$ through analytically differentiable functions. This is the case for $r(x;\boldsymbol{\nu})$, and typically also for the auxiliary likelihood ratio. The loss also depends on the reconstructed coefficient functions $\widehat{\delta}$. However this dependence is purely parametric and the parameters of the $\widehat{\delta}$ networks are fixed at their optimal values, opportunely determined in a previous training as described in Sect. 2.3. After training, τ is obtained as minus two times the minimal loss owing to Eq. (24).

Our strategy to evaluate τ is a relatively straightforward extension of the one developed in Refs. [1,2]. In the absence of nuisance parameters, namely in the limit where $r(x;\boldsymbol{\nu})$ is independent of $\boldsymbol{\nu}$ and identically equal to one, the loss in Eq. (25) reduces to the one of Refs. [1,2], plus the auxiliary log likelihood ratio that carries all the dependence on $\boldsymbol{\nu}$ and can be minimized independently. The latter term however cancels in the test statistic t when subtracting the correction term Δ (18) and the results of Refs. [1,2] are recovered in the absence of nuisances, as it should be.

2.5 Asymptotic formulae

We now discuss the actual feasibility of a frequentist hypothesis test based on our variable t (8). The generic problem with frequentist tests stems from the determination of the distribution of the t variable in the null hypothesis, $P(t|H_0)$, out of which the p-value of the observed data is extracted. If the null hypothesis is a simple one, this can be obtained rigorously by running toy experiments, with a procedure that is computationally demanding but not unfeasible, especially if one does not target probing the extreme tail of the t distribution. If instead the null hypothesis $H_0 = \mathrm{R}_{\boldsymbol{\nu}}$ is composite as in this case, due to the nuisances, and if $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ (and in turn the p-value) depends on the value of $\boldsymbol{\nu}$, the problem becomes extremely hard as one should in principle run toy experiments and compute $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ for each value of $\boldsymbol{\nu}$. Indeed in frequentist statistics there is no notion of probability for the parameters. Consequently each value of $\boldsymbol{\nu}$ defines an equally "likely" hypothesis in the null hypotheses set $\mathrm{R}_{\boldsymbol{\nu}}$. We can thus quantify the level of incompatibility of the data with the null hypothesis only by defining the p-value as the maximum p-value that is obtainable by a scan over the $\boldsymbol{\nu}$ parameters in their entire allowed range. Since this is not feasible, the only option is to employ a suitably-designed test statistic variable, such that $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ is independent of $\boldsymbol{\nu}$ to a good approximation.

The considerations above are deeply rooted in the standard treatment of nuisance parameters. They actually constitute the very reason for the choice, in LHC analyses [24], of a specific Maximum Likelihood ratio test statistic, whose distribution is in fact independent of $\boldsymbol{\nu}$ in the asymptotic limit where the number of observations is large. Specifically, $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ approaches a χ^2 distribution with a number of degrees of freedom equal to the number of free parameters in the "numerator" hypothesis $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$, minus the number of parameters of the "denominator" hypothesis $\mathrm{R}_{\boldsymbol{\nu}}$, owing to the Wilks–Wald Theorem [15,16]. In a regular model-dependent search [24], the number of degrees of freedom of the χ^2 equals the number of free parameters of the new physics model that is being searched for (i.e., the so-called "parameters of interest"). The exact same asymptotic result applies in our case because our test statistic is also defined and rigorously computed as a Maximum Likelihood ratio. Its distribution in the null hypothesis will thus be independent of $\boldsymbol{\nu}$ and approach the χ^2. The number of degrees of freedom

is given in this case by the number of trainable parameters $\mathbf{w}$ of the neural network.

As already stressed in Sect. 1, however, asymptotic formulae such as the Wilks–Wald Theorem only hold in the limit of an infinitely large data set and therefore they offer no guarantee that $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ will resemble a χ^2 (and be independent of $\boldsymbol{\nu}$) in concrete analyses where the dataset has a finite size. At fixed dataset size, whether this is the case or not depends on the complexity (or expressivity) of the parameter-dependent hypothesis that is being compared with the data. When fitted by the likelihood maximization, an extremely flexible hypothesis will adapt its free parameters to reproduce (overfit) the observed data points individually. Therefore the value of t that results from the maximization can be driven by low-statistics portions of the dataset and thus violate the asymptotic condition even if the total size of the dataset is large. The expressivity of our hypothesis is driven by the architecture (number of neurons and layers) of the neural network $f(x;\mathbf{w})$, and by the other hyper-parameters (a weight clipping, in our implementation) that regularize the network preventing it from developing overly sharp features. We can thus enforce the validity of the asymptotic formula, i.e. ensure that $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ is close to a χ^2 and independent of $\boldsymbol{\nu}$, by properly selecting the neural network hyper-parameters.

For the selection of the hyper-parameters according to the χ^2 compatibility criterion we proceed as in Ref. [2], where this criterion had been already introduced on a more heuristic basis, unrelated with nuisance parameters. We generate toy datasets following the central-value hypothesis $\mathrm{R}_{\mathbf{0}}$, we compute t and we compare its empirical distribution with a χ^2 with as many degrees of freedom as the number of parameters of the neural network. We select the largest neural network architecture and the maximal weight clipping for which a good level of compatibility is found. Notice that whether or not a given neural network model is sufficiently "simple" to respect the asymptotic formula is conceptually unrelated with the presence of nuisance parameters. Furthermore our goal is to show that the presence of nuisances does not affect the distribution of t. Therefore when we enforce the χ^2 compatibility, with the strategy outlined above, we compute t as if nuisance parameters were absent. After the model is selected, based on the Wilks–Wald Theorem we expect that the distribution of t will be a χ^2 with the same number of degrees of freedom even in the presence of nuisance parameters. This can be verified by recomputing the distribution of t, including this time the effect of nuisances, on the $\mathrm{R}_{\mathbf{0}}$ toys and on new toy samples generated according to $\mathrm{R}_{\boldsymbol{\nu}}$ with different values $\boldsymbol{\nu} \neq \mathbf{0}$ of the nuisance parameters. Explicit implementations of this procedure, and confirmations of the validity of the asymptotic formulae, will be described in Sects. 3 and 4.

The Wilks–Wald Theorem also enables us to develop a qualitative understanding of the interplay between the τ and Δ terms in the determination of t (Eq. (11)). Both τ (Eq. (12)) and Δ (Eq. (13)) are Maximum Likelihood log-ratios, with the simple hypothesis $\mathrm{R}_{\mathbf{0}}$ playing the role of the denominator hypothesis. Therefore τ and Δ are also distributed as a χ^2_d with d degrees of freedom, if the data follow the $\mathrm{R}_{\mathbf{0}}$ hypothesis itself. In the case of τ, d is the number of neural network parameters plus the number of nuisance parameters. The number of degrees of freedom of Δ is instead given by the number of nuisance parameters. The test statistic t, whose value emerges from a cancellation between τ and Δ, has d equal to the number of neural network parameters, as previously discussed. The cancellation is not severe in this case, because the number of nuisance parameters is typically smaller than the number of neural network parameters. Namely the values of τ and Δ for each individual toy will not be, on average, much larger that $t = \tau - \Delta$. Suppose instead that the data follow $\mathrm{R}_{\boldsymbol{\nu}}$ with some $\boldsymbol{\nu} \neq \mathbf{0}$. This hypothesis belongs to the numerator (composite) hypothesis in the definitions of τ and Δ. The Wilks–Wald Theorem predicts in this case non-central χ^2 distributions [15], with increasingly large non-centrality parameters as we increase the distance between $\boldsymbol{\nu}$ and $\mathbf{0}$. Therefore when we compute $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ with larger and larger $\boldsymbol{\nu}$, the τ and Δ distributions shift more and more to the right and their typical value over the toys becomes large. The typical value of t is instead given by the number of neural network parameters, because t follows a central χ^2 distribution independently of $\boldsymbol{\nu}$. A sharp correlation between τ and Δ will thus engineer a delicate cancellation on toys generated with very large values of the nuisance parameters. The occurrence of the cancellation amplifies the uncertainties in the calculation of τ and Δ that emerge (dominantly) from the imperfect modeling of the $\delta(x)$ coefficient functions. Obtaining a χ^2 for the distribution of t is thus increasingly demanding at large $\boldsymbol{\nu}$, as we will see more quantitatively in Sects. 3 and 4.

2.6 New physics in auxiliary measurements or in control regions

The step we took in Eq. (5) of postulating the absence of new physics in the auxiliary data deserves further comments. In regular model-dependent searches for new physics the alternative hypothesis H_1 is a physical model that accounts for new phenomena in addition to the SM ones. One can thus assess whether or not these new phenomena can manifest themself in the auxiliary data. If they do not, Eq. (5) is justified. The situation is different in model-independent searches. On one hand, there is no way to tell if Eq. (5) holds because the new physics model is not given. On the other hand, in our framework we are always free to postulate Eq. (5). In a model-dependent search Eq. (5) could be wrong, in our case it is a restriction on the set of new physics models that we are testing.

Still it is interesting to discuss how the model-independent strategy that we are constructing would react to the presence of new physics effects in the auxiliary data. New (or mismodeled) effects in auxiliary data could in general reduce the sensitivity of the test to new physics, however it is not obvious that this reduction will be significant. Consider the extreme case in which new physics is absent from the dataset of interest, and is present only in the auxiliary measurements. The new physics effects make the true auxiliary likelihood function different from the postulated one, $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})$. Therefore, in the likelihood maximization, the $\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})$ term will push $\boldsymbol{\nu}$ to values that are different from the true values of the nuisance parameters. This will occur both in the maximization of the $\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D},\mathcal{A})$ and of the $\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D},\mathcal{A})$ likelihoods. For these incorrect values of the nuisance parameters, $n(x|\mathrm{R}_{\boldsymbol{\nu}})$ does not provide a good description of the distribution of the data of interest $\mathcal{D}$. Therefore the maximal likelihood of $\mathrm{R}_{\boldsymbol{\nu}}$ will be small, due to the mismatch between the data and the Reference distribution estimated from the "signal-polluted" auxiliary dataset. The $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ hypothesis instead possesses enough flexibility to adapt $n(x|\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}})$ according to the data of interest, thanks to the flexibility of the neural network (4). The likelihood of $\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}$ will thus possess a high maximum, in the configuration where $\boldsymbol{\nu}$ maximizes the auxiliary likelihood and the neural network accounts for the discrepancy between the x distribution at that value of $\boldsymbol{\nu}$ and the true x distribution at the true value of the nuisance parameters. This can enable our test to reveal a tension of the data with the Reference Model even in this limiting configuration, as we will see happening in Sect. 3.5 in a simple setup. New physics effects in the auxiliary data might thus not spoil the potential to achieve a discovery. On the other hand, they would complicate its interpretation.

Similar considerations hold for possible new physics contaminations in the Reference dataset $\mathcal{R}$ employed for training. These contaminations emerge if $\mathcal{R}$ has a data-driven origin, and if new physics affects the distribution of the data control region. Since the control region data are transferred to the region of interest by assuming the validity of the Reference Model, the net effect is a mismatch between the true distribution of x in the (central-value) Reference Model and the actual distribution of the instances of x in the Reference sample. As for auxiliary measurements, new physics in control regions does not necessarily spoil the sensitivity to new physics. Indeed our test is sensitive to generic departures of the observed data distribution with respect to the distribution of the Reference dataset. Departures which are due to a mismodeling of the Reference induced by new physics in the control region, rather than to new physics in the data of interest, could still be seen. Our strategy would instead completely loose sensitivity if new physics affects the control region and the data of interest in the exact same way, because in this case there would be strictly no difference between the distribution of the data and the one of the Reference dataset.

3 Step-by-step implementation

The present section describes the detailed implementation of our strategy and its validation in a simple case study that will serve as an explanatory example throughout the presentation of the algorithm. In particular, we consider a one-dimensional feature $x \in [0, \infty)$ with exponentially falling distribution in the Reference hypothesis. We assume that our knowledge of the Reference hypothesis is not perfect and that our lack of knowledge is described by a two-dimensional nuisance parameters vector $\boldsymbol{\nu} = (\nu_{\mathrm{N}}, \nu_{\mathrm{S}})$. The two parameters account, respectively, for the imperfect knowledge of the normalization of the distribution (i.e., of the total number of expected events $\mathrm{N}(\mathrm{R}_{\boldsymbol{\nu}}) \equiv e^{\nu_{\mathrm{N}}}\mathrm{N}(\mathrm{R}_{\mathbf{0}})$) and of a multiplicative "scale" factor (defined by $x = x_{\mathrm{meas.}} = e^{\nu_{\mathrm{S}}} x_{\mathrm{true}}$) in the measurement of x. The Reference Model distribution of x reads

$$n(x|\mathrm{R}_{\boldsymbol{\nu}}) = n(x|\mathrm{R}_{\nu_{\mathrm{N}},\nu_{\mathrm{S}}}) = \mathrm{N}(\mathrm{R}_{\mathbf{0}})\, \exp\left[-x\, e^{-\nu_{\mathrm{S}}} - \nu_{\mathrm{S}} + \nu_{\mathrm{N}}\right], \tag{26}$$

with the total number of expected events in the central-value hypothesis, $\mathrm{N}(\mathrm{R}_{\mathbf{0}})$, fixed at $\mathrm{N}(\mathrm{R}_{\mathbf{0}}) = 2\,000$. As discussed in Sect. 2.2, the central-value Reference hypothesis $\mathrm{R}_{\mathbf{0}}$ is defined to be at the point $(\nu_{\mathrm{N}}, \nu_{\mathrm{S}}) = (0, 0)$ in the nuisances' parameter space. We have parametrized the normalization, $e^{\nu_{\mathrm{N}}}$, and the scale factor, $e^{\nu_{\mathrm{S}}}$, so that they are positive in the entire real plane spanned by $(\nu_{\mathrm{N}}, \nu_{\mathrm{S}})$.

We suppose that the normalization and the scale nuisances are measured independently using an auxiliary set of data $\mathcal{A}$. The estimators of the measurements central values are denoted as $\widehat{\nu}_{\mathrm{N}} = \widehat{\nu}_{\mathrm{N}}(\mathcal{A})$ and $\widehat{\nu}_{\mathrm{S}} = \widehat{\nu}_{\mathrm{S}}(\mathcal{A})$. We assume that these estimators are unbiased and Gaussian-distributed with standard deviations σ_{N} and σ_{S}. The auxiliary likelihood log-ratio thus reads

$$2\,\log\left[\frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right] = -\left(\frac{\widehat{\nu}_{\mathrm{N}} - \nu_{\mathrm{N}}}{\sigma_{\mathrm{N}}}\right)^2 + \left(\frac{\widehat{\nu}_{\mathrm{N}}}{\sigma_{\mathrm{N}}}\right)^2 - \left(\frac{\widehat{\nu}_{\mathrm{S}} - \nu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)^2 + \left(\frac{\widehat{\nu}_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)^2. \tag{27}$$

It should be noted that $\widehat{\nu}_{\mathrm{N}}$ and $\widehat{\nu}_{\mathrm{S}}$ are statistical variables, owing to their dependence on the auxiliary data $\mathcal{A}$. Therefore we must let them fluctuate when generating the simulated experiments (toys) that we employ to validate the algorithm. Namely, denoting as $\boldsymbol{\nu}^* = (\nu_{\mathrm{N}}^*, \nu_{\mathrm{S}}^*)$ the true value of the nuisance parameter vector, the estimators $\widehat{\nu}_{\mathrm{N}}$ and $\widehat{\nu}_{\mathrm{S}}$ are thrown from Gaussian distributions with standard deviations σ_{N} and σ_{S}, centered ν_{N}^* and ν_{S}^*, respectively. This mimics the statistical fluctuations of the auxiliary data $\mathcal{A}$, out of which the estimators $\widehat{\nu}_{\mathrm{N,S}}$ are derived in the actual experiment. The true

value of the nuisance parameters $\boldsymbol{\nu}^*$ is unknown, and the validation of the method consists in verifying that the distribution of the test statistic is independent on $\boldsymbol{\nu}^*$. We will verify this on toy datasets generated with $\nu^*_{\mathrm{N,S}}$ at the central value ($\nu^*_{\mathrm{N,S}} = 0$), and at plus and minus one standard deviation ($\nu^*_{\mathrm{N,S}} = \pm\sigma_{\mathrm{N,S}}$).

The rest of this section is structured as follows. In Sect. 3.1 we describe the selection of the neural network model and regularization parameters based on the χ^2 compatibility criterion introduced in the previous section (and in Ref. [2]), and in particular in Sect. 2.5. Next, in Sect. 3.2, we illustrate the reconstruction of the coefficient functions that model the dependence of the Reference Model distribution on the nuisance parameters, following Sect. 2.3. In Sect. 3.3 we present our implementation of the calculation of the test statistic as in Sect. 2.4. In Sect. 3.4 we validate our strategy by verifying the asymptotic formulae of Sect. 2.5 and in turn the independence of the distribution of the test statistic on the true value of the nuisance parameters. Finally, in Sect. 3.5 we study the sensitivity to putative "new physics" signals that distort the distribution of x relative to the Reference Model expectation in Eq. (26). It should be emphasized that this latter study has a merely illustrative purpose. All the steps that are needed to set up our strategy, from the model selection to the evaluation of the distribution of the test statistic, are performed based exclusively on knowledge of the Reference Model and not on putative new physics signals, as appropriate for a model-independent search strategy.

While presented in the context of a simple univariate problem that is rather far from a realistic LHC data analysis problem, the technical implementation of all the steps described in the present section is straightforwardly applicable to more complex situations. The application to a more realistic problem will be discussed in Sect. 4.

3.1 Model selection

The first step towards the implementation of our strategy is to select the hyper-parameters of the neural network model "$f(x;\mathbf{w})$", which we employ to parametrize possible new physics (or Beyond the SM, BSM) effects as in Eq. (4). We restrict our attention to fully-connected feedforward neural networks, with an upper bound on the absolute value of each weight and bias. The upper limit is set by a weight clipping regularization parameter that needs to be selected. The other hyper-parameters are the number of hidden layers and of neurons per layer that define the neural network architecture.

According to the general principles outlined in Sect. 2.5, the model selection results from two competing principles. The first one is that the model should have the highest complexity that can be handled by the available computational resources in a reasonable amount of time. This maximizes the model's capability to fit complex departures from the Reference Model expectation, making it sensitive to the largest possible variety of putative new physics signals. On the other hand, the model should be simple enough for the distribution of the associated test statistic to be in the asymptotic regime, given the finite amount of training data. This condition is enforced by monitoring the compatibility with the χ^2 asymptotic formula for the test statistic distribution.

As explained in Sect. 2.5, the χ^2 compatibility condition that underlies the selection of the neural network hyperparameters will be enforced in the limit where the nuisance parameters do not affect the distribution of the variable x or, equivalently, in the limit where the auxiliary measurements of the nuisance parameters are infinitely accurate (i.e., $\sigma_{\mathrm{N,S}} \to 0$). It is easy to see from the results of Sect. 2, or from Refs. [1,2], that the test statistic in this limit becomes

$$\bar{t}(\mathcal{D}) = 2 \max_{\mathbf{w}} \log\left[\frac{\mathcal{L}(H_{\mathbf{w}}|\mathcal{D})}{\mathcal{L}(R_0|\mathcal{D})}\right] = -2 \min_{\mathbf{w}} \left\{\bar{L}\left[f(\cdot;\mathbf{w});\right]\right\}. \quad (28)$$

The minimization is performed by training the network f with the loss function

$$\bar{L}\left[f(\cdot;\mathbf{w})\right] = -\sum_{x\in\mathcal{D}} \left[f(x;\mathbf{w})\right] + \sum_{\mathrm{e}\in\mathcal{R}} w_{\mathrm{e}} \left[e^{f(x_{\mathrm{e}};\mathbf{w})} - 1\right]. \quad (29)$$

The asymptotic distribution of $\bar{t}$ is a χ^2 with a number of degrees of freedom which is equal to the number of trainable parameters of the neural network. The χ^2 compatibility of a given neural network model will be monitored by generating toy instances of the dataset $\mathcal{D}$ in the $\mathbf{R_0}$ hypothesis, running the training algorithm on each of them, computing the empirical probability distribution of $\bar{t}$ and comparing it with the χ^2.

