MIT Open Access Articles

Anomaly-aware summary statistic from data batches

Anomaly-aware summary statistic from data batches

G. Grosso 🔾

*The NSF AI Institute for Artificial Intelligence and Fundamental Interactions,
Cambridge, MA, U.S.A.*

*Laboratory for Nuclear Science, Massachusetts Institute of Technology,
Cambridge, MA, U.S.A.*

*School of Engineering and Applied Sciences, Harvard University,
Cambridge, MA, U.S.A.*

E-mail: gaia.grosso@cern.ch

ABSTRACT: Signal-agnostic data exploration based on machine learning could unveil very subtle statistical deviations of collider data from the expected Standard Model of particle physics. The beneficial impact of a large training sample on machine learning solutions motivates the exploration of increasingly large and inclusive samples of acquired data with resource efficient computational methods. In this work we consider the New Physics Learning Machine (NPLM), a multivariate goodness-of-fit test built on the Neyman-Pearson maximum-likelihood-ratio construction, and we address the problem of testing large size samples under computational and storage resource constraints. We propose to perform parallel NPLM routines over batches of the data, and to combine them by locally aggregating over the data-to-reference density ratios learnt by each batch. The resulting data hypothesis defining the likelihood-ratio test is thus shared over the batches, and complies with the assumption that the expected rate of new physical processes is time invariant. We show that this method outperforms the simple sum of the independent tests run over the batches, and can recover, or even surpass, the sensitivity of the single test run over the full data. Beside the significant advantage for the offline application of NPLM to large size samples, the proposed approach offers new prospects toward the use of NPLM to construct anomaly-aware summary statistics in quasi-online data streaming scenarios.

Contents

1 Introduction

Numerous experimental tests have confirmed the Standard Model as the best description of the particle physics world, reaching levels of precision up to few parts per thousand. New physical processes, if detectable, should therefore manifest as very subtle statistical deviations of the data distributions with respect to the Standard Model expectations. Increasing the phase space acceptance and the statistics of the analysed datasets could unveil interesting data structures that were missed so far. Along with the regular signal-specific statistical analyses, signal-agnostic anomaly detection tools have grown of interest in the high energy physics community, and machine learning has proved to be a crucial ingredient to scale their discovery power.

Anomaly detection broadly refers to the problem of identifying data patterns that do not align with their expected behaviours. Events can be interpreted as anomalous for multiple reasons; either because their appearance in the dataset is rare, or because they manifest out of the expected data support. Moreover, anomalies can emerge as a collective behaviour affecting the statistical model describing the data occurrence in the experiment. In this work we focus

on collective anomalies, namely phenomena that rise as an unexpected deformation of the data probability distribution. These kind of anomalies are tackled by means of statistical anomaly detection tools. Statistical anomaly detection refers to the problem of recognising and quantifying distributional deviations of a dataset from its nominal expected behaviour. Besides its use in new physics searches, statistical anomaly detection finds a wide spectrum of applications in experimental particle physics. The most striking examples are data quality monitoring of experimental setups at data taking time, and validation of simulated samples.

When a tractable model of the data nominal behaviour is available, statistical anomaly detection reduces to a goodness-of-fit (GoF) test: a test statistic is defined according to a notion of similarity between the data distribution and the nominal model, and the level of compatibility is quantified in terms of the test statistic p-value with respect to the test distribution in nominal conditions. However, in most of the applications of statistical anomaly detection in high energy physics, a tractable model for the nominal data distribution is not available and needs to be replaced by a dataset of reference (denoted as $\mathcal{R}$ in this work), simulated according to the nominal model or built in a data-driven fashion. In these cases, the problem of statistical anomaly detection is solved as a two-sample-test, introducing a test statistic that measures the distance between two samples.

High energy physics, and in particular collider data, pose extremely challenging problems for statistical anomaly detection due to the level of precision required to test the Standard Model. For instance, the accuracy of the statistical models employed in the process of scientific discovery is a particularly delicate matter that requires powerful testing tools. Similarly, the lack of new interesting tensions with the Standard Model hypothesis after the Higgs boson discovery seem to indicate that, if detectable, new physics should manifest at the LHC as a very rare process, whether in time or in phase space. Anomaly detection algorithms must therefore be able to capture very subtle anomalous behaviours, and they should be run over the most inclusive set of data allowed by the computational and methodological constraints. An extensive review of GoF algorithms routinely adopted in the high energy physics community can be found in ref. [1].

Machine learning (ML) has provided significant improvements to the statistical tools employed in data analysis at the LHC, allowing to deal with increasingly complex data structures, and exploiting a wider amount of the physics information produced by the experiments. In the specific context of GoF, machine learning based algorithms have opened the path to multivariate statistical analyses, able to capture mismodeling also in the correlations between variables ref. [2].

In absence of constraining inductive biases, as it is the case in unknown anomalous signals searches, the power of machine learning strategies resides in their ability to learn from the training examples coming directly from the experimental data. However, the latter could be either scarce or completely lacking of anomalous events due to their rarity. The impact of machine learning is therefore expected to become more and more evident by scaling the training sample size and, as a consequence, the rate of collected anomalies. While beneficial to the training, larger sample sizes introduce computational and storage constraints, that need to be addressed. In this work, we focus on the New Physics Learning Machine (NPLM) algorithm, an ML-empowered goodness-of-fit method based on the Neyman-

Pearson hypothesis test (refs. [3–7]), and we address the problem of how to test large datasets under computational and storage constraints. The approach presented in this paper proposes parallel computing of the test over batches of the original sample and a final aggregation strategy that combines the local approximation of the data-to-reference ratios learnt by each instance of the algorithm. We outline three operational approaches for different resource constraints settings, aiming at maximally exploiting the available experimental data both in offline and quasi-online configurations.

We apply the split-aggregation approach to a one-dimensional toy model and two realistic LHC datasets of increasing complexity. The former is a dimuon final state characterised by 5 observables and sample size of $O(10^4)$ events, and a 24-dimensional problem defined by the tri-momenta of eight objects selected at the L1 trigger of the CMS experiment with sample size of $O(10^6)$ events.

Our numerical experiments for one-dimensional and five-dimensional problems show that the aggregation strategy can fully recover or, in same cases, even outperform the single batch implementation, thanks to the effect of regularization against statistical noise that the aggregation strategy introduces.

Furthermore, the studies conducted on the 24-dimensional problem show the discovery potential of the various approaches in high-dimensional and large statistic scenarios.

The paper is organised as follows. In section 2 we describe the NPLM algorithm and the proposed batch solution for handling large sample sizes. In section 3 we study the performances of the split-aggregation approach on the one-dimensional and five-dimensional problems and we compare them with the original implementation of NPLM. In section 4, we apply the split-aggregation strategy to a large size dataset simulating the CMS L1 trigger, showing the scalability of the algorithm to more complex scenarios. We conclude in section 6 discussing our main findings and future research directions.

2 Algorithm

In this section we briefly summarise the main features of the NPLM algorithm and we provide the statistical foundations for the new proposed split-aggregation solution.

2.1 New Physics Learning Machine (NPLM)

New Physics Learning Machine (NPLM) is a ML-based signal-agnostic strategy to detect and quantify statistically significant deviations of the observed data $\mathcal{D}$ from a defined reference model R ref. [3]. The NPLM method is inspired by the classical approach to hypothesis testing based on the log-likelihood-ratio ref. [8]

$$t(\mathcal{D}) = 2 \max_{\mathbf{w}} \sum_{x \in \mathcal{D}} \log \frac{\mathcal{L}(x|\mathrm{H}_{\mathbf{w}})}{\mathcal{L}(x|\mathrm{R})}, \tag{2.1}$$

where the two compared hypotheses are the reference, R, and a composite alternative, $\mathrm{H}_{\mathbf{w}}$, whose nature is not known a priori. A model $f_{\mathbf{w}}(x)$ with trainable parameters $\mathbf{w}$ is defined on the space of the data x to parametrize the density distribution of the data in the family of alternatives

$$n(x|\mathrm{H}_{\mathbf{w}}) = n(x|\mathrm{R})e^{f_{\mathbf{w}}(x)}. \tag{2.2}$$
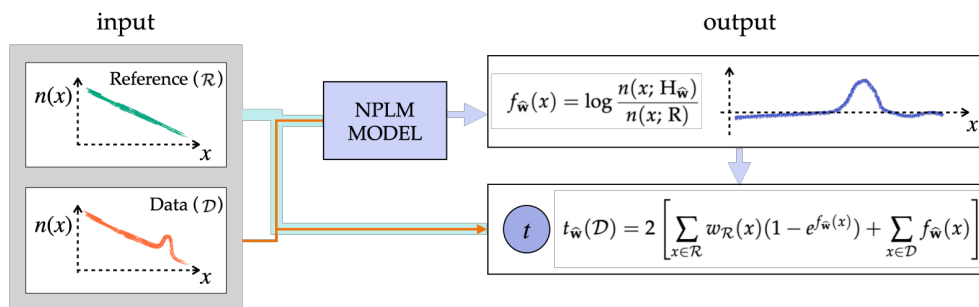
Figure 1. NPLM training over a single dataset. Schematic representation of a single run of the NPLM algorithm. The NPLM model, a kernel method in this work, takes as input two datasets, a reference sample ($\mathcal{R}$) and a data sample $\mathcal{D}$, and it fit via maximum-likelihood the log-ratio of their densities $f(x; \widehat{\mathbf{w}})$. The resulting model is evaluated in-sample on $\mathcal{D}$ and $\mathcal{R}$ to compute the test statistic t.

Since the total number of events in x is a Poissonian random variable whose expectation depends on the theory model, the test statistic in eq. (2.1) is computed as the extended log-likelihood-ratio, which in terms of $f_{\mathbf{w}}$ is written as

$$t(\mathcal{D}) = 2 \max_{\mathbf{w}} \left[\mathrm{N}(\mathrm{R}) - \mathrm{N}(\mathrm{H}_{\mathbf{w}}) + \sum_{x \in \mathcal{D}} f_{\mathbf{w}}(x) \right], \tag{2.3}$$

with $\mathrm{N}(\mathrm{R})$ and $\mathrm{N}(\mathrm{H}_{\mathbf{w}})$ the expectation values of the total number of events in $\mathcal{D}$ under the reference and the alternative hypotheses respectively. The maximisation problem described by eq. (2.3) can be solved by a machine learning task

$$t_{\mathbf{w}}(\mathcal{D}) = -2 \min_{\mathbf{w}} L[f_{\mathbf{w}}], \tag{2.4}$$

where the loss function

$$L[f_{\mathbf{w}}] = \left[\sum_{x \in \mathcal{R}} w_x (e^{f_{\mathbf{w}}(x)} - 1) - \sum_{x \in \mathcal{D}} f_{\mathbf{w}}(x) \right], \tag{2.5}$$

takes as input two training datasets: the dataset of interest $\mathcal{D}$, and a sample of reference $\mathcal{R}$ drawn according to the reference hypothesis and weighted by w_x to match the experimental luminosity. The weight encodes the information about the cross section of the physics process at the origin of each event, and hence it is assigned at the per event level and, generally, it could be different for various data points. Thus, the machine learning task learns the optimal values of the trainable parameters of f, $\widehat{\mathbf{w}}$, to approximate the log-density-ratio of the two samples

$$f_{\widehat{\mathbf{w}}}(x) = \log \frac{n(x|\mathrm{H}_{\widehat{\mathbf{w}}})}{n(x|\mathrm{R})} \tag{2.6}$$

A schematic representation of the NPLM training task is provided in figure 1

As for any test statistic, the value of the test obtained for $\mathcal{D}$ has to be compared to the distribution of the test under the reference hypothesis, $p(t|\mathrm{R})$, and the p-value is finally used as a GoF metric. We build $p(t|\mathrm{R})$ empirically, running pseudo-experiments on datasets that

are generated according to the R hypothesis. The model $f_{\mathbf{w}}(x)$ could be a neural network as in refs. [3–5], or it could be built with kernel methods refs. [6, 9]. In this work we employ kernel methods since they have been showed to be less time demanding than neural networks. Details about the NPLM training strategy based on kernel methods are postponed to section 2.4.

It should be mentioned that the test statistic in eq. (2.4) has good statistical properties (i.e. its null distribution approximates a χ^2 in the asymptotic regime) only if the number of events representing the reference hypothesis, $N_{\mathcal{R}}$ is large enough that the statistical fluctuations affecting the reference sample are negligible with respect to the one affecting the data. We refer the reader to ref. [4] and ref. [6] for a detailed discussion on the impact of the reference sample size on the NPLM test statistic properties.

