Multiple testing for signal-agnostic searches for new physics with machine learning

Citation: Grosso, G., Letizia, M. Multiple testing for signal-agnostic searches for new physics with machine learning. Eur. Phys. J. C 85, 4 (2025).

As Published: https://doi.org/10.1140/epjc/s10052-024-13722-5

Publisher: Springer Berlin Heidelberg

Persistent URL: https://hdl.handle.net/1721.1/157946

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Eur. Phys. J. C (2025) 85:4
https://doi.org/10.1140/epjc/s10052-024-13722-5

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Experimental Physics

Multiple testing for signal-agnostic searches for new physics with machine learning

Gaia Grosso[1,2,3], **Marco Letizia**[4,5,a]

[1] NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Cambridge, MA, USA
[2] MIT Laboratory for Nuclear Science, Cambridge, MA, USA
[3] School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[4] MaLGa-DIBRIS, University of Genoa, Genoa, Italy
[5] INFN, Sezione di Genova, Genoa, Italy

Received: 25 September 2024 / Accepted: 17 December 2024

Abstract In this work, we address the question of how to enhance signal-agnostic searches by leveraging multiple testing strategies. Specifically, we consider hypothesis tests relying on machine learning, where model selection can introduce a bias towards specific families of new physics signals. Focusing on the New Physics Learning Machine, a methodology to perform a signal-agnostic likelihood-ratio test, we explore a number of approaches to multiple testing, such as combining p-values and aggregating test statistics. Our findings show that it is beneficial to combine different tests, characterised by distinct choices of hyperparameters, and that performances comparable to the best available test are generally achieved, while also providing a more uniform response to various types of anomalies. This study proposes a methodology that is valid beyond machine learning approaches and could in principle be applied to a larger class model-agnostic analyses based on hypothesis testing.

1 Introduction

After decades of experimental results that contributed to the development and confirmation of the Standard Model of particle physics (SM), we are in a phase in which no compelling theoretical prediction is guiding experimental searches. It is therefore important to develop model-independent analyses that are potentially sensitive to new physics effects not necessarily predicted by any specific Beyond the Standard Model (BSM) scenario. This is an extraordinary difficult task given the complexity of collider data and the fact that new physics can manifest itself as a deviation from the SM predictions in infinitely many ways. Moreover, these effects are expected to be extremely rare (poor signal-to-background ratio) and/or hidden (uncommon observables).

Several proposals for partial model-independent analyses have been applied to experimental data. Early instances, such as those in [1–5], were based on simplifying assumptions about the way new physics effects could appear in the measurements and, as a consequence, they were limited to a selection of interesting final states. Crucially, these methodologies focused on theoretically motivated high-level features to reduce dimensionality and resorted to traditional statistical techniques.

Recently, machine learning (ML) has been leveraged to design flexible and multivariate data-driven tests, further enhancing signal-agnostic strategies. Various approaches have been proposed over the past few years (the reader can find an exhaustive review in [6]), some of which have already been applied to experimental data (see [7] and [8]). Despite their potential, the adoption of these techniques introduces new challenges, particularly in understanding how model selection can impact sensitivity and bias the analysis towards specific signal hypotheses.

Here, we address this topic considering as a case study the model introduced in [9], where classifiers based on efficient kernel methods [10] are used to design a multivariate and unbinned likelihood-ratio test in which the alternative hypothesis is derived from data. This idea (dubbed The New Physics Learning Machine, NPLM for brevity) was initially presented in [11] using neural networks. The approach to hyperparameters selection proposed in [9] is based on a mix of statistical and heuristic criteria that was shown to work well empirically on a number of benchmarks (see also [12,13]). However, it is not guaranteed that the result-

[a] e-mail: marco.letizia@edu.unige.it (corresponding author)

Published online: 04 January 2025

ing model has optimal performance or a uniform response over a wide range of possible deviations from the reference expectation. It is then natural to ask whether it is possible to explore more principled and controlled approaches to model selection with the purpose of improving sensitivity as well as robustness.

In this work, we propose to improve model-independent searches by exploiting the idea of multiple testing [14,15]. Instead of selecting a single test based on a specific learning model, we propose to define multiple ones characterised by different choices of hyperparameters, implicitly defining different alternative hypotheses, and combine their outputs into a single meta-analysis while accounting for the *look-elsewhere effect*, commonly known as the *multiple comparison problem* in the statistic literature. Similar strategies have been proposed in recent studies [16,17], in the context of a class of kernel-based two-sample tests known as maximum mean discrepancy. We compare different methods and show that this framework results in a more uniform response over new physics effects of different nature, and that the achieved performance is comparable or close to the best model, which is not known a priori in real use cases.[1] Our findings suggest that exploiting multiple testing can be a powerful tool to enhance model-agnostic searches for new physics, beyond the challenge of hypeparameter tuning in ML-based strategies.

The paper is organised as follows. In Sect. 2 we recall the underlying statistical framework and revise the NPLM approach to hypothesis testing in its implementation based on kernel methods, with a focus on the model selection pipeline. In Sect. 3 we review the multiple testing problem, explore some approaches to address it and introduce how to integrate them in the NPLM methodology. Section 4 is dedicated to numerical experiments. Concluding remarks are given in Sect. 5.

2 The search for new physics as a signal-agnostic hypothesis test

Let us consider a set $\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}$ of realisations of a random variable $x \in \mathcal{X} \subseteq \mathbb{R}^d$ representing experimental measurements, independent and identically distributed according to an unknown true distribution $p_{\text{true}}(x)$. We call *reference distribution*, $p(x|R)$, the distribution of the events as predicted by a reference model R (in our case, the SM). Additionally, in high energy collider physics, the number of collected events $N_\mathcal{D}$ is also a random variable, following a Poisson distribution characterised by a true expected value $N(\text{true})$. We name $N(R)$ the expected number of collected events as predicted by the reference model. Ideally, the analysis should also be sensitive to discrepancies between these two values. This can be formalised by introducing the quantity

$$n(x|\cdot) \;=\; N(\cdot)p(x|\cdot), \tag{1}$$

namely the probability density normalised to the associated expected number of events for any given physical theory. The goal of a signal-agnostic test is to determine whether $p(x|R)$ is a good description of the data without introducing alternative models. In statistical terms, this can be framed as a *goodness-of-fit* (GoF) test. However, the reference distribution is commonly not available in closed form. In this work, we consider the situation in which a *reference sample* $\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}$ can be obtained by simulations or with measurements from a control region. The problem is then to assess the goodness of the population-level null hypothesis

$$H_0 : n_{\text{true}}(x) = n(x|R) \tag{2}$$

from finite data, by comparing $\mathcal{D}$ with $\mathcal{R}$. A task of this kind is commonly known in statistics as a *two-sample test*. In this framework, the alternative hypothesis is simply the negation of the null

$$H_1 : n_{\text{true}}(x) \neq n(x|R). \tag{3}$$

In order to have an accurate description of the reference distribution, we assume that $N_\mathcal{R} \gg N_\mathcal{D}$.