We first discuss how to select the weight clipping regularization parameter for a given architecture of the neural network. We consider for illustration, in the simple univariate example at hand, a network with four nodes in the hidden layer (and one-dimensional input and output). We refer to this architecture as (1, 4, 1), for brevity. This network has a total of 13 trainable parameters, therefore the target $\bar{t}$ distribution is a χ^2_{13} with 13 degrees of freedom. We generated a Reference sample $\mathcal{R}$, with $\mathrm{N}_{\mathcal{R}} = 200\,000 = 100\,\mathrm{N}(\mathbf{R_0})$ entries, following the $\mathbf{R_0}$ distribution of the variable x as given by Eq. (26) for $\nu_{\mathrm{N,S}} = 0$. The sample is unweighted, therefore the weights in the sample are all equal and $w_{\mathrm{e}} = \mathrm{N}(\mathbf{R_0})/\mathrm{N}_{\mathcal{R}} = 0.01$. We also generate 400 toy instances of the dataset $\mathcal{D}$ in the same hypothesis. The number of instances of x in $\mathcal{D}$, $\mathcal{N}_{\mathcal{D}}$, is thrown from a Poisson distribution with mean $\mathrm{N}(\mathbf{R_0}) = 2\,000$ in accordance with the $\mathbf{R_0}$ expectation. For different values of the weight clipping parameter, ranging from 1 to 100, we train the neural network with the loss in Eq. (29) and we compute $\bar{t}(\mathcal{D})$ on the toy datasets using Eq. (28). The empirical $P(\bar{t}|\mathbf{R_0})$ distributions obtained in this

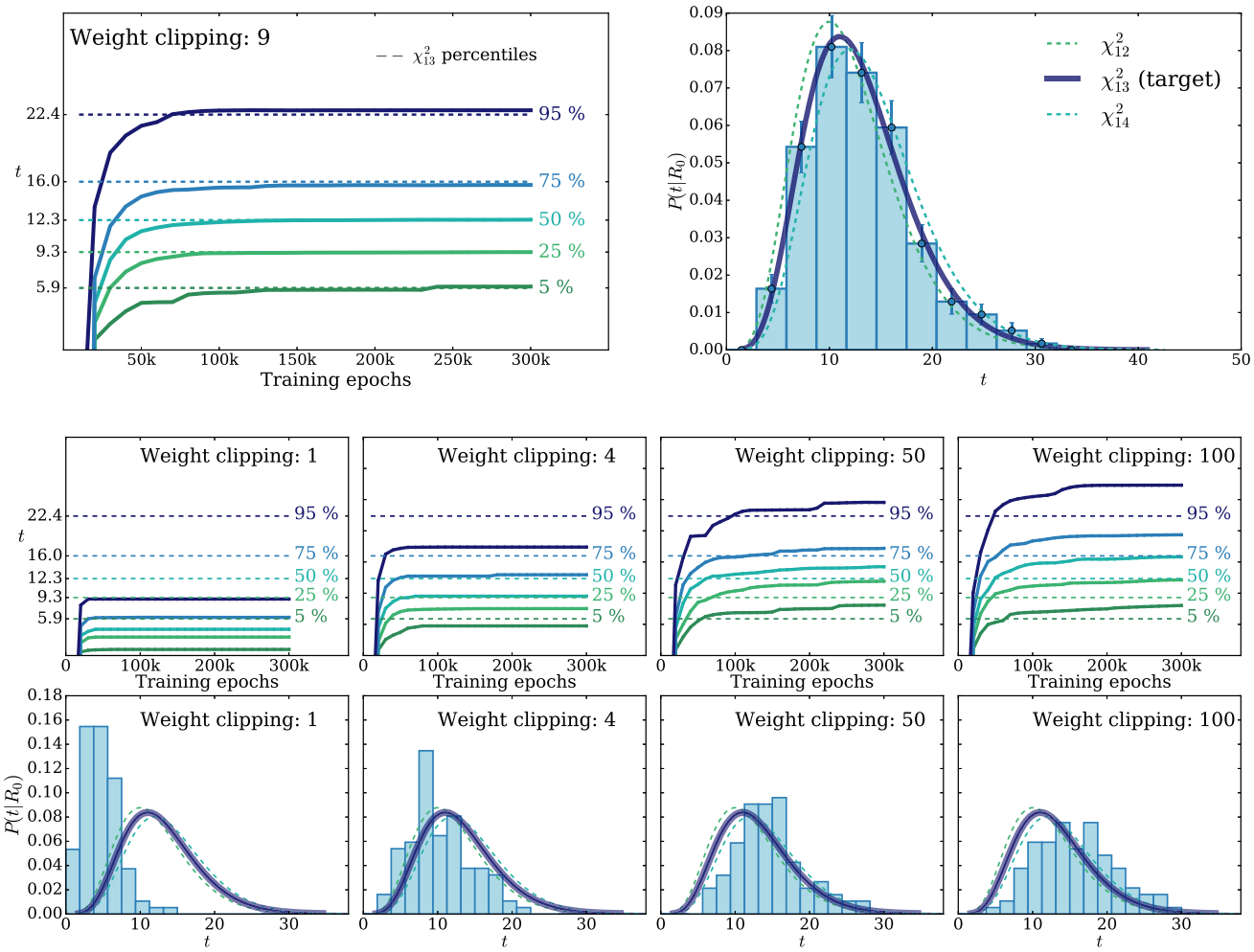


Fig. 1 Empirical distributions of $\bar{t}$ after 300 000 training epochs for different values of the weight clipping parameter, compared with the χ^2_{13} distribution expected in the asymptotic limit for the (1, 4, 1) network. The evolution during training of the $\bar{t}$ distribution percentiles, compared with the χ^2_{13} expectation, is also shown. Only 100 toy datasets are employed to produce the results shown in the figure, except for the ones for weight clipping equal to 9 where all the 400 toys are used

way after 300 000 training epochs, and some of its percentiles as a function of the number of epochs, are reported in Fig. 1.

We see that for large values of the weight clipping parameter the distribution sits slightly to the right of the target χ^2 with 13 degrees of freedom. Furthermore the training is not stable and significant changes in the $\bar{t}$ percentiles (especially the 95% one) occur even after 150 000 epochs. Very small values of the weight clipping make the distribution stable with training, but push it lower than the χ^2_{13} expectation. A good compatibility is instead obtained for intermediate values of the weight clipping parameter. We see that a weight clipping equal to 9 reproduces the χ^2_{13} formula quite accurately.

The strategy to find the value of the weight clipping parameter that best complies with the χ^2 compatibility criterion can be refined and optimized. We can start from one small and one large value of the weight clipping, for which we expect that the distribution of $\bar{t}$ will, respectively, undershoot and overshoot the χ^2 expectation, and compute $\bar{t}$ by running the training algorithm on a limited number n of toy datasets. The average of $\bar{t}$ over the n toys will be below (above) the mean of the target χ^2 distribution (i.e., 13, in the case at hand) for the small (large) value of the weight clipping. We thus obtain a window of values where the optimal weight clipping sits, which can be further narrowed by applying a standard root finding algorithm on the average $\bar{t}$ compared with the expected mean. Clearly the average $\bar{t}$ will be affected by a relatively large error if n is small. Therefore after a few iterations of the root finding algorithm, it will become compatible with the expected mean, preventing us from further restricting the weight clipping compatibility window.

Rather than looking at the compatibility of the average, a more powerful compatibility test should be employed at this stage in order to pick up the optimal weight clipping value

Table 1 Kolmogorov–Smirnov p-value and average $\bar{t}$ (minus the expected mean of 13) for the (1, 4, 1) network trained over samples of 40, 100 and 400 toy datasets, for different values of the weight clipping regularization parameter

Weight clipping	40 toys		100 toys		400 toys	
	KS p-value	$\langle \bar{t} \rangle - 13$	KS p-value	$\langle \bar{t} \rangle - 13$	KS p-value	$\langle \bar{t} \rangle - 13$
35	0.10	1.0 ± 0.7	$< 10^{-5}$	2.6 ± 0.6		
15	0.09	2.0 ± 1.5	0.01	1.5 ± 0.7		
11	0.36	1.0 ± 0.8	0.01	1.2 ± 0.5		
10	0.86	0.6 ± 0.9	0.56	0.9 ± 0.6	0.78	0.4 ± 0.3
9	0.68	0.5 ± 0.9	1.0	0.0 ± 0.5	0.93	0.0 ± 0.3
8	0.44	0.3 ± 0.7	0.53	-0.4 ± 0.4	0.42	-0.3 ± 0.2
7	0.40	-0.6 ± 0.8	0.21	-0.8 ± 0.4	0.02	-0.7 ± 0.2
4	0.11	-1.4 ± 0.7	$< 10^{-5}$	-2.7 ± 0.4		

inside the window. Furthermore this test should be sensitive to the entire shape of the distribution and not only to its central value. One can consider for instance a Kolmogorov–Smirnov (KS) test and maximize, in the window, the p-value for the compatibility with the target χ^2 of the empirical $\bar{t}$ distribution.[5]

It is advantageous to implement the strategy described above using a rather small number n of toy datasets, because training could become computationally demanding in realistic applications of our strategy. On the other hand, if n is small the KS compatibility test has limited power, leaving space for considerable departures from the target χ^2 of the true distribution of $\bar{t}$, even with the value of the weight clipping that has been selected as "optimal". A more accurate determination of the optimal weight clipping could however be obtaining by increasing n and repeating the previous optimization step. Clearly at this stage one could restrict to the much narrower window obtained at the end of the previous step, and benefit from the previous determination of the optimal weight clipping in order to speed up the convergence. The entire procedure could be further repeated with an even larger n, until a certain compatibility goal is achieved. For instance, one might require a KS p-value larger than some threshold, at the optimal weight clipping point, with a relatively large number n (say, 400) of toy experiments.

The results reported in Table 1 illustrate the weight clipping optimization strategy described above for the (1, 4, 1) network in the univariate problem under consideration. Actually a systematic optimization strategy is not needed to deal with the simple problem at hand, because training is sufficiently fast to test many points in the weight clipping parameter space with a large number of toys. Furthermore the departures from the χ^2 of the empirical $\bar{t}$ distribution are rather mild, as shown by Fig. 1, in a rather wide range of weight clipping values. We will instead need the optimization strategy in order to deal with the more realistic five-features problem of Sect. 4 where training is longer and the distribution is more sensitive to the weight clipping parameter.

Up to now we have considered a single architecture, and found one choice of the weight clipping parameter that ensures a good level of χ^2 compatibility. According to general principles of model selection, we should now switch to more complex architectures, with more neurons and/or hidden layers, aiming at selecting the most complex network that respects the asymptotic formula and that can be practically handled by the available computational resources. We saw in Ref. [2] that computational considerations play an important role in the selection, however the univariate problem at hand is not sufficiently demanding to illustrate this aspect. Indeed we have found χ^2-compatible networks with up to one hundred neurons, which are clearly an overkill for the univariate problem. Therefore, we will not describe the process of architecture optimization in the univariate example and postpone the discussion to a more realistic context in Sect. 4. The (1, 4, 1) network, with weight clipping equal to 9, will be employed in the rest of the present section.

3.2 Learning nuisances

We now turn to the problem of learning the effect of the nuisance parameters on the distribution of the variable x, following the methodology described in Sect. 2.3. In the simple univariate problem at hand, we have access to the distribution in closed form (Eq. (26)), and in turn to the exact analytic expression for the log distribution ratio

$$\log r(x; \boldsymbol{\nu}) = \log \frac{n(x|\mathrm{R}_{\boldsymbol{\nu}})}{n(x|\mathrm{R}_{\mathbf{0}})} = \nu_{\mathrm{N}} + x\,(1 - e^{-\nu_{\mathrm{S}}}) - \nu_{\mathrm{S}}. \quad (30)$$

The dependence on the normalization nuisance ν_{N} is trivial and it can be incorporated analytically, both in the univariate problem and in realistic analyses. The dependence on the scale nuisance ν_{S} is more complex, and not analytically

[5] Other compatibility tests can be adopted as well. For instance, one could minimize pdf-distance metrics such as the Kullback–Leibler divergence or the Earth Mover distance.

available in realistic problems. We thus approximate it by a Taylor series as in Eq. (14). Namely we define

$$\log \widehat{r}(x;\boldsymbol{\nu}) = \nu_{\rm N} + \nu_{\rm S}\,\widehat{\delta}_1(x) + \frac{1}{2}\nu_{\rm S}^2\,\widehat{\delta}_2(x) + \cdots . \tag{31}$$

Truncations of the $\nu_{\rm S}$ series at the first and at the second order will be considered in what follows.

We model each $\widehat{\delta}_a(x)$ coefficient function (with a ranging from 1 to the desired order of the series truncation in Eq. (31)) with fully-connected (1, 4, 1) neural networks with ReLU activation functions, trained with the loss function in Eq. (17). The training samples $\mathrm{S}_1(\boldsymbol{\nu}_i)$ and $\mathrm{S}_0(\boldsymbol{\nu}_i)$ contain 20 000 events each. The events in $\mathrm{S}_1(\boldsymbol{\nu}_i)$ are thrown according to the probability distribution of x in the $\mathrm{R}_{\mathbf{0}}$ hypothesis. The ones in $\mathrm{S}_0(\boldsymbol{\nu}_i)$ are thrown according to the $\mathrm{R}_{\boldsymbol{\nu}}$ hypothesis at selected points $\boldsymbol{\nu}_i = (0, \nu_{{\rm S},i})$ in the nuisance parameters space. The choice of the $\nu_{{\rm S},i}$ values used for training has a considerable impact on the quality of the reconstruction of the $\widehat{\delta}_a(x)$ functions. They should be such as to expose the dependence of the distribution ratio on each monomial of the expansion. For instance, when dealing with the quadratic approximation one would employ a relatively small value of $\nu_{\rm S}$, for which the linear term dominates, in order to learn $\widehat{\delta}_1(x)$, and a relatively large one for the reconstruction of $\widehat{\delta}_2(x)$. At least one additional value of $\nu_{\rm S}$ would be needed in order to go to the cubic order. This value would be taken even larger, namely in the regime where the quadratic approximation starts becoming insufficient and the dependence of the distribution ratio on the cubic term plays a role. Employing a redundant set of $\nu_{{\rm S},i}$'s (for instance, 4 points rather than 2 at the quadratic order) is beneficial. In general it is convenient to pick up the $\nu_{{\rm S},i}$'s in pairs of opposite sign, symmetric around the origin.

The set of $\nu_{{\rm S},i}$'s that duly captures all the terms in the Taylor expansion can be determined by inspecting the dependence on $\nu_{\rm S}$ of the distribution integrated in bins, and identifying the points on the $\nu_{\rm S}$ axis where a change of regime (say, from linear to quadratic) is observed. This is illustrated in Fig. 2, where we plot the dependence on $\nu_{\rm S}$ of $\log N_{\rm b}(\nu_{\rm S})/N_{\rm b}(0)$, with $N_{\rm b}$ the integral of the distribution in selected bins of the variable x. The points represent the true value of the log ratio as obtained from the distribution in Eq. (26). The dot-dashed, dashed and continuous lines are the fit to these points with polynomials of order 1, 2 and 4, respectively. More precisely the first-order polynomial fit only employs the points in the interval $\nu_{\rm S} \in [-0.1, 0.1]$, the second-order one employs the range $\nu_{\rm S} \in [-0.3, 0.3]$, while the fourth-order polynomial fit is performed on all the points. Compatibly with Eq. (30), we see that the behavior is almost exactly linear when x is very small. Considerable departures from linearity are instead present, for bigger x, when $\nu_{\rm S}$ is as large as 0.3 in absolute value. Based on these plots, for training the linear order we selected the set of values $\nu_{{\rm S},i} \in \{\pm 0.05, \pm 0.1\}$, for which the linear approximation is valid.[6] The set $\nu_{{\rm S},i} \in \{\pm 0.05,\ \pm 0.3\}$ was instead employed for the quadratic order approximation. The figure also suggests that the quadratic order truncation in Eq. (31) should be sufficient to model the dependence of $\log r(x;\boldsymbol{\nu})$ on $\nu_{\rm S}$ in the entire phase-space of x, at least if we limit ourselves to the range $\nu_{\rm S} \in [-0.6, 0.6]$.

The quality of the reconstruction of the log-ratio is displayed in Fig. 3 for the two different polynomial orders (linear and quadratic) that we have considered for the truncation of the series in Eq. (31). The exact analytic log-ratio in Eq. (30) is represented as dashed lines, to be compared with the reconstructed ratio reported as empty dots. The different colors correspond to different values of $\nu_{\rm S}$. As expected, the first-order truncation is accurate only if $\nu_{\rm S}$ is small. The accuracy improves with the quadratic truncation, for which the reconstructed log-ratio is essentially identical to the exact log-ratio. It should be kept in mind that, as explained in Sect. 2.3, the $\nu_{\rm S}$ range where an accurate reconstruction is needed depends on the allowed range of variability of $\nu_{\rm S}$, namely on its standard deviation $\sigma_{\rm S}$. From the figure we see that the linear polynomial modeling is adequate only if $\sigma_{\rm S}$ is below around 0.3, while with the quadratic one $\sigma_{\rm S}$ could be as large as 0.6.[7] The figure also reports the binned prediction for the log-ratio, as obtained from the quartic fit to $\log N_{\rm b}(\nu_{\rm S})/N_{\rm b}(0)$ previously described and displayed in Fig. 2. In realistic examples where the analytic log-ratio is not available, the binned prediction can be employed to monitor the quality of the reconstruction provided by the $\widehat{\delta}_a(x)$ networks. A more stringent test of the accuracy of the distribution log-ratio approximation, connected with the final validation of our strategy and its robustness to nuisances, will be discussed in Sect. 3.4.

3.3 Computing the test statistic

We finally have at our disposal all the ingredients to compute the test statistic $t(\mathcal{D}, \mathcal{A})$. This consists of the τ term, subtracted by the correction Δ. We now illustrate the evaluation of the two terms in turn, as implemented in the TensorFlow [25] package. The implementation is schematically represented in Fig. 4, and the corresponding code is available at [26].

[6] Obviously the linear approximation is also valid for even smaller $\nu_{\rm S}$. However, very small $\nu_{{\rm S},i}$'s reduces the impact of the nuisance parameters of the distribution of $\mathrm{S}_0(\boldsymbol{\nu}_i)$ relative to the one of $\mathrm{S}_1(\boldsymbol{\nu}_i)$, making training harder. The best option is to employ the largest $\nu_{{\rm S},i}$'s for which the linear approximation is still satisfactory.

[7] If $\sigma_{\rm S}$ was even larger, additional polynomial orders would be needed up to the point where the convergence of the Taylor series breaks down. After that, the only way to proceed would be to replace the Taylor series with a more rapidly convergent expansion of the log-ratio, if it exists, or to employ a parametric neural network [18].

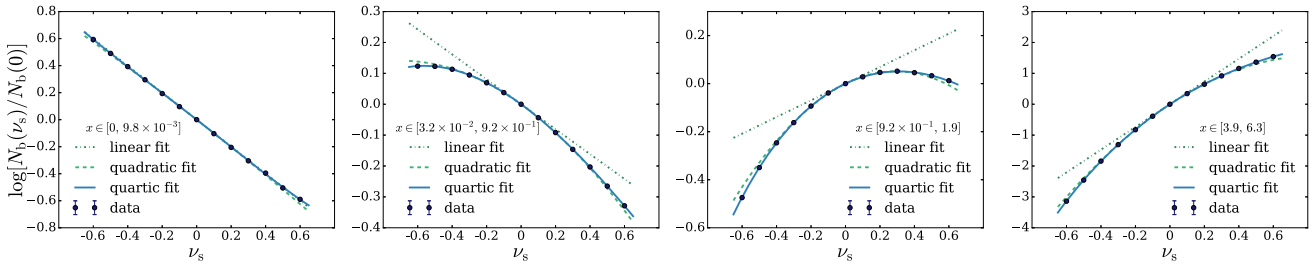


Fig. 2 The dependence on ν_S of $\log N_b(\nu_S)/N_b(0)$ in selected bins. The dots represent the true value of the log-ratio. The linear, quadratic and quartic fits are performed using a subset of the true values points as explained in the main text

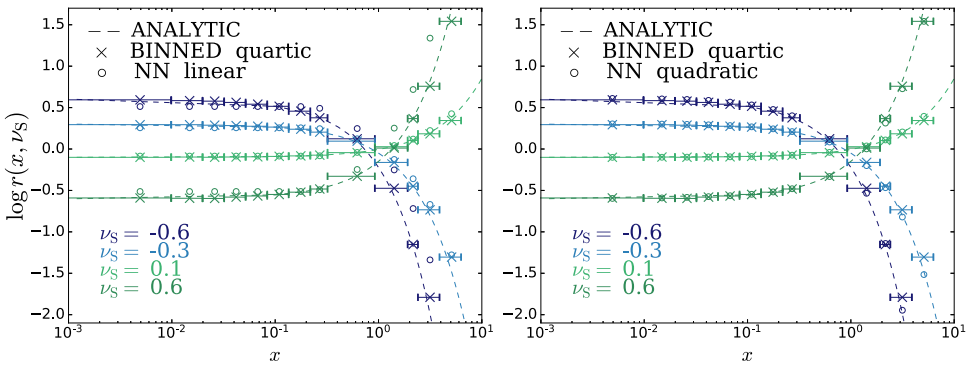


Fig. 3 The reconstructed distribution log-ratio (empty dots) for different values of ν_S, compared with the exact log-ratio and with the fourth-order binned approximation described in the main text. The two panels correspond to truncations of the series in Eq. (31) at linear and at the quadratic order

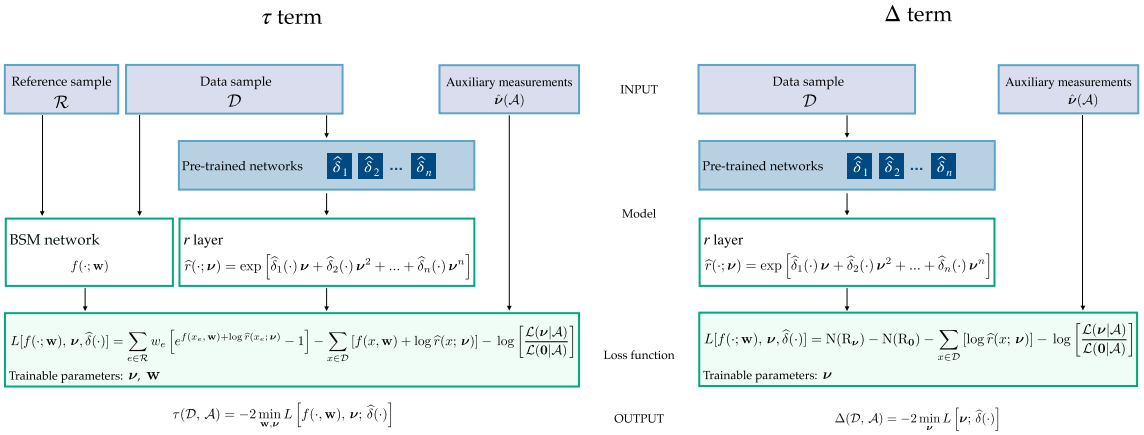


Fig. 4 Schematic representation of the TensorFlow implementation of our algorithm

As described in Sect. 2.4, computing τ requires the simultaneous optimization of the parameters $\mathbf{w}$ of the neural network model $f(x; \mathbf{w})$ (dubbed "BSM network" in the figure) and of the nuisance parameters $\boldsymbol{\nu}$. The loss function is the one of Eq. (25). It depends on $\boldsymbol{\nu}$ through the distribution ratio r, or more precisely through its estimate $\widehat{r}(x; \boldsymbol{\nu})$ as in Eq. (16). The estimated $\widehat{r}$ ratio is implemented as a TensorFlow "λ-layer" (denoted as "r layer" in the figure) that takes as input the output of the $\widehat{\delta}$ networks and builds the required polynomial function of $\boldsymbol{\nu}$. Notice that the parameters of the $\widehat{\delta}$ networks are "fixed" parameters during training, namely they are not optimized. Indeed, the $\widehat{\delta}$ networks have been trained at a previous stage of the implementation, as described in Sect. 3.2. The evaluation of τ thus proceeds as shown in the left panel of Fig. 4. The inputs are the Reference sample, the (observed or toy) Data and the central value of the auxiliary likelihood $\widehat{\boldsymbol{\nu}}(\mathcal{A})$. Notice that $\widehat{\boldsymbol{\nu}}(\mathcal{A}) = \mathbf{0}$ by construction in the true experiment, but it fluctuates in the toy experiments as discussed at the beginning of the present section. As in the figure, the Reference data feed only the BSM network, while the Data feed both the BSM and the r-layer, after passing through the pre-trained $\widehat{\delta}$ networks. The loss function takes as input

the BSM network, the r-layer output and $\widehat{\boldsymbol{\nu}}(\mathcal{A})$, which enters in the auxiliary term of the Likelihood. The only trainable parameters are the ones of the BSM network and of the r-layer, namely $\mathbf{w}$ and $\boldsymbol{\nu}$. The loss at the end of training, times -2, produces the τ term.