Moreover, in realistic particle physics scenarios uncertainties affect the knowledge of the reference hypothesis requiring the introduction of nuisance parameters, defining families of transformations of the reference distribution that should be seen as not anomalous. A solution to extend the NPLM strategy to account for systematic errors has been introduced in ref. [5]. In our studies, we will consider systematic uncertainties as negligible and leave the treatment of systematic uncertainties within the batch approach for future studies. Finally, often times simulations are not accurate enough to provide a good reference model. In those circumstances, data driven techniques need to be implemented to build $\mathcal{R}$. The problem of building an accurate reference model R is a distinct problem and out of the scope of this work. Comment on possible solutions to this problem in section 4. In the next section we introduce the batching idea to address the problem of large sample size.

2.2 Learning New Physics from batches

Let's assume that the experimental data $\mathcal{D}$ are acquired in N_{b} batches $\mathcal{D}_i$

$$\mathcal{D} = \cup_i \mathcal{D}_i \,, \tag{2.7}$$

and the size N_i of each batch is a Poissonian random variable whose expectation is a function of the integrated data acquisition time.[1] In addition, let's assume that each batch $\mathcal{D}_i$ can be stored in a buffer long enough time to undergo the NPLM training task.

Since the proton-proton collisions are independent events, the optimal test of the full dataset according to Neyman and Pearson would be the sum of the log-likelihood-ratio test performed over each single batch, with alternative hypothesis set to the True model (T) underlying the data:

$$t_{id}(\mathcal{D} = \cup_i \mathcal{D}_i) = 2 \sum_{i=1}^{N_{\mathrm{b}}} \log \frac{\mathcal{L}(\mathcal{D}_i|\mathrm{T})}{\mathcal{L}(\mathcal{D}_i|\mathrm{R})} \,. \tag{2.8}$$

Clearly, the True hypothesis T is not known a priori and we want to exploit the NPLM strategy to estimate it. The solution we propose in this work is to run separate instances of the NPLM algorithm over different batches and recombine the information encoded in

---

[1]More precisely, the expected number of observations for a given physics process (H) is a function of its physics cross section ($\sigma(H)$), the integrated luminosity (L), and the acceptance (a) and efficiency (ϵ) of the experimental setup: $N(H) = \sigma(H) \times L \times a \times \epsilon$.

the log-density-ratio functions

$$f_{\widehat{\mathbf{w}}, i}(x) = \log \frac{n(x;\, \mathrm{H}_{\widehat{\mathbf{w}}}^i)}{n(x;\, \mathrm{R})} \tag{2.9}$$

learnt by the different instances. Recalling the parametrization defined in eq. (2.2), for each learnt model $f_{\mathbf{w}, i}$ we can write an estimate of the density ratio between the data batch and the reference $\mathcal{R}$

$$n(x;\, \mathrm{H}_{\widehat{\mathbf{w}}}^i) = n(x;\, \mathrm{R})e^{f_{\widehat{\mathbf{w}}, i}(x)}\,. \tag{2.10}$$

Each density $n(x;\, \mathrm{H}_{\widehat{\mathbf{w}}}^i)$ can be interpreted as an approximated view of the same True physical model. If the learning process outputs an accurate enough representation of T, then eq. (2.8) can be straightforward approximated summing the likelihood-ratio-tests output by NPLM for the separate batches

$$t_{\mathrm{SUM}}^{N_{\mathrm{b}}}(\mathcal{D}) = \sum_{i=1}^{N_{\mathrm{b}}} t(\mathcal{D}_i) = 2 \sum_{i=1}^{N_{\mathrm{b}}} \log \frac{\mathcal{L}(\mathcal{D}_i|\mathrm{H}_{\mathbf{w}}^i)}{\mathcal{L}(\mathcal{D}_i|\mathrm{R})} = 2 \sum_{i=1}^{N_{\mathrm{b}}} \left[\sum_{x \in \mathcal{R}} w_x(1 - e^{f_{i,\widehat{\mathbf{w}}}(x)}) + \sum_{x \in \mathcal{D}_i} f_{i,\widehat{\mathbf{w}}}(x) \right]\,. \tag{2.11}$$

This approach produces a powerful test if the exact shape and normalization of the signal is captured by all models. However, when the signal is very rare its impact in a small size data sample can be overlooked by the model. We can exploit two simple assumptions to provide a better solution. First, we assume that the new physics source is static, namely its occurence rate is uniform over time. Second, we assume that biases affecting single batch views are due to statistical noise whose impact over time is on average null. Under these assumptions, averaging the density-ratio terms learnt in different batches provide a single alternative hypothesis $\mathrm{H}_{\mathbf{w}}^{N_{\mathrm{b}}}$ shared among batches, that helps to enhance systematic discrepancies, interpreted as signals, while suppressing random ones. The density of the data under $\mathrm{H}_{\mathbf{w}}^{N_{\mathrm{b}}}$ would then have the following form:

$$n(x;\, \mathrm{H}_{\mathbf{w}}^{N_{\mathrm{b}}}) = \frac{1}{N_{\mathrm{b}}} \sum_{i=1}^{N_{\mathrm{b}}} n(x;\, \mathrm{H}_{\mathbf{w}}^i)\,. \tag{2.12}$$

By expressing the model of the data density distribution in each batch in terms of the set of functions $\{f_{\mathbf{w}, i}\}_{i=1}^{N_{\mathrm{b}}}$, we can define an "aggregated" model of the log-ratio of the densities as

$$F^{N_{\mathrm{b}}}(x;\, \mathbf{W}) = \log \frac{n(x;\, \mathrm{H}_{\mathbf{w}}^{N_{\mathrm{b}}})}{n(x;\, \mathrm{R})} = \log \left[\frac{1}{N_{\mathrm{b}}} \sum_{i=1}^{N_{\mathrm{b}}} e^{f_i(x;\, \mathbf{w})} \right]\,, \tag{2.13}$$

and use it to compute the log-likelihood-ratio test statistics over the set of batches:

$$t_{\mathrm{AGGR}}^{N_{\mathrm{b}}}(\mathcal{D}) = 2 \sum_{i=1}^{N_{\mathrm{b}}} \log \frac{\mathcal{L}(\mathcal{D}_i|\mathrm{H}_{\mathbf{w}}^{N_{\mathrm{b}}})}{\mathcal{L}(\mathcal{D}_i|\mathrm{R})} = 2 \sum_{i=1}^{N_{\mathrm{b}}} \left[\sum_{x \in \mathcal{R}} w_x(1 - e^{F_{\mathbf{W}}^{N_{\mathrm{b}}}(x)}) + \sum_{x \in \mathcal{D}_i} F_{\mathbf{W}}^{N_{\mathrm{b}}}(x) \right]\,. \tag{2.14}$$

Batching solutions for computationally intensive calculations with kernel methods have already been proposed in the literature for supervised approaches (see for instance ref. [10]). In ref. [10] the authors show theoretical guarantees for recovering the performances of the full sample training solution, provided that the number of splitting is not too large.

2.3 Operational strategies in resource constrained settings

Despite the advantage provided by splitting the dataset in batches, the computational requirements for the algorithm execution might still pose significant challenges to its application in realistic scenarios. In fact, computing $t_{\text{AGGR}}^{N_{\text{b}}}$ requires evaluating $F_{\mathbf{W}}^{M}(x)$ over each element x in the total set of acquisitions $\mathcal{D} = \cup_i \mathcal{D}_i$. This means that all data batches have to be stored for the whole duration of the NPLM algorithm execution, and the memory locations at which the models are stored need to be accessible by all batches. Moreover, the execution time largely varies depending on several factors concerning the algorithm setup, such as the size of the batches, the choice of input variables, but also to the computational resources available and the way these are dispatched. The possibility of running NPLM tasks in parallel on multiple GPUs provides, for instance, the potential to significantly reduce the computational time.

On the other hand, if the storage resources are limited, it could happen that not all the data can be saved long enough for the NPLM routines to be all completed. In this situation, the test in eq. (2.14) cannot be evaluated exactly. Two alternatives can then be contemplated. The first one is to run NPLM over all the batches and save the final models $\{f_{\mathbf{w},i}\}_{i+1}^{N_{\text{b}}}$ to estimate $\text{H}_{\mathbf{w}}^{N_{\text{b}}}$, but running the test only on a subset of N_{test} batches that could be stored:

$$t_{\text{AGGR}}^{N_{\text{b}},N_{\text{test}}}(\mathcal{D}) = 2 \sum_{i=1}^{N_{\text{test}}} \log \frac{\mathcal{L}(\mathcal{D}_i|\text{H}_{\mathbf{w}}^{N_{\text{b}}})}{\mathcal{L}(\mathcal{D}_i|\text{R})} = 2 \sum_{i=1}^{N_{\text{test}}} \left[\sum_{x \in \mathcal{R}} w_x(1 - e^{F_{\mathbf{w}}^{N_{\text{b}}}(x)}) + \sum_{x \in \mathcal{D}_i} F_{\mathbf{W}}^{N_{\text{b}}}(x) \right] . \quad (2.15)$$

The second alternative is considering eq. (2.10) as the limit of a smart multidimensional binning approach, where the bins centres are the elements of the reference sample $\mathcal{R}$. In this limit, each bin has counting equal to the unity under the reference hypothesis and value $e^{F_{\mathbf{w}}^{N_{\text{b}}}(x)}$ under the alternative. A "saturated" binned likelihood-ratio test [11] can then be computed:

$$t_{\text{SAT}}^{N_{\text{b}}}(\mathcal{R}) = 2 \sum_{x \in \mathcal{R}} w_x \left[(1 - e^{F_{\mathbf{w}}^{N_{\text{b}}}(x)}) + e^{F_{\mathbf{w}}^{N_{\text{b}}}(x)} \log F_{\mathbf{W}}^{N_{\text{b}}}(x) \right] . \quad (2.16)$$

Notice that the test does not depend on the data $\mathcal{D}$ directly, but only via $F_{\mathbf{W}}^{N_{\text{b}}}$, that has been learnt exploiting the dataset $\mathcal{D}$ in batches. Moreover, the test does not compute the likelihood-ratio of the data, but it rather evaluates the distance between the density of the reference and that of the alternative hypothesis on the data points belonging to the reference sample $\mathcal{R}$. Eq. (2.16) is a powerful metric when the model $F_{\mathbf{W}}^{N_{\text{b}}}$ is able to capture the signal shape well enough. Its performances deteriorates for signal events that lay in rare regions of the input space, or whose shape is too complex compared to the expressivity of the family of universal approximators used to define the models $\{f_{\mathbf{w},i}\}$. An alternative approach of online data summarization based on machine learning was proposed in ref. [12]. There are two main conceptual differences between the approach presented in ref. [12] and the one presented in the current paper. The first is that the authors of ref. [12] propose to learn directly the probability distribution of the data, while via NPLM we propose to learn the ratio of the data density with respect to a reference. The second difference concerns the training strategy. The authors of ref. [12] suggest to train a unique model (specifically, a normalizing flow) feeding the acquired data batches sequentially. Conversely, in the present work we propose to perform a complete and independent training for each batch, thus preventing catastrophic forgetting.

In the experiments reported in section 4 we consider three representative resource constrained scenarios (depicted in figure 2):

1. **No resource constraints (NPLM-ALL):** all batches are used both for training and for testing. The final test is the one obtained aggregating over all models as in eq. (2.14).

2. **Long term storage constraint (NPLM-ONE):** there are enough computational resources to train multiple instances of the algorithm in a quasi-online fashion but only one batch is available for testing. For this configuration, the final test is the one in eq. (2.15), obtained aggregating over all models and testing on one batch ($N_{\text{test}} = 1$).

3. **Saturated test (NPLM-SAT):** all data are temporarily available for a quasi-online training of several instances of the algorithm but they cannot be stored and tested. In this case, we only test the distance between the density model of the data and that of the reference distribution by computing the saturated test as in eq. (2.16).

2.4 Time-efficient NPLM implementation with non-parametric models

To harness the execution time of a single instance of the NPLM algorithm we consider the kernel-based implementation of NPLM presented in ref. [6], and applied in ref. [9], which exploits the FALKON library [13], a modern solver for kernel methods that leverages several algorithmic solutions developed in refs. [13–15] to allow for their use in large scale problems. The model used in FALKON is a Nystrom approximated Kernel method defined as a finite weighted sum of kernel functions

$$f_{\mathbf{w}}(x) = \sum_{i=1}^{M} w_i k(x;\, x_i,\, \sigma_{\text{FLK}}) \,. \tag{2.17}$$

The kernel function, $k(x;\, x_i,\, \sigma_{\text{FLK}})$, adopted in this implementation is a Gaussian kernel

$$k(x;\, x_i,\, \sigma_{\text{FLK}}) = \exp\left(-\frac{\|x - x'\|^2}{2\sigma_{\text{FLK}}^2} \right) \,, \tag{2.18}$$

where the hyper parameter σ_{FLK} denotes the Gaussian width.