A two-sample hypothesis test requires to choose a test statistic, namely a function

$$t : \mathcal{X}^{N_\mathcal{D}} \times \mathcal{X}^{N_\mathcal{R}} \to \mathbb{R} \tag{4}$$

that maps the measured data and the reference sample to a measure of their compatibility defined as a real number $t_{\text{obs}} = t(\mathcal{R}, \mathcal{D})$. To establish the statistical significance of the outcome of the test, a p-value is computed. This quantity is the probability, under the null hypothesis, of observing values that are at least as extreme as the measured ones

$$p_{\text{obs}} = P(t \geq t_{\text{obs}}|H_0). \tag{5}$$

The observed p-value is then compared to a predefined threshold $\alpha \in [0, 1]$, representing the highest acceptable rate of false positives associated with the test, i.e. the probability of rejecting the null hypothesis if true. A discovery is claimed if $p_{\text{obs}} < \alpha$. These probabilities can be mapped to *Z-scores* using the following expression

$$\mathrm{Z} = \Phi^{-1}(1 - p), \tag{6}$$

where Φ^{-1} is the quantile function of a standard Gaussian distribution. Different tests are compared by evaluating their

[1] The code to reproduce our results can be found in https://github.com/mletizia/multiple-testing-nplm.

power, namely their rate of true positives at the critical value t_α

$$\alpha = P(t > t_\alpha | H_0), \tag{7}$$

$$\text{power} = P(t \geq t_\alpha | H_1). \tag{8}$$

Given α, the best test is the one maximising the power with a false positive rate at most equal to α. In order to be able to estimate the rate of true positive, the alternative hypothesis H_1 needs to be realised concretely. In this work, we consider a number of scenarios from the HEP literature [11,13,18,19]. However, when performing a two-sample test on real measurements, the result of the analysis would be reported as the observed p-value defined in Eq. (5).

2.1 The NPLM methodology

NPLM is an approach to signal-agnostic hypothesis testing based on machine learning that aims at approximating the maximum-likelihood-ratio test as defined by Neyman and Pearson [20]. It is based on the idea of introducing a local deformation of the reference distribution (as defined in Eq. (1))

$$n_w(x) = e^{f_w(x)} n(x|R), \tag{9}$$

with $\mathcal{F} = \{f_w\}$ a rich family of functions parametrised by w. In [9] and in this work we consider kernel methods, for which the function f_w is expresses as the following weighted sum

$$f_w(x) = \sum_{i=1}^{N} w_i k_\sigma(x, x_i), \tag{10}$$

with the parameters w to be selected from data and $N = N_\mathcal{R} + N_\mathcal{D}$ the total number of data points. Specifically, we use a Gaussian kernel

$$k_\sigma(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right), \tag{11}$$

where σ is the kernel width, a hyperparameter. The resulting space of functions allows to approximate any continuous function given enough data. This approach is powerful but limited by large computational requirements. To solve this problem we use Falkon [10], a modern solver for large-scale kernel methods which replaces Eq. (10) with

$$f_w(x) = \sum_{i=1}^{M} w_i k_\sigma(x, x_i), \tag{12}$$

where $\{\tilde{x}_1, ..., \tilde{x}_M\}$ are called Nyström centres and are sampled uniformly at random from the input data, with M a hyperparameter. The corresponding solution can be shown to be with high probability as accurate as the exact one (see [21] and references therein). In practice, the optimal parameters $\hat{w}$ are learned from data with a supervised classifier trained to separate $\mathcal{R}$ from $\mathcal{D}$ by minimising the following empirical risk

$$\frac{1}{N} \sum_i \ell(y_i, f_w(x_i)) + \lambda R(f_w), \tag{13}$$

based on a weighted logistic loss

$$\begin{aligned} \ell(y, f_w(x)) = \; & (1 - y) \frac{N(R)}{N_\mathcal{R}} \log\left(1 + e^{f_w(x)}\right) \\ & + y \log\left(1 + e^{-f_w(x)}\right), \end{aligned} \tag{14}$$

with $y = 0$ if $x \in \mathcal{R}$ and $y = 1$ if $x \in \mathcal{D}$. This loss can be shown (see [9]) to have the correct target function

$$f_{\hat{w}}(x) \approx f^*(x) = \arg\min_f \mathbb{E}\left[\ell(y, f(x))\right] = \log \frac{n_{\text{true}}(x)}{n(x|R)}, \tag{15}$$

meaning that the desired function (in this case the ratio of the data-generating densities) is recovered in the limit of infinite data. The second term in Eq. (13) is a regularisation term

$$R(f_w) = \sum_{ij} w_i w_j k_\sigma(x_i, x_j). \tag{16}$$

constraining the complexity of the model. The problem defined in Eq. (13) is then solved by an approximate Newton method, as discussed in detail in [10].

At the end of training, the model is evaluated in-sample on the whole dataset with the following metric

$$t_{\text{obs}}(\mathcal{D}) = -2\left[\frac{N(R)}{N_\mathcal{R}} \sum_{x \in \mathcal{R}} \left(e^{f_{\hat{w}}(x)} - 1\right) - \sum_{x \in \mathcal{D}} f_{\hat{w}}(x)\right], \tag{17}$$

which is derived from the extended likelihood-ratio (see [9,11,22]). To simplify the notation, we omit the dependence of Eq. (17) on the reference sample $\mathcal{R}$ and on the learned parameters $\hat{w}$. This method allows to leverage the Neyman–Pearson approach to hypothesis testing with a data driven alternative hypothesis, without the need to specify it a priori. The connection between goodness-of-fit tests and the Neyman–Pearson construction at the core of NPLM was discussed earlier in [23] and more recently in [13]. The latter contribution also include comparisons with other standard metrics and methods commonly used in statistics and machine learning, such as the binned χ^2 test, the

Kolmogorov–Smirnov test, the aurea under the ROC curve and classifier two-sample tests [24].

2.1.1 Model selection

Falkon possesses three main hyperparameters: the number of centres M, the kernel width σ and the regularisation parameter λ. These are tuned only on reference data to avoid biases toward specific anomalous features that might be present in the measurements $\mathcal{D}$. Following [9,12], they are selected as follows:

- The Gaussian width σ is selected as the 90th percentile of the pairwise distance among reference-distributed data points. Heuristics of this type are common for kernel methods, see for instance [25].
- To achieve optimal statistical bounds and preserve performance, the number of centres M must be at least be of order $\sqrt{N}$, as discussed in [26]. Studies presented in [9] suggest that values close to the number of data points $N_{\mathcal{D}}$ in the measurements work well but can be reduced for a faster training.
- The regularisation parameter λ is kept as small as possible while maintaining a stable training [26].

As a consequence of these criteria, we consider the kernel width as the main hyperparameter that regulates the complexity of the model and sets the typical scale of the problem. Indeed, it is easy to show that if σ is small the model tends to overfit while, if large, it behaves as a linear model.[2] In the context of two-sample testing, the specific choice of σ has a crucial impact on the families of alternative hypotheses that are effectively explored by the test, as we will discuss in Sects. 3 and 4 (see also [16,17]).

2.1.2 Single test at fixed hyperparameters

Given a particular set of hyperparameters $\theta^* = (M^*, \sigma^*, \lambda^*)$, a single test proceeds as follows. As a first step, the model is trained on the reference sample $\mathcal{R}$ and the measurements $\mathcal{D}$, returning the value of the observed test statistic $t_{\text{obs}} = t_{\text{obs}}(\mathcal{D})$, as given by Eq. (17). Next, the distribution of the test statistic under the null hypothesis $p(t|H_0)$ is estimated empirically. There are different ways to do it. We consider here the scenario in which the reference model can be sampled at will via simulations. Therefore, we re-train the NPLM model from scratch on the reference sample $\mathcal{R}$ and multiple ($N_{\text{toys}}^{(H_0)}$) reference-distributed samples $\mathcal{D}_i^{(R)}$, mimicking measurements in the absence of new physics. Each test returns a value $t_i = t(\mathcal{D}_i^{(R)})$. The collection of test statistics $\{t_i\}_{i=1}^{N_{\text{toys}}^{(H_0)}}$ is used to empirically estimate the p-value as (see [27])

$$\hat{p}_{\text{obs}} = \frac{1}{N_{\text{toys}}^{(H_0)} + 1} \left[\sum_{i=1}^{N_{\text{toys}}^{(H_0)}} \mathbb{1}(t_i - t_{\text{obs}}) + 1 \right], \tag{18}$$

where $\mathbb{1}(x)$ is the Heaviside step function, which is zero when $x < 0$ and one otherwise. It is worth stressing that the result of the test is implicitly conditioned on the selected hyperparameters.