The evaluation of the Δ term, depicted on the right panel of Fig. 4, follows the strategy described in Sect. 2.3. It has been implemented in TensorFlow employing the same building blocks used for the evaluation of τ, apart from the BSM network that does not participate in the evaluation of Δ. The Reference dataset is similarly not employed at this step. The loss function is merely given by minus the argument of the maximum in Eq. (18), so that Δ is the minimal loss at the end of training, times -2. For the evaluation of Δ, the parameters $\boldsymbol{\nu}$ of the r-layer are the only ones to be optimized by the training algorithm.

The TensorFlow modules described above are also employed for the preliminary steps of the algorithm described in Sects. 3.1 and 3.2. In the latter, the $\widehat{\delta}$ networks are trained using the loss function in Eq. (29) and the relevant datasets. In the former step, namely the selection of the BSM network hyper-parameters, the r-layer and the $\widehat{\delta}$ networks are not employed and the loss function is replaced with the one, in Eq. (29), where the effect of nuisance parameters is not taken into account.

3.4 Validation

As previously emphasized, it is vital for the applicability of our strategy that the distribution $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ of the test statistic is nearly independent of $\boldsymbol{\nu}$. This is ensured in line of principle by the asymptotic formulae described in Sect. 2.5. Verifying in practice the validity of the asymptotic formulae is thus the crucial validation step, which we will perform by computing the empirical $P(t|\mathrm{R}_{\boldsymbol{\nu}})$ distribution on toy experiments. Toy datasets are generated according to the $\mathrm{R}_{\boldsymbol{\nu}}$ hypothesis, at different points $\boldsymbol{\nu} = \boldsymbol{\nu}^{\boldsymbol{*}} = (\nu_{\mathrm{N}}^*, \nu_{\mathrm{S}}^*)$ of the nuisances' parameter space. Each toy dataset $\mathcal{D}$ is accompanied by one instance of the nuisance parameters estimators $\widehat{\boldsymbol{\nu}} = (\widehat{\nu}_{\mathrm{N}}, \widehat{\nu}_{\mathrm{S}})$. As explained at the beginning of the present section, the estimators are thrown as Gaussians with standard deviations $\sigma_{\mathrm{N,S}}$ centered at $\nu_{\mathrm{N,S}}^*$. They appear in the auxiliary likelihood log-ratio as in Eq. (27).

We start by setting $\sigma_{\mathrm{N}} = \sigma_{\mathrm{S}} = 0.15$, and from central-value nuisance parameters $(\nu_{\mathrm{N}}^*, \nu_{\mathrm{S}}^*) = (0, 0)$, obtaining the results on the left panel of Fig. 5. The plot shows the empirical τ distribution in green and, in blue, the distribution of $t = \tau - \Delta$. In spite of the fact that the toys are generated according to the central-value Reference hypothesis, which is the same hypothesis under which we enforced compatibility with the χ^2_{13} by choosing the weight clipping parameter in Sect. 3.1 (see Fig. 1), the distribution of τ is slightly different from the χ^2_{13}. This is not surprising because the χ^2-compatibility was enforced on the variable $\bar{t}$ (28), which does not account for the presence of nuisances and is different from τ. The distribution of τ is instead quite close to the χ^2 with a number of degrees of freedom equal to 15, which is the number of parameters of the neural network plus the number of nuisance parameters. This is compatible with the asymptotic expectation as discussed in Sect. 2.5. Again compatibly with the asymptotic formulae, we see in the figure that the distribution of $t = \tau - \Delta$ is instead a χ^2 with 13 degrees of freedom.

The left panel of Fig. 5 provides a first confirmation of the validity of the asymptotic formula for $P(t|\mathrm{R}_{\boldsymbol{\nu}})$, thought not a particularly striking one because the τ distribution is not vastly different from the one of t, meaning that the correction term Δ does not play an extremely significant role in this case. A more interesting result is obtained when setting ν_{N}^* or ν_{S}^* one σ away from the central value, as shown in the four plots in the right panel of the figure. In this case, as expected from the asymptotic formulae, the τ distribution is radically different from the one of t. It is expected to follow a non-central χ^2 with a non-centrality parameter that is controlled by the departure of the true values of the nuisances from the central values. The correction term Δ has a big impact on the distribution of t, bringing it back to the expected χ^2_{13}. The effect is due to a strong correlation between the τ and Δ distribution over the toys, which engineers a cancellation in $t = \tau - \Delta$.

A more quantitative and systematic validation of the compatibility of t with the χ^2_{13} can be obtained by computing the Kolmogorov–Smirnov test p-value as in Sect. 3.1. The results are reported in Table 2. The "w/o correction" columns report the p-value obtained by comparing the distribution of τ (i.e., without the Δ correction term) with the χ^2_{13}. The "w/ correction" columns report the p-value for the distribution of t, including the correction. The table contains the results obtained for $\sigma_{\mathrm{N,S}} = 0.15$, as well as those for lower values of the nuisances' standard deviations $\sigma_{\mathrm{N,S}} = 0.10,\ 0.05$.

The above results establish the validity of the asymptotic formulae when the standard deviation of the nuisance parameters is of order 15% or less. Notice that it is increasingly simple to deal with smaller standard deviations (i.e., with more precisely measured nuisances), merely because when $\boldsymbol{\nu}$ is small the ratio $\widehat{r}(x; \boldsymbol{\nu})$ approaches 1 becoming independent of $\boldsymbol{\nu}$, regardless of the accuracy with which it is reconstructed by the $\widehat{\delta}_a(x)$ networks. Consequently the maximization over $\mathbf{w}$ in τ (24) tends to decouple from the maximization over $\boldsymbol{\nu}$ and the cancellation between τ and Δ in the determination of t is guaranteed. On the contrary, larger standard deviations are more difficult to handle. Indeed, as explained in Sect. 2.5, larger values of $\boldsymbol{\nu}$ push the τ distribution away from the target χ^2, forcing the correction term to engineer an increasingly delicate cancellation. This enhances the impact of all the imperfections that are present in the implementation of

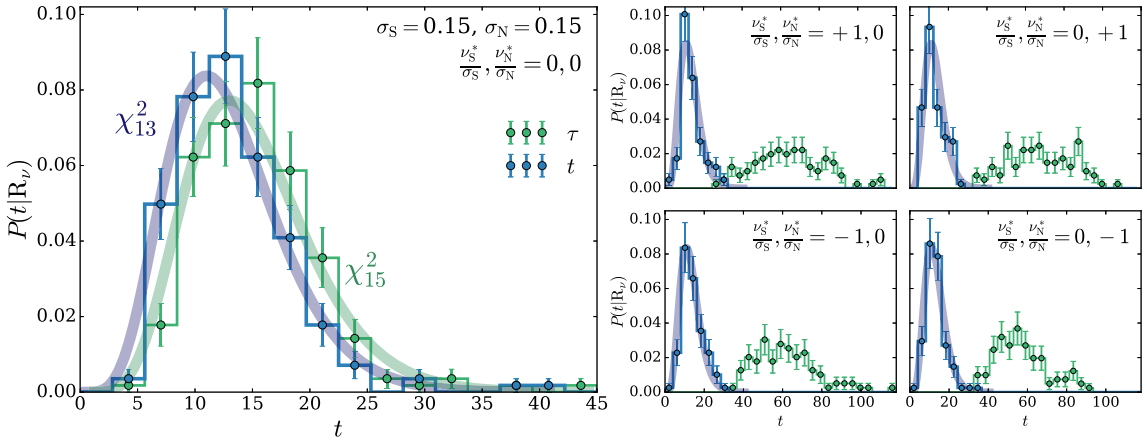


Fig. 5 The empirical distribution of τ (in green) and of t (in blue) computed by 100 toy experiments performed in the R_ν hypothesis at different points in the nuisances' parameters space. The χ^2_{13} distribution is reported in blue in all the plots. The χ^2_{15} distribution is shown in green on the left plot

Table 2 Kolmogorov–Smirnov p-value for the compatibility of the τ ("w/o correction" columns) and of the t ("w/ correction" columns) distributions with the χ^2_{13}. The KS test is based 100 toy experiments performed in the R_ν hypothesis at different points in the nuisance parameters space

$\left(\frac{\nu_S}{\sigma_S}, \frac{\nu_N}{\sigma_N}\right)$	$\sigma_S = 5\%, \sigma_N = 5\%$		$\sigma_S = 10\%, \sigma_N = 10\%$		$\sigma_S = 15\%, \sigma_N = 15\%$	
	KS p-value		KS p-value		KS p-value	
	w/o correction	w/ correction	w/o correction	w/ correction	w/o correction	w/ correction
$(0, 0)$	$< 10^{-5}$	0.33	$< 10^{-5}$	0.49	$< 10^{-5}$	0.08
$(+1, 0)$	$< 10^{-5}$	0.41	$< 10^{-5}$	0.55	$< 10^{-5}$	0.72
$(0, +1)$	$< 10^{-5}$	0.80	$< 10^{-5}$	0.86	$< 10^{-5}$	0.45
$(-1, 0)$	$< 10^{-5}$	0.33	$< 10^{-5}$	0.88	$< 10^{-5}$	0.37
$(0, -1)$	$< 10^{-5}$	0.47	$< 10^{-5}$	0.82	$< 10^{-5}$	0.36

the algorithm, and in particular of the ones related with the quality of the reconstruction of $\widehat{r}$ that is achieved by the $\widehat{\delta}_a(x)$ networks. The results presented up to now (namely, Fig. 5 and Table 2) are obtained by employing the linear-order reconstruction for $\log \widehat{r}$. The good observed level of compatibility with the asymptotic formula thus shows that the linear-order reconstruction is sufficiently accurate in order to deal with $\sigma_{N,S} \leq 15\%$. However the accuracy is expected to become insufficient for larger $\sigma_{N,S}$, owing to the considerable departures of the exact $\log r$ from linearity described in Sect. 3.2.

We illustrate this aspect by computing the empirical t distribution for $\sigma_{N,S} = 0.6$ and setting $(\nu_N^*, \nu_S^*) = (0, -0.6)$.[8] The result reported in the left panel of Fig. 6 employ the linear-order approximation of $\log \widehat{r}$. The ones in the middle panel are obtained with the quadratic order approximation while the exact $\log r$ (30) is employed in the right panel. The figure shows that the linear-order approximation is insufficient, while a good compatibility with the target χ^2_{13} is found with the quadratic approximation and with the exact log-ratio.

A similar test performed with $(\nu_N^*, \nu_S^*) = (0, +0.6)$ produced however a non-satisfactory level of compatibility as shown on the left panel of Fig. 7. The reason is that for positive and relatively large $\nu_S^* = +0.6$, the scale factor $e^{\nu_S} \simeq 1.8$ is considerably larger than one and pushes the Reference Model distribution (26) towards large x. Therefore, toy data generated with positive and large ν_S^* can often display instances of x that fall in a region that is not populated by the Reference sample. The "new physics" network f identifies these instances as highly anomalous, since they do not have any counterpart in the Reference sample, producing outliers in the τ distribution and in turn in the one of t. An illustration of this behavior is displayed on the right panel of the figure. For the toy experiment under consideration, the large observed $t = 217$ is due to the data points at $x \gtrsim 13$, which falls well above the largest instance of x ($\simeq 11$) that is present in the Reference sample. Such problematic outliers with no counterpart in the Reference sample can not occur if ν_S^* is sufficiently small, such that the $n(x|R_{\nu^*})$ distribution is similar to the central-vale $n(x|R_0)$ distribu-

[8] We set $\nu_N^* = 0$ for this study because non-vanishing values of ν_N^* are easy to deal with, since the dependence of r on the normalization nuisance is known exactly.

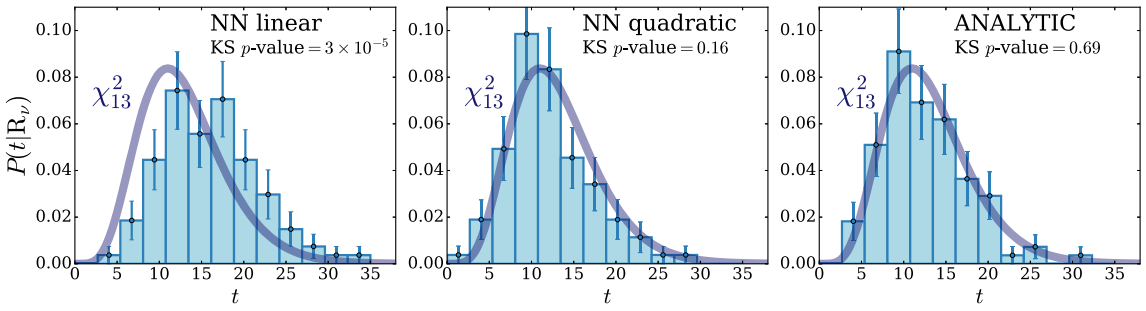


Fig. 6 The empirical distribution of t computed with 100 toy experiments for $(\nu_{\mathrm{N}}^*, \nu_{\mathrm{S}}^*) = (0, -0.6)$. Increasingly accurate modelings of $\log \widehat{r}(x; \boldsymbol{\nu})$ are employed in the three panels, namely the linear- and quadratic-order approximations and the analytic log-ratio in Eq. (30)

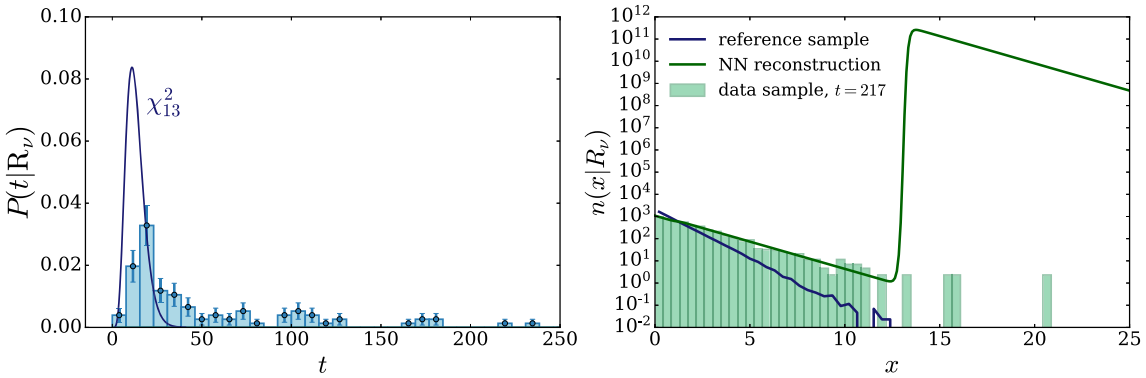


Fig. 7 Left panel: the empirical distribution of t computed with 100 toy experiments for $(\nu_{\mathrm{N}}^*, \nu_{\mathrm{S}}^*) = (0, 0.6)$. Right panel: neural network reconstruction of the x variable distribution (using Eqs. (4) and (26)) of a single toy experiment for which the test statistic output is an outlier ($t \simeq 217$)

tion according to which the Reference sample is generated, because the Reference sample is more abundant (100 times, in the case at hand) than the data. But they can occur if, as for $|\nu_{\mathrm{S}}^*| = 0.6$, the nuisance parameters are so large that they modify the central-value distribution at order one and, as for $\nu_{\mathrm{S}}^* = +0.6$, they push it towards phase-space regions that are particularly rare in the central-value hypothesis. This potential issue should be kept in mind when dealing with nuisance parameters that are poorly constrained by the auxiliary measurements. Similar problems occur in traditional analyses, whenever the reference control sample statistics is insufficient. A typical mitigation of this effect is obtained binning the dataset with larger binwidths on distribution tails. For our method, which in its generic formulation does not make use of bins, possible solutions are either to restrict the variables to a region that is well-populated by the available Reference sample, or to produce a Reference sample that populates the tail of the features distribution more effectively. Further discussion on this point is postponed to Sect. 4.2, where we will see the same issue emerging again in a more realistic context.

3.5 Sensitivity to new physics

We conclude the discussion of the univariate example by testing its sensitivity to putative new physics effects. We consider three New Physics (NP) scenarios that foresee, respectively, the presence of a resonant bump in the tail of the x distribution, a non-resonant enhancement and a resonant peak in the bulk of the distribution. Following Ref. [1], we consider

NP$_1$: a peak in the tail of the exponential Reference distribution, modeled by a Gaussian

$$n(x|\mathrm{NP}_{1;\boldsymbol{\nu}}) = n(x|\mathrm{R}_{\boldsymbol{\nu}}) + \mathrm{N}_1 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x}_1)^2}{2\sigma^2}}, \tag{32}$$

with $\bar{x}_1 = 6.4$, $\sigma = 0.16$ and $\mathrm{N}_1 = 10$.

NP$_2$: a non resonant effect in the tail of the Reference distribution

$$n(x|\mathrm{NP}_{2;\boldsymbol{\nu}}) = n(x|\mathrm{R}_{\boldsymbol{\nu}}) + \mathrm{N}_2 \frac{x^2}{2} e^{-x}, \tag{33}$$

with $\mathrm{N}_2 = 180$.

NP$_3$: a peak in the bulk, again modeled by a Gaussian shape

$$n(x|\mathrm{NP}_{3;\boldsymbol{\nu}}) = n(x|\mathrm{R}_{\boldsymbol{\nu}}) + \mathrm{N}_3 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x}_3)^2}{2\sigma^2}}, \tag{34}$$

with $\bar{x}_3 = 1.6$, $\sigma = 0.16$ and $\mathrm{N}_3 = 90$.

All our putative new physics scenarios give a positive contribution to the Reference distribution. As such, they can be interpreted as an additional "signal" component in the distribution of the data, on top of the "background" Reference distribution. This is obviously not necessary for our method, which can equally well be sensitive to new physics effects that interfere quantum-mechanically with the Reference Model producing a non-additive contribution. Also notice that we decided not to include nuisance parameters in the new physics term, which is thus assumed to be perfectly known. Also this assumption is not crucial for the sensitivity since a modeling of the signal is not required in our method. Nuisance parameters related to the signal come at play whenever one wants to interpret the outcome of the method as a bound on the theoretical parameters of a specific scenario.