To approximate the log-ratio between the distribution of the $\mathcal{D}$ and $\mathcal{R}$ datasets the training exploits a weighted binary cross-entropy loss

$$L(f_{\mathbf{w}}) = \sum_{x \in \mathcal{R}} w_x \log\left[1 + e^{f_{\mathbf{w}}(x)} \right] + \sum_{x \in \mathcal{D}} \log\left[1 + e^{-f_{\mathbf{w}}(x)} \right] \,, \tag{2.19}$$

with a L2 regularisation term. The NPLM method works under the assumption that the reference sample is large enough to be a good representation of the R hypothesis, which typically leads to an imbalanced dataset with $N_{\mathcal{R}} > N_{\mathcal{D}}$. The weight w_x, is thus introduced to balance the two classes' contributions.
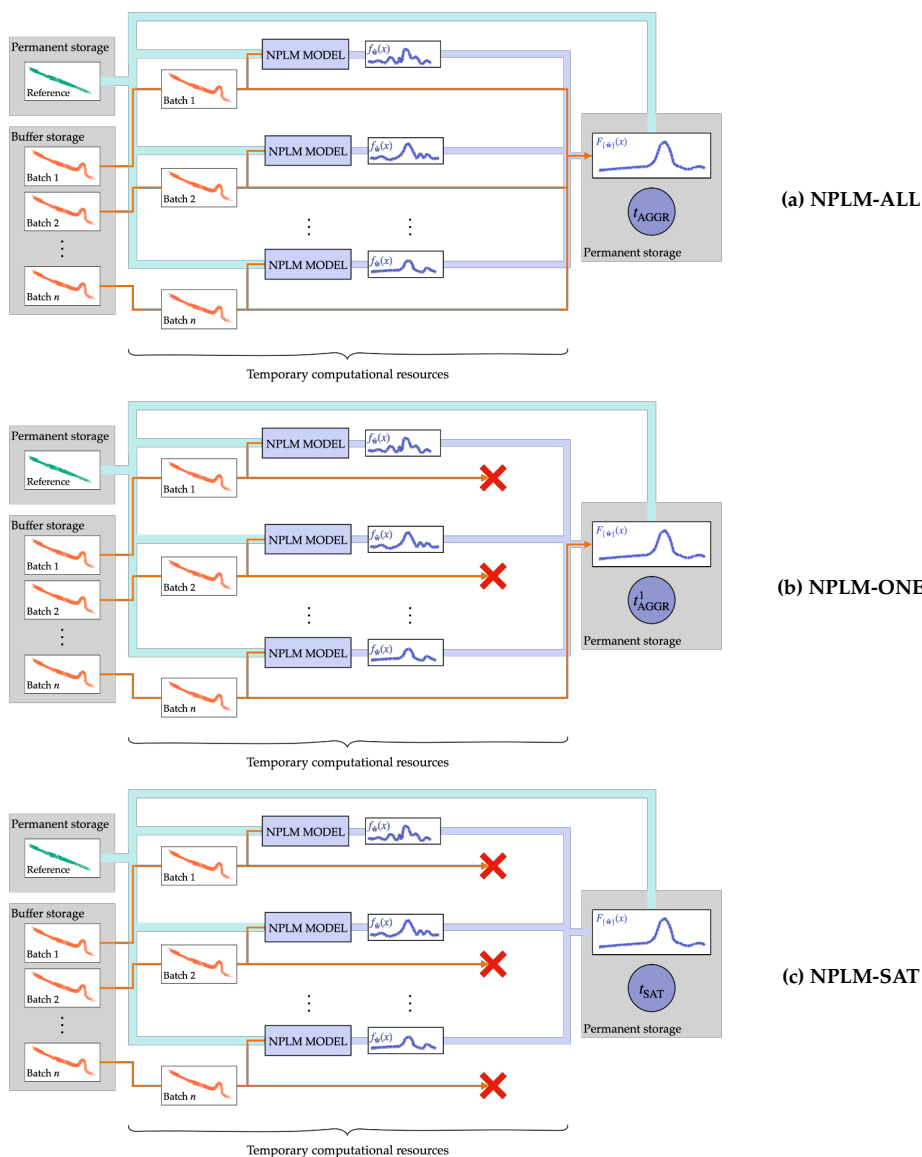
Figure 2. Operational strategies in resource constrained scenarios. Schematic representation of the three scenarios considered in this work. The experimental data are collected in batches in a storage unit or buffer (orange distributions) and used to train multiple instances of the NPLM model (purple squares). In the NPLM-ALL scenario (a), all data batches and trained models are sent downstream to compute the aggregated test statistic in eq. (2.14). In the NPLM-ONE scenario (b), all the models are aggregated but only one batch is sent downstream for testing using eq. (2.15) with $N_{\text{test}} = 1$. Finally, in the NPLM-SAT scenario (c), only the models are stored and can be used to test using eq. (2.16).

Hyper parameters choice. The hyper parameters choice for a given family of universal approximators defines the "modes" of the data that the NPLM algorithm is able to capture. For the kernel-methods considered in this work, there are three hyper parameters that need to be defined: the number of centres M, the Gaussian width σ, and the regularisation parameter λ. In ref. [6] it has been showed that there exists a "valley" of (M, σ, λ) configurations that guarantee the convexity of the loss function, arbitrarily small prediction error and reasonable convergence time. However, no further indication can be provided to enhance the sensitivity to statistical anomalies. Different hyper parameters choices could be optimal to capture different types of data deviations from the reference distribution. For instance, small values of the Gaussian kernels width, σ_{FLK}, would allow to better capture localised deviations, like narrow resonances, while larger values would be better suited to represent effects affecting the distribution on a wider range, like scale mismatches. If the nature of the anomaly is unknown and no good prior is available, then the most inclusive way of testing the data is combining multiple choices of the hyper-parameters configurations. In this work we fix M and λ following the prescription proposed in ref. [6]. Namely, we chose $M = c\sqrt{N}$ with $c \sim O(1)$ such that the training time is reasonable given the available computational resources, and we choose λ as small as possible, provided no instabilities are encountered.

For the Gaussian width, σ_{FLK}, we adopt the approach recently proposed in ref. [16] rather than limiting to a single value we consider five values corresponding to the 5, 25, 50, 75 and 95 % quantiles of the pairwise distance distribution between the elements in the reference set. We combine the results obtained with different values of σ_{FLK} by selecting the minimum:

$$P = \min_{\{\sigma_{\text{FLK}}\}} (p_{\sigma_{\text{FLK}}}) \tag{2.20}$$

It is worth noticing that the authors of ref. [10] comment on the impact of regularisation on the batched solutions, highlighting how the correct level of generalisation can be recovered if the trainings run on the batches is under-regularised. The NPLM algorithm is a signal-agnostic strategy and therefore an optimal regularisation level cannot be determined prior to testing the data. Nevertheless, we observe the benefit of the aggregation as a regularisation in our studies. Details are reported in section 3.

3 Numerical experiments

In this section we study the properties of the batch-based strategy outlined in section 2.3 over two proof-of-concept applications. We start considering a one-dimensional problem characterised by a smoothly falling reference distribution (**EXPO-1D**) for which the optimal test statistic according to Neyman and Pearson can be computed analytically and used as a target. We then move to the multidimensional problem of a dimuon final state observed at collider experiments (**DIMUON-5D**). A detailed description of the datasets' properties and signal benchmarks is given in appendix A.1. The details about the algorithm implementation, execution time and resource consumption are given in section 5. Throughout this section, we will summarise our findings in terms of power curves representing the true positives rate of different tests as a function of the false positives rate α. To help the interpretation, we

transform α to the corresponding Z-score (Z_α in the plots), defined as

$$Z_\alpha = \Phi^{-1}(1 - \alpha), \tag{3.1}$$

where Φ^{-1} is the quantile of the standard normal distribution. For a given signal benchmark, the best performing test is the one which exhibits the highest values of the power at a fixed Z-score.

3.1 Improving regularisation and sensitivity by aggregating over batches

We start by studying the impact of the number of batches $N_{\rm b}$ on the power of the NPLM-ALL aggregated test statistic (eq. (2.14)). We run the NPLM algorithm for different values of $N_{\rm b}$ and we compare the power curves of the aggregated test (eq. (2.14)) and the simple sum of tests (eq. (2.11)) resulting by combining them. We run the experiments on the various signal benchmarks reported in table 2 for the **EXPO-1D** dataset and on table 3 for the **DIMUON-5D** dataset. We report in figure 4 the power curves for one signal benchmark of the **EXPO-1D** dataset, and in figure 5 the power curves for one signal benchmark of the **DIMUON-5D** dataset. Interestingly, the aggregated test shows similar or higher power as the number of batches increases. Conversely, we observe that the simple sum of tests suffers from significant power degradation as the number of batches increases. A possible explanation for the full power recover can be found in the in-sample nature of the NPLM algorithm. Each trained model is by construction induced to (over)fit fluctuations in the data batch distribution within the constraints imposed by regularisation, even when those fluctuations are not statistically significant in the single batch. Combining the outcome of each batch training at the level of the functional forms of the models allows therefore to restore the suite of information carried along by each model, and exploit it to gain sensitivity to very rare signals. As for the observed increased power, we suggest that the inductive bias introduced by averaging over the batches has a beneficial role in the model regularisation, which lead to higher expressivity without a dramatic increasing of the effective degrees of freedom of the test, hence producing lower p-values. Additional studies to investigate this matter are left for future work.

An illustrative example of the aggregation idea is given in figure 3. The panels represent the data distribution (black histograms) and reference distribution (light blue histograms) for the one-dimensional **EXPO-1D** dataset, when the data are split in four batches. A rare Gaussian signal is injected in the tail of the exponentially falling distribution. The light green solid lines represent the models learnt in each training. On the right hand side of the figure, we report four panels showing the individual batches and the relative trained models. The models follow the statistical fluctuations of the training sample producing an imprecise view of the true model of the data. On the left hand side, the four models are combined according to eq. (2.13) to produce the aggregated version, showed in dark green. It is easy to notice the improved stability of the aggregated model with respect to the ones trained on single batches. The observed behaviour is consistent across all the studied signal benchmarks for both datasets. Additional power plots for the aggregated and simple-sum tests for different values of $\sigma_{\rm FLK}$ on different datasets and signal benchmarks are reported in appendix A.2.
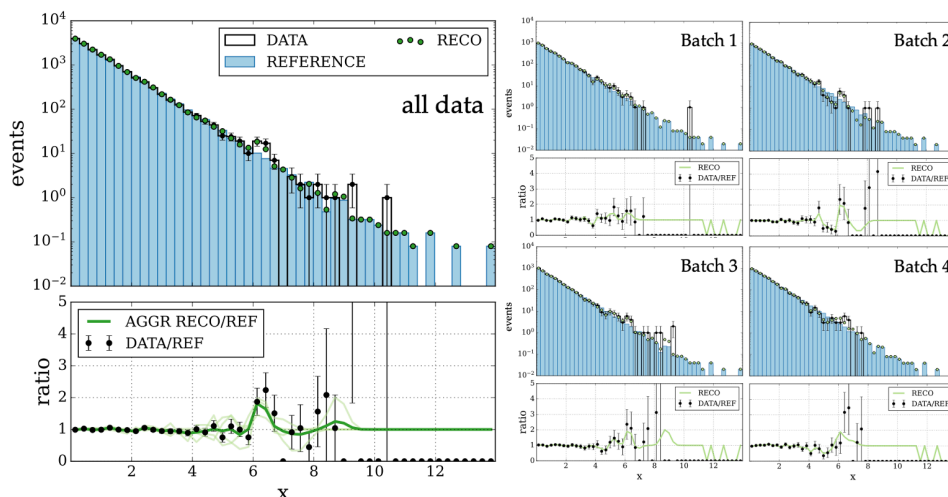
Figure 3. Understanding the regularising effect of the batching strategy. Visual representation of the densities functions reconstructed by the NPLM models trained on different batches (light green lines), and after combining them following eq. (2.13) (dark green line in the main panel). The reconstruction after combination is more robust to statistical fluctuations and approximates the true signal model better than each single one.
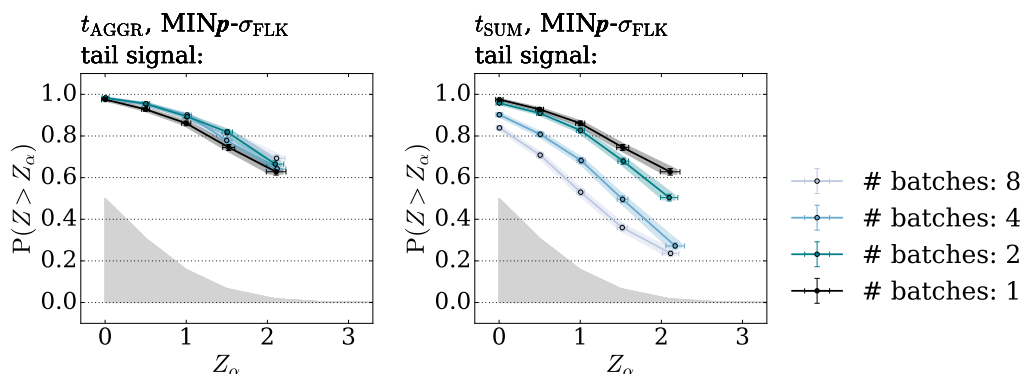

Figure 4. NPLM-ALL. EXPO-1D — localised signal in the tail. Power curves for the aggregated test t_{AGGR} (left panel) and the simple sum of tests t_{SUM} (right panel) at different number of batches. Data splitting improves the performances of t_{AGGR}, while degrades t_{SUM}.

