Authors: Anna Stakia, Tommaso Dorigo, Giovanni Banelli, Daniela Bortoletto, Alessandro Casa, Pablo de Castro, Christophe Delaere, Julien Donini, Livio Finos, Michele Gallinaro, Andrea Giammanco, Alexander Held, Fabricio Jiménez Morales, Grzegorz Kotkowski, Seng Pei Liew, Fabio Maltoni, Giovanna Menardi, Ioanna Papavergou, Alessia Saggio, Bruno Scarpa, Giles C. Strong, Cecilia Tosciri, João Varela, Pietro Vischia, Andreas Weiler

Title: Advances in Multi-Variate Analysis Methods for New Physics Searches at the Large Hadron Collider

Advances in Multi-Variate Analysis Methods for New Physics Searches at the Large Hadron Collider

Anna Stakia [a,b], Tommaso Dorigo [c,*], Giovanni Banelli [d], Daniela Bortoletto [e], Alessandro Casa [f,g], Pablo de Castro [c,h], Christophe Delaere [i], Julien Donini [j], Livio Finos [k], Michele Gallinaro [l], Andrea Giammanco [i], Alexander Held [a,m,n], Fabricio Jiménez Morales [j,o], Grzegorz Kotkowski [f], Seng Pei Liew [d], Fabio Maltoni [i], Giovanna Menardi [f], Ioanna Papavergou [p], Alessia Saggio [i,q], Bruno Scarpa [f,r], Giles C. Strong [l,c,h], Cecilia Tosciri [e,s], João Varela [l], Pietro Vischia [l,t,i], Andreas Weiler [d]

[a] European Organization for Nuclear Research (CERN), Geneva, Switzerland
[b] Institute of Nuclear and Particle Physics, National Centre for Scientific Research 'Demokritos', Athens, Greece
[c] Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Italy
[d] Physics Department, Technical University of Munich, Garching, Germany
[e] Department of Physics, University of Oxford, UK
[f] Department of Statistical Sciences, University of Padova, Italy
[g] School of Mathematics and Statistics, University College Dublin, Ireland
[h] Department of Physics, University of Padova, Italy
[i] Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, Louvain-la-Neuve, Belgium
[j] Laboratoire de Physique de Clermont, Université Clermont Auvergne, IN2P3/CNRS, Clermont-Ferrand, France
[k] Department of Developmental Psychology and Socialization, University of Padova, Italy
[l] Laboratório de Instrumentação e Física Experimental de Partículas, LIP Lisbon, Portugal
[m] Department of Physics, University of British Columbia, Canada
[n] Department of Physics, New York University, USA
[o] Laboratoire Leprince-Ringuet, CNRS, École polytechnique, Institut Polytechnique de Paris, Palaiseau, France
[p] Department of Physics, National and Kapodistrian University of Athens, Greece
[q] Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany
[r] Department of Mathematics, University of Padova, Italy
[s] Department of Physics, University of Chicago, USA
[t] Department of Physics, Universidad de Oviedo, Spain

ARTICLE INFO

ABSTRACT

Between the years 2015 and 2019, members of the Horizon 2020-funded Innovative Training Network named "AMVA4NewPhysics" studied the customization and application of advanced multivariate analysis methods and statistical learning tools to high-energy physics problems, as well as developed entirely new ones. Many of those methods were successfully used to improve the sensitivity of data analyses performed by the ATLAS and CMS experiments at the CERN Large Hadron Collider; several others, still in the testing phase, promise to further improve the precision of measurements of fundamental physics parameters and the reach of searches for new phenomena. In this paper, the most relevant new tools, among those studied and developed, are presented along with the evaluation of their performances.

* Corresponding author.
E-mail address: dorigo@pd.infn.it (T. Dorigo).

1. Introduction

1.1. Background

Over forty quadrillion proton–proton collisions were produced by the CERN Large Hadron Collider (LHC) [1] at the centre of the ATLAS [2] and CMS [3] detectors since the start of LHC operations in 2009. The data samples produced by the reconstruction of the resulting detector readouts allowed those two experiments to vastly expand our knowledge of matter and interactions at the shortest distance scales. Besides delivering a much awaited discovery of the Higgs boson in 2012 [4,5], the two giant multi-purpose experiments published hundreds of precision measurements of fundamental physics constants and searches for new phenomena which previous experiments could not be sensitive to [6,7].

The intrinsic complexity of the collected data and the intent to fully exploit the information they yielded on subnuclear phenomena significantly increased the need of experimentalists to optimize their information extraction procedures by employing the most performant multivariate analysis methods. A concurrent rise in the development of modern machine learning (ML) techniques and the increasing degree of their application to scientific research enabled the LHC experiments to achieve that goal, by squeezing more information from their datasets and improving the quality of their scientific output.