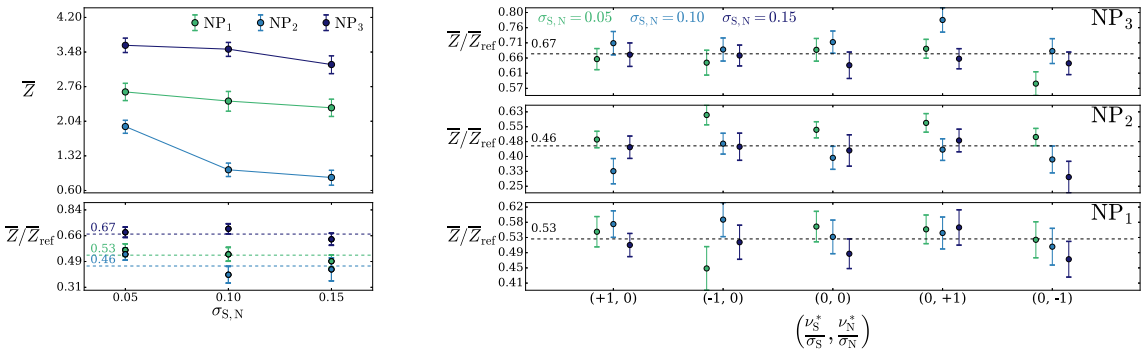
We quantify the potential of our strategy to detect departures from the Reference Model, if one of the three $\mathrm{NP}_{1,2,3}$ models is present in the data, in terms of the median Z-score $\overline{Z}$ obtained by running our algorithm on toy datasets generated according to the $n(x|\mathrm{NP})$ distribution. For each NP-hypothesis toy we repeat the exact same operations we described in Sect. 3.3 to obtain the test statistic t, in the exact same configuration (architecture, weight clipping, etc.) we used in Sect. 3.4 for validation on the Reference-hypothesis toy datasets. The linear-order reconstruction of $\log \widehat{r}$ is employed for the modeling of the nuisance parameters effect. We saw in Sect. 3.4 that this modeling is sufficiently accurate if we limit our analysis to the regime $\sigma_{\mathrm{N,S}} \leq 15\%$. The value of t on each NP toy is compared with the χ^2_{13} distribution and converted to a p-value by exploiting the asymptotic formulae we verified Sect. 3.4. For each $\mathrm{NP}_{1,2,3}$ new physics scenario, the median p-value is computed using 100 NP toy datasets, obtaining $\overline{Z} = \Phi^{-1}(1-p)$, with Φ the cumulative of the Standard Gaussian. The results are reported in Fig. 8 under multiple assumptions ($\sigma_{\mathrm{N,S}} = 5,\ 10,\ 15\%$) for the nuisance parameters standard deviations and for different choices ($\nu^*_{\mathrm{N,S}} = 0,\ \pm\sigma_{\mathrm{N,S}}$) of the true values of the nuisance parameters that underly (through the $\mathrm{R}_{\boldsymbol{\nu}}$ component of $n(x|\mathrm{NP})$) the generation of the NP toys.

The figure also reports a "reference" median Z-score $\overline{Z}_{\mathrm{ref}}$, that quantifies the sensitivity of a model-dependent data analysis strategy targeted and optimized for the detection of each individual NP hypothesis. A model-dependent search is necessarily more powerful than a model-independent one for the detection of the NP signal it is designed for. Correspondingly, $\overline{Z}_{\mathrm{ref}}$ must be significantly larger than $\overline{Z}$ by consistency and the two quantities should not be compared directly. As in Refs. [1,2], we use $\overline{Z}_{\mathrm{ref}}$ to quantify how "difficult" or "easy" the $\mathrm{NP}_{1,2,3}$ signals are to detect in absolute terms, and we report the ratio $\overline{Z}/\overline{Z}_{\mathrm{ref}} < 1$ as a measure of the degradation in sensitivity of our model-independent strategy relative to dedicated searches.

As a "reference" model-dependent search strategy we consider a hypothesis test based on the profile likelihood ratio, and more precisely on the test statistic "q_0" for the discovery of positive signals defined in Ref. [24]. Namely, we extend the NP hypothesis by a "signal strength" parameter $\mu \geq 0$ that rescales $\mathrm{N}_i \to \mu\,\mathrm{N}_i$ (for $i = 1, 2, 3$) in Eqs. (32)–(34). Denoting as $\widehat{\mu}$ the value of the signal strength parameter that maximizes the likelihood of the NP hypothesis, and $\widehat{\boldsymbol{\nu}}$ the maximum in the nuisances' space, we define

$$q_0 = -2\,\log \frac{\max_{\boldsymbol{\nu}} \mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D},\mathcal{A})}{\mathcal{L}(\mathrm{NP}_{i;\widehat{\boldsymbol{\nu}};\widehat{\mu}}|\mathcal{D},\mathcal{A})}, \tag{35}$$

if $\widehat{\mu} > 0$, and we set $q_0 = 0$ otherwise. In the equation, $\mathcal{L}$ denotes the extended likelihood constructed as in



(a) **(b)**

Fig. 8 The median Z-score ($\overline{Z}$) obtained with our model-independent strategy, compared to the median reference Z-score ($\overline{Z}_{\mathrm{ref}}$) of a model-dependent search (see the main text) optimized for each of the three new physics scenarios in Eqs. (32)–(34). The left panel **a** shows the dependence of $\overline{Z}$ on the nuisance parameter uncertainties, and the mild dependence of the ratio $\overline{Z}/\overline{Z}_{\mathrm{ref}}$. The left panel is obtained with NP toys generated with central-value nuisance parameters $\nu^*_{\mathrm{N,S}} = 0$. The right panel **b** displays $\overline{Z}/\overline{Z}_{\mathrm{ref}}$ under multiple assumptions are for the nuisance parameters uncertainties ($\sigma_{\mathrm{N,S}} = 5,\ 10,\ 15\%$) and for the true values ($\nu^*_{\mathrm{N,S}} = 0,\ \pm\sigma_{\mathrm{N,S}}$) of the nuisance parameters. The error bars quantify the statistical uncertainties (on 100 toys) in the determination of the median

Sect. 2, exploiting the analytic knowledge of the new physics distributions provided by Eqs. (32)–(34). The "numerator" hypothesis $\mathrm{R}_{\boldsymbol{\nu}}$ coincides by construction with the NP hypothesis at $\mu = 0$. The distribution of q_0 under the Reference (numerator) hypothesis is known in the asymptotic limit. We can thus associate a p-value to the value of q_0 that is obtained on each NP toy data set. The median p-value over the toys provide the median Z-score [24]

$$\overline{Z}_{\mathrm{ref}} = \mathrm{median}\left[\sqrt{q_0}\right]. \tag{36}$$

The physical interpretation of the results on the left panel of Fig. 8 is quite straightforward. The sensitivity to the resonant new physics scenarios $\mathrm{NP}_{1,3}$ is not affected by the presence of nuisances, because the nuisance parameters we are considering can not produce deformations of the Reference distribution that mimic a resonant peak. On the contrary, the scale nuisance parameter can mimic non-resonant new physics and indeed the sensitivity to NP_2 considerably deteriorates as $\sigma_{\mathrm{N,S}}$ increases. The same behavior is observed for the model-dependent $\overline{Z}_{\mathrm{ref}}$, as well as for the sensitivity $\bar{Z}$ of our model-independent strategy. Indeed, we see that the $\overline{Z}/\overline{Z}_{\mathrm{ref}}$ ratio is quite stable under the variation of $\sigma_{\mathrm{N,S}}$. This confirms the existence of a direct correlation, as in previous studies [1,2], between the sensitivity of our model-dependent strategy and the "absolute degree of detectability" of the new physics scenario, as quantified by the sensitivity of a model-dependent search. A further confirmation of this correlation is provided by the right panel of the figure.

Before concluding this section, it is interesting to consider a fourth scenario for new physics, which does not manifest itself in the variable of interest "x", but rather in the auxiliary measurements that constrain the nuisance parameters. As discussed in Sect. 2.6, our strategy is not necessarily blind to this type of effects. Consider a situation where the estimator for the scale nuisance parameter, $\widehat{\nu}_{\mathrm{S}}(\mathcal{A})$, is biased due to new physics by an amount $\Delta\nu_{\mathrm{S}} = 5\,\sigma_{\mathrm{S}}$. Since we do not know about this bias, our auxiliary likelihood remains the one in Eq. (27), but $\widehat{\nu}_{\mathrm{S}}(\mathcal{A})$ in reality is not distributed around the true ν_{S}^*, but around $\nu_{\mathrm{S}}^* + \Delta\nu_{\mathrm{S}}$. In order to generate toy experiments that describe this scenario, one has to take $\nu_{\mathrm{S}}^* + \Delta\nu_{\mathrm{S}}$ as the central-value for the generation of the toy $\widehat{\nu}_{\mathrm{S}}$ values while using the true ν_{S}^* for the generation of the x toy datasets. The mismatch, on average, between $\widehat{\nu}_{\mathrm{S}}$ and the value of ν_{S} that truly underlies the x variable distribution can lead to the detection of new physics as explained in Sect. 2.6. For $\sigma_{\mathrm{S}} = 15\%$ we find sensitivities

$(\frac{\nu_{\mathrm{S}}^*}{\sigma_{\mathrm{S}}}, \frac{\nu_{\mathrm{N}}^*}{\sigma_{\mathrm{N}}})$	$(0,0)$	$(+1,0)$	$(0,+1)$	$(-1,0)$	$(0,-1)$
$\overline{Z}$	$2.87^{+0.16}_{-0.15}$	$3.53^{+0.12}_{-0.11}$	$3.04^{+0.14}_{-0.14}$	$3.22^{+0.14}_{-0.14}$	$3.31^{+0.14}_{-0.14}$

4 Two-body final state

In the previous section we described the practical implementation of our strategy and its validation in a very simple univariate toy problem. We now turn to a slightly more complex setup, which is inspired by the realistic problem of model-independent new physics searches in two-body final states at the LHC (see Ref. [2]). While not yet a complete LHC analysis, the setup that we study in the present section is at a similar scale of complexity, and it poses novel challenges with respect to the univariate problem. We will show how to deal with them, aiming at providing the reader with useful indications on how to handle the various technical aspects that might show up in realistic physics analysis contexts.

A two-body final state can be characterized in terms of the five kinematical features $p_{T,1(2)}$, $\eta_{1(2)}$ and $\Delta\phi_{12} = \phi_1 - \phi_2$, with p_T, η and ϕ the transverse momentum, the pseudorapidity and the azimuthal angle of the individual particles.[9] The particles are p_T-ordered, namely $p_{T,1} > p_{T,2}$. Data are supposed to be selected by requiring the two particles to have same flavor and opposite sign, but this information is not retained at this stage. We do not specify sharply the nature of the final state objects. In the typical cases we have in mind, these are either muons, electrons or τ leptons reconstructed by the detector. On the other hand, the same construction could be applied to objects with similar resolution, e.g., trading electrons for photons or taus for jets. The kinematical distributions would be quite different in the different cases, however we do not expect these differences to impact the technical viability of our strategy, which we aim at demonstrating. The total cross-section of the process would be also different. However we can compensate this adjusting the assumed integrated luminosity of the dataset, making the total number of expected events $\mathrm{N}(\mathrm{R}_{\mathbf{0}})$ roughly equal in the various cases. Therefore, for our purpose the only relevant difference between muons, electrons and τ final states resides in the increasingly large systematic uncertainties that affect the corresponding SM predictions. Since larger uncertainties are more difficult to handle, as outlined in the previous section, it is instructive to investigate these three scenarios.

Owing to the previous discussion, we ignore the difference in the distributions of the different final states and we model all of them as opposite-sign muons. Namely, the central-value Reference distribution $n(x|\mathrm{R}_{\mathbf{0}})$ is the same in all cases, and it corresponds to the SM simulation of $pp \to \mu^+\mu^- + X$ at the 13 TeV LHC obtained with `MadGraph5` [27] at LO, with extra jets matching and using `Pythia6` [28] and `Delphes3` [29] for parton showering and detector simula-

[9] Two additional variables such as the total transverse momentum and the pseudorapidity of the two-particles system could be included in order to enhance the sensitivity of the analysis to the production of the two particles in association with hard objects.

tion. The data samples we employ for the analysis are the ones described in Ref. [2] and can be downloaded from Zenodo [30]. We consider two Gaussian nuisance parameters ν_N and ν_S describing, as in the previous section, the uncertainty on the event yield normalization and on the scale factor in the measurement of the transverse momenta. We adopt a simple modeling of the normalization uncertainties by a global (phase-space-independent) factor with standard deviation $\sigma_N = 2.5\%$, corresponding to the uncertainty of the luminosity measurement. Since the normalization nuisance parameter can be incorporated analytically in the likelihood, as we have discussed, it is essentially trivial to deal with it.

The scale factor, on the other hand, affects the input variable distributions in a non-trivial manner. Furthermore, the uncertainty in its determination widely depends on the nature of the particle. We consider three representative scenarios, having in mind the specific case of CMS[10]:

- muon-like: for the CMS experiment, the uncertainty on the muon momentum scale is very small due to the combined information of the inner tracker and the dedicated muon detectors. Based on Ref. [31], we set the uncertainty to a typical value $\sigma_S^{(b)} = 5\times10^{-4}$ for central muons with $|\eta| < 2.1$ (barrel region) and $\sigma_S^{(e)} = 15 \times 10^{-4}$ for $|\eta| \geq 2.1$ (endcaps region). Here and in the following cases we ignore the dependence of the uncertainty on the particle transverse momentum for simplicity, but a generalization in this direction is straightforward.
- electron-like: the momentum reconstruction for electrons is instead based on the combination of the inner tracker information and the energy deposit in the electromagnetic calorimeter. The LHC pileup makes the trajectory reconstruction harder while for the energy reconstruction from the calorimeter information one has to consider the energy loss through bremsstrahlung in the detector material before the calorimeter is reached. The resulting uncertainty is then typically [32] an order of magnitude worse than the one affecting the muons. We here consider an error of $\sigma_S^{(b)} = 3 \times 10^{-3}$ and $\sigma_S^{(e)} = 9 \times 10^{-3}$.
- τ-like: tau leptons decay in the CMS detector and their 4-momenta has to be reconstructed starting from the decay products; the information of all sub-detectors is combined to reconstruct all the particles produced in the collision events in the so called ParticleFlow algorithm [33]. For hadronically decaying taus the energy scale uncertainty was found to be always better than 3%; here we simply assume an error on the τ-lepton momentum reconstruction of 3×10^{-2} for both the barrel and the endcaps regions, independently of the magnitude of the momentum [34].

[10] Our assumptions loosely apply also to the case of the ATLAS detector.

In all cases, we treat the effects on the barrel and endcaps regions as fully correlated and we employ a single nuisance parameter ν_S to describe both. Specifically, ν_S is the scale uncertainty in the barrel, with standard deviation $\sigma_S \equiv \sigma_S^{(b)}$.

The Monte Carlo samples for non-central values ($\nu_S \neq 0$) of the scaling nuisance parameters, needed for the implementation and the validation of our strategy, are obtained by reprocessing the di-muon dataset with the transformation $p_{T,1(2)}^{(b,e)} \to \exp\left(\nu_S \sigma_S^{(b,e)}/\sigma_S^{(b)}\right) p_{T,1(2)}^{(b,e)}$, which acts differently on the barrel and endcaps regions. After the transverse momenta rescaling, we apply acceptance cuts $p_{T,2(1)} > 20$ GeV, as well as a lower threshold on the di-body invariant mass of $M_{12} > 100$ GeV, in order to exclude the resonant peak associated with the Z boson production. Indeed, if included the Z peak would dominate the composition of the data sample by several orders of magnitude, and our analysis would effectively turn into a search for new physics at the Z-pole. We thus exclude the Z peak for a more broad exploration of the two-body phase space. The invariant mass cut will have to be raised to 120 GeV in the τ-like scenario. As we will discuss, this is because Z-pole events contamination of the signal-region enhances the effect of scale uncertainties to a non-manageable level at low invariant mass. A similar analysis could also be repeated below the Z mass, as done by the CMS and the LHCb experiments, exploiting real-time analysis techniques [35,36]. We do not discuss this case here.

In what follows we describe the implementation of our model-independent search strategy on a dataset whose integrated luminosity corresponds to $N(R_{\mathbf{0}}) = 8\,700$ expected events in the signal region defined by the acceptance and the 100 GeV invariant mass cut. In the case of opposite-sign muons, this number of events would correspond to an integrated luminosity of around 0.35 fb^{-1}. The expected event yield in the non-central Reference hypothesis, $N(R_{\boldsymbol{\nu}})$, is computed with the same integrated luminosity, duly taking into account the normalization nuisance factor e^{ν_N}, and the effect of the scale nuisance ν_S on the selection cuts efficiency. A higher integrated luminosity, of 1.1 fb^{-1}, is considered in the τ-like scenario in order to maintain $N(R_{\mathbf{0}})$ as large as (specifically, $N(R_{\mathbf{0}}) = 8\,400$) in the other scenarios compensating for the higher invariant mass cut.

Finally, in all scenarios we apply an upper cut $p_{T,1(2)} < 1$ TeV. The phase space region excluded by this cut is populated, for the luminosity we are considering, with a probability as low as 10^{-5} in the Reference model. Therefore it has essentially no impact on the analysis and on its sensitivity to new physics, also in light of the fact that the mere observation of a few events in the region excluded by the cut would constitute a discovery. On the other hand, it is technically important to set some upper cut (though extremely mild, as in this case) in order to strictly avoid the presence in toy datasets of high-p_T outliers, falling in a region that is too

rare to be populated even in the Reference sample.[11] Indeed, we will see that our strategy would overreact to such outliers, similarly to what we discussed in Sect. 3.4 in the univariate example.

4.1 Model selection

The first step in our strategy implementation is the selection of a suitable neural network model "$f(x;\mathbf{w})$", and of its weight-clipping regularization parameter, for the BSM network (see Fig. 4). The principles underlying the selection, and its technical implementation, are described in detail in Sect. 3.1 for the univariate example. However the choice of the weight clipping parameter turns out to be more delicate for the multivariate analysis under examination. We believe that this is due to the enhanced sensitivity to the statistical fluctuations of the training sample, which in turn stems from two reasons. First, the sparsity of data in more dimensions unavoidably favors overfitting, to be mitigated with a more aggressive weight clipping. Second, in the current study we will employ a Reference sample size that is only 5 times larger than $\mathrm{N}(\mathrm{R_0})$, namely $\mathrm{N}_{\mathcal{R}} = 5\,\mathrm{N}(\mathrm{R_0}) \simeq 40\,000$, to be compared with $\mathrm{N}_{\mathcal{R}} = 100\,\mathrm{N}(\mathrm{R_0})$ in the univariate case. This choice, which obviously enhances the statistical fluctuations of the Reference sample, was made in order to validate our strategy in a realistic context where an extremely abundant Reference sample might (possibly because of the resources needed to run the full detector simulation) not be available.[12]

In the same spirit, the results of the present section are obtained (if not specified otherwise) using a single Monte Carlo sample of 3.6 million unweighted events in total, generated with mild acceptance requirements. Each toy dataset was obtained by random sampling around (up to Poisson fluctuations) 200 000 events in the original sample. After the events are selected according to these requirements, the desired average number $\mathrm{N}(\mathrm{R_0})$ of toy events is found. The Reference dataset employed for the training of each toy experiment was obtained by sampling 1 million events from the original data, out of the remaining 3.4 million. This way of proceeding is different from the one we adopted in the univariate example, where each toy and the corresponding Reference sample were generated independently. Clearly, this procedure dictated by the constraints of our limited computational power, is not ideal as it introduces unwanted correlation among the toys. Since we sample with probability $2\times10^5/3.6\times10^6 = 1/18$, we can still reasonably regard the different toys as independent if we generate around 100 of them (but not more). The Reference samples are instead quite correlated because we extract 1 million points out of 3.4 million only. However there is no conceptual need for Reference samples being uncorrelated across toys. Indeed, we described the conceptual role played by the Reference sample, in Sect. 2.4, under the implicit assumption that only one such sample is available for the training of all the toys. The only condition on the Reference sample is $\mathrm{N}_{\mathcal{R}}/\mathrm{N}(\mathrm{R_0}) \gg 1$. We are assuming here that $\mathrm{N}_{\mathcal{R}}/\mathrm{N}(\mathrm{R_0}) = 5$ suffices. This assumption has been validated by verifying the stability of the training outcome of individual toys under re-sampling of the Reference sample. Further cross-checks of this and other aspects, including the approximate independence of the toys, have been performed using a second independent 3.6 million points sample. In addition, the results of the present section concerning the tuning of the weight clipping and the hyperparameters optimization have been reproduced using this second sample.

In light of the items discussed above, it is important to study model selection in detail for the two-body final state problem outlining the differences with the univariate case results presented in Sect. 3.1. This is the purpose of the present section.