3.2 Improving anomaly detection on a single batch exploiting quasi-online learning

The results presented in section 3.1 suggest that the aggregation improves the quality of the density-ratio modelling. This motivates the use of NPLM in a quasi-online setup to extract powerful information from data even when only a fraction of the latter can be eventually stored (storage constrained scenario). This is always the case at collider experiments, where trigger algorithms are designed to filter the data. We study the impact of aggregating the modelling over multiple batches on the power of the NPLM test for a single data batch (NPLM-ONE scenario). For the **EXPO-1D** dataset, the data are split in eight batches, all of them are analysed via NPLM but only one is saved for testing. For the **DIMUON-5D**
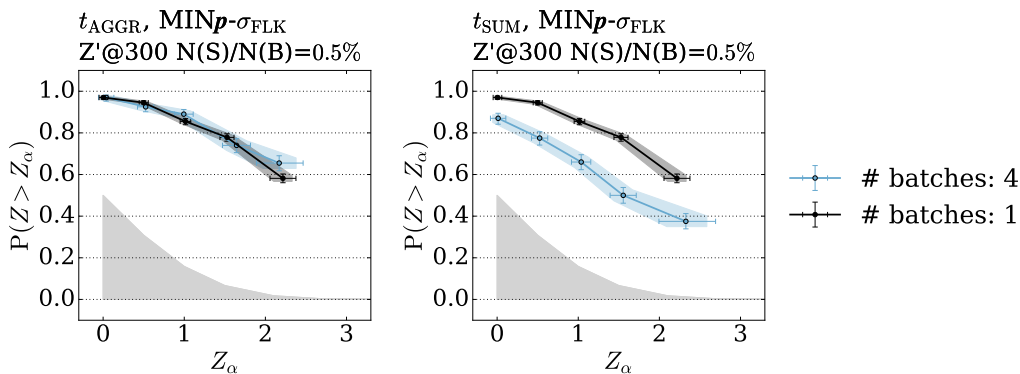
Figure 5. NPLM-ALL. DIMUON-5D — Z' resonance at 300 GeV. Power curves for t_{AGGR} (left panel) and t_{SUM} (right panel) at different number of batches. Data splitting improves the performances of t_{AGGR}, while degrades t_{SUM}.
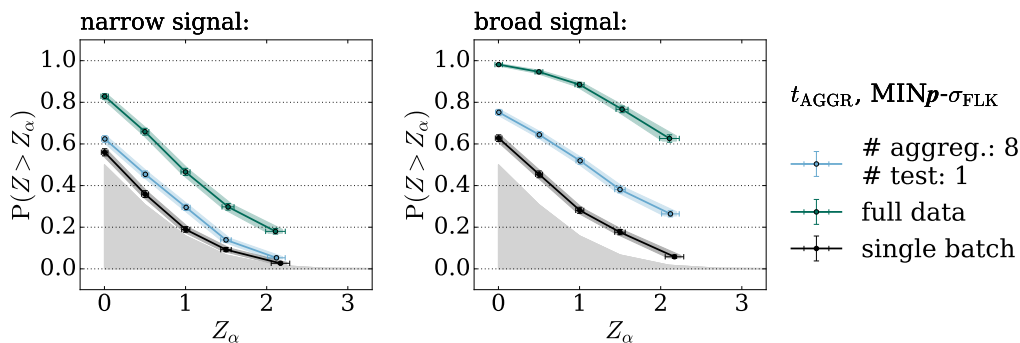

Figure 6. NPLM-ONE. EXPO-1D. Power curves for t_{AGGR} when evaluated over one single data batch (light blue line) compared with the power of the test learnt and evaluated over the full dataset (green line) and one data batch (black line). The aggregation improves the accuracy over the learnt alternative model enhancing the sensitivity with respect to the single batch test. The two panels show the results of our tests for two signal benchmarks, the narrow peak (left panel) and the broad peak (right panel).

dataset, instead, the data are split in four batches. Figure 6 shows the results of our study for the narrow (left panel) and broad (right panel) signal benchmarks in the **EXPO-1D** dataset, whereas figure 7 shows the results for two different injections of a Z' resonant signal with invariant mass of 300 GeV, considered in the **DIMUON-5D** dataset. While not competitive with the ideal scenario of unlimited storage availability in which the full dataset can be tested, either in one batch (green line) or in multiple batches without loosing power, the power curve of our experiment in the NPLM-ONE setup (lightblue line) shows significant improvement over the simple use of NPLM on a single batch (black line). Similar behaviours are found for different signal benchmarks in both the univariate and multivariate datasets. Additional plots are reported in appendix A.3.
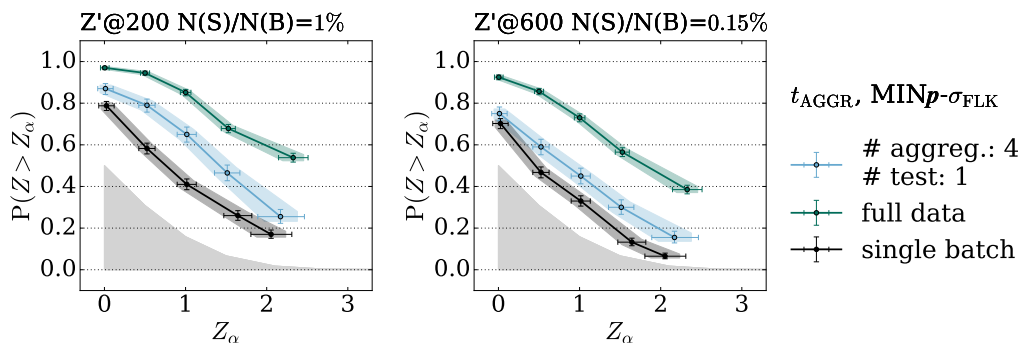
Figure 7. NPLM-ONE. DIMUON-5D. Power curves for t_{AGGR} when evaluated over one single data batch (light blue line) compared with the power of the test learnt and evaluated over the full dataset (green line) and one data batch (black line). The aggregation improves the accuracy over the learnt alternative model enhancing the sensitivity with respect to the single batch test. The two panels show the results of our tests for two signal benchmarks, the a Z' resonance in the bulk of the mass spectrum (left panel) and a Z' resonance in the tail of the mass spectrum (right panel).
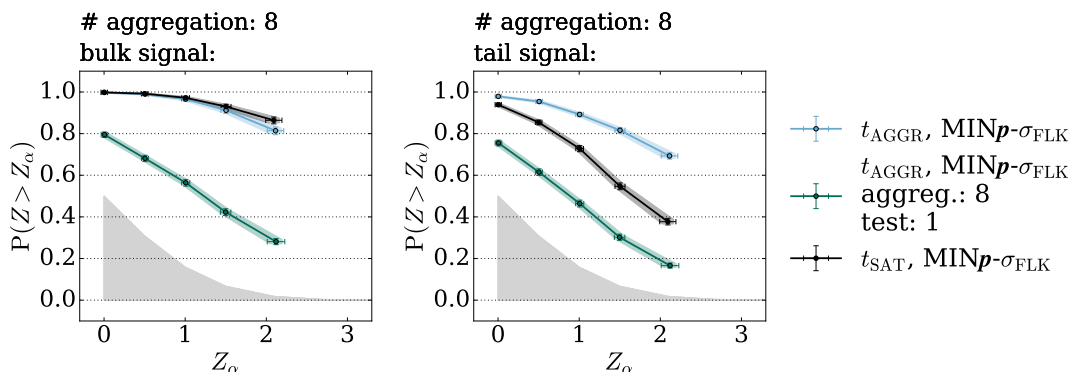
Figure 8. NPLM-SAT. EXPO-1D. Comparison of the aggregated and saturated test statistic power curves for a localized signal in the bulk (left panel) and in the tail (right panel). We compare the power curve of the saturated test (black line) with that of the aggregated test evaluated over th e full dataset (light blue line), or on one batch only (green line). The experiments have been performed aggregating over 8 batches.

3.3 Anomaly-preserving data compression via likelihood-ratio

Finally, we consider the extreme scenario in which data cannot be permanently stored but are only available quasi-online for a limited time window. In this case, the proposed solution is storing the NPLM models $\{f_{\mathbf{w},i}\}_{i+1}^{N_{\mathrm{b}}}$ as compressed representations of the data batches, and build the saturated likelihood-ratio test described in eq. (2.16). The saturated test is not properly a likelihood-ratio test because the likelihood is not directly evaluated on the data points in $\mathcal{D}$. This in principle could lead to inefficiencies. A trivial failure example, for instance, is that of a signal localised in a low density region of the reference data distribution. Since the likelihood-ratio reconstructed by F is only evaluated in the data points of the reference sample $\mathcal{R}$, if the shape of the signal is underestimated then the test would not be sensitive to it. This intuition can be verified running numerical experiments of different nature. Figure 8 shows
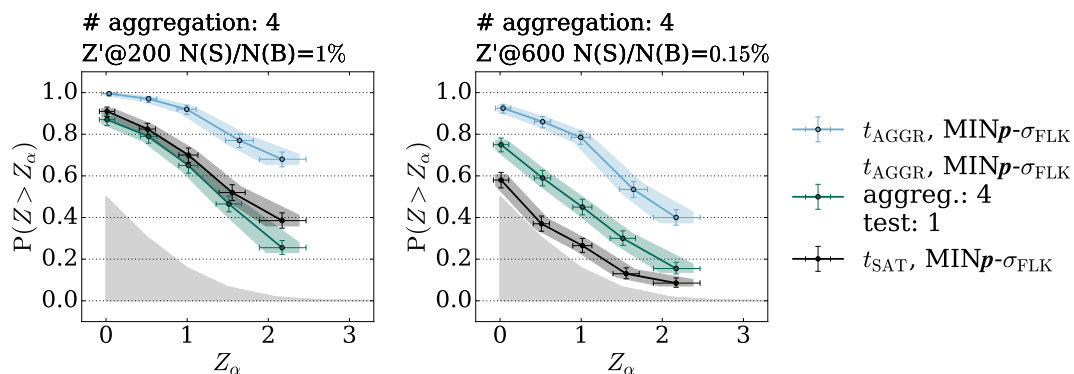
Figure 9. NPLM-SAT. DIMUON-5D. Comparison of the aggregated and saturated test statistic power curves for a Z' resonant signal in the bulk (left panel) and in the tail (right panel) of the mass spectrum considered in this work. We compare the power curve of the saturated test (black line) with that of the aggregated test evaluated over the full dataset (light blue line), or on one batch only (green line). The experiments have been performed aggregating over 4 batches.

two representative experiments run on the **EXPO-1D** dataset. In the figure we compare the power curve of the saturated test (black line) with that of the aggregated test evaluated over the full dataset (light blue line), or on one batch only (green line). For the signal located in the bulk of the reference distribution (left panel), the saturated test has comparable power to the aggregated test evaluated over the full dataset. Instead, for the Gaussian signal in the tail of the reference distribution (right panel), the power of the saturated test deteriorates, as fewer reference data points are available in the signal region to properly quantify the discrepancy. Similar results are found for the **DIMUON-5D** dataset (see figure 9). Additional figures for the remaining signal benchmarks are reported in appendix A.4.