3 Multiple tests for robust detection

3.1 The multiple testing problem

In two-sample testing, one is interested in determining whether the null hypothesis that two samples are drawn from the same probability distribution can be rejected. The alternative hypothesis is the negation of the null and no assumption is made about how the data-generating distributions might differ. In practice, a specific test statistic has to be chosen to formulate a concrete procedure and this will in general bias the test towards specific hypotheses. For example, both the Kolmogorov-Smirnov and the Anderson–Darling tests [28] are viable options for a non-parametric test. However, the latter is more sensitive to discrepancies in the tails of the distributions. It is therefore logical to explore the possibility of conducting multiple tests to enhance the likelihood of detection.

The problem of multiple testing, also known as the *look-elsewhere effect* in the HEP literature (see [29–31]), arises in this type of scenarios. Each individual test outputs a p-value. Naively, it would be ideal to simply retain the test returning the smallest p-value, associated with the highest detected degree of discrepancy. However it is not correct to simply compare the smallest p-value to the desired false positive rate α. Indeed, it is crucial to take into account the fact that we are (at least implicitly) testing different hypotheses simultaneously, resulting in an increased possibility of having at least one false detection among the collection of considered tests (see [32]). To address this problem, several methods to combine tests into a single meta-test have been explored in the literature. In common settings, all tests are designed to be sensitive to a specific signal of interest, i.e. they all share the same alternative hypothesis, and are applied to sets of independent measurements. The reader can find in [33] an overview of the most common approaches based on combining p-values, and theoretical arguments on their optimality given a specific class of alternative hypothesis.

[2] It is however worth highlighting that, in general, all three hyperparameters act as regularisers.

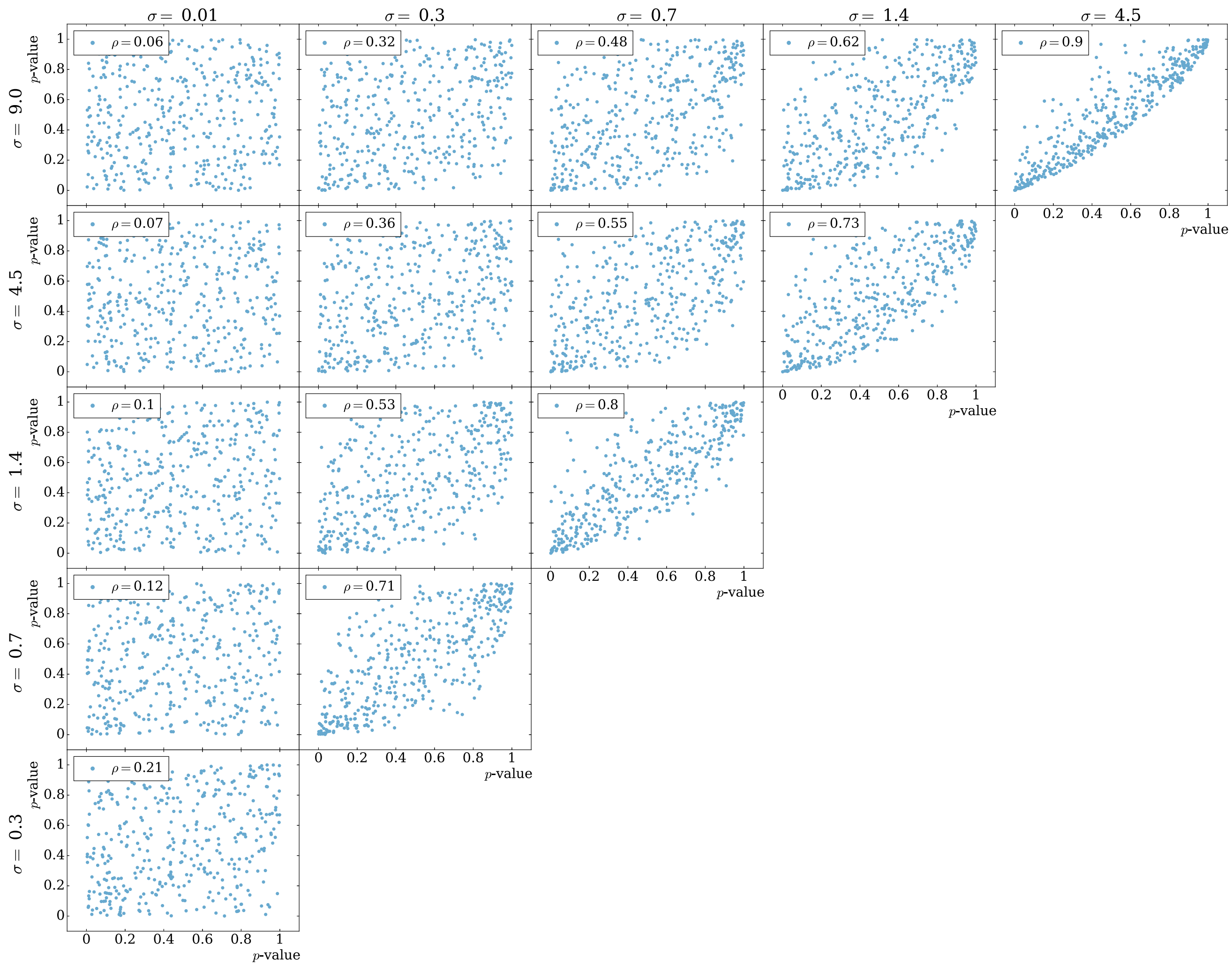

Fig. 1 EXPO-1D – Corner plots showing correlations between the p-values obtained from different tests in the background-only hypothesis. The Pearson's correlation (ρ) is reported in the legend

In this work, we are interested in the case in which multiple tests are performed to explore different hypotheses (i.e. different types of new physics signals) on the same set of measurements, hence with a potentially high degree of correlation. An example of this scenario can be found in [34], in the context of common goodness-of-fit tests in one dimension. Here, we employ multiple testing strategies to reduce the bias in the anomaly detection task caused by specific hyperparameter choices in the machine learning model powering the NPLM test.

3.2 Designing multiple tests for NPLM

As elaborated at the end of 2.1.1, the choice of the kernel width σ in Eq. (11) introduces a bias towards specific families of anomalous effects potentially present in the measurements, with respect to the reference prediction (see also Sect. 4.2.1 and Fig. 1). Following the previous discussion, it would then be ideal to consider multiple NPLM tests defined by different values of σ to explore various types of alternative hypotheses on the same measurements.

We proceed by choosing a set of unique values $\Sigma = \{\sigma_i\}_{i=1}^n$, defining the following set of tests

$$T = \{t^{(\sigma_i)} | \sigma_i \in \Sigma\}_{i=1}^n, \tag{19}$$

while the other hyperparameters of the method, M and λ, are kept fixed. In this regard, it is important to realise that values of σ that are close will give rise to highly correlated tests, while far apart values will result in less correlated tests.

Additionally, since the test considered here is based on a learning model, it will be highly adaptive to the data, potentially increasing correlation among tests. To ensure a more uniform performance across different anomalous scenarios and decrease correlation, few well-separated values of σ are preferable. Following a standard practice in kernel methods (see for instance [25]) and similarly to the original proposal presented in [9], we select them as percentiles of the distribution of the pairwise distance in a set of reference-distributed data points, after proper feature rescaling. This provides an estimate of the relevant scales in the problem. However, it might be beneficial to also include larger values to consider possible long-range effects. The number of tests $n = |\Sigma|$ is a free parameter of the algorithm. In choosing it, one should keep in mind that performing an extensive number of tests is computationally more demanding, although this cost can in principle be amortised with adequate distributed computing strategies. On the other hand, this could ultimately have a negative impact on the sensitivity if, as n grows, the rate of false positives increases faster than the rate of true positives.