In a previous study [2] of the same dataset we found that a (5, 5, 5, 5, 1) network with 3 hidden layers of 5 nodes each (for a total of 96 degrees of freedom) returns a distribution for the test statistic $\bar{t}$ which is well compatible with the target χ^2_{96} distribution, for an appropriate choice of the weight clipping parameter.[13] The weight clipping selection is performed with the algorithm described in Sect. 3.1, which iteratively reduces the window of potentially viable values of the weight clipping parameter. The last step of the selection process, where the window is already as small as the $[2.1,\ 2.2]$ interval, is illustrated in Fig. 9. A comparison with Fig. 1 and Table 1 immediately reveals a number of differences between the univariate and the multivariate case. First of all, the empirical $\bar{t}$ distribution is much more sensitive to the weight clipping. Values of the weight clipping that differ from the optimal one (of 2.16) at the second digit produce distributions that are appreciably different from the target χ^2_{96}, while in the univariate case good compatibility with the χ^2_{13} was observed in a quite wide range of weight clipping. Moreover, the stabilization of the distributions with a reasonable degree of compatibility is observed only after 500 000 training epochs or more, while 100 000 epochs were sufficient in the univariate case. For the problem at hand, such large number of epochs requires a few hours CPU time.[14]

[11] In a real-life situation, the value of this upper threshold would be set just above the highest p_T value observed in data.

[12] As a side remark, we acknowledge the importance of a reliable fast simulation to make it feasible to generate very large reference datasets. To this purpose, it would be crucial to explore the use of analysis-specific deep-learning based data augmentation techniques (as in Ref. [37]), in conjunction with the speed up of event generators [38].

[13] In this section we employ the concepts, terminology and notations introduced in Sect. 3.1. In particular, $\bar{t}$ is the test statistic in the absence of nuisances, defined by Eq. (28).

[14] The training time required for a given architecture clearly depends on the problem. In particular it is proportional to the number of

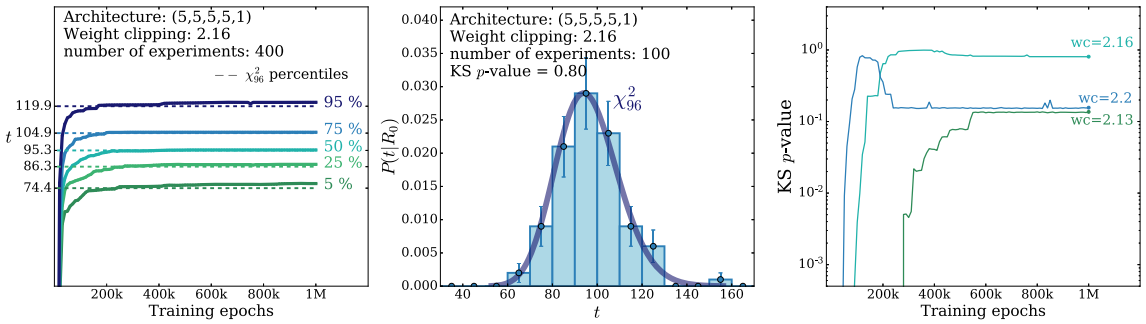


Fig. 9 Left panel: Percentiles of the empirical $\bar{t}$ distribution for the (5, 5, 5, 5, 1) network, with 100 toys, as a function of the number of training epochs for the optimal value (2.16) of the weight clipping parameter. Middle: The distribution after 1 million epochs. Right: The evolution during training of the KS p-value for different values of the weight clipping

No further studies were made in Ref. [2] on the choice of the model architecture. On one hand, this is justified by the fact that identifying one single χ^2-compatible configuration is sufficient for the applicability of our strategy. On the other hand, of the many configurations that potentially satisfy this requirement one should select the most complex model, because more expressive networks have more potential to fit putative BSM effects, enhancing the sensitivity of the search. There is not a unique notion of complexity for neural network models. Complexity can, for instance, be enhanced by increasing the number of hidden layers or the number of nodes per layer or, alternatively, by introducing more sophisticated activation functions and connection maps. It is hard to reduce such concepts to a unique scalar metric. One simple way to proceed would be to count the number of trainable parameters, but this would not discriminate between models with different architectures. In our study we restrict our attention to fully connected feedforward neural networks, with the same number of nodes at each layer. Different architectures are thus characterized by two parameters, namely the number of hidden layers and the number of nodes per layer, i.e. the depth and the width of the network. In what follows we explore this two-dimensional architectures space in slices of depth, trying to identify the maximum number of nodes that, for fixed number of layers, can be made compatible with the target χ^2 distribution for an appropriate choice of the weight clipping parameter.

The conceptual criteria for model selection discussed above must be combined with practical considerations, taking into account the available computational resources that limit the complexity of the model we can concretely handle. With "computational resources" we refer both to the memory required to store the model and its gradients during training, and to the training time needed to get a stable solution. For models with a good level of compatibility with the target χ^2 distribution, we sharply define a solution as "stable" by requiring the KS p-value not to vary more than 10% for at least 100 000 epochs. The memory is not a limiting factor. It does not exceed around 1 GB even for the most complex models we have considered. The training time is instead considerable, because of the large number of epochs that is typically required. For the present study we consider a neural network model "manageable" when a stable training (on a single toy dataset) takes less than 6 hours CPU time. This threshold takes into account the need of repeating training on many toys (we use 100 toys to establish χ^2-compatibility), of performing a scan on the value of the weight clipping parameter that ensures compatibility, and of exploring different architectures. One should notice that our procedure offers parallelization opportunities by running toy experiments in parallel. Because of this, and having at hand a large-size cluster of CPUs (CERN `lxplus` cluster) and a handful of GPUs, we found it convenient to run in parallel many time-consuming toys on CPUs as opposed of running a few fast toys on GPUs.

Based on the above considerations, we identified the (5, 50, 1) network as the most complex viable model among those with a single hidden layer. The last step of the weight clipping selection process is illustrated in Fig. 10. The observed behaviour is similar to the one of Fig. 9 in terms of the sensitivity to the weight clipping and of the number of epochs required for training. The (5, 50, 1) network has many more parameters (351 versus 96) than the (5, 5, 5, 5, 1) one, but all concentrated in one layer. These two aspects combined make the training time somewhat longer, but still within the boundary of 6 hours CPU time that defines our computational threshold. Increasing the number of neurons of the network would further increase the training time, therefore the (5, 50, 1) model is selected among the one-layer architectures. Among the architectures with two hidden layers, we selected by similar considerations (see Fig. 11) the (5, 10, 10, 1) network.

Footnote 14 continued
training points which in turn, keeping the ratio $N_{\mathcal{R}}/N(\mathbf{R_0}) = 5$ fixed, scales with the number of expected events.

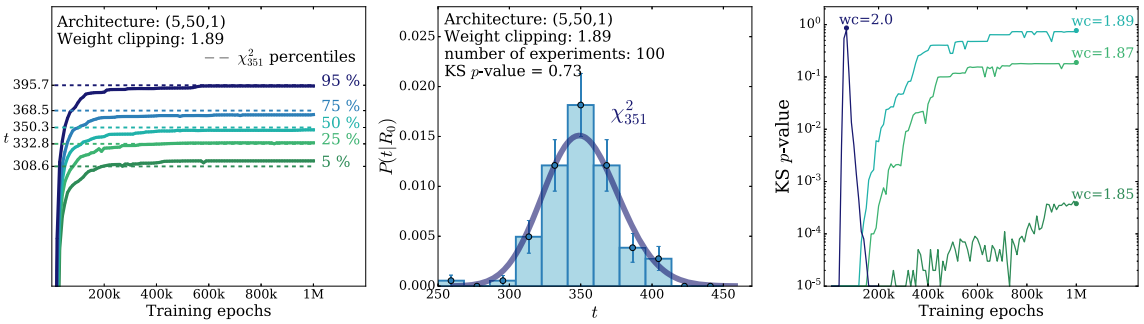


Fig. 10 Same as Fig. 9, but for the (5, 50, 1) architecture

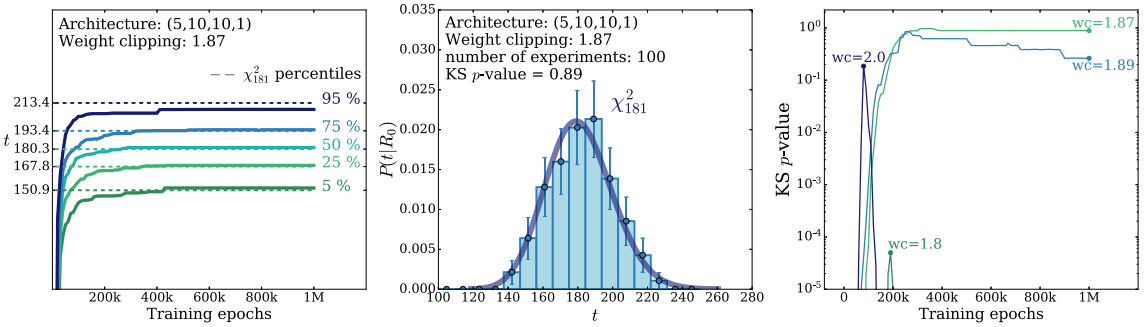


Fig. 11 Same as Fig. 9, but for the (5, 10, 10, 1) architecture

Table 3 Summary of the weight clipping tuning results for the architectures considered in this section

# of layers	Latent size	dof	Weight clipping	Training epochs	KS p-value
1	5	36	1	100k	0.33
	50	351	1.89	1M	0.90
2	6	85	1.8	200k	0.81
	7	106	1.84	200k	1.00
	10	181	1.87	1M	0.99
3	**5**	96	2.16	1M	0.89
	10	291	–	–	–

We also tested other architectures, with the results summarized in Table 3. For networks with 1 (2) hidden layers and less than 50 (10) neurons, we could easily tune the weight clipping parameter obtaining a good level of compatibility with the target χ^2. The number of epochs that safely ensures convergence, reported in the table, decreases with the network size as expected, and training becomes computationally less demanding. Networks with more neurons are beyond our computational threshold as previously explained. A 3 layers network with 10 neurons was also considered, but the weight clipping tuning could not be achieved, because of the behaviour displayed in Fig. 12. If the weight clipping is small, training is stable but the $\bar{t}$ distribution strongly undershoots the target χ^2. By raising the weight clipping the distribution moves to the right, but it is not stable even after one million epochs. More training time would be needed to establish if, for instance, the configuration with weight clipping equal to 1.9 will eventually converge to the target χ^2. Since this goes beyond our computational threshold, the (5, 10, 10, 10, 1) network has to be discarded. We thus retained the (5, 5, 5, 5, 1) network, in the 3-layers class. We did not consider networks with four or more layers because we expect, in light of these results, that for these networks we would be obliged to use less than 5 (the number of features) neurons in the hidden layers, entailing dimensionality reduction. In summary, the only architectures to be considered for further studies are (5, 50, 1), (5, 10, 10, 1) and (5, 5, 5, 5, 1). We will refer to them as Model 1, 2 and 3 respectively.

4.2 Learning nuisances and validation

Our next task is to model the effect of nuisance parameters on the distribution log-ratio $\log r(x, \boldsymbol{\nu})$. This is a rather straightforward application of the methodology of Sect. 2.3,

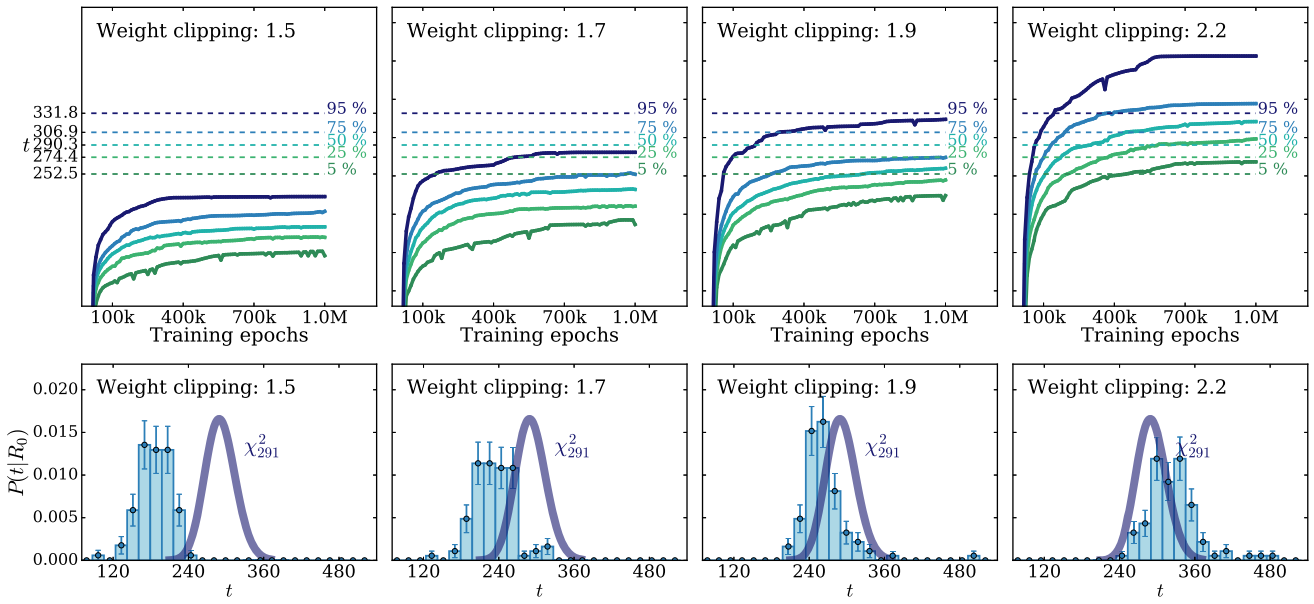


Fig. 12 The percentiles of the empirical $\bar{t}$ distribution as a function of the training epochs (top row) and the distribution of the empirical $\bar{t}$ distribution after 1M training epochs (bottom row) for the (5, 10, 10, 10, 1) network at different values of the weight clipping

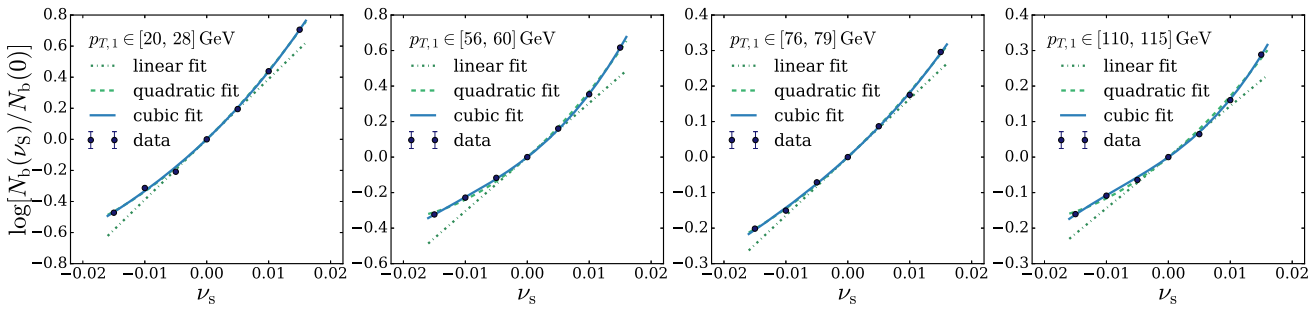


Fig. 13 The dependence on ν_S of $\log N_b(\nu_S)/N_b(0)$ in selected bins of the transverse momentum distribution. The dots represent the true value of the log-ratio. The linear, quadratic fits are performed using a subset of the true values points within ± 0.01. The quartic one also considers points at ± 0.015

only slightly more computationally demanding than the one presented in Sect. 3.2 for the univariate problem. The normalization nuisance ν_N contributes linearly to the log-ratio, we thus incorporate it analytically in the reconstructed $\log \widehat{r}(x; \boldsymbol{\nu})$, as in Eq. (31). The effect of the scale nuisance ν_S is reconstructed locally in the five-dimensional space of features by means of two neural networks $\widehat{\delta}_{1,2}(x)$ that parametrize the Taylor expansion of the log-ratio up to quadratic order, again as in Eq. (31). The $\nu_{S,i}$ values used for training were selected by studying the effect of the scale uncertainty nuisance on the features distribution, like in Fig. 13. The figure shows the dependence on ν_S of the expected number of events in selected bins of the transverse momentum of the leading lepton ($p_{T,1}$). The scale uncertainty in the endcaps region has been taken 3 times the one in the barrel, as appropriate for the muon and electron scenarios defined at the beginning of this section. The result is expressed as a function of the scale in the barrel, ν_S. The uncertainty σ_S will be set to 5×10^{-4} and to 3×10^{-3} in the muon and electron scenarios, respectively. We see that the dependence is quadratic to a good approximation in the interval $\nu_S \in [-0.02, 0.02]$, which comfortably covers the range that is relevant for the electron scenario up to more than 3 sigma (and even more for the muon-like one). Training points $\nu_{S,i} = \{\pm 1.5 \times 10^{-3}, \pm 1.5 \times 10^{-2}\}$ are selected as a reasonable choice which exposes the $\widehat{\delta}$ networks both to the linear and to the quadratic component of the likelihood log-ratio. The validity of this choice was confirmed by also inspecting the nuisance dependence of other kinematical variables.

Five hidden layers with 10 neurons (and ReLU activation functions) was identified as a viable architecture for the $\widehat{\delta}$ networks. The training samples $S_0(\boldsymbol{\nu}_i)$ were obtained using half of the original 3.6 million sample. After the selection requirements are applied, they consist of around 80 000 events for each value of ν_i. The S_1 sample, with central-value nuisance,

was provided by the remaining 1.8 million events, weighted by a factor of 4 in order to compensate for the presence of the four $S_0(\boldsymbol{\nu}_i)$ non-central-value samples. For training we applied an early stopping criterion based on the quality of the log-ratio reconstruction achieved by the networks. The quality of the reconstruction was monitored by plots like the one in Fig. 14 and also by testing the capability of the $\widehat{\delta}$ networks to reabsorb the effect of non-central nuisances in the test statistic distribution. Good performances were obtained with 2 000 epochs. A mild overfitting was observed training longer.

In order to test the accuracy of the log-ratio reconstruction, we use the reconstructed $\widehat{r}(x;\boldsymbol{\nu})$ to re-weight the Monte Carlo sample with central-value nuisances, and we compare the predictions for the binned distribution log-ratio (in p_T bins), as obtained by this re-weighting, with those obtained using non-central-value samples. Figure 14 shows good agreement, for ν_S in the range that is relevant to cover the muon- and electron- like scenarios.

The most stringent cross-check of the quality of the log-ratio reconstruction is however provided by the final validation of the whole strategy, that consists in verifying the independence on the nuisance parameters of the distribution of the test statistic, $P(t|\mathrm{R}_{\boldsymbol{\nu}})$. Indeed, as emphasized in previous sections (see in particular Sect. 3.4), the emergence of a χ^2 distribution for the test statistic $t = \tau - \Delta$, with the appropriate number of degrees of freedom, provides a highly non-trivial test of all aspects of the algorithm implementation, ranging from the selection of the BSM network hyperparameters (which affects τ) to the accuracy of the log-ratio reconstruction (which affects both the τ and the Δ terms). In Figs. 15 and 16 we display some of the validation plots that have been produced in order to verify the independence of the test statistic distribution on the true values $\boldsymbol{\nu}^* = (\nu_N^*, \nu_S^*)$ of the nuisance parameters. A summary of the results is provided in Table 4, covering the three neural network models (1, 2 and 3) selected in Sect. 4.1 for the BSM network, and in the electron-scenario for the scale uncertainty. The KS p-value is typically low in the "w/o correction" columns, showing that the presence of nuisances impacts the distribution of τ significantly. The asymptotic formula for the distribution of $t = \tau - \Delta$ is recovered by the inclusion of the Δ term, as shown by the higher p-values in the "w/o correction" columns.

In summary, we have demonstrated the possibility to deal with a level of uncertainties that corresponds to the electron-like scenario, as defined at the beginning of Sect. 4. Trivially (since lower uncertainties are easier to manage), the same holds in the muon-like setup. The larger uncertainty that is foreseen in the τ-like scenario is instead more difficult to manage, and deserves an extensive dedicated discussion, which is the subject of the following section.

4.3 The $\boldsymbol{\tau}$-like scenario

The first difficulty we encounter in the τ-like scenario is the wild dependence of the distribution on the scale nuisance parameter, displayed in Fig. 17. The effect is due to the migration of events from the Z-peak to the signal region defined by the invariant mass cut $M_{12} > 100$ GeV. Since the Z-peak events are overly abundant, even a small correction to the Z-peak rejection efficiency (of order $\sigma_S = 3 \times 10^{-2}$ in the τ-like scenario) affects at order one the distribution in the signal region. Our current setup is only capable to deal with relatively small distortions, for which the Taylor expansion in Eq. (31) is justified. Therefore we do not even try to study the τ-like scenario in the entire signal region $M_{12} > 100$ GeV, but rather consider a harder cut $M_{12} > 120$ GeV that mitigates the Z-peak migration effects. Figure 18 shows that the effects of the nuisance are still sizable in this region, but moderate enough to justify the expansion in ν_S up to the quadratic order. The harder invariant mass cut reduces the expected number of events by a factor of around 3. We compensate by raising the luminosity as discussed at the beginning of this section, in order to maintain $\mathrm{N}(\mathrm{R}_{\mathbf{0}}) = 8\,400$ similar to the one of the muon- and electron-like setups. We also want to maintain a similar proportion between $\mathrm{N}(\mathrm{R}_{\mathbf{0}})$ and the total number of Monte Carlo events employed in the analyses. We must thus use three samples with 3.6 million events (for a total of 10.8 millions) each before cuts.