4 Scaling up to big data LHC scenarios

In this section we apply the batch-aggregation approach to a larger scale dataset, emulating a typical data stream acquired at the CMS experiment with a one light lepton (electron or muon) requirement at the first level of data filtering (L1 trigger). The dataset, initially introduced in refs. [17, 18] and later adapted for anomaly detection [19] is publicly available on Zenodo [20–24]. It consists of simulated events for the Standard Model (SM) and four beyond the Standard Model signatures that we report below:[2]

- $A \to 4l$: a neutral scalar boson (A) decaying to two off-shell Z bosons, each forced to decay to two leptons;

- $h^\pm \to \tau^\pm \nu$: a charged scalar boson ($h^\pm$) decaying to a τ lepton and a neutrino;

- $h^0 \to \tau^+ \tau^-$: a scalar boson (h^0) decaying to two τ leptons;

- $LQ \to b\tau$: a leptoquark (LQ) decaying to a b quark and a τ lepton.

---

[2]Additional information about the dataset can be found in ref. [19].

The dataset contains a list of 19 physics objects (4 muons, 4 electrons, 10 jets and the event missing transverse energy) for each collision event. Each object is characterised by three kinematic variables and a class label, for a total of 76 features. We run the batched version of NPLM over the 24 dimensional problem defined by considering the kinematic features of eight out of the nineteen objects: the missing transverse energy, the first two most energetic muons, the first two most energetic electrons, and first three most energetic jets. We consider a data sample $\mathcal{D}$ of 1 million SM events, on top of which signal events of the order of few parts per mill are injected to test the algorithm sensitivity. The marginal distributions over the input features for the Standard Model background and the four signal benchmarks are reported in figures 12 and 13 in appendix A.1.

The reference sample $\mathcal{R}$ is taken to be 10 million SM events (ten times larger than the full $\mathcal{D}$ sample), and the $\mathcal{D}$ sample is split in 5 batches ($N_b = 5$). As for the kernel methods hyper parameters, we set $M = 10\,000$ and $\lambda = 10^{-6}$. For sake of time, we study the performances of the model for a single value of the kernel width σ_{FLK}, corresponding to the 90% quantile of the distribution of pair-wise distance between SM data points. We run the three versions of the algorithm presented in section 3: train and test over all batches (NPLM-ALL), train over all and test only one batch (NPLM-ONE), and the saturated test on the model obtained averaging over all batches (NPLM-SAT).

Figure 10 shows the median Z-score reached by the three approaches for a 0.2% and 0.4% signal injection of the four different signal benchmarks. For completeness, we also show the sensitivity of the algorithm if only a fifth of the luminosity, namely one batch, is exploited by NPLM. The results of our tests confirm the main finding already outlined in section 3. First, the NPLM-ONE approach exhibits increased performances with respect to the application of NPLM to a single batch, remarking that the aggregation via average proposed in this work allows to unveil sensitive information about the signal signature, even when the full set of data is not available at testing time. Moreover, when all the data are available at testing time (NPLM-ALL), the NPLM algorithm can take advantage of the full sample statistics to make discovery possible. This is the case for the $A \to 4l$ and $h^{\pm} \to \tau^{\pm}\nu$ signal benchmarks, where a 2 per mill signal injection (corresponding to a 2 sigma global normalization effect in our studies), can be raised to evidence (for $A \to 4l$) or discovery (for $h^{\pm} \to \tau^{\pm}\nu$) level of significance.[3] For the $A \to 4l$ and $h^{\pm} \to \tau^{\pm}\nu$ signal benchmarks, we observe impressive reaches for the NPLM-SAT approach as well. This result aligns with what observed in the one-dimensional and five-dimensional experiments, namely that NPLM-SAT is as successful as NPLM-ALL when the signal support is mainly in the bulk of the reference sample support, as it is the case for the signal benchmarks considered here (see marginal distributions reported in figures 12 and 13). Additionally, this result points out the potential of the NPLM split-aggregation method as a tool to compute *anomaly-preserving summary statistics*, allowing to

---

[3]The Z-score values reported in figure 10 are computed empirically up to 3. For larger values of the default NPLM, NPLM-ALL and NPLM-ONE, we rely on the asymptotic χ^2 distribution of the test statistic under the null hypothesis (the compatibility of the NPLM test with a χ^2 for kernel methods was previously studied in refs. [6, 9]). For NPLM-SAT we don't observe the emergence of an asymptotic χ^2. However the null distribution has a good agreement with a normal distribution. We therefore use the normal asymptotic to extrapolate an estimate of Z-scores above 3. Points above 3 in the figure should therefore be taken more as a qualitative trend than an accurate estimate.
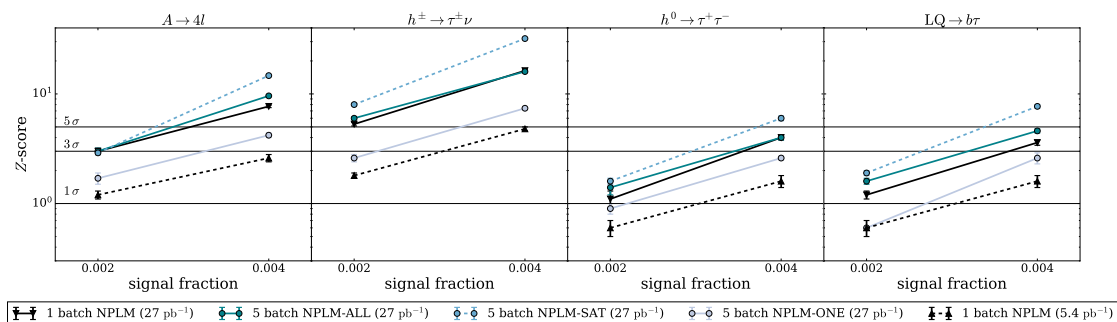
Figure 10. CMS-L1-24D Summary plot. Significance of the tests in terms of the median Z-score as a function of the fraction of signal injection. We split the dataset in 5 batches and run NPLM-ALL (train and test over 5 batches), NPLM-ONE (train over 5 batches and test only one batch), and NPLM-SAT. For comparison, we report the performance of the NPLM strategy run without batching over the full luminosity $(27\,\mathrm{pb}^{-1})$ and over the over a fifth of it $(5.4\,\mathrm{pb}^{-1})$, corresponding to the luminosity of one of the batches used for NPLM-ALL, NPLM-ONE and NPLM-SAT. The total number of background events sum to 1 million. The inspected signal fractions are of the order of few per mill. Each panel corresponds to a different signal benchmark.

save information about data that are only temporarily available and eventually discarded by the data acquisition system. We conclude this section with commenting on the performances of NPLM on the $h^{\pm} \to \tau^{\pm}\nu$ and LQ $\to b\tau$ signals, that are found to be poorer. We believe that the reason could reside in the choice of the features adopted in these studies — only a subset of the original nineteen physics objects is retained for this analysis. Moreover, the algorithm performances are sensitive to the size of the reference sample $\mathcal{R}$ — our studies suggest that the larger the sample the more accurate the result becomes, though more computationally demanding. Finally, one peculiar feature of this dataset is the so called "zero padding", namely filling with zeros the entries corresponding to specific objects in the event that are not observed. The zero padding gives rise to sharp features that project great part of the data on lower dimensional surfaces of the input manifold. The problem of variable size events could be solved choosing a different data representations, model architecture, or by mapping the data to a lower dimensional embedding. These interesting directions are out of the scope of this paper and are left for future work.

5 Computational resources and training time

In this section, we report the details of the algorithm implementation, highlighting the typical execution time and computational resources needed for the numerical experiments studied in this work.

Execution time. The execution time of a single instance of the NPLM algorithm based on kernel methods depends on several factors: the total number of data points, the number of kernels M, the kernel width σ_{FLK}, the smoothness parameter λ, and on the hardware resources employed. Our implementation relies on the Falkon library [13], that allows to speed up the fitting time by optimally exploiting GPUs [6]. We report in table 1 the typical time range for a single batch execution for the hyper parameters choice adopted in our experiments.

dataset	M	σ_{FLK}	λ	$N_{\mathcal{R}}$	$N_{\mathcal{D}}$ (all)	Time (single batch)
EXPO-1D	1k	[0.1, 0.3, 0.7, 1.4, 3.0]	10^{-6}	200k	16k	0.5–1.5 s
DIMUON-5D	10k	[0.8, 1.9, 3.0, 4.4, 6.6]	10^{-6}	40k	8k	3–12 s
L1CMS-24D	10k	[10.5]	10^{-6}	10M	1M	40–70 s

Table 1. Model hyper parameters and execution time. Unless stated otherwise, all instances of the NPLM algorithm for each dataset follow the settings described in this table.

The variance within experiments on the same dataset is mainly due to the σ_{FLK} value, smaller values require a higher number of iterations of the Falkon algorithm to reach convergence. The main difference in execution time between the univariate and multivariate cases is due to the model size, being the number of kernels used in the five-dimensional problems a factor 10 larger than in the one-dimensional case. The training time is not significantly affected by the batch size, as the reference sample size is dominating the overall size of the training sample. For large samples and complex models the execution time increases exponentially with the dataset size (see studies reported in ref. [6]), and the benefit of batching the dataset in terms of execution time become more striking. This has been observed, for instance, when studying the L1CMS-24D.

Storage resources. Aggregating over batches requires both temporary and permanents computational and storage resources. The temporary resources are used to train each single instance of the algorithm and should be sufficient to store the data batch input to the model, and the model itself. Importantly, the model location should be accessible by all data batches to allow estimating $F^{N_{\text{b}}}$ according to eq. (2.13). The reference sample $\mathcal{R}$ and the values that the aggregated model $F^{N_{\text{b}}}$ takes in the test points should instead be stored in a permanent location. The reference sample is a static input to the NPLM algorithm and is common to all training instances.[4] Together with the test data, it is used to asses the value of the test in all three versions described in section 3. The amount of memory resources to be allocated for the aggregated model depends on the number of data points required by the test. For the test in eq. (2.14), the aggregated function $F^{N_{\text{b}}}$ is evaluated over all data points and reference points, requiring a vector of length $N_{\mathcal{R}} + N_{\mathcal{D}}$. For the test in eq. (2.15), only a subset $N_{\text{test}} < N_{\text{b}}$ of batches is used for evaluation and therefore the vector length reduces to $N_{\mathcal{R}} + \frac{N_{\text{test}}}{N_{\text{b}}} N_{\mathcal{D}}$. Finally for the saturated test in eq. (2.16), the aggregated function is only evaluated over the reference set, requiring a $N_{\mathcal{R}}$ length vector only. For the first two cases, the storage resources needed scale linearly with the data taking time, while in the last case they are constant. The implications of information loss on the sensitivity of the different summary statistics are discussed in section 3.2 and 3.3.

---

[4]As previously stated, this work doesn't cover the case of an imperfect reference, namely the case in which the reference model is affected by systematic uncertainties and its optimal configuration given the experimental condition is not constant. We leave this case for future work.

6 Conclusions

In this work we explore the possibility of scaling the use of NPLM to test large size data samples in presence of variable computational and storage resource constraints. We propose to split the dataset in batches and run parallel instances of the NPLM algorithm on each of them. The final results are combined averaging the density-ratio functions learnt by each instance of the NPLM algorithm. We apply this strategy to both univariate and multivariate benchmarks representative of typical data distributions in particle physics experiments. Our experiments show that the split-aggregation strategy preserves or even surpass the sensitivity power of the NPLM test applied to the full sample (original implementation). A possible explanation for the success can be found in the core design of the NPLM training strategy, aiming at the maximum-likelihood fit of the model to the input data without seeking generalisation, and thus obtaining a representation of the true model underneath the data that over-fits the specific sample used for training. While over-fitting is commonly evaded in machine learning algorithms, here it assumes a crucial role in preserving data structures that do not generalise, either because they are statistical fluctuations or because they are especially rare traits of the data. The ability of preserving rare structures is what makes the batching approach for NPLM successful. Subtle hints of anomalous events in different batches are recombined into a statistically significant signal, suggesting that the density-ratio function learnt by the NPLM algorithm for each batch could be a good candidate to build an *anomaly-preserving summary statistic*.