3.3 Aggregation methods

We explore various options to combine tests based on the existing literature, and discuss the benefits and disadvantages of each of them given the design choices outlined in Sect. 3.2. Specifically, we consider the following meta-test statistics:

min-*p* Introduced in [35], the meta-test statistic is defined as the smallest individual p-value

$$p_{\text{min}} = -\log \min_{\sigma \in \Sigma} p^{(\sigma)}. \tag{20}$$

prod-*p* Following [36], the meta-test statistic is defined as the log-scaled product of the individual p-values

$$p_{\text{prod}} = -\sum_{\sigma \in \Sigma} \log p^{(\sigma)}. \tag{21}$$

avg-*p* Similarly to [37], the individual p-values are averaged as

$$p_{\text{avg}} = -\frac{1}{n} \sum_{\sigma \in \Sigma} p^{(\sigma)}. \tag{22}$$

smax-*t* Inspired by [17], the test statistics are combined directly via the following smooth maximum function

$$t_{\text{smax}} = T \log \frac{1}{n} \sum_{\sigma \in \Sigma} e^{t^{(\sigma)}/T}, \tag{23}$$

where $T \in \mathbb{R}_{>0}$ plays the role of a temperature. For $T \to 0$ it corresponds to the maximum, while for $T \to \infty$ it reduces to the arithmetic mean (see [38]). We fix $T = 1$, unless specified otherwise.

HB The Holm–Bonferroni method [39] is a standard approach for correcting p-values to address the multiple comparison problem in statistics that is uniformly more powerful than the Bonferroni correction. It proceeds by first sorting the observed p-values in decreasing order. Then, starting from the largest one, each p-value is compared with an associated adjusted threshold, according to the following formula

$$p^{(\sigma_i)} \leq \frac{\alpha}{n + 1 - i}, \tag{24}$$

where the index i runs over the ordered set (lowest-to-highest) of p-values $\{p^{(\sigma_1)} \leq p^{(\sigma_2)} \leq \cdots \leq p^{(\sigma_n)}\}$ and n is the total number of tests. The null hypothesis is rejected as soon as one p-value satisfies the inequality.

Choosing the optimal method without prior knowledge on the type of signal potentially in the measured data is generally not a solvable problem for composite hypotheses. However, the specificities of these tests can be used as a guide to isolate the most promising options.

The avg-p method assigns uniform weights to all the tests. This is generally a good choice if the tests are expected to perform similarly. This is not necessarily the case for the NPLM set of tests, since the choice of Σ is made such that the overlap between families of alternatives is small.

The log-scaled product of p-values (prod-p) allows to direct the combination focus toward the smallest p-values. This can be a good choice if a subgroup of the tests performs well relative to the others, as it allows to enhance their contributions to the sum.

The minimum over p-values (min-p) is intuitively the best solution if a specific test is expected to perform significantly better than the others.

The typical values of the NPLM test statistic strongly depend on the complexity of the model. In particular, for any given set of data, $t^{(\sigma_1)} > t^{(\sigma_2)}$ if $\sigma_1 < \sigma_2$, as also observed in previous studies [9,11,18]. The result of combining tests via smax-t is therefore equivalent to selecting the test statistic with smallest value of σ. Therefore, we do not expect this strongly biased strategy to work well in our study.

Estimating the level of correlation among NPLM tests is thus crucial to identify the best aggregation method. A signal-agnostic strategy to address this task is to inspect the pairwise correlation under the null hypothesis, i.e. when detecting statistical fluctuations in background-only samples $\mathcal{D}_i^{(R)}$. We will give practical examples within the scope of our numerical experiments in Sect. 4. Finally, it is worth emphasizing how each of these aggregation methods, besides the HB approach, does not require any explicit correction to the p-values to account for the look-elsewhere effect. However, the latter still manifest itself in the fact that while more agnostic tests generally deliver a more uniform response to different alternative hypotheses, a loss in power can occur with respect to tests that are fine tuned to specific scenarios.

4 Numerical results

4.1 Methodology

This section is dedicated to comparing the different approaches to multiple testing outlined in Sect. 3.3. They can be classified into three categories: combining p-values (min-p, prod$p$, avg-p), aggregating test statistics (smax-t) and adjusting p-values (HB). We utilize three benchmarks from the high-energy physics literature on signal-agnostic searches and anomaly detection [11,13,18,19] with minor modifications. Each benchmark is defined by a reference distribution, characterising the null hypothesis, and different types of new physic signals, characterising different alternative hypotheses. When possible, we vary certain parameters to alter the alternative, such as the width of a resonance or the number of signal events, to explore more diverse scenarios and increase the validity of our study.

For each benchmark and each value of the kernel width $\sigma \in \Sigma$, we proceed with the following steps:

Selection of α We set the false positive rate α or its corresponding Z-scores, as defined in Eq. (6). Specifically, we consider $Z_\alpha = 3$, the standard threshold in high-energy physics for evidence of a signal, and $Z_\alpha = 2$.

Estimation of test distributions We estimate the distribution of the NPLM test statistic under the null hypothesis $p(t^{(\sigma)}|H_0)$, as outlined in Sect. 2.1.2, and under each alternative hypothesis $p(t^{(\sigma)}|H_1)$. The empirical distribution of the test is obtained repeating the test $N_{\text{toys}}^{(H_{0/1})}$ times on statistically independent samples that are drawn from the true data-generating distributions whenever available. Alternatively, a bootstrap-based approach is employed, resampling with replacement from a large dataset.

Calculation of smax-t The smax-t test statistic is computed directly from the test statistic values using Eq. (23), separately for the null and alternative hypotheses.

Computation of p-values For each toy sample drawn from the null or alternative hypotheses, we compute the p-value $p^{(\sigma)}$ as defined in Eq. (18). When estimating the null distribution, the test sample itself is excluded and the empirical p-value is computed with respect to the remaining $N_{\text{toys}}^{(H_0)} - 1$ samples.

Application of the HB method The output of the HB method is computed directly from the p-values associated with each sample as in Eq. (24) using the predefined threshold α.

Determination of critical values The threshold α is translated into a critical value t_α based on the $N_{toys}^{(H_0)}$ test statistic values under the null hypothesis. The critical value corresponds to the highest value at an empirical quantile not exceeding $1-\alpha$. This conservative approach ensures that the actual false positive rate is not larger than α.

Evaluation of meta-test performance The performance of meta-tests (min-p, prod$p$, avg-p, smax-t) is assessed by computing their statistical power for each benchmark and alternative hypothesis. Power, defined by Eq. (7), is estimated empirically using:

$$\widehat{\text{power}}_\alpha = \frac{1}{N_{\text{toys}}^{(H_1)}} \sum_{i=1}^{N_{\text{toys}}^{(H_1)}} \mathbb{1}(t_i - t_\alpha). \tag{25}$$

4.2 Datasets and hyperparameters

4.2.1 EXPO-1D

In this univariate benchmark (see also [9,11,13]), we consider a reference model given by an energy spectrum that decays exponentially, described by the following density

$$n(x|\text{R}) = N(R)e^{-x}\,, \tag{26}$$

where the expected number events in the reference hypothesis is $N(R) = 2000$. The reference sample is composed of $N_{\mathcal{R}} = 100\,N(R)$ events. We consider the following parametrised alternative hypothesis

$$n_{\text{true}}(x) = n(x|\text{R}) + N(S)\frac{1}{\sqrt{2\pi}\sigma_{\text{NP}}}\exp\left[-\frac{(x-\bar{x}_{\text{NP}}^2)}{2\sigma_{\text{NP}}^2}\right], \tag{27}$$

representing a Gaussian peak with mean $\bar{x}_{\text{NP}}$ and standard deviation σ_{NP}, on top of the reference background. In our tests, we vary both parameters and the average number of injected new physics events $N(S)$ to establish the performance of the method. We evaluated the null hypothesis with $N_{\text{toys}}^{(H_0)} = 4000$ and each alternative hypothesis with $N_{\text{toys}}^{(H_1)} = 2000$.