It is straightforward to repeat in this new setup all the steps described in the previous section. In particular the three neural network architectures identified in Sect. 4.1 are still viable up to a mild retuning of the weight clipping parameter. However, validation is more delicate because of the stronger impact of systematics uncertainties on the distribution of τ. As discussed in Sect. 2.5 and verified in Sect. 3.4 in the univariate example, we expect that a higher accuracy is required in the computation of τ and of Δ in order to properly capture the cancellation that takes place in the test statistics $t = \tau - \Delta$. We observe that different levels of accuracy are required to validate the three neural network models, depending on the sensitivity of each model to the sparsity of input features. In particular, Model 3 (with 3 hidden layers) turns out not to be particularly sensitive, and its validation does not pose any particular issue, even if the KS compatibility p-values for non-central nuisances (see Fig. 19) are somewhat lower than those we found in the previous section for the muon- and electron-like scenarios. For Model 1 and 2, instead, the compatibility with the target χ^2 is remarkably low, especially if ν_S^* is positive. The exact same asymmetric behavior was found in Sect. 3.4 in the univariate example, and attributed (see Fig. 6) to the fact that positive scale variations push the data to the extreme tail of the Reference model distribution, which is not populated in the Reference sample. The same effect was found to be responsible for the behavior

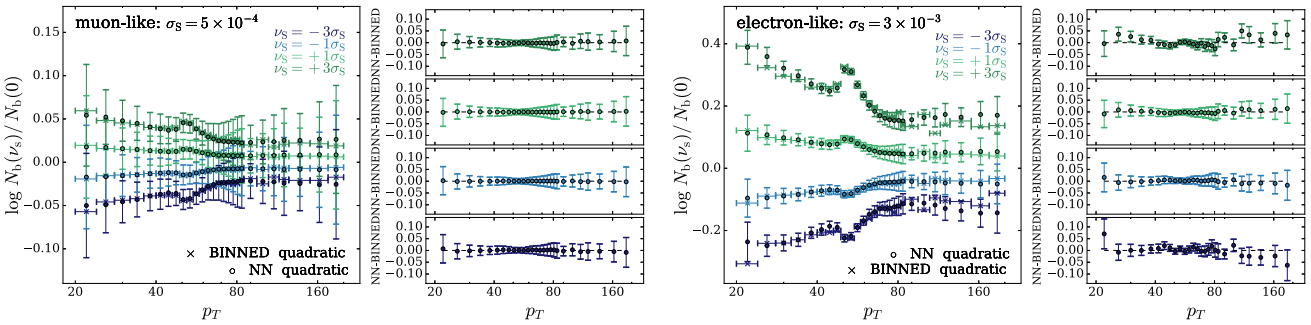


Fig. 14 The reconstructed distribution log-ratio (dots) for different values of ν_S, compared with the quadratic binned approximation. The two panels cover the ranges of ν_S that are relevant for the muon- and electron like scenarios respectively

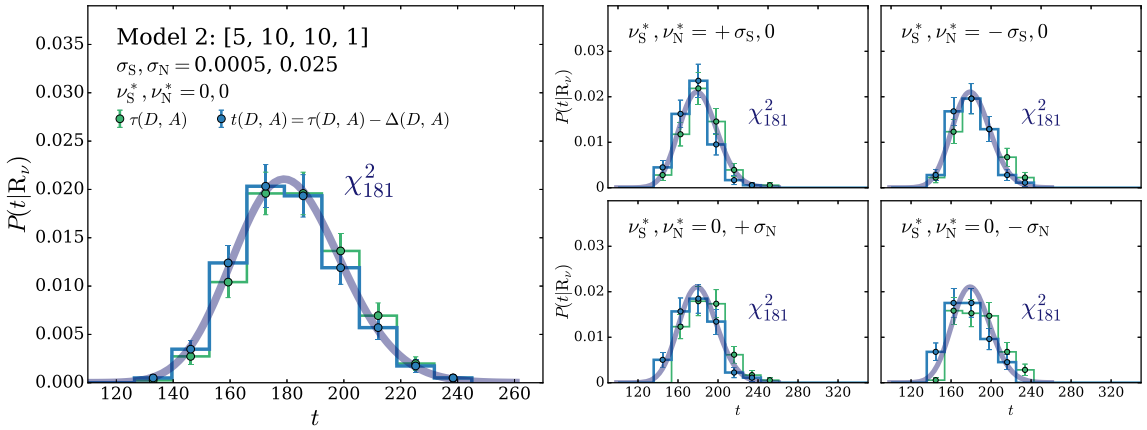


Fig. 15 The empirical distribution of τ (in green) and of t (in blue) computed by 100 toy experiments performed in the R_ν hypothesis at different points in the nuisance parameters space for the muon-like regime. The χ^2_{181} distribution is reported in blue in all the plots

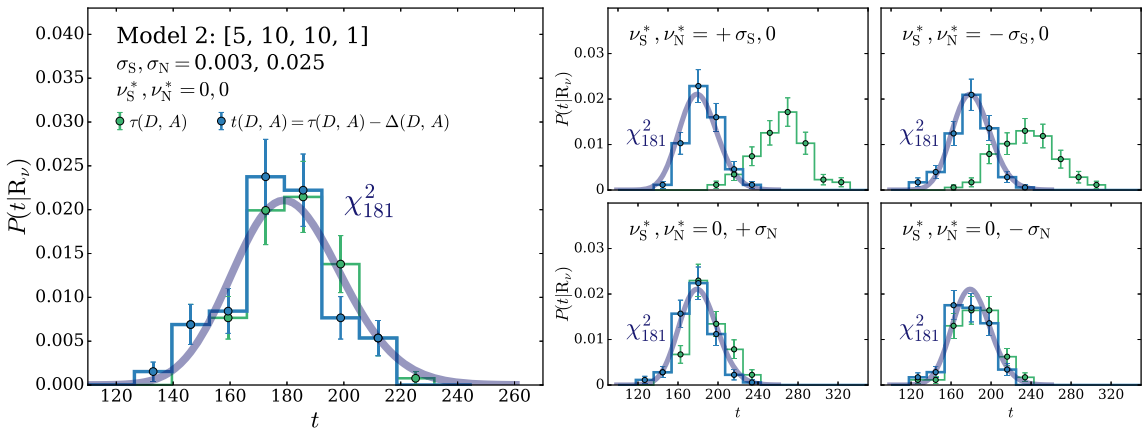


Fig. 16 Same as Fig. 15, but for the electron-like regime

we observe in the present setup. Indeed we could check the presence of extreme outliers in the trained neural network output, localized in a transverse momentum region that is not populated in the Reference sample.

Since the problem is due to lack of Reference data in the tail, a way out could be to add statistics to the Reference sample, which however is computationally costly. Certainly feasible with the computing power of a large experiment but beyond our capabilities. A more efficient solution is instead to enrich the Reference sample with a new Monte Carlo sample with a cut on the transverse momenta at generation-level. We then generate 200′000 events with 200 GeV generation-level cut on the minimal leading p_T (plus basic acceptance cuts), and we further cut it at 250 GeV on the reconstructed momenta. We add such events with appropriate weights to the original 10.8 million sample, and we remove the original events with $p_T > 250$ GeV. The so-obtained weighted sample is then employed to generate Reference samples, and

Table 4 Kolmogorov–Smirnov p-value for the compatibility of the τ ("w/o correction" columns) and of the t ("w/ correction" columns) distributions with the target χ^2 distribution for model 1, 2, 3 in the electron-like regime. The KS test is based 100 toy experiments performed in the R_ν hypothesis at different points in the nuisance parameters space

$\left(\frac{\nu_S^*}{\sigma_S}, \frac{\nu_N^*}{\sigma_N}\right)$	Model 1		Model 2		Model 3	
	KS p-value		KS p-value		KS p-value	
	w/o correction	w/ correction	w/o correction	w/ correction	w/o correction	w/ correction
(0, 0)	0.59	0.86	0.082	0.10	0.02	0.03
(+1, 0)	$< 10^{-5}$	0.02	$< 10^{-5}$	0.05	$< 10^{-5}$	0.18
(0, +1)	0.0002	0.58	0.002	0.18	0.11	0.13
(−1, 0)	$< 10^{-5}$	0.17	$< 10^{-5}$	0.83	$< 10^{-5}$	0.20
(0, −1)	0.24	0.71	0.09	0.24	0.002	0.06

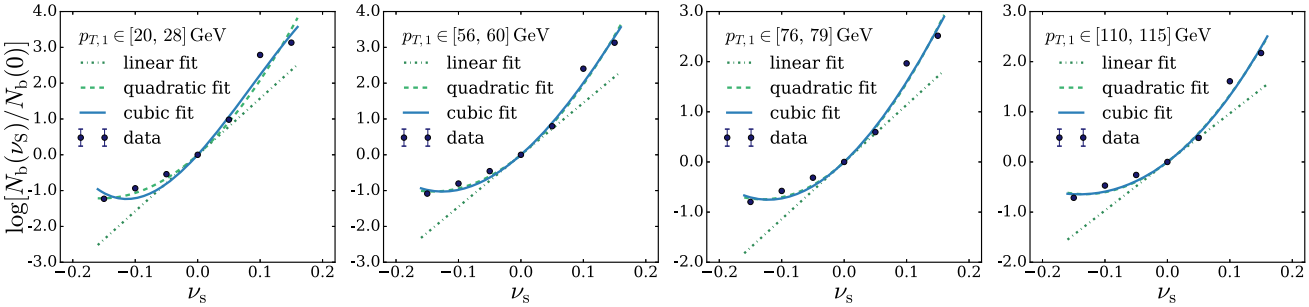


Fig. 17 The dependence on ν_S of $\log N_b(\nu_S)/N_b(0)$ in selected bins of the transverse momentum distribution for $M_{12} > 100$ GeV. The dots represent the true value of the log-ratio. The linear and quadratic fits are performed using a subset of the true values points within ± 0.1; the cubic one also considers two additional points at ± 0.15

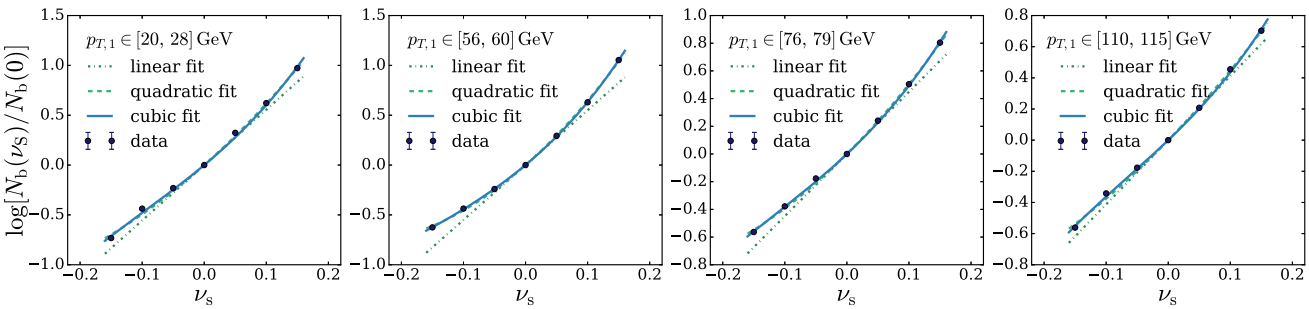


Fig. 18 Same as Fig. 17, but for $M_{12} > 120$ GeV (lower panel)

the toy data by hit-or-miss unweighting. It is also used for the training of the $\widehat{\delta}$ networks, improving the quality of the distribution ratio reconstruction in the high-p_T tail.

The usage of the enriched sample allows us to validate Model 3 with higher KS p-value, as shown in Fig. 20. Furthermore it eliminates the outliers in the neural network output and drastically ameliorates the χ^2-compatibility of Model 1 and 2. On the other hand, a satisfactory validation of Model 1 and 2 requires a further improvement of the sample. By increasing the number of events in the high p_T tail from 200′000 to 400′000, good results are found for the validation of Model 2, shown in Fig. 21. Figure 22 shows that good compatibility can be obtained for Model 1 as well, but only with 600′000 high-p_T events. The improvement can be traced back to the more accurate reconstruction of the nuisance coefficient functions δ, which can be monitored by comparing the left panels of the two figures.

4.4 Sensitivity to new physics

We conclude the section on the two-body final state experiments presenting some examples of the algorithm performances to detect New Physics in the data. For definiteness, we chose the Model 3 architecture to perform the sensitivity tests. We consider two new physics benchmark scenarios: [15]

[15] The same benchmarks are employed in Ref. [2], to which the reader is referred for additional details.

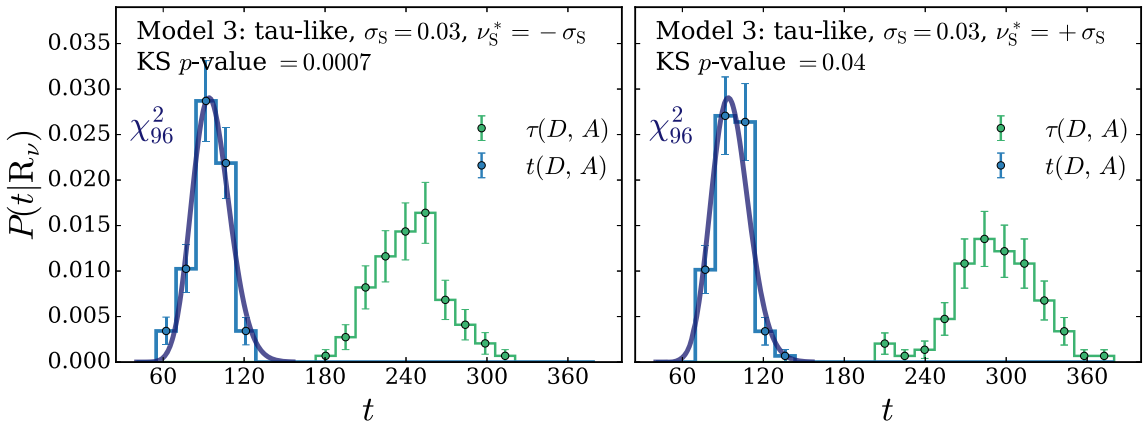


Fig. 19 The empirical distribution of τ (in green) and of t (in blue) computed by 100 toy experiments performed for Model 3 in the R_ν hypothesis at $\nu_S = -1$ (left side) and $\nu_S = +1$ (right side) for the τ-like regime before enriching the reference sample in the region of high transverse momentum

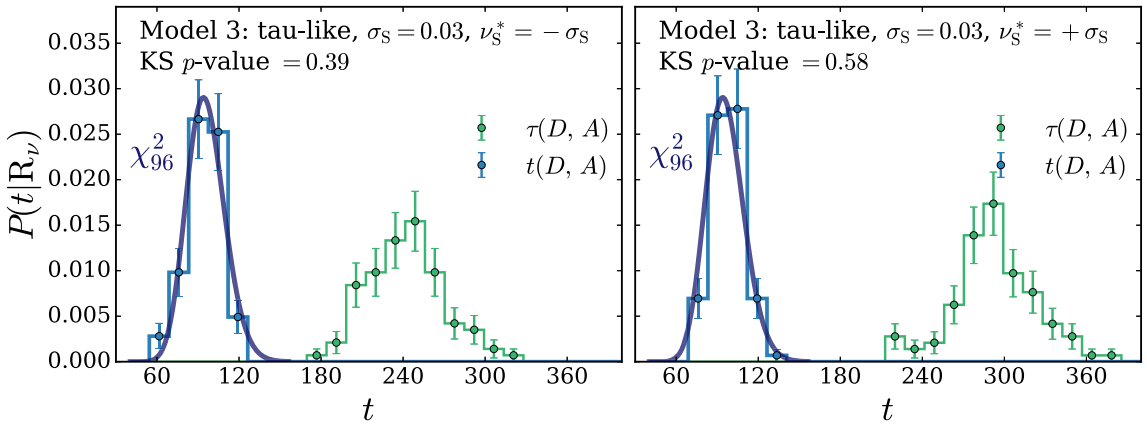


Fig. 20 Same as Fig. 19, but after enriching the reference sample in the region of high transverse momentum with 200′000 additional events

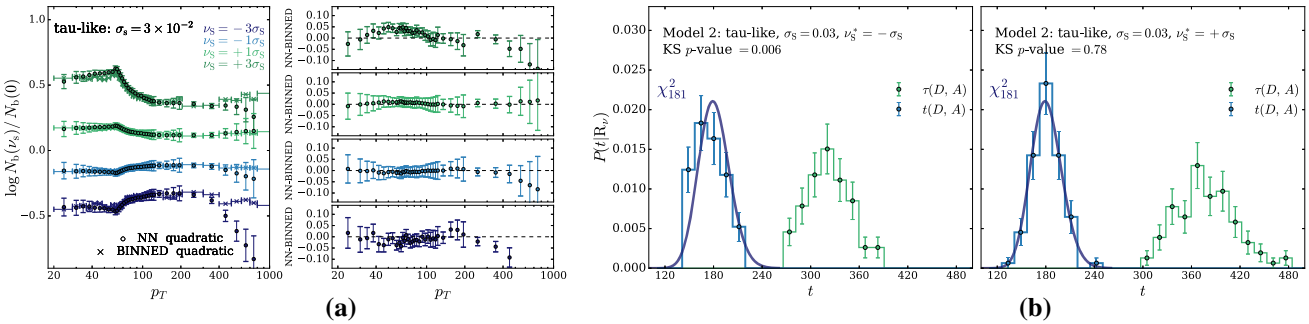


Fig. 21 Left side: the reconstructed distribution log-ratio (dots) for different values of ν_S, compared with the quadratic binned approximation. Right side: the empirical distribution of τ (in green) and of t (in blue) computed by 100 toy experiments performed for Model 2 in the R_ν hypothesis at $\nu_S = -1$ (left side) and $\nu_S = +1$ (right side) for the τ-like regime. Both plots have been obtained enriching the reference sample in the region of high transverse momentum with 400′000 additional events

- Z' scenario: a new vector boson with the same couplings to SM fermions as the SM Z boson and mass of 300 GeV;.
- EFT scenario: a non-resonant effect due to a dimension-6 4-fermion interaction

$$\frac{c_W}{\Lambda} J^a_{L\mu} J^{\mu}_{La} \tag{37}$$

where J^{μ}_{La} is the $\mathrm{SU(2)_L}$ SM current, the energy scale Λ is fixed at 1 TeV and the Wilson coefficient c_W determines the coupling strength.

Both benchmarks are studied in the three regimes of systematic uncertainties considered so far and the median

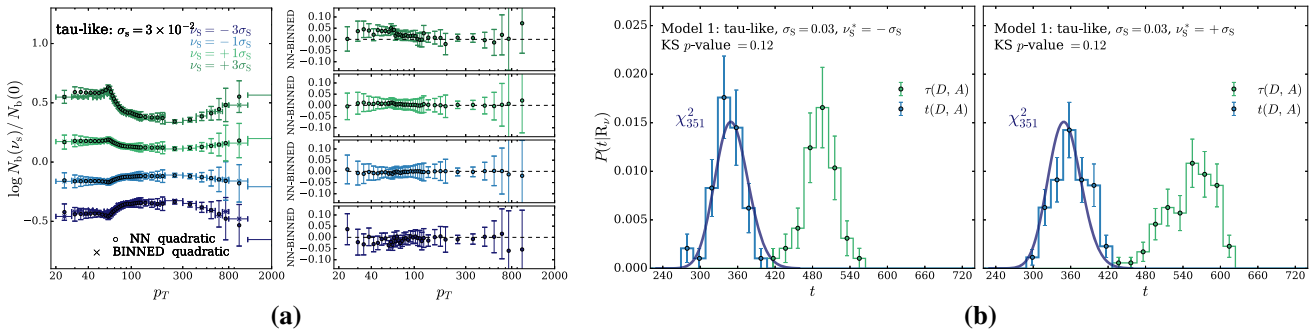


Fig. 22 Same as Fig. 21, but for Model 1 and 600′000 additional events in the region of high transverse momentum

observed Z-score ($\overline{Z}$) is compared with a median reference Z-score ($\overline{Z}_{\text{ref}}$). As in Sect. 3.5, the reference Z-score is defined as a model dependent measure of the significance, performed by assuming that the specific new physics model is known a priori. As a first approximation, in both scenarios a model dependent analysis would select the two-body invariant mass as the variable of interest. We thus compute the test statistic in Eq. (8) by binning the two-body invariant mass and studying the effects of the nuisance parameters and that of the signals in each bin. Notice that the upper cut on the transverse momentum, which we employ in our analysis, is not applied at this stage. For the SM hypothesis the dependence on the momentum scale nuisance parameter ν_S is approximated by a quadratic polynomial, whereas for the Z' signal we use a quartic one. We call $N(S)$ the total number of expected Z' events, and we introduce a global exponential factor to describe the normalization uncertainty. Namely, we parametrize the number of events expected in each bin as

$$\hat{n}_i^{(Z')}(N(S), \nu_S, \nu_N) = [(a_{0i} + a_{1i}\nu_S + a_{2i}\nu_S^2) + N(S)\,(b_{0i} + b_{1i}\nu_S + b_{2i}\nu_S^2 + b_{3i}\nu_S^3 + b_{4i}\nu_S^4)] \cdot e^{\nu_N}. \tag{38}$$

For the EFT instead, the number of events in each bin depends quadratically on the Wilson coefficient c_W, while the dependence on ν_S on the New Physics term (i.e., on the linear and quadratic c_W terms) can be safely ignored. Therefore, we have

$$\hat{n}_i^{(\text{EFT})}(c_W, \nu_S, \nu_N) = (a_{0i} + a_{1i}^{\nu_S}\nu_S + a_{2i}^{\nu_S}\nu_S^2 + a_{1i}^{c_W} c_W + a_{2i}^{c_W} c_W^2) \cdot e^{\nu_N}. \tag{39}$$

The numerical a and b coefficients in the above equations where determined by a fit to the Monte Carlo simulations in each bin.