In the above context are set the activities of AMVA4NewPhysics, an Innovative Training Network funded through the Marie-Skłodowska Curie Actions of the European Union Horizon 2020 program. The network, which operated from September 2015 to August 2019, saw the participation of about fifty researchers and students from nine beneficiary nodes among European research institutes and universities,[1] in addition to nine academic and non-academic partners in Europe and the United States.[2] The network, while keeping as its primary goal excellence in training of a cohort of Ph.D. students, conducted cutting-edge research and fostered the use of advanced multivariate analysis methods in physics data analysis for the ATLAS and CMS experiments at the CERN LHC [8].

The four main pillars, upon which most of the studies performed within AMVA4NewPhysics were based, comprise:

1. The customization and optimization of advanced Statistical Learning tools for the precise measurement of Higgs boson properties;
2. The development of new Statistical Learning algorithms to increase the sensitivity of physics analyses targeting model-specific and aspecific searches for new physics;
3. The improvement of the Matrix Element Method through the addition of new tools that extend its applications;
4. The development of new Statistical Learning algorithms for use in high-energy physics (HEP) analyses, ranging from data modelling methods to anomaly detection methods in model-independent searches.

In this paper we summarize some of the research outcomes that resulted from work performed by AMVA4NewPhysics members in the four pillars defined above, highlighting the importance of the results for future studies at the LHC and beyond.

1.2. Plan of this document

The structure of this document follows loosely the order of the four pillars defined above; however, new tools belonging to the fourth one are in some cases described earlier, where they find their most relevant research application. We start in Section 2 where we describe a detailed study of the performance of deep neural networks applied to the complex task of distinguishing a signal of Higgs boson decays to tau lepton pairs from competing backgrounds; the study focuses on the most performant strategies by leveraging information from a competitive effort (the Kaggle 'HiggsMLChallenge'). This is followed in Section 3 by a description of multivariate methods applied to the extraction of the Higgs pair-production signal: an innovative technique for the precise data-driven modelling of multi-jet backgrounds in the search of the $HH \to b\bar{b}b\bar{b}$ process performed by CMS on Run 2 data, and neural-network studies for the extraction of the $HH \to b\bar{b}\tau\tau$ signal in future high-luminosity LHC running conditions.

In Section 4 we describe the development of a high-performance method for identifying the flavour of the parton originating a hadronic jet in CMS data; the resulting algorithms are now among the crucial ingredients for a wide class of analysis tasks, ranging from high-sensitivity measurements of Higgs boson properties, to top quark precision measurements, and to wide-reach searches for new physics signatures in collider data. Section 5 is then devoted to describing improvements achieved on the Matrix-Element Method, which is a complex multi-dimensional calculation that approximates the likelihood function to extract SM parameters from the observed data, and resulting applications to Higgs boson searches in ATLAS and CMS data.

Section 6 focuses on the new methods we designed to search for new physics in LHC data in model-aspecific ways through the identification of anomalous regions of the feature space of the observed datasets; the section also includes description of a technique developed to improve inference on the presence of new physics signals in invariant mass distributions, and its expected performance in searches for high-mass resonances decaying to jet pairs. Section 7 describes an innovative study of particle showers in the ATLAS forward electromagnetic calorimeter, aimed at the production of a fast simulation of the complex physics processes detected by that instrument. In Section 8 we offer an outlook of future studies targeting the end-to-end optimization of data analyses aimed at

---

[1] The involved beneficiary nodes were the Italian Institute for Nuclear Physics and the University of Padova (Italy), the University of Oxford (England), the Université catholique de Louvain (Belgium), the Université Clermont Auvergne (France), the Laboratório de Instrumentação e Física Experimental de Partículas, Lisbon (Portugal), the CERN laboratories, the Technische Universitat Munchen (Germany), and the Institute for Accelerating Systems and Applications (Greece).

[2] The network included as academic partners the Universidad de Oviedo (Spain), the University of California Irvine (USA), the École Polytechnique Fédérale de Lausanne (Switzerland), the University of British Columbia (Canada), the National and Kapodistrian University of Athens (Greece); and as non-academic partners the Mathworks Company (Massachusetts, USA), SDG group Milano (Italy), B12 (Belgium), and YANDEX (Russia).

the loss-less extraction of information from multi-dimensional datasets such as those common in HEP problems, and describe an algorithm we developed for that task. We finally offer some concluding remarks and summary of our review of AMVA4NewPhysics contributions to LHC data analysis in Section 9.