Hyperparameters Being an illustrative benchmark, we select values for M and λ that result in faster training times than those reported in [13]. We use $M = 1000$, $\lambda = 10^{-6}$ and $\Sigma = \{0.01, 0.3, 0.7, 1.4, 4.5, 9\}$. The first five values correspond to the 0.01, 0.25, 0.50, 0.75 and 0.99 quantiles;

Table 1 EXPO 1D. $P(Z > 3)$ – probability of observing $Z \geq 3$ for different types of new physics signals, as represented in Eq. (27). The last value of σ follows from the prescription from the original proposal [9]. Bold values highlight the best performances

N(S)	7	18	13	10	90
$\bar{x}_{\mathrm{NP}}$	4	4	4	6.4	1.6
σ_{NP}	0.01	0.16	0.64	0.16	0.16
$\sigma = 0.01$	0.0028 ± 0.0008	0.0010 ± 0.0006	0.0005 ± 0.0004	0.0001 ± 0.0001	0.029 ± 0.004
$\sigma = 0.3$	**0.012 ± 0.002**	0.107 ± 0.007	0.008 ± 0.002	0.246 ± 0.009	0.65 ± 0.01
$\sigma = 0.7$	0.006 ± 0.001	**0.123 ± 0.007**	**0.011 ± 0.002**	**0.36 ± 0.01**	**0.70 ± 0.01**
$\sigma = 1.4$	0.004 ± 0.001	0.078 ± 0.006	**0.012 ± 0.002**	0.29 ± 0.01	0.54 ± 0.01
$\sigma = 4.5$	0.0023 ± 0.0007	0.020 ± 0.003	**0.011 ± 0.002**	0.098 ± 0.007	0.28 ± 0.01
$\sigma = 9.0$	0.0028 ± 0.0008	0.018 ± 0.003	**0.012 ± 0.002**	0.075 ± 0.006	0.24 ± 0.01
$\sigma = 2.3$ [9]	0.0023 ± 0.0007	0.044 ± 0.005	0.013 ± 0.002	0.028 ± 0.004	0.36 ± 0.01
Min-p	**0.008 ± 0.001**	**0.103 ± 0.007**	0.007 ± 0.002	**0.32 ± 0.01**	**0.66 ± 0.01**
Prod-p	0.005 ± 0.001	0.083 ± 0.006	**0.012 ± 0.002**	0.26 ± 0.01	0.65 ± 0.01
Avg-p	0.006 ± 0.001	0.049 ± 0.005	0.011 ± 0.002	0.068 ± 0.006	0.50 ± 0.01
smax-t	0.0028 ± 0.0008	0.0010 ± 0.0006	0.0005 ± 0.0004	0.0001 ± 0.0001	0.029 ± 0.004
HB	0.0020 ± 0.0007	0.0495 ± 0.0007	0.0035 ± 0.0013	0.231 ± 0.011	0.546 ± 0.009

the last value is chosen as twice the value of the 0.99 quantile. Correlation among tests can be studied and estimated in a signal-agnostic way by inspecting the pairwise correlations under the null hypothesis, as depicted in Fig. 1. Here, the panels closer to the diagonal show the correlation between tests that are closer in the σ space. As anticipated in Sect. 3, the correlation is higher for nearby tests.

4.2.2 MUMU-5D

This five dimensional dataset (introduced in [18]) is composed of simulated LHC collision events producing two opposite charged muons in the final state ($pp \to \mu^+\mu^-$) at a center-of-mass energy of 13 TeV. [3] The features are the transverse momenta and pseudorapidities of the two muons, and their relative azimuthal angle, i.e.,
$x = [p_{T1}, p_{T2}, \eta_1, \eta_2, \Delta\phi]$. We consider two types of new physics contributions: the first one is a new vector boson (Z') for which we study different mass values ($m_{Z'} = 180, 200$ and 600 GeV); the second one is a non-resonant signal obtained by adding a four-fermion contact interaction to the Standard Model Lagrangian for which the Wilson coefficient c_W determines the coupling strength. We fix $N(R) = 2 \times 10^4$ expected events in the reference hypothesis and the size of the reference sample is $N_{\mathcal{R}} = 10^5$. Also in this case, we vary the number of expected signal events $N(S)$. We evaluated the null hypothesis with $N_{\mathrm{toys}}^{(H_0)} = 2000$ and each alternative hypothesis with $N_{\mathrm{toys}}^{(H_1)} = 1000$.

Hyperparameters We selected $M = 10^4$, $\lambda = 10^{-6}$ and $\Sigma = \{0.31, 1.19, 1.79, 2.49, 4.23, 8.0\}$. The first five values correspond to the 0.01, 0.25, 0.50, 0.75 and 0.99 quantiles; the last value is chosen as approximately twice the 0.99 quantile. We evaluated the null hypothesis with $N_{\mathrm{toys}}^{(H_0)} = 2000$ and each alternative hypothesis with $N_{\mathrm{toys}}^{(H_1)} = 1000$.

4.2.3 LHCO-6D

The LHC Olympic dataset is a widely utilised benchmark for resonant anomaly detection proposed as a challenge by [19]. The dataset, available on Zenodo [40], consists of LHC collision events with two jets in the final state. The Standard Model background consists of QCD events while the signal is modelled as a resonant W' decaying into two massive particles X and Y, with $X \to qq$ and $Y \to qq$. The W', X, and Y masses are 3.5 TeV, 500 GeV and 100 GeV respectively. Both constituents level and jet level information is provided for each event. In this application we focus on six high level observables describing the dijet system: the dijet invariant mass, the mass of the leading jet, the difference between the two jets masses, the angular separation between the two jets, and the 2-subjettiness ratios for both jets ($\tau_{2,1}^{J1}$ and $\tau_{2,1}^{J2}$). Events are required to have at least one R = 1.0, pseudrapidity $|\eta| < 2.5$, and transverse momentum $p_{\mathrm{T}}^{J} > 1.2$ TeV. Since most of the applications concerning this dataset rely on a bump-hunt approach with sliding window on the dijet invariant mass, we focus our test on one single mass window, corresponding to the signal region ($3.1 \leq m_{J1,J2} \leq 3.7$ TeV). In this selection, the expected number of background events is approximately $N(R) = 121{,}000$, on top of which we inject an average of $N(S) = 333$ signal events. We evaluated the null hypothesis with $N_{\mathrm{toys}}^{(H_0)} = 2000$ and each alternative hypothesis with $N_{\mathrm{toys}}^{(H_1)} = 2000$.