Moreover, the new NPLM split-aggregation strategy offers a solution to offline signal-agnostic analyses of collider data, often characterised by large sample sizes. It is beneficial both in terms of training time and sensitivity performances thanks to its effect of *regularization* against statistical fluctuations, able in some cases to increase the significance of the observed discrepancy. It should be mentioned that a rigorous treatment of the systematic uncertainties affecting the reference sample in this context is crucial. A way to address systematic uncertainties in the original NPLM implementation was proposed in ref. [5] for neural network based models. In the split-aggregation approach additional challenges arise from the fact that each batch has access only to "local" information about uncertainties. An accurate global assessment of the systematic effects via nuisance parameters estimation needs therefore the design of a successful aggregation strategy over the batches (see for instance ref. [25]). The extension of the approach proposed in ref. [5] to handle systematic uncertainties in the split-aggregation NPLM strategy is subject of ongoing study and left for future work.

The potential application of the new batch-based NPLM strategy extends beyond the offline analysis, to quasi-online signal-agnostic analyses of streamed data that are only temporarily available. Online approaches to data analyses allow to inspect collider data at the experiments prior the stage of filtering applied by the triggers, opening the way to the exploration of phase space regions never analysed before, at significantly high rates. Efforts in this direction are ongoing at the LHC experiments. The CMS experiment is investigating new approaches to analyse the experimental data at the collision rate of 40 MHz with focus on the muon detector chambers and the electromagnetic calorimeters [26, 27]. Moreover, the LHCb experiment has recently upgraded to a triggerless data readout at a rate of roughly 30 MHz, with partial event reconstruction that allows to reduce the rate down to 1 MHz prior the final

trigger stage [28]. In this work we proposed two possible solutions to run the NPLM algorithm over a continuous stream of data under storage and computational resources constraints. In complete absence of permanent storage, we propose to train NPLM over multiple batches and test the aggregated density model following the saturated likelihood-ratio approach (NPLM-SAT). In case of limited storage, in addition to NPLM-SAT we propose to train NPLM over multiple batches and evaluate the test only on one batch available for long term storage (NPLM-ONE). Experiments run under constrained storage scenarios (NPLM-ONE) show that performing the aggregation over batches improves the modelling of the signal increasing the power of the NPLM test applied to a single batch. As for the saturated test, we observe sensitivity powers comparable to those of the NPLM algorithm applied to the full statistics when the signal lays on high density regions, while the power is only partially recovered for signals localised in the tails. In both cases, the sensitivity is expected to scale as the statistics of the processed data increases. Online analysis is a promising avenue for anomaly detection. The smart use of machine learning algorithms can leverage large statistics to efficiently navigate unexplored data and recognise novel rare processes of unexpected nature or location. The studies presented in this work are a first step in the direction of investigating online or quasi-online solutions for statistical anomaly detection with the NPLM algorithm. Detailed studies on scalability and uncertainties quantification in typical LHC data streaming scenarios are left for future work.

Acknowledgments

A   Supplementary materials

A.1   Datasets details

EXPO 1D. This simple univariate setup introduced in ref. [3] and further studied in refs. [5, 7] represents an energy or transverse-momentum spectrum that falls exponentially. Such type of distributions are fairly common in collider physics experiments. Studying GoF techniques in this setup is thus illustrative of some of the challenges associated with the search for new physics at these experiments. The density distribution under the reference model is defined as

$$n(x|\mathrm{R}) = \mathrm{N(R)}\, e^{-x}\,, \tag{A.1}$$

where N(R) denotes the number of expected events in the dataset. In our studies we set N(R) at 16k events for the full dataset $\mathcal{D}$, and we consider batches of size 8k (2 batches), 4k (4 batches) or 2k (8 batches) events. To test the model ability to detect features of various

		μ	σ	N(S)/N(R)	Z_{id} (full batch)
Gaussian peaks	bulk	1.6	0.16	$1.5 \cdot 0^{-2}$	4.9
	broad	4	0.64	$6.5 \cdot 0^{-3}$	4.2
	narrow	4	0.01	$1.5 \cdot 0^{-3}$	4.8
	tail	6.4	0.16	$1.5 \cdot 0^{-3}$	4.0
Tail excess	excess			$1.5 \cdot 0^{-2}$	4.5

Table 2. Summary of the 1D signal benchmarks.
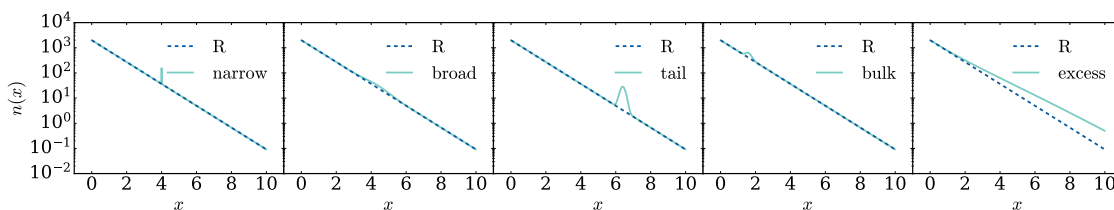
Figure 11. Expo-1D signal benchmarks.

narrowness and in different locations, we consider four Gaussian signals

$$n(x|G_{\mu,\sigma}) = N(S)\frac{1}{\sqrt{2\pi}\sigma_{NP,i}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,, \tag{A.2}$$

and an excess in the tail defined by

$$n(x|E) = \frac{N(S)}{2}x^2 e^{-x}\,. \tag{A.3}$$

The data distribution under the alternative hypotheses has therefore the following form

$$n(x|H_i) = n(x|R) + n(x|S)\,, \tag{A.4}$$

with $S = \{G_{\mu,\sigma}, E\}$. A visual representation of the data distribution under the various signal scenarios is presented in figure 11. For this toy model, the ideal test statistic according to Neyman-Pearson can be computed analytically as

$$t_{id}^S(\mathcal{D}) = 2\log\left[e^{-N(S)}\prod_{x\in\mathcal{D}}\frac{n(x|S)+n(x|R)}{n(x|R)}\right]\,. \tag{A.5}$$

The fraction of signal injection over reference events, N(S)/N(R), the Z-score[5] for the ideal test (Z_{id}), as well as the location μ and scale σ of the Gaussian signals are reported in table 2.

Running the NPLM algorithm requires a reference sample ($\mathcal{R}$) to compare the data $\mathcal{D}$ with. Our $\mathcal{R}$ sample is composed of $N_{\mathcal{R}} = 200k$ events, sampled according to the R hypothesis. It should be noticed that this studied is conducted under the assumption that we can sample synthetic data from the R hypothesis, which is perfectly known, at will.

---

[5]The Z-score is defined as the quantile of a standard normal distribution whose survival function matches the p-value p:

$$Z = \Phi^{-1}(1-p)$$

.

	Mass (GeV)	Σ (GeV)	N(S)/N(R)	Z_{id} (full batch)
Z' scenarios	600	16	$6 \cdot 10^{-4}$	3
		16	$1.5 \cdot 10^{-3}$	7.1
	300	8	$2.5 \cdot 10^{-3}$	5
		8	$5 \cdot 10^{-3}$	8.8
	200	5	$5 \cdot 10^{-3}$	5.1
		5	$1 \cdot 10^{-2}$	9.1
	$c_w(\text{TeV}^{-2})$		N(D)/N(R)	
EFT scenarios	1.0		1.002	3.7

Table 3. Summary of the MUMU signal benchmarks. The signal yields, N(S), and data yileds, N(D), are the ones expected for the luminosity of one batch in our experiments configuration.

Dimuon 5D. The second benchmark considered in this work consists of a set of Monte Carlo simulated proton-proton collisions happening at 13 TeV at the LHC, with two opposite charged muons in the final state ref. [29]. In this dataset, each event is represented by five variables describing the kinematics of the dimuon system: the transverse momentum of each muon ($p_{T,1(2)}$), their pseudorapidities ($\eta_{1(2)}$), and the relative azimuthal angle between the two objects ($\Delta\phi_{12} = \phi_1 - \phi_2$). We focus on final states with transverse momenta greater than 20 GeV, pseudorapidities lower than 2.1 in absolute value, and with invariant mass of the dimuon system larger than 60 GeV. The dominant background process in this configuration is the Drell-Yan, that for sake of simplicity we consider as the only source of background. In our experiments the total number of expected events in the SM background hypothesis is around N(R) = 8000, after the acceptance selections are applied. We then study the impact of splitting the data in four batches and combining them via the batch-based strategy proposed in section 3. As signal benchmarks, we considered a set of Z' bosons with variable mass and width and a set of EFT scenarios described by the dimension-6 4-fermions Lagrangian $\frac{c_W}{\Lambda} J^a_{L\mu} J^\mu_{La}$, with variable Wilson coefficients. Details on the signal benchmarks are given in table 3.

CMS-L1 24D. As supplementary material we report the marginal distribution of the SM dataset and the four signal benchmarks over the 24 input variables considered in this work (figures 12, 13). The narrow peak at zero present in all plots is the effect of "zero padding" (see main text for more comments).

A.2 Improving regularization and sensitivity by aggregating over batches

In this appendix we provide additional plots showing the power curves of the aggregation method NPLM-ALL proposed in this work, both for the univariate **EXPO-1D** dataset (figure 14) and for the multivariate **MUMU-5D** one (figures 15 and 16). The power of the method is compared with a simple sum of tests (details in sections 3 and 3.1).
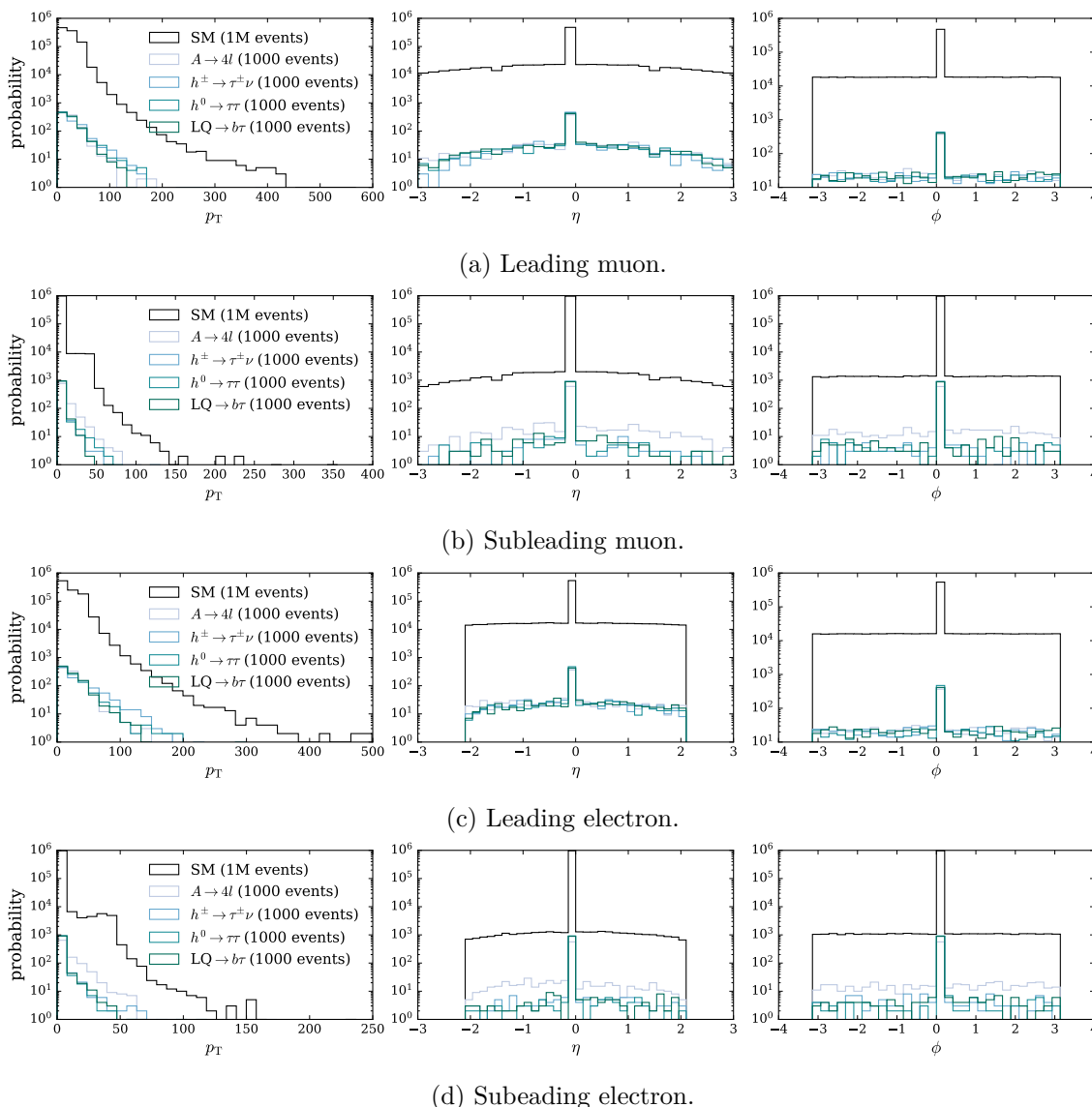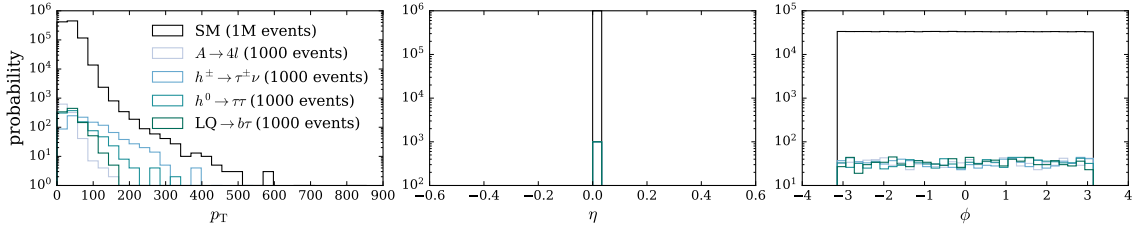
Figure 12. Input features for the CMS-L1-24D dataset (panel 1). In each row we show the transverse momentum, pseudorapidity and azymuthal angle of an object.