Denoting collectively as "μ" the signal strengths in the two scenarios, namely $\mu = N(S)$ or $\mu = c_W$, respectively, the binned log-likelihood reads (up to an irrelevant additive constant)

$$\log \mathcal{L}(\mu, \nu_S, \nu_N|\mathcal{D}, \mathcal{A}) = \sum_{i \in \text{bins}} n_i \log[\hat{n}_i(\mu, \nu_S, \nu_N)] - N(\mu, \nu_S, \nu_N) + \log \mathcal{L}(\mathbf{0}|\mathcal{A}), \tag{40}$$

where n_i denotes the number of observed events in the i-th bin. The binned log-likelihood is then used to compute the test statistic

$$t_{\text{ref}}(\mathcal{D}, \mathcal{A}) = 2\frac{\max_{\mu,\boldsymbol{\nu}}[\log\, \mathcal{L}(\mu, \nu_S, \nu_N|\mathcal{D}, \mathcal{A})]}{\max_{\boldsymbol{\nu}}[\log\, \mathcal{L}(0, \nu_S, \nu_N|.\mathcal{D}, \mathcal{A})]}. \tag{41}$$

The reference Z-score is finally obtained by throwing toy experiments in the new physics hypothesis and computing the p-value of the median of the empirical test statistic distribution. In the regimes considered for this work, the counts per bin are always greater than 4. Therefore it is legitimate to assume the asymptotic behavior for the distribution of the test statistic under the null (SM) hypothesis to be valid, and compute the p-value with respect to a χ^2_1. The asymptotic behavior has been verified by running the procedure on SM-distributed toys.

Figure 23 shows the algorithm performances in the muon-like and electron-like regimes. The setup is the one described at the beginning of this section, with an effective luminosity (set by assuming the cross-section of the di-muon process) of 0.35 fb^{-1} and a cut on the two-body invariant mass at 100 GeV, which leads to approximately $N(\mathbf{R_0}) = 8\,400$ expected SM events in the search region. For the Z' scenario we inject a number of signal events which is Poisson-distributed around the expected value $N(S) = 120$, which is around 1% of $N(\mathbf{R_0})$. Whereas for the EFT scenario we generate a Monte Carlo sample with Wilson coefficient set to 1 TeV^{-2}, which increases the total cross section only at the 2 per mille level. Figure 23 shows that muon- and electron-like systematics do not affect appreciably the sensitivity of our method, nor the sensitivity of the model-dependent analysis strategy that we take as reference.

The results in the τ-like regime are presented in Fig. 24. As previously explained the effective luminosity is now set

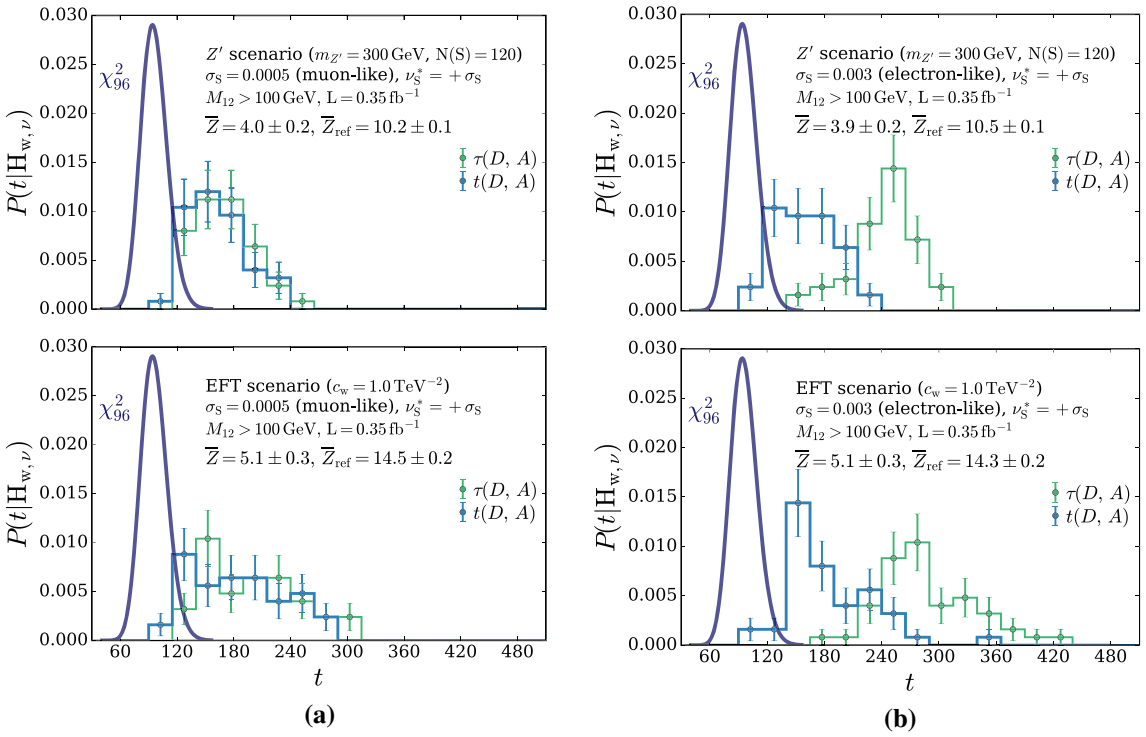


Fig. 23 Sensitivity to two New Physics scenarios in the muon-like (**a**) and electron-like (**b**) regimes. The upper panels show the sensitivity of the method to the presence of a Z' ($m_{Z'} = 300$ GeV, N(S) = 120) resonance in the two leptons invariant mass. The lower panels show the sensitivity of the method to a non resonant effect due to a dimension-6 4-fermions interaction (EFT scenario, $c_W = 1.0\,\text{TeV}^{-2}$). In all panels the true value of the scale nuisance parameter is assumed to be 1 standard deviation above the central value

to 1.1 fb^{-1} and the cut on the two-body invariant mass is moved to 120 GeV. Since the data integrated luminosity is now a factor 3 larger than what used in the previous cases, the sensitivity to new physics improves making the previous benchmark models visible with overly high significance. In order to define realistically challenging benchmarks we thus reduce the Z' cross-section such that $\text{N}(S) = 210 < 3 \cdot 120$, while in the EFT scenario we lower the Wilson coefficient to $c_W = 0.25\,\text{TeV}^{-2}$. In order to asses the role of systematics, we compare the τ-like setup to an idealized experiment where the uncertainties are negligible (specifically, $\sigma_S = \sigma_N = 1 \times 10^{-4}$). We observe a slight degradation of the sensitivity due to the uncertainties, but only in the case of the EFT new physics scenario, as expected because the resonant Z' signal can not be mimicked by systematics effects.

We conclude that our strategy to deal with systematic uncertainties, on top of being robust against false positives as verified in the previous sections, maintains a remarkably high sensitivity to putative new physics effects. The observed mild sensitivity loss due to uncertainties, when present, is perfectly in line with the degradation of the model-dependent reference analysis performances, signally that the sensitivity lowers because the new physics signal is genuinely harder to see and not because of an intrinsic limitation of our model-independent method. Furthermore, the results of the present section confirm the weak dependence on the specific type of new physics, claimed in our previous works [1,2], of the ratio $\overline{Z}/\overline{Z}_{\text{ref}}$. This is shown in Fig. 25 by summarizing the performances we have obtained at different luminosities, systematic uncertainties regime and new physics scenarios. In all the experiments our reach is a factor ~ 2.7 lower than the reference Z-score.

5 Conclusions and outlook

We have proposed and validated a strategy for model-independent new physics searches that duly takes into account the imperfect knowledge of the Reference model predictions. The methodology is robustly based on the canonical Maximum Likelihood ratio treatment of uncertainties as nuisance parameters for hypothesis testing, which emerges as a completely natural and conceptually straightforward extension of the basic framework we proposed and developed in Refs. [1,2]. Our findings open the door to real analysis applications, where a "New-Physics-Learning" Machine (NPLM) inspects the LHC data in search from departures from the Standard Model, with no bias on the nature and the origin of the putative discrepancy. The proposed method is an end-to-

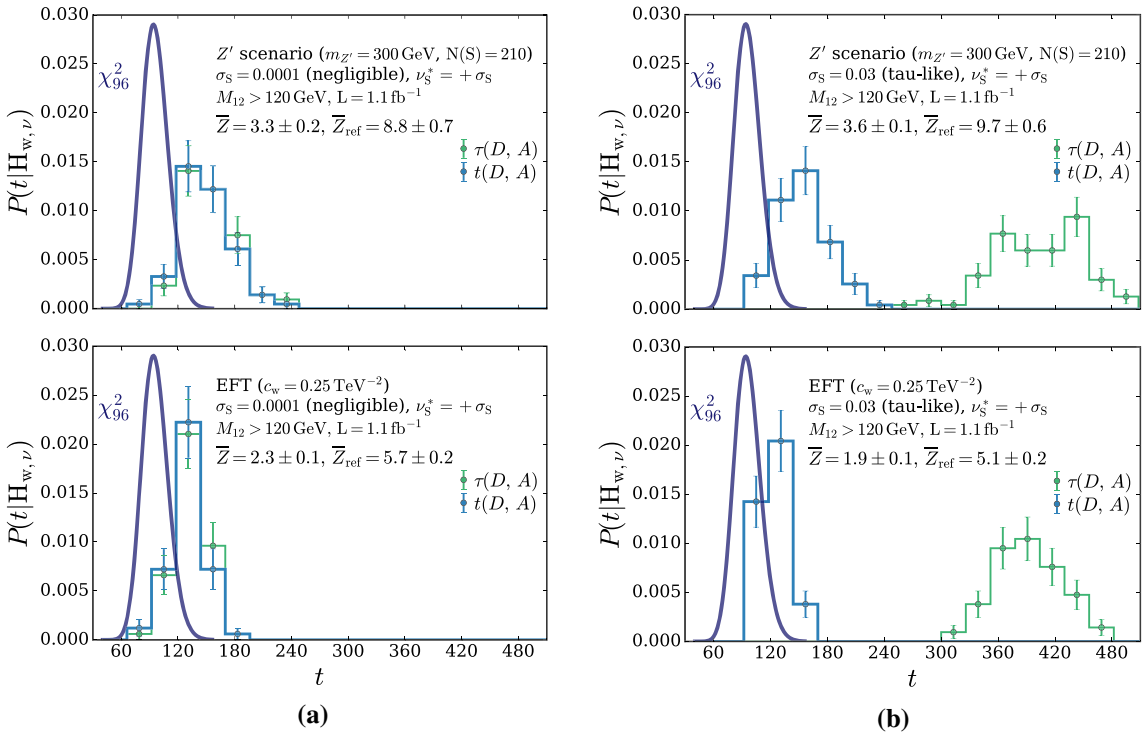


Fig. 24 Sensitivity to two New Physics scenarios in the case of negligible uncertainties (**a**) and τ-like regime (**b**). The upper panels show the sensitivity of the method to the presence of a Z' ($m_{Z'} = 300$ GeV, N(S) = 210) resonance in the two leptons invariant mass. The lower panels show the sensitivity of the method to the EFT scenario, with $c_W = 0.25\,\mathrm{TeV}^{-2}$). In all panels the true value of the scale nuisance parameter is assumed to be 1 standard deviation above the central value

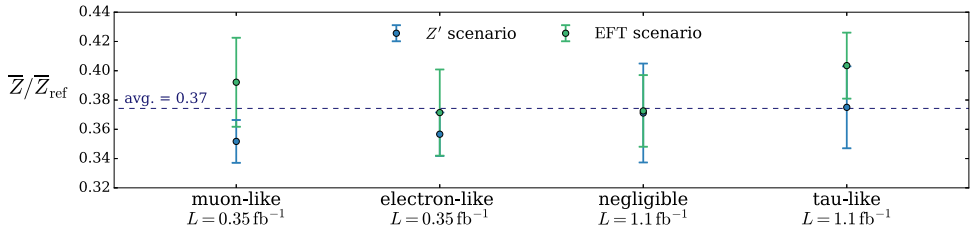


Fig. 25 Summary of the sensitivity of our method, relative to the sensitivity of dedicated model-dependent searches, to selected New Physics benchmark models. The relative performances depend neither on the New Physics model nor on the assumed scenario for systematic uncertainties

end statistical analysis, ultimately returning a p-value that quantifies the level of discrepancy of the data and the Standard Model hypothesis. Moreover it returns the trained neural network, which can be exploited for a first characterization of the discrepancy. This will pave the way to dedicated model-dependent analyses of the discrepant data set, that will eventually unveil the nature of the discovered new physics.

The detailed study of the method in real LHC analyses will be essential in order to identify possible implementation issues, which might require further developments of the NLPM strategy itself or methodological advances in related domains. Based on the studies performed in the present paper we can anticipate interesting directions for future developments:

1. The need of a statistically accurate enough (large or "smart") Reference sample. We have seen in the study of the two-body final state example how a limited Reference to Data ratio $\mathrm{N}_{\mathcal{R}}/\mathrm{N}(\mathrm{R_0}) = 5$ (to be compared with $\mathrm{N}_{\mathcal{R}}/\mathrm{N}(\mathrm{R_0}) = 100$ in the univariate problem) poses a number of technical difficulties, ranging from an enhanced sensitivity to the weight clipping parameter (see Sect. 4.1) to possible validation failures (see Sects. 3.4 and 4.3) due to Data outliers in regions that are not populated in the Reference sample. Raising the Reference sample statistic is not the only way to address these issues. Our results in Sect. 4.3 suggest that with a suitably weighted Reference one might obtain the same

effect without increasing $N_{\mathcal{R}}$, i.e. without impacting the training execution time.

2. The generation of Reference-distributed toys. Our strategies for model selection and validation heavily rely on the availability of toy datasets, namely sets of unweighted data that mimic the outcome of the real experiment under the Standard Model hypothesis. Generating a large set of toys requires, in the first place, a large enough sample of Standard Model data. The potential issue is, as for item 1, that such large sample might not be available or it might be computationally too demanding to be generated. Furthermore, if the Standard Model data are weighted, producing unweighted events with the hit-or-miss technique can be highly inefficient in the presence of large weights, and conceptually impossible if some of the weights are negative as it is the case for simulations at Next-to-Leading order.
3. Accurate learning of nuisance effects. We have seen in Sects. 3.4 and 4.3 that an accurate reconstruction of $\log r(x;\boldsymbol{\nu})$ is essential and that higher accuracy is needed for those nuisance parameter that impact the distribution of τ more considerably. On the other hand the accuracy could be limited by an insufficient statistical accuracy of the data used for training the $\widehat{\delta}$ networks. Moreover when the dependence on the nuisance parameters is not a small correction to the central-value distribution, such that it can not be Taylor-expanded in $\boldsymbol{\nu}$, we expect that learning $\log r(x;\boldsymbol{\nu})$ might become more demanding.
4. Training execution time. The time needed for training the "BSM" network is considerable, and entails (see Sect. 4.1) a computational constraint on the maximal neural network complexity that we can handle. The time obviously increases with $N_{\mathcal{R}}$, potentially posing an obstruction to the data statistics we can handle, at fixed $N_{\mathcal{R}}/N(R_{\mathbf{0}})$, or to $N_{\mathcal{R}}$ itself, which on the other hand we might need to take large as per item 1.

It should be noted that items 1 and 3, as well as item 4, are not absolute obstructions to the applicability of the NPLM strategy. They rather limit the integrated luminosity of the data (i.e., $N(R_{\mathbf{0}})$) that our algorithm can handle. Indeed item 1 can be addressed by lowering $N(R_{\mathbf{0}})$, and item 3 as well because the impact of systematic uncertainties on the analysis is relatively smaller if the data statistics is lower. On the other hand an upper limit on $N(R_{\mathbf{0}})$ does not prevent us from employing the full data luminosity for the analysis. One could indeed split the data in several independent datasets, run NPLM on each and combine statistically the corresponding p-values. However, this necessarily entails a reduced sensitivity to new physics effects.

We also see that most of the items listed above are not specific of the NPLM methodology. In particular the availability of sufficient samples of Standard Model data is a generic need of any LHC analysis, which will become more pressing with the high data statistics of the HL-LHC. Similarly, the generation of toy datasets is in principle a need for any un-binned analysis that can not rely on asymptotic formulas. Finally, learning the effect of nuisance parameters is methodologically identical to (and directly relevant for) the regression on the distribution dependence on parameters of interest, which is being studied extensively for other applications such as inference on new physics parameters. Potential limitations related with the training time are instead obviously specific of the NPLM methods. It is not excluded that the training time could be substantially reduced by a better choice of the training algorithm or of its implementation, which is an aspect we did not investigate in great detail so far. A more radical solution is to trade neural networks with non-parametric Kernel models, which are radically faster to train [39]. See Ref. [40] for an implementation of the NPLM strategy based on kernel models.

NPLM aims at the detection of unexpected manifestations of new physics, therefore its design and optimization should not be based on its sensitivity to specific new physics models. On the other hand, it would be interesting to perform an extensive study of the sensitivity to a variety of putative new physics models, possibly displaying exotic or unconventional signatures. On top of assessing the effectiveness of the strategy, this analysis might suggest new general model-independent criteria for the design of the method. Furthermore, it could clarify if and how the selection of the neural network model impacts the sensitivity. Investigations in these directions are left to future work.

In summary, NPLM emerges as a promising option for the development of a new kind of model-independent new physics searches. The extensive deployment of this type of analyses might play a vital role in experimental programs where, like at the LHC, increasingly rich experimental data are accompanied by an increasingly blurred theoretical guidance in their interpretation. Furthermore, designing NPLM analyses and addressing the corresponding challenges might trigger developments in event generation and in likelihood-free inference techniques, with broader implications on LHC physics.

Acknowledgements M.P. and G.G. are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement no. 772369). A.W. acknowledges support from the Swiss National Science Foundation under contract 200021-178999 and PRIN grants 2017FMJFMW.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: The data used for the paper (as a tar.gz file) is uploaded and available on Zenodo: https://zenodo.org/record/4442665.]

Funded by SCOAP[3].

Appendix A: Model-independent strategies

In the categorisation of model-independent LHC searches, using either machine learning techniques or more traditional statistical methods, one should first of all keep "Anomaly Detection" strategies, like the ones in Refs. [41–49], distinct from the other methods. Anomaly Detection algorithms aim at detecting outliers in the data, namely at classifying events (i.e., instances of x in $\mathcal{D}$) that are "rare" in the dataset. The great advantage of this strategy is that it relies marginally or does not rely at all on the availability of a Reference dataset because the notion of "rarity" can in principle be extracted exclusively from the observed data. The disadvantage is that its sensitivity is limited to those manifestations of new physics that take place in regions of the phase-space that are weakly populated in the Reference Model. Consider for illustration a univariate feature variable x and a new physics model producing a peak in that variable. The model can be detected if the peak falls in the extreme tail of the Reference Model distribution, because the events falling in the tail will be identified as rare and thus selected by the algorithm. If instead the peak emerges in the bulk of the Reference Model distribution, the events originating from resonance production will not be selected because they are not rare in the dataset. The sensitivity to new physics in this case can only emerge from inspecting the observed dataset as a whole, rather than searching for individual anomalous events. Similar considerations apply to realistic multivariate problems where the presence of the resonance also affects the distribution of other variables. New physics will be detected only if the resonance effects are pronounced in a region that is rare in the Reference Model distribution of these variables.

Furthermore, it should be kept in mind that the selection of "rare" events that is achieved by Anomaly Detection algorithms is only the first step of a search for new physics. The second one is to assess whether or not the observed number of selected rare events signals the presence of new physics. The first step can be achieved purely based on the observed data, but the second one requires comparing with the predictions of the Reference Model.

Anomaly Detection methods are thus more limited in scope, both in terms of new physics targets and methodologically, since they constitute only one step of the new physics search. Therefore they should be kept distinct from the strict model-independent strategies as we defined them in Sect. 1.