[3] Data available at https://zenodo.org/record/4442665

Table 2 EXPO 1D. $P(Z > 2)$ – Probability of observing $Z \geq 2$ for different types of new physics signals, as represented in Eq. (27). The last value of σ follows from the prescription from the original proposal [9]. Bold values highlight the best performances

N(S)	7	18	13	10	90
$\bar{x}_{\text{NP}}$	4	4	4	6.4	1.6
σ_{NP}	0.01	0.16	0.64	0.16	0.16
$\sigma = 0.01$	0.039 ± 0.003	0.046 ± 0.005	0.023 ± 0.003	0.025 ± 0.003	0.25 ± 0.01
$\sigma = 0.3$	**0.083 ± 0.004**	0.35 ± 0.01	0.053 ± 0.005	0.49 ± 0.01	0.881 ± 0.008
$\sigma = 0.7$	0.072 ± 0.004	**0.37 ± 0.01**	0.076 ± 0.006	**0.66 ± 0.01**	**0.913 ± 0.007**
$\sigma = 1.4$	0.052 ± 0.003	0.28 ± 0.01	**0.082 ± 0.006**	0.59 ± 0.01	0.820 ± 0.009
$\sigma = 4.5$	0.037 ± 0.003	0.166 ± 0.008	0.080 ± 0.006	0.37 ± 0.01	0.63 ± 0.01
$\sigma = 9.0$	0.0304 ± 0.003	0.121 ± 0.007	0.074 ± 0.006	0.29 ± 0.01	0.58 ± 0.01
$\sigma = 2.3$ [9]	0.039 ± 0.003	0.207 ± 0.009	0.079 ± 0.006	0.48 ± 0.01	0.69 ± 0.01
min-p	0.063 ± 0.004	0.31 ± 0.01	0.066 ± 0.006	**0.58 ± 0.01**	0.877 ± 0.007
prod-p	**0.065 ± 0.004**	**0.32 ± 0.01**	**0.085 ± 0.006**	**0.58 ± 0.01**	**0.897 ± 0.007**
Avg-p	0.063 ± 0.004	0.230 ± 0.009	0.082 ± 0.006	0.34 ± 0.01	0.835 ± 0.009
smax-t	0.039 ± 0.003	0.046 ± 0.005	0.023 ± 0.003	0.025 ± 0.003	0.25 ± 0.01
HB	0.046 ± 0.003	0.25 ± 0.01	0.047 ± 0.005	0.52 ± 0.01	0.85 ± 0.01

Table 3 MUMU 5D – $P(Z > 3)$ for different types of new physics signals. The last value of σ follows from the prescription from the original proposal [9]

Test	Z' M = 180 GeV W = 0.02 GeV	Z' M = 300 GeV W = 15 GeV	Z' M = 600 GeV W = 30 GeV	EFT $c_w = 1.5 \times 10^{-6}$
$\sigma = 0.31$	0.007 ± 0.003	0.004 ± 0.002	0.0010 ± 0.0008	0.0010 ± 0.0008
$\sigma = 1.19$	**0.096 ± 0.009**	0.10 ± 0.01	0.006 ± 0.002	0.017 ± 0.004
$\sigma = 1.79$	0.065 ± 0.008	0.11 ± 0.01	0.012 ± 0.003	0.026 ± 0.005
$\sigma = 2.49$	0.036 ± 0.006	0.11 ± 0.01	0.027 ± 0.005	0.053 ± 0.007
$\sigma = 4.23$	0.037 ± 0.006	**0.13 ± 0.01**	**0.066 ± 0.008**	0.13 ± 0.01
$\sigma = 8.0$	0.023 ± 0.004	0.068 ± 0.008	0.056 ± 0.007	**0.22 ± 0.01**
$\sigma = 3.0$ [9]	0.031 ± 0.005	0.13 ± 0.01	0.044 ± 0.006	0.092 ± 0.009
min-p	0.065 ± 0.008	0.16 ± 0.01	**0.057 ± 0.007**	**0.23 ± 0.01**
prod-p	0.089 ± 0.009	**0.18 ± 0.01**	0.028 ± 0.005	0.083 ± 0.009
Avg-p	**0.14 ± 0.01**	0.15 ± 0.01	0.035 ± 0.006	0.098 ± 0.009
smax-t	0.007 ± 0.003	0.004 ± 0.002	0.0010 ± 0.0008	0.0010 ± 0.0008
HB	0 ± 0	0 ± 0	0 ± 0	0 ± 0

Table 4 MUMU 5D – $P(Z > 2)$ for different types of new physics signals. The last value of σ follows from the prescription from the original proposal [9]

Test	Z' M = 180 GeV W = 0.02 GeV	Z' M = 300 GeV W = 15 GeV	Z' M = 600 GeV W = 30 GeV	EFT $c_w = 1.5 \times 10^{-6}$
$\sigma = 0.31$	0.11 ± 0.01	0.042 ± 0.006	0.023 ± 0.005	0.036 ± 0.006
$\sigma = 1.19$	**0.30 ± 0.01**	0.35 ± 0.02	0.047 ± 0.007	0.11 ± 0.01
$\sigma = 1.79$	**0.30 ± 0.01**	0.41 ± 0.02	0.11 ± 0.01	0.16 ± 0.01
$\sigma = 2.49$	0.25 ± 0.01	**0.42 ± 0.02**	0.19 ± 0.01	0.25 ± 0.01
$\sigma = 4.23$	0.23 ± 0.01	0.41 ± 0.02	0.25 ± 0.01	0.32 ± 0.01
$\sigma = 8.0$	0.19 ± 0.01	0.31 ± 0.01	**0.29 ± 0.01**	**0.52 ± 0.02**
$\sigma = 3.0$ [9]	0.25 ± 0.02	**0.42 ± 0.02**	0.24 ± 0.02	0.30 ± 0.02
min-p	0.32 ± 0.01	0.47 ± 0.02	**0.28 ± 0.01**	**0.53 ± 0.02**
prod-p	**0.38 ± 0.02**	**0.53 ± 0.02**	0.23 ± 0.01	0.38 ± 0.02
Avg-p	0.37 ± 0.02	0.46 ± 0.02	0.18 ± 0.01	0.31 ± 0.01
smax-t	0.11 ± 0.01	0.042 ± 0.006	0.023 ± 0.005	0.036 ± 0.006
HB	0.255 ± 0.003	0.40 ± 0.01	0.22 ± 0.01	0.46 ± 0.02

Table 5 LHCO-6D – $P(Z > 2)$. The last value of σ follows from the prescription from the original proposal [9]

Test	$N(S) = 333$
$\sigma = 1.2$	0.28 ± 0.01
$\sigma = 2.5$	0.28 ± 0.01
$\sigma = 3.2$	$\mathbf{0.29 \pm 0.01}$
$\sigma = 3.9$	0.28 ± 0.01
$\sigma = 6.1$	0.219 ± 0.009
$\sigma = 12.2$	0.175 ± 0.009
$\sigma = 4.6$ [9]	0.28 ± 0.01
Min-p	0.36 ± 0.01
Prod-p	$\mathbf{0.37 \pm 0.01}$
Avg-p	$\mathbf{0.37 \pm 0.01}$
smax-t	0.28 ± 0.01
HB	0.29 ± 0.01

Hyperparameters We select $M = 853$, $\lambda = 10^{-6}$ and $\Sigma = \{1.2, 2.5, 3.2, 3.9, 5.1, 6.1, 12.2\}$. The first five values correspond to the 0.01, 0.25, 0.50, 0.75 and 0.99 quantiles; the last value is chosen as twice the 0.99 quantile.

4.3 Results

Tables 1, 2, 3, 4, 5 and 6 summarise the power of the various meta-tests described in Sect. 3 for all the benchmarks in this study. They report the probability of observing a Z-score greater or equal to 3, in other words the chances of finding evidence for the signal, and the probability of observing a Z-score greater or equal to 2. In the upper part of the tables, we show the sensitivity of the NPLM test for each individual value in Σ. In the middle part we show the performance of the standard NPLM approach presented in [9]. In the bottom part we report the performance of the various aggregation strategies introduced in Sect. 3.3. All entries in the table are endowed with uncertainties computed as the 68% Clopper-Pearson [41] confidence interval.