## A.3 Impact of aggregation on a single batch test

We provide in figure 17 additional plots showing the power curves of the aggregation method NPLM-ONE proposed in this work. The power of the method is compared with a simple sum of tests (details in sections 3 and 3.2).
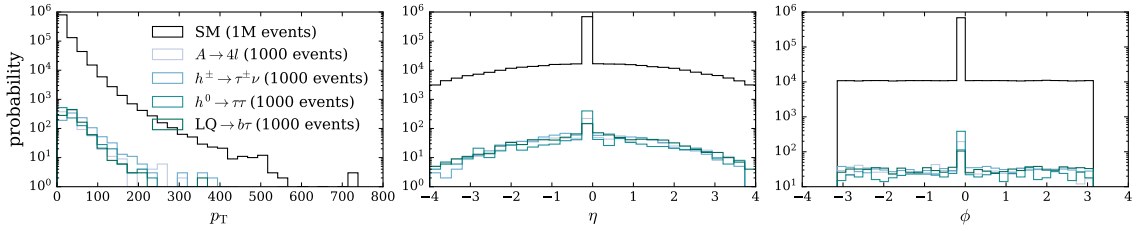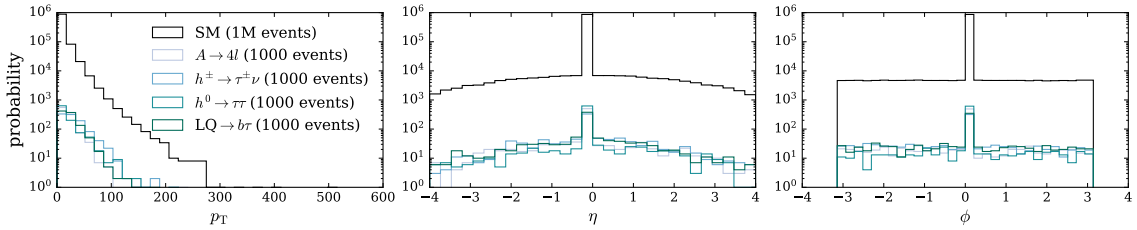
A.4 Saturated test statistic

In this appendix we provide additional plots showing the power curves of the aggregation method NPLM-SAT proposed in this work (see figures 18 and 19). The power of the method is compared with a simple sum of tests (details in sections 3 and 3.3).

(a) Missing transverse energy.
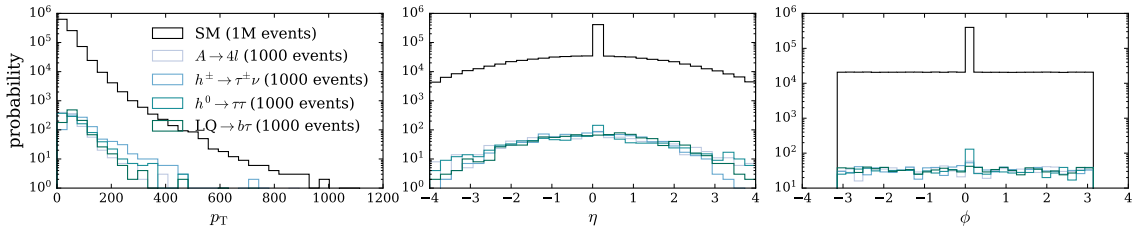
(b) Leading jet.

(c) Second leading jet.

(d) Third leading jet.

Figure 13. **Input features for the CMS-L1-24D dataset (panel 2).** In each row we show the transverse momentum, pseudorapidity and azymuthal angle of an object.
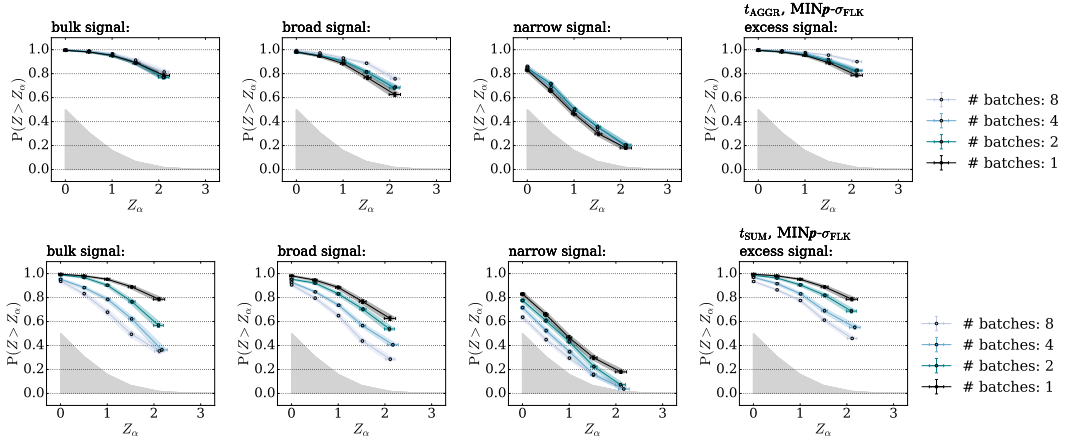
Figure 14. EXPO-1D, NPLM-ALL. Power curves for t_{AGGR} (top row) and t_{SUM} (bottom row) at different number of batches. Each column shows a different signal benchmark. Data splitting improves the performances of t_{AGGR}, while degrades t_{SUM}.
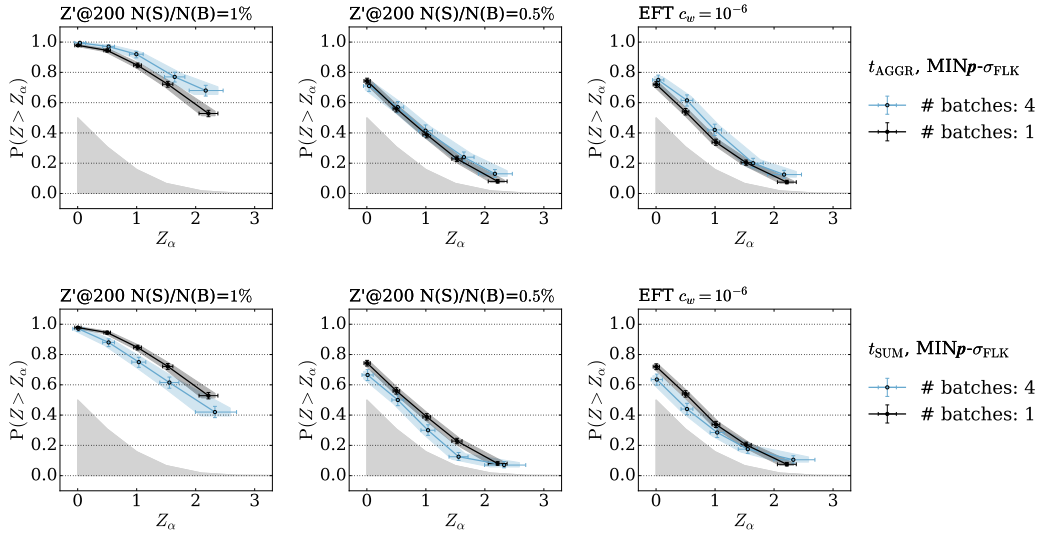
Figure 15. MUMU-5D, NPLM-ALL (panel 1). Power curves for t_{AGGR} (top row) and t_{SUM} (bottom row) at different number of batches. Each column shows a different signal benchmark. Data splitting maintains or improves the performances of t_{AGGR}, while degrades t_{SUM}.
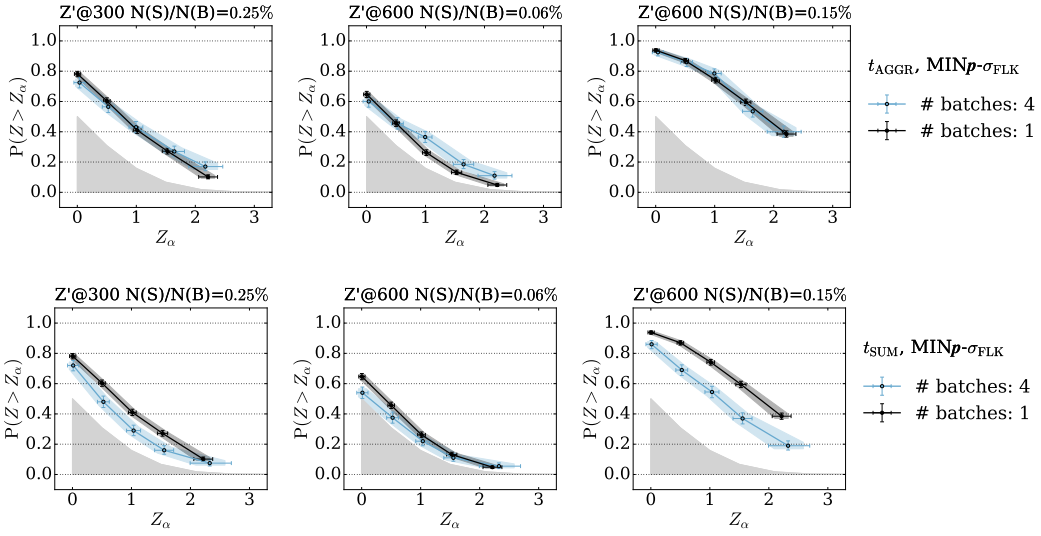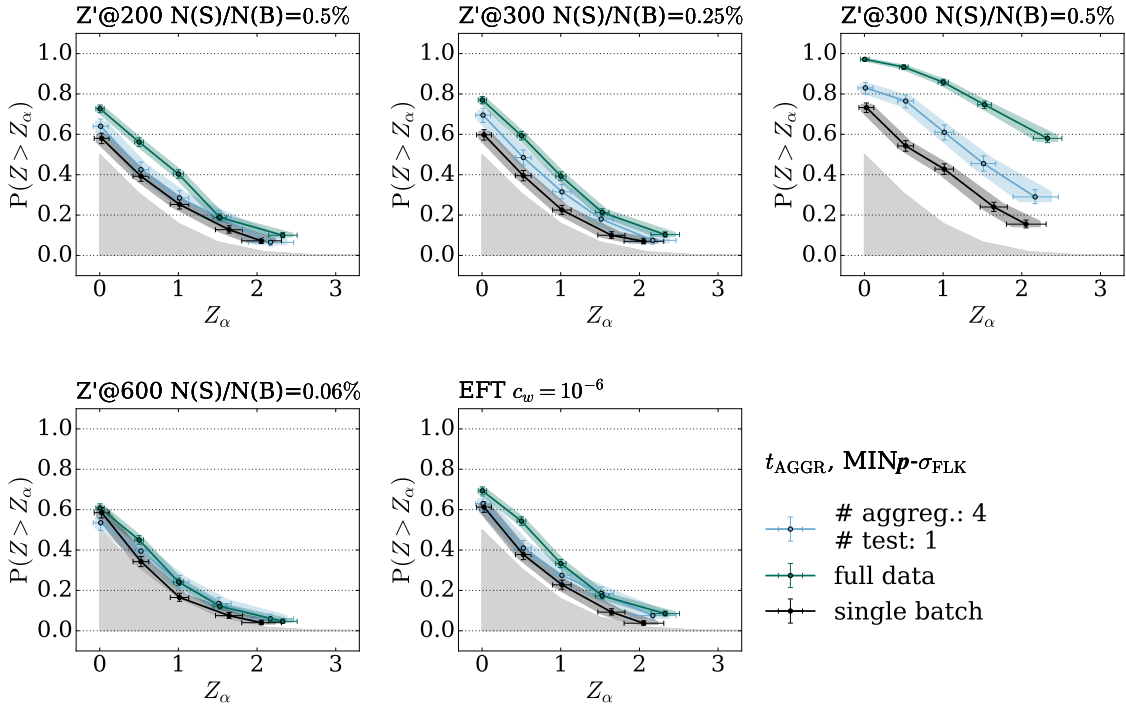
Figure 16. MUMU-5D, NPLM-ALL (panel 2). Power curves for t_{AGGR} (top row) and t_{SUM} (bottom row) at different number of batches. Each column shows a different signal benchmark. Data splitting maintains or improves the performances of t_{AGGR}, while degrades t_{SUM}.