It should be noted that even the strict (in our definition) model-independent strategies [1,2,50–67] are not "fully" model-independent because they rely on the prior choice of the feature variables x and possibly of the region X ($x \in X$) of the phase-space where they are considered to be of interest for the analysis. A fully model-independent test should ideally employ the entire ATLAS and CMS raw datasets (i.e., the collection of detector hits) before the reconstruction of high-level objects, which is however unfeasible particularly because the trigger selection would be still an implicit bias. On the other hand it is not hard to identify physically motivated features and search regions where new physics could emerge. For instance new heavy particles and short-distance interactions would generically show up in final states with high-level reconstructed SM particles or jets with high transverse momentum. New light particles might instead be found in jets and alter their inner structure, and/or produce anomalous tracks to be found in the analysis of detector data of even lower level, etc. By restricting to the corresponding datasets one could perform new physics searches that are still way more model-independent than the search for one postulated new physics model. Clearly if many such model-independent searches were actually performed one should in principle include an estimate of the look-elsewhere effect in the assessment of the statistical significance of a putative excess. However this effect would be present, and to a larger extent, also in an agnostic interpretation of the regular model-dependent LHC searches.

In order to proceed to a finer characterization of the strict model-independent methods it is important to clarify the role and the origin of the Reference dataset $\mathcal{R}$, which is an essential element of these strategies. The availability of sufficiently accurate SM background predictions is obviously a potential concern for any LHC analysis, and in particular it is a concern for the deployment of model-independent methods. In order to emphasize this issue, it was proposed in Ref. [68] to accompany the regular notion of (signal) model-independence with the one of "background model independence" and to treat the two notions on equal footing.

As mentioned above, to perform a new physics search up to the quantification of an excess significance it is necessary to have a trustable Reference Model. The need of a Reference prediction is conceptually unavoidable in any strategy to search for "new" phenomena, as it provides the necessary notion of "old" phenomena. In absence of a Reference Model one can only perform the first step of a new physics search, as discussed for Anomaly Detection methods. We can identify

"rare" events, but we can not say whether they are present in the dataset because of new or old physics.

The need of a trustable Reference prediction is clearly not a new conclusion, it is a common requirement of any model-dependent or model-independent search ever performed. Our interpretation of the emphasis on "background model independence" given in [68] is the quest for methods with a built-in data-driven estimate of the background. We do not consider this aspect relevant, for two reasons. First, because it very commonly happens in concrete LHC final states that a data-driven background estimate is not available, and Monte Carlo simulations need to be employed for at least one of the dominant components of the background. We thus need a method that can also employ Monte Carlo background simulation, rather than being limited to data-driven estimates. Second, because there are strategies, like ours, that indeed work both with first-principle and with data-driven background estimates. Therefore it is possible, and convenient, to keep the background estimate problem separate from the development of the search strategy itself.

The regular notion of (signal) model-independence is also subject to caveats as detailed above. But is still possible to classify and rank different methodologies by their "degree of model-independence" as we did for Anomaly Detection, by trying to figure out which type of new physics signals they might or might not be sensitive to. From this viewpoint one would rank BUMPHUNTER [69] and similar strategies [68,70–76], which target resonant signals in a pre-specified variable, lower than methods with a broader target [1,2,50–67,77–80]. On the other hand, one should not employ these generic considerations to tell which one is the "right" strategy to pursue. That depends on aspects that are specific of the final state (features set and phase-space region) one is willing to explore, such as the availability (or not) of a trustable Reference sample and the actual perspectives of progress in the characterization of the data relative to more standard analysis techniques. This is why, ultimately, all these methods should be tried with data and their complementarities exploited to extract the most out of the LHC datasets.

The aim of this Appendix was to introduce some elements for the classification of model-independent strategies, not to provide an exhaustive overview of the field. Strategies like those in Ref. [81] and Ref. [82] would require a more in-depth exposition as they do not fully fall in any of our categories. The approach in Ref. [81] is to train a multi-categories classifier to tag multiple SM processes and specific putative new physics signals, with the idea that if used on the data this classifier will be sensitive to new physics models not used in training, which can be verified with numerical experiments. In Ref. [82], one employs machine learning techniques to cluster the data into categories that correspond, in the Reference hypothesis, to the different components of the background. A new category can emerge in the presence of new physics. This approach is conceptually interesting as an Anomaly Detection (dubbed "Collective Anomaly Detection" in Ref. [83]) performed on the entire dataset rather than on individual events.

References

1. R.T. D'Agnolo, A. Wulzer, Learning new physics from a machine. Phys. Rev. D **99**, 015014 (2019). https://doi.org/10.1103/PhysRevD.99.015014 arXiv:1806.02350
2. R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Learning multivariate new physics. Eur. Phys. J. C **81**, 89 (2021). https://doi.org/10.1140/epjc/s10052-021-08853-y arXiv:1912.12155
3. 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
4. L.-G. Xia, QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics. Nucl. Instrum. Methods A **930**, 15 (2019). https://doi.org/10.1016/j.nima.2019.03.088 arXiv:1810.08387
5. C. Englert, P. Galler, P. Harris, M. Spannowsky, Machine learning uncertainties with adversarial neural networks. Eur. Phys. J. C **79**, 4 (2019). https://doi.org/10.1140/epjc/s10052-018-6511-8 arXiv:1807.08763
6. V. Estrade, C. Germain, I. Guyon, D. Rousseau, Systematic aware learning—a case study in High Energy Physics. EPJ Web Conf. **214**, 06024 (2019). https://doi.org/10.1051/epjconf/201921406024
7. J.M. Clavijo, P. Glaysher, J. Jitsev, J.M. Katzy, Adversarial domain adaptation to reduce sample bias of a high energy physics event classifier*. Mach. Learn. Sci. Tech. **3**(1), 015014 (2022). https://doi.org/10.1088/2632-2153/ac3dde
8. A. Ghosh, B. Nachman, D. Whiteson, Uncertainty-aware machine learning for high energy physics. Phys. Rev. D **104**, 056026 (2021). https://doi.org/10.1103/PhysRevD.104.056026 arXiv:2105.08742
9. A. Ghosh, B. Nachman, A cautionary tale of decorrelating theory uncertainties. arXiv:2109.08159
10. M. Neal, Radford, Computing likelihood functions for high-energy physics experiments when distributions are defined by simulators with nuisance parameters
11. P. De Castro, T. Dorigo, INFERNO: inference-aware neural optimisation. Comput. Phys. Commun. **244**, 170 (2019). https://doi.org/10.1016/j.cpc.2019.06.007 arXiv:1806.04743
12. S. Wunsch, S. Jörger, R. Wolf, G. Quast, Optimal statistical inference in the presence of systematic uncertainties using neural network optimization based on binned Poisson likelihoods with nuisance parameters. Comput. Softw. Big Sci. **5**, 4 (2021). https://doi.org/10.1007/s41781-020-00049-5 arXiv:2003.07186
13. Particle Data Group Collaboration, P. Zyla et al., Review of particle physics. PTEP **2020**, 083C01 (2020). https://doi.org/10.1093/ptep/ptaa104
14. J. Neyman, E.S. Pearson, On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. A **231**, 289 (1933). https://doi.org/10.1098/rsta.1933.0009
15. S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60 (1938). https://doi.org/10.1214/aoms/1177732360
16. A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. **54**, 426 (1943). https://doi.org/10.2307/1990256
17. K. Cranmer, J. Pavez, G. Louppe, Approximating likelihood ratios with calibrated discriminative classifiers. arXiv:1506.02169

18. P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, D. Whiteson, Parameterized neural networks for high-energy physics. Eur. Phys. J. C **76**, 235 (2016). https://doi.org/10.1140/epjc/s10052-016-4099-4 arXiv:1601.07913
19. J. Brehmer, G. Louppe, J. Pavez, K. Cranmer, Mining gold from implicit models to improve likelihood-free inference. Proc. Natl. Acad. Sci. **117**, 5242 (2020). https://doi.org/10.1073/pnas.1915980117 arXiv:1805.12244
20. J. Brehmer, F. Kling, I. Espejo, K. Cranmer, MadMiner: machine learning-based inference for particle physics. Comput. Softw. Big Sci. **4**, 3 (2020). https://doi.org/10.1007/s41781-020-0035-2 arXiv:1907.10621
21. S. Chen, A. Glioti, G. Panico, A. Wulzer, Parametrized classifiers for optimal EFT sensitivity. JHEP **05**, 247 (2021). https://doi.org/10.1007/JHEP05(2021)247 arXiv:2007.10356
22. S. Chen, A. Glioti, G. Panico, A. Wulzer, Boosted likelihood learning from event re-weighting, to appear (2021)
23. S. Chen, A. Glioti, G. Panico, A. Wulzer, Learning systematic uncertainties, to appear (2021)
24. G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics. Eur. Phys. J. C **71**, 1554 (2011). https://doi.org/10.1140/epjc/s10052-011-1554-0 arXiv:1007.1727
25. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro et al., TensorFlow: large-scale machine learning on heterogeneous systems (2015)
26. G. Grosso, New physics learning machine (NPLM): package, 11 (2021). https://github.com/GaiaGrosso/NPLM_package
27. J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer et al., The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. JHEP **07**, 079 (2014). https://doi.org/10.1007/JHEP07(2014)079 arXiv:1405.0301
28. T. Sjostrand, S. Mrenna, P.Z. Skands, PYTHIA 6.4 physics and manual. JHEP **05**, 026 (2006). https://doi.org/10.1088/1126-6708/2006/05/026 arXiv:hep-ph/0603175
29. DELPHES 3 Collaboration, J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens et al., DELPHES 3, a modular framework for fast simulation of a generic collider experiment. JHEP **02**, 057 (2014). https://doi.org/10.1007/JHEP02(2014)057. arXiv:1307.6346
30. G. Grosso, R.T. D'Agnolo, M. Pierini, A. Wulzer, M. Zanetti, NPLM: learning multivariate new physics (2021). https://doi.org/10.5281/zenodo.4442665
31. CMS Collaboration, A.M. Sirunyan et al., Performance of the CMS muon detector and muon reconstruction with proton–proton collisions at $\sqrt{s} = 13$ TeV. JINST **13**, P06015 (2018). https://doi.org/10.1088/1748-0221/13/06/P06015. arXiv:1804.04528
32. CMS Collaboration, V. Khachatryan et al., Performance of electron reconstruction and selection with the CMS detector in proton–proton collisions at $\sqrt{s} = 8$ TeV. JINST **10**, P06005 (2015). https://doi.org/10.1088/1748-0221/10/06/P06005. arXiv:1502.02701
33. CMS Collaboration, A.M. Sirunyan et al., Particle-flow reconstruction and global event description with the CMS detector. JINST **12**, P10003 (2017). https://doi.org/10.1088/1748-0221/12/10/P10003. arXiv:1706.04965
34. CMS Collaboration, S. Chatrchyan et al., Performance of tau-lepton reconstruction and identification in CMS. JINST **7**, P01001 (2012). https://doi.org/10.1088/1748-0221/7/01/P01001. arXiv:1109.6034
35. CMS Collaboration, A.M. Sirunyan et al., Search for a narrow resonance lighter than 200 GeV decaying to a pair of muons in proton–proton collisions at $\sqrt{s} =$ TeV. Phys. Rev. Lett. **124**, 131802 (2020). https://doi.org/10.1103/PhysRevLett.124.131802. arXiv:1912.04776
36. LHCb Collaboration, R. Aaij et al., Search for dark photons produced in 13 TeV pp collisions. Phys. Rev. Lett. **120**, 061801 (2018). https://doi.org/10.1103/PhysRevLett.120.061801. arXiv:1710.02867
37. C. Chen, O. Cerri, T.Q. Nguyen, J.-R. Vlimant, M. Pierini, Data augmentation at the LHC through analysis-specific fast simulation with deep learning. arXiv:2010.01835
38. K. Hagiwara, J. Kanzaki, Q. Li, N. Okamura, T. Stelzer, Fast computation of MadGraph amplitudes on graphics processing unit (GPU). Eur. Phys. J. C **73**, 2608 (2013). https://doi.org/10.1140/epjc/s10052-013-2608-2 arXiv:1305.0708
39. G. Meanti, L. Carratino, L. Rosasco, A. Rudi, Kernel methods through the roof: handling billions of points efficiently (2020). arXiv preprint. arXiv:2006.10350
40. M. Letizia, G. Losapio, M. Rando, G. Grosso, L. Rosasco, Efficient kernel methods for model-independent new physics searches. [NeurIPS ML4PS 2021 146]
41. A. Blance, M. Spannowsky, P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches. JHEP **10**, 047 (2019). https://doi.org/10.1007/JHEP10(2019)047 arXiv:1905.10384
42. O. Knapp, O. Cerri, G. Dissertori, T.Q. Nguyen, M. Pierini, J.-R. Vlimant, Adversarially learned anomaly detection on CMS open data: re-discovering the top quark. Eur. Phys. J. Plus **136**, 236 (2021). https://doi.org/10.1140/epjp/s13360-021-01109-4 arXiv:2005.01598
43. T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, T. Golling, Variational autoencoders for anomalous jet tagging. arXiv:2007.01850
44. T.S. Roy, A.H. Vijay, A robust anomaly finder based on autoencoders. arXiv:1903.02032
45. T. Heimel, G. Kasieczka, T. Plehn, J.M. Thompson, QCD or what? SciPost Phys. **6**, 030 (2019). https://doi.org/10.21468/SciPostPhys.6.3.030 arXiv:1808.08979
46. O. Cerri, T.Q. Nguyen, M. Pierini, M. Spiropulu, J.-R. Vlimant, Variational autoencoders for new physics mining at the large hadron collider. JHEP **05**, 036 (2019). https://doi.org/10.1007/JHEP05(2019)036 arXiv:1811.10276
47. M. Farina, Y. Nakai, D. Shih, Searching for new physics with deep autoencoders. Phys. Rev. D **101**, 075021 (2020). https://doi.org/10.1103/PhysRevD.101.075021 arXiv:1808.08992
48. J.A. Aguilar-Saavedra, J.H. Collins, R.K. Mishra, A generic anti-QCD jet tagger. JHEP **11**, 163 (2017). https://doi.org/10.1007/JHEP11(2017)163 arXiv:1709.01087
49. J.A. Aguilar-Saavedra, Anomaly detection from mass unspecific jet tagging. Eur. Phys. J. C **82**(2), 130 (2022). https://doi.org/10.1140/epjc/s10052-022-10058-w
50. D0 Collaboration, B. Abbott et al., Search for new physics in eμX data at DØ using SLEUTH: a quasi-model-independent search strategy for new physics. Phys. Rev. D **62**, 092004 (2000). https://doi.org/10.1103/PhysRevD.62.092004. arXiv:hep-ex/0006011
51. D0 Collaboration, B. Abbott et al., A quasi-model-independent search for new high p_T physics at D\O. Phys. Rev. Lett. **86**, 3712 (2001). https://doi.org/10.1103/PhysRevLett.86.3712. arXiv:hep-ex/0011071
52. D0 Collaboration, V.M. Abazov et al., A quasi model independent search for new physics at large transverse momentum. Phys. Rev. D **64**, 012004 (2001). https://doi.org/10.1103/PhysRevD.64.012004. arXiv:hep-ex/0011067
53. D0 Collaboration, V.M. Abazov et al., Model independent search for new phenomena in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV. Phys. Rev. D **85**, 092015 (2012). https://doi.org/10.1103/PhysRevD.85.092015. arXiv:1108.5362
54. H1 Collaboration, A. Aktas et al., A general search for new phenomena in ep scattering at HERA. Phys. Lett. B

602, 14 (2004). https://doi.org/10.1016/j.physletb.2004.09.057. arXiv:hep-ex/0408044
55. H1 Collaboration, F.D. Aaron et al., A general search for new phenomena at HERA. Phys. Lett. B **674**, 257 (2009). https://doi.org/10.1016/j.physletb.2009.03.034. arXiv:0901.0507
56. CDF Collaboration, T. Aaltonen et al., Model-independent and quasi-model-independent search for new physics at CDF. Phys. Rev. D **78**, 012002 (2008). https://doi.org/10.1103/PhysRevD.78.012002. arXiv:0712.1311
57. CDF Collaboration, T. Aaltonen et al., Global search for new physics with 2.0 fb^{-1} at CDF. Phys. Rev. D **79**, 011101 (2009). https://doi.org/10.1103/PhysRevD.79.011101. arXiv:0809.3781
58. CMS Collaboration, Model unspecific search for new physics in pp collisions at $\sqrt{(s)} = 7$ TeV. CMS-PAS-EXO-10-021
59. CMS Collaboration, MUSIC—an automated scan for deviations between data and Monte Carlo simulation. CMS-PAS-EXO-08-005
60. CMS Collaboration, MUSiC, a model unspecific search for new physics, in pp collisions at $\sqrt{s} = 8$ TeV. CMS-PAS-EXO-14-016
61. CMS Collaboration, A.M. Sirunyan et al., MUSiC: a model unspecific search for new physics in proton–proton collisions at $\sqrt{s} = 13$ TeV. arXiv:2010.02984
62. ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{(s)} = 7$ TeV
63. ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV
64. ATLAS Collaboration, A model independent general search for new phenomena with the ATLAS detector at $\sqrt{s} = 13$ TeV
65. ATLAS Collaboration, M. Aaboud et al., A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment. Eur. Phys. J. C **79**, 120 (2019). https://doi.org/10.1140/epjc/s10052-019-6540-y. arXiv:1807.07447
66. J. Alwall, M.-P. Le, M. Lisanti, J.G. Wacker, Model-independent jets plus missing energy searches. Phys. Rev. D **79**, 015005 (2009). https://doi.org/10.1103/PhysRevD.79.015005 arXiv:0809.3264
67. T. Dorigo, M. Fumanelli, C. Maccani, M. Mojsovska, G.C. Strong, B. Scarpa, RanBox: anomaly detection in the copula space. arXiv:2106.05747
68. B. Nachman, D. Shih, Anomaly detection with density estimation. Phys. Rev. D **101**, 075042 (2020). https://doi.org/10.1103/PhysRevD.101.075042 arXiv:2001.04990
69. G. Choudalakis, On hypothesis testing, trials factor, hypertests and the BumpHunter, in *PHYSTAT 2011*, 1 (2011). arXiv:1101.0390
70. E.M. Metodiev, B. Nachman, J. Thaler, Classification without labels: learning from mixed samples in high energy physics. JHEP **10**, 174 (2017). https://doi.org/10.1007/JHEP10(2017)174 arXiv:1708.02949
71. J.H. Collins, K. Howe, B. Nachman, Anomaly detection for resonant new physics with machine learning. Phys. Rev. Lett. **121**, 241803 (2018). https://doi.org/10.1103/PhysRevLett.121.241803 arXiv:1805.02664
72. J.H. Collins, K. Howe, B. Nachman, Extending the search for new resonances with machine learning. Phys. Rev. D **99**, 014038 (2019). https://doi.org/10.1103/PhysRevD.99.014038 arXiv:1902.02634
73. A. Andreassen, B. Nachman, D. Shih, Simulation assisted likelihood-free anomaly detection. Phys. Rev. D **101**, 095004 (2020). https://doi.org/10.1103/PhysRevD.101.095004 arXiv:2001.05001
74. K. Benkendorfer, L.L. Pottier, B. Nachman, Simulation-assisted decorrelation for resonant anomaly detection. Phys. Rev. D **104**(3), 035003 (2021). https://doi.org/10.1103/PhysRevD.104.035003
75. O. Amram, C.M. Suarez, Tag N' train: a technique to train improved classifiers on unlabeled data. JHEP **01**, 153 (2021). https://doi.org/10.1007/JHEP01(2021)153 arXiv:2002.12376
76. ATLAS Collaboration, G. Aad et al., Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector. Phys. Rev. Lett. **125**, 131801 (2020). https://doi.org/10.1103/PhysRevLett.125.131801. arXiv:2005.02983
77. M. Kuusela, T. Vatanen, E. Malmi, T. Raiko, T. Aaltonen, Y. Nagai, Semi-supervised anomaly detection—towards model-independent searches of new physics. J. Phys. Conf. Ser. **368**, 012032 (2012). https://doi.org/10.1088/1742-6596/368/1/012032 arXiv:1112.3329
78. A. De Simone, T. Jacques, Guiding new physics searches with unsupervised learning. Eur. Phys. J. C **79**, 289 (2019). https://doi.org/10.1140/epjc/s10052-019-6787-3 arXiv:1807.06038
79. P. Chakravarti, M. Kuusela, J. Lei, L. Wasserman, Model-independent detection of new physics signals using interpretable semi-supervised classifier tests. arXiv:2102.07679
80. K.T. Matchev, P. Shyamsundar, J. Smolinsky, A quantum algorithm for model independent searches for new physics. arXiv:2003.02181
81. S.E. Park, D. Rankin, S.-M. Udrescu, M. Yunus, P. Harris, Quasi anomalous knowledge: searching for new physics with embedded knowledge. JHEP **21**, 030 (2020). https://doi.org/10.1007/JHEP06(2021)030 arXiv:2011.03550
82. A. Casa, G. Menardi, Nonparametric semisupervised classification for signal detection in high energy physics. arXiv:1809.02977
83. V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey. ACM Comput. Surv. (CSUR) **41**, 1 (2009). https://doi.org/10.1145/1541880.1541882