The results obtained with single values of σ highlight the dependency of the NPLM test outcome on the specific choices of kernel width and signal benchmark. Narrow peaks, like the one reported in the first columns of Tables 1 and 3, are better detected by small values of σ, whereas large values are preferable to detect broad peaks, like the one reported in the third columns. We report in Figs. 2 and 3 examples of the full power curve of each individual test, showing how the sensitivity changes according to σ and follows different trends depending on the signal.

Our studies show that it is beneficial to combine multiple tests. With the exception of smax-t, that corresponds to systematically selecting the test with the smallest width, the other methods return powers that are comparable with or larger than the original kernel-based NPLM proposal in [9] and often competitive with the best overall test, which would be hard to identify a priori in real analyses. We observe that min-p is the most balanced choice across multiple signal scenarios. Indeed, it is the one that gives the best results in most cases and when it does not, its failure is not catastrophic. The HB method can fail when the signal is hard to detect as the corrected threshold in Eq. (24) can become too conservative. The advantage of using min-p becomes particularly evident when a specific individual test performs better. This can be seen in the $m_{Z'} = 600$ signal and, more strongly, in the EFT case, where the best individual test is the one with the largest σ with a clear trend. We also observe that prod-p gives good results if a subset of tests performs similarly well as in the third column of Table 1. The method avg-p is instead performing well when the there is not a strong separation between the performance of the individual tests, as shown in Tables 5 and 6. These results confirm the intuitions discussed in Sect. 3.3.

Table 6 LHCO-6D – $P(Z > 3)$. The last value of σ follows from the prescription from the original proposal [9]

Test	$N(S) = 333$
$\sigma = 1.2$	0.0763 ± 0.006
$\sigma = 2.5$	$\mathbf{0.103 \pm 0.007}$
$\sigma = 3.2$	0.088 ± 0.006
$\sigma = 3.9$	0.048 ± 0.005
$\sigma = 6.1$	0.046 ± 0.005
$\sigma = 12.2$	0.054 ± 0.005
$\sigma = 4.6$ [9]	0.051 ± 0.005
Min-p	0.089 ± 0.006
Prod-p	0.092 ± 0.006
Avg-p	$\mathbf{0.130 \pm 0.007}$
smax-t	0.076 ± 0.006
HB	0.0045 ± 0.0015

5 Conclusions

In this paper we investigate the problem of model-selection in ML-based solutions for signal-agnostic searches. By focusing on the NPLM goodness-of-fit test, we show how hyperparameter tuning can introduce biases towards specific signal hypotheses.

We propose to mitigate this effect by performing multiple tests, characterised by different hyperparameters, on the same set of experimental measurements, and combining them into a meta-test in a way that is robust against the look-elsewhere effect. By adopting this approach, we turn a potential limitation of the kernel-based NPLM method into a feature that allows for a more inclusive analysis. We show that this strategy improves over the baseline proposal in [9] and we observe

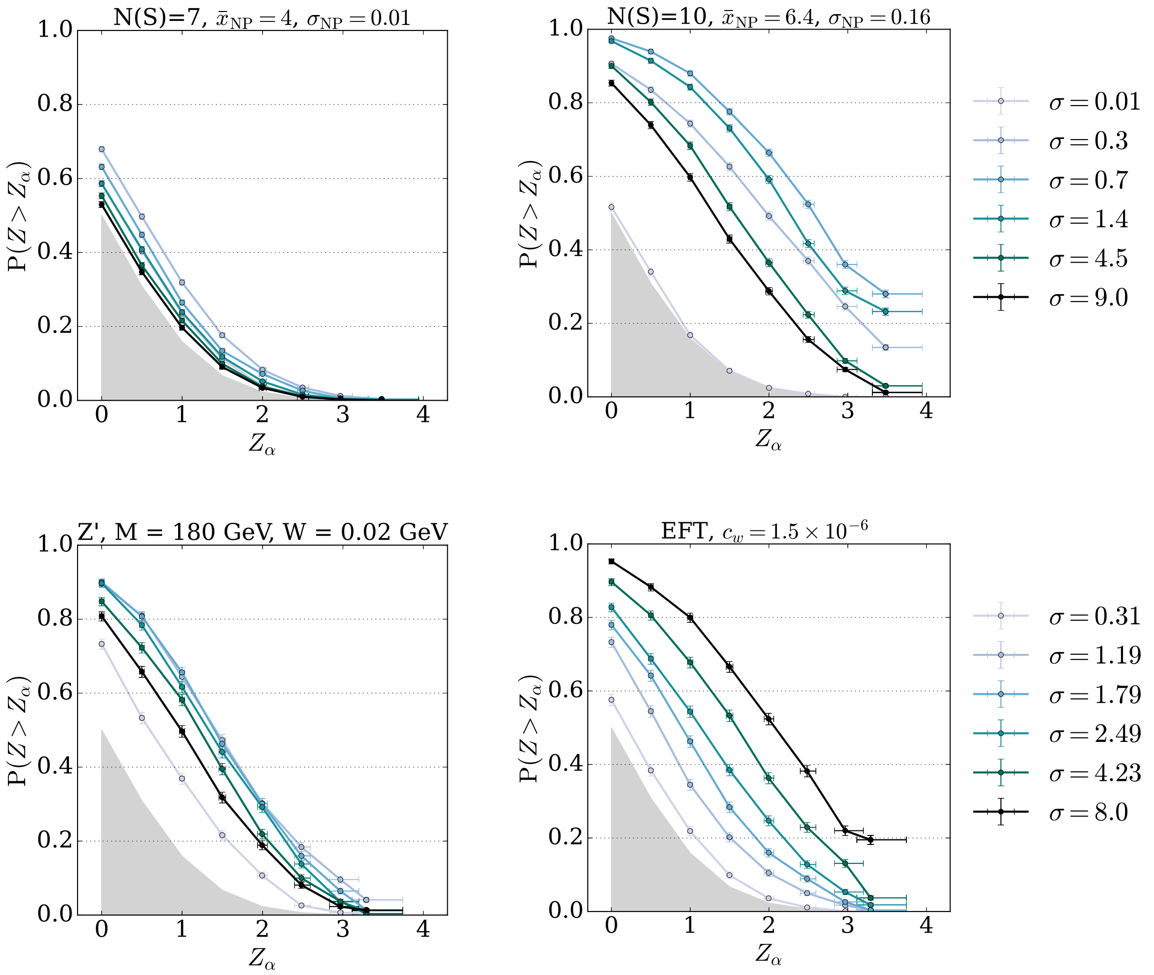

Fig. 2 EXPO-1D and MUMU-5D – Illustrative examples of power curves for NPLM tests performed with different choices of σ. The upper left and right hand panels show the power curves for a narrow signal and a signal in the tail respectively in the univariate scenario. The bottom panels illustrate the same effect for the five dimensional dataset in the case of a Z' vector boson and a non-resonant signal. The grey-filled area represents the region with no detection power

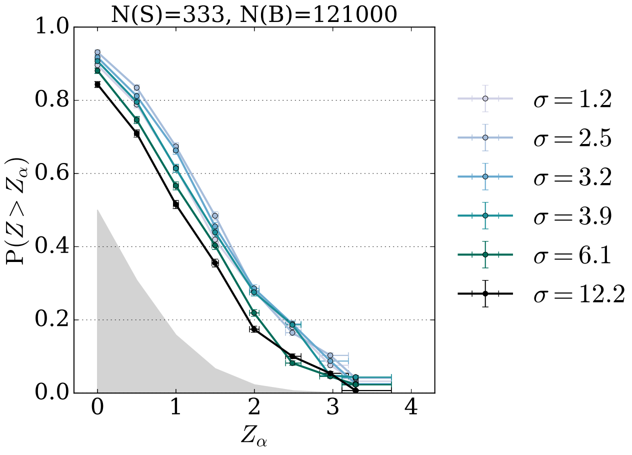

Fig. 3 LHCO-6D. Power curves for different choices of σ (shades of green lines), compared with the one for the min-p aggregation (black line). The grey-filled area represents the region with no detection power

a more uniform response across multiple signal scenarios. In particular, we show that combining individual p-values by selecting the smallest value (the min-p approach) is the most effective method, especially for signals that are hard to detect. This approach involves increased computational requirements as multiple tests have to be performed in place of a single one. However, this cost could be mitigated by an appropriate parallelised strategy.