Figure 17. NPLM-ONE. DIMUON-5D. Power curves for the aggregated test t_{AGGR} when evaluated over one single data batch (light blue line) and over the full dataset (green line), compared with the power of the test learnt and evaluated over one data batch (black line). The aggregation improves the accuracy over the learnt alternative model enhancing the sensitivity with respect to the single batch test. The two panels show the results of our tests for two signal benchmarks, the a Z' resonance in the tail of the mass spectrum (left panel) and a EFT scenario (right panel).
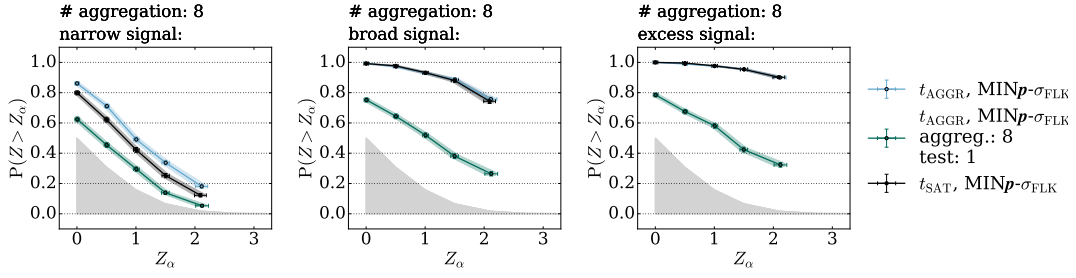
Figure 18. NPLM-SAT. EXPO-1D. Comparison of the aggregated and saturated test statistic power curves for a narrow gaussian signal (left panel), a broad gaussian signal (central panel), and a non resonant excess in the tail (right panel) of the exponential falling reference distribution. The experiments have been performed splitting the dataset in 8 batches. We show in black the power curve of the saturated test (NPLM-SAT), in lightblue the power curve of the aggregated test evaluated over the full dataset (NPLM-ALL), and in green the one of the aggregated test evaluated over one batch (NPLM-ONE).
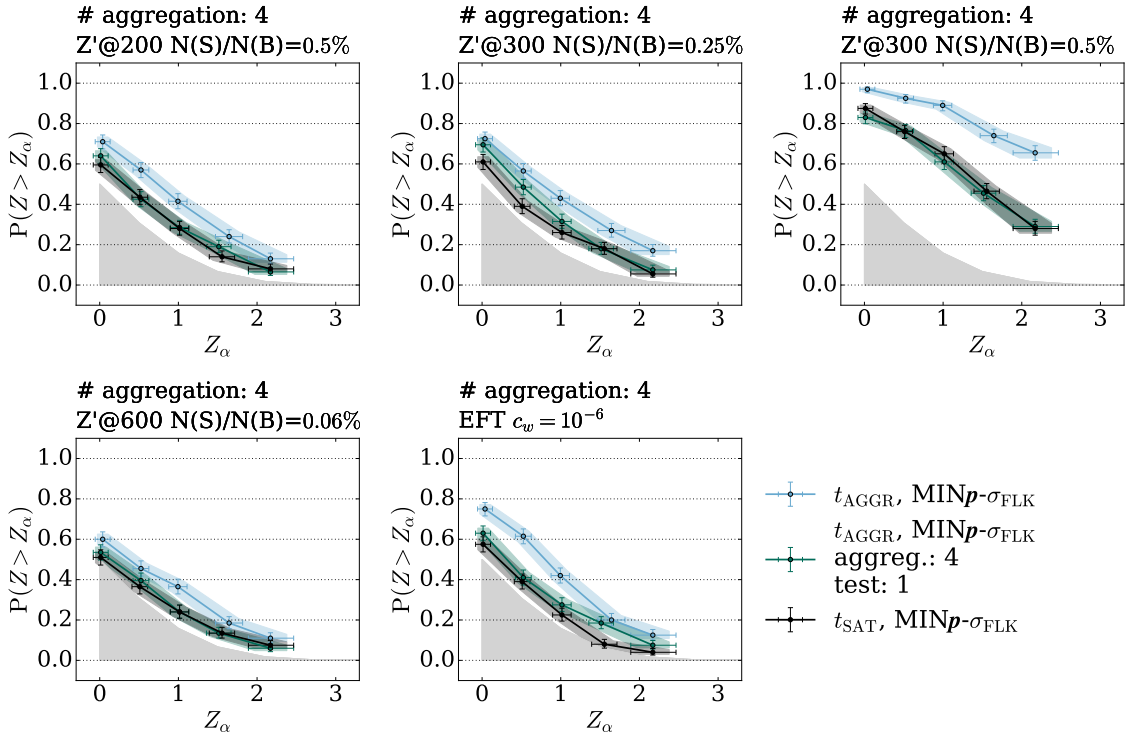
Figure 19. NPLM-SAT. DIMUON-5D. Comparison of the aggregated and saturated test statistic power curves for a Z' resonant signal in the bulk (left panel) and in the tail (right panel) of the mass spectrum considered in this work. The experiments have been performed splitting the dataset in 4 batches.

Data Availability Statement. This article has associated data in a data repository.

Code Availability Statement. This article has no associated code or the code will not be deposited.

References

[1] M. Williams, *How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics*, 2010 *JINST* **5** P09004 [`arXiv:1006.3019`] [ɪɴSPIRE].

[2] C. Weisser and M. Williams, *Machine learning and multivariate goodness of fit*, [`arXiv:1612.07186`] [ɪɴSPIRE].

[3] R.T. D'Agnolo and A. Wulzer, *Learning New Physics from a Machine*, *Phys. Rev. D* **99** (2019) 015014 [`arXiv:1806.02350`] [ɪɴSPIRE].

[4] R.T. D'Agnolo et al., *Learning multivariate new physics*, *Eur. Phys. J. C* **81** (2021) 89 [`arXiv:1912.12155`] [ɪɴSPIRE].

[5] R.T. d'Agnolo et al., *Learning new physics from an imperfect machine*, *Eur. Phys. J. C* **82** (2022) 275 [`arXiv:2111.13633`] [ɪɴSPIRE].

[6] M. Letizia et al., *Learning new physics efficiently with nonparametric methods*, *Eur. Phys. J. C* **82** (2022) 879 [`arXiv:2204.02317`] [ɪɴSPIRE].

[7] G. Grosso, M. Letizia, M. Pierini and A. Wulzer, *Goodness of fit by Neyman-Pearson testing*, *SciPost Phys.* **16** (2024) 123 [`arXiv:2305.14137`] [ɪɴSPIRE].

[8] J. Neyman and E.S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, *Phil. Trans. Roy. Soc. Lond. A* **231** (1933) 289 [ɪɴSPIRE].

[9] G. Grosso et al., *Fast kernel methods for data quality monitoring as a goodness-of-fit test*, *Mach. Learn. Sci. Tech.* **4** (2023) 035029 [`arXiv:2303.05413`] [ɪɴSPIRE].

[10] Y. Zhang, J.C. Duchi and M.J. Wainwright, *Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates*, *J. Machine Learning Res.* **16** (2015) 3299 [`arXiv:1305.5029`].

[11] R.D. Cousins, *Lectures on Statistics in Theory: Prelude to Statistics in Practice*, [`arXiv:1807.05996`] [ɪɴSPIRE].

[12] A. Butter et al., *Ephemeral Learning — Augmenting Triggers with Online-Trained Normalizing Flows*, *SciPost Phys.* **13** (2022) 087 [`arXiv:2202.09375`] [ɪɴSPIRE].

[13] G. Meanti, L. Carratino, L. Rosasco and A. Rudi, *Kernel methods through the roof: handling billions of points efficiently*, *Adv. Neural Inf. Process. Syst.* **33** (2020) 14410 [`arXiv:2006.10350`].

[14] A. Rudi, L. Carratino and L. Rosasco, *FALKON: An Optimal Large Scale Kernel Method*, *Adv. Neural Inf. Process. Syst.* **30** (2017) [`arXiv:1705.10958`].

[15] U. Marteau-Ferey, F. Bach and A. Rudi, *Globally Convergent Newton Methods for Ill-conditioned Generalized Self-concordant Losses*, *Adv. Neural Inf. Process. Syst.* **32** (2019) [`arXiv:1907.01771`].

[16] G. Grosso and M. Letizia, *Multiple testing for signal-agnostic searches of new physics with machine learning*, `arXiv:2408.12296` [ɪɴSPIRE].

[17] O. Cerri et al., *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, *JHEP* **05** (2019) 036 [`arXiv:1811.10276`] [ɪɴSPIRE].

[18] O. Knapp et al., *Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark*, *Eur. Phys. J. Plus* **136** (2021) 236 [`arXiv:2005.01598`] [ɪɴSPIRE].

[19] E. Govorkova et al., *LHC physics dataset for unsupervised New Physics detection at 40 MHz*, *Sci. Data* **9** (2022) 118 [`arXiv:2107.02157`] [ɪɴSPIRE].

[20] T. Aarrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini and K.A. Woźniak, *Unsupervised New Physics detection at 40 MHz: Training Dataset*, (June, 2021) [`DOI:10.5281/zenodo.5046389`].

[21] T. Aarrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini and K.A. Woźniak, *Unsupervised New Physics detection at 40 MHz: $A \to 4$ leptons Signal Benchmark Dataset*, (Oct., 2022) [`DOI:10.5281/zenodo.7152590`].

[22] T. Aarrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini and K.A. Woźniak, *Unsupervised New Physics detection at 40 MHz: $LQ \to b$ tau Signal Benchmark Dataset*, (Oct., 2022) [`DOI:10.5281/zenodo.7152599`].

[23] T. Aarrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini and K.A. Woźniak, *Unsupervised New Physics detection at 40 MHz: $h^0 \to$ tau tau Signal Benchmark Dataset*, (Oct., 2022) [`DOI:10.5281/zenodo.7152614`].

[24] T. Aarrestad, E. Govorkova, J. Ngadiuba, E. Puljak, M. Pierini and K.A. Woźniak, *Unsupervised New Physics detection at 40 MHz: $h+ \to$ tau nu Signal Benchmark Dataset*, (Oct., 2022) [`DOI:10.5281/zenodo.7152617`].

[25] L. Heinrich, S. Mishra-Sharma, C. Pollard and P. Windischhofer, *Hierarchical Neural Simulation-Based Inference Over Event Ensembles*, `arXiv:2306.12584` [ɪɴSPIRE].

[26] CMS collaboration, *A 40 MHz Level-1 trigger scouting system for the CMS Phase-2 upgrade*, *Nucl. Instrum. Meth. A* **1047** (2023) 167805 [ɪɴSPIRE].

[27] CMS collaboration, *40 MHz triggerless readout of the CMS Drift Tube muon detector*, 2024 *JINST* **19** C02050 [ɪɴSPIRE].

[28] R. Aaij et al., *Allen: A high level trigger on GPUs for LHCb*, *Comput. Softw. Big Sci.* **4** (2020) 7 [`arXiv:1912.09161`] [ɪɴSPIRE].

[29] G. Grosso, R.T. D'Agnolo, M. Pierini, A. Wulzer and M. Zanetti, *Nplm: Learning multivariate new physics*, (Jan., 2021) [`DOI:10.5281/zenodo.4442665`].