This work represents a further step towards building unbiased machine learning tools for anomaly detection and hypothesis testing in the context of collider experiments. From this perspective, the strategy proposed in this study goes beyond the NPLM approach and could be tested to combine methods for new physics searches that have been designed to be sensitive to specific families of signals, including those that are not based on ML.

In conclusion, our study indicates that the impact of model selection on sensitivity can be leveraged to enhance interpretability, particularly for machine learning models with a limited number of hyperparameters that can be connected to physical priors. This is an interesting direction that we leave for future developments.

Acknowledgements The authors would like to thank Louis Lyons for his decisive support and pivotal suggestions, and Arthur Gretton and his collaborators for the stimulating conversations. M.L. acknowledges the financial support of the European Research Council (Grant SLING 819789). G.G. acknowledges the financial support of the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). Computations in this paper were partially run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Data Availability Statement My manuscript has associated data in a data repository. [Authors' comment: The datasets analysed during the current study are available in the NPLM: Learning Multivariate New Physics and LHC Olympics 2020 repositories, https://zenodo.org/record/4442665 and https://zenodo.org/records/4536377 respectively.]

Code Availability Statement My manuscript has associated code/software in a data repository. [Authors' comment: The code generated during the current study is available in the following repository https://github.com/mletizia/multiple-testing-nplm.]

Funded by SCOAP^3.

References

1. 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)
2. T. Aaltonen et al., Model-independent and quasi-model-independent search for new physics at CDF. Phys. Rev. D **78**, 012002 (2008)
3. G. Choudalakis, On hypothesis testing, trials factor, hypertests and the BumpHunter. in *PHYSTAT 2011*, vol. 1 (2011)
4. 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**(2), 120 (2019)
5. A.M. Sirunyan et al., MUSiC: a model-unspecific search for new physics in proton-proton collisions at $\sqrt{s} = 13$ TeV. Eur. Phys. J. C **81**(7), 629 (2021)
6. V. Belis, P. Odagiu, T.K. Aarrestad, Machine learning for anomaly detection in particle physics. Rev. Phys. **12**, 100091 (2024)
7. Model-agnostic search for Dijet resonances with anomalous jet substructure in proton–proton collisions at $\sqrt{s}$ = 13 TeV. (2024)
8. 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**(13), 131801 (2020)
9. M. Letizia, G. Losapio, M. Rando, G. Grosso, A. Wulzer, M. Pierini, M. Zanetti, L. Rosasco, Learning new physics efficiently with nonparametric methods. Eur. Phys. J. C **82**(10), 879 (2022)
10. G. Meanti, L. Carratino, L. Rosasco, A. Rudi, Kernel methods through the roof: handling billions of points efficiently. Adv. Neural Inf. Process. Syst. **33**, 14410–14422 (2020)
11. R.T. D'Agnolo, A. Wulzer, Learning new physics from a machine. Phys. Rev. D **99**(1), 015014 (2019)
12. G. Grosso, N. Lai, M. Letizia, J. Pazzini, M. Rando, L. Rosasco, A. Wulzer, M. Zanetti, Fast kernel methods for data quality monitoring as a goodness-of-fit test. Mach. Learn. Sci. Technol. **4**(3), 035029 (2023)
13. G. Grosso, M. Letizia, M. Pierini, A. Wulzer, Goodness of fit by Neyman–Pearson testing. SciPost Phys. **16**, 123 (2024)
14. E.L. Lehmann, J.P. Romano, *Testing Statistical Hypotheses* (2022)
15. L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer Science & Business Media, Berlin, 2013)
16. A. Schrab, I. Kim, M. Albert, B. Guedj, A. Gretton, B. Laurent, MMD aggregated two-sample test. (2023)
17. F. Biggs, A. Schrab, A. Gretton, MMD-fuse: learning and combining kernels for two-sample testing without data splitting. Adv. Neural Inf. Process. Syst. **36**, 75151–75188. Curran Associates, Inc. (2023)
18. R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Learning multivariate new physics. Eur. Phys. J. C. **81**(1), 89 (2021)
19. G. Kasieczka et al., The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics. Rep. Prog. Phys. **84**(12), 124201 (2021)
20. J. Neyman, E.S. Pearson, On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. Roy. Soc. Lond. A. **231**(694–706), 289–337 (1933)
21. A. Rudi, L. Rosasco, Generalization properties of learning with random features. Adv. Neural Inf. Process. Syst. (2017). arXiv:1602.04474 [stat.ML]
22. R. Barlow, Extended maximum likelihood. Nuclear Instrum. Methods Phys. Res. Sect. A Acceler. Spectrom. Detect. Assoc. Equip. **297**(3), 496–506 (1990)
23. S. Baker, R.D. Cousins, Clarification of the use of Chi square and likelihood functions in fits to histograms. Nucl. Instrum. Methods **221**, 437–442 (1984)
24. D. Lopez-Paz, M. Oquab, Revisiting classifier two-sample tests. in *International Conference on Learning Representations* (2017)
25. A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test. J. Mach. Learn. Res. **13**(1), 723–773 (2012)
26. A. Rudi, R. Camoriano, L. Rosasco, Less is more: Nyström computational regularization. Adv. Neural Inf. Process. Syst. 28 (2015). arXiv:1507.04717 [stat.ML]
27. B.V. North, D. Curtis, P.C. Sham, A note on the calculation of empirical p values from Monte Carlo procedures. Am. J. Hum. Genet. **71**(2), 439–441 (2002)
28. D.A. Darling, The Kolmogorov–Smirnov, Cramer–Von mises tests. Ann. Math. Stat. **28**(4), 823–838 (1957)
29. E. Gross, O. Vitells, Trial factors for the look elsewhere effect in high energy physics. Eur. Phys. J. C **70**, 525–530 (2010)
30. L. Lyons, Open statistical issues in particle physics. Ann. Appl. Stat. **2**(3), 887–915 (2008)
31. L. Demortier, P values and nuisance parameters. in *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, pp. 23–33 (2007)

32. R.O. Kuehl, *Design of Experiments: Statistical Principles of Research Design and Analysis. Statistics Series* (Duxbury/Thomson Learning, 2000)
33. N.A. Heard, P. Rubin-Delanchy, Choosing between methods of combining-values. Biometrika **105**(1), 239–246 (2018)
34. W. Rolke, Supplemental studies for simultaneous goodness-of-fit testing. **7** (2020). arXiv:2007.04727
35. L.H.C. Tippett et al., The methods of statistics. in *The Methods of Statistics* (1931)
36. R.A. Fisher, Statistical methods for research workers, in *Breakthroughs in Statistics: Methodology and Distribution*. (Springer, Berlin, 1970), pp.66–70
37. E.S. Edgington, An additive method for combining probability values from independent experiments. J. Psychol. **80**(2), 351–363 (1972)
38. K. Asadi, M.L. Littman, An alternative softmax operator for reinforcement learning. in *International Conference on Machine Learning* (PMLR, 2017), pp. 243–252
39. S. Holm, A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**(2), 65–70 (1979)
40. G. Kasieczka, B. Nachman, D. Shih, R&D dataset for LHC Olympics 2020 anomaly detection challenge. (2019)
41. C.J. Clopper, E.S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26**(4), 404–413 (1934)