2. Supervised Classification Methods for the Search of Higgs Boson Decays to Tau Lepton Pairs

2.1. Background

The groundbreaking discovery of the Higgs boson in 2012 by the ATLAS and CERN experiments at the LHC [4,5] led to the award of the 2013 Nobel Prize in Physics to P.W. Higgs and F. Englert; the Scottish and Dutch awardees must indeed be commended for their visionary theoretical predictions, which had to wait for almost five decades to be experimentally confirmed. While concluding a long quest for the origin of electroweak symmetry breaking, that scientific milestone initiated a new era of large-scale precision measurement studies and new physics searches related to the Higgs boson and its properties. Curiously, the year 2012 also marks an important milestone for machine learning, as in that year deep neural networks reached paradigm-changing performance in the benchmark problem of image classification [9]. It is thus not a surprise to observe that, from that year onwards, HEP data analysis withstood a boom of applications of ML-based techniques, aimed at optimizing the experimental output of their measurements and searches.

A particularly significant activity in that context was represented by the 'HiggsMLChallenge' competition on Kaggle [10]. That competition, held in 2014, brought together thousands of participants, both belonging to the HEP collaborations most interested in the specific application object of the challenge, as well as academic and non academic participants with background in computer science. Besides that success, the challenge managed to achieve the set goal of understanding which were the most performant machine learning techniques in discriminating the Higgs boson decay to a pair of tau leptons from the various background processes, and at the same time allowed to introduce new promising methods and tools to HEP research and to the broader scientific community. The complexity of the classification task, combined with the high expertise behind the best proposed solutions, made the HiggsMLChallenge competition a benchmark against which to evaluate and compare different ML approaches for supervised classification, as well as to gauge their applicability on HEP datasets. Triggered by the interest of the challenge and the derived conclusions, we conducted a thorough study published in Ref. [11], whose results are summarized *infra*. In parallel, the new Lumin [12] software package was developed to provide implementations of the investigated methods, using PyTorch [13] as the underlying tensor library.

2.2. Challenge details and datasets

The data used in the competition were constituted by information on all particles produced in proton–proton collisions simulated under the 2012 LHC run conditions (a centre-of-mass energy of 8 TeV and a typical instantaneous luminosity of 10^{34} cm^{-2} s^{-1}), which was fed through a simulation of the ATLAS detector, and from which, after applying state-of-the-art reconstruction algorithms, a set of 30 high-level physics observables were derived per simulated collision. The signal was constituted by direct production of a Higgs boson followed by its decay into a pair of τ leptons, $pp \to H + X \to \tau^+\tau^- + X$ (where X denotes any additional produced particles) with the subsequent mixed decay of the τ lepton pair, $\tau^+\tau^- \to e^+(\mu^+)\nu_{e(\mu)}\overline{\nu_\tau} + \tau_h^- \nu_\tau$ (or to the charge-conjugate final state), where τ_h denotes the hadronic decay products of one of the tau leptons. These simulated collision events thus contained a semi-hadronically decaying tau lepton and in addition a reconstructed muon or electron, plus at least three unobserved neutrinos. The background sample consists of three major processes: $Z \to \tau^+ + \tau^-$, $t\bar{t}$, and $W +$jets. The simulated samples include labels identifying the originating process-class (signal or background) of each event, as well as weights normalizing the various contributing processes to the target integrated luminosity.

2.2.1. Data preprocessing

The training and testing datasets [14] consist of 250,000 and 550,000 events, respectively, each a mixture of signal and backgrounds. At the time of the competition, process-identifying labels were only supplied for the training dataset, and the testing dataset was split into two parts: the public test set, for which scores (see Section 2.2.2, *infra*) were supplied to the participant after every submission; and the private test set, for which scores were supplied for the solution selected by the participant once the competition ended, and on which the competition was judged. The analysis discussed *infra* attempted to reproduce the challenge conditions, by developing the model based on scores on the public dataset, and then checking the performance of the final model on the private dataset.

Dataset features consist of low-level information and high-level information, the latter calculated via (non-)linear combinations of the low-level features or from hypothesis-based fitting procedures. Additionally, events carry a weight to normalize the datasets to a fixed integrated luminosity. Since this weight carries information about the event process (signal or background), weights were not publicly available for the testing datasets, and they cannot be used as an input feature in developing the models. Ref. [14] provides a compact summary of the dataset features.

2.2.2. Scoring metric

The performance of a solution classifying testing data as belonging to the signal or background classes is measured using the so-called 'approximate median significance' (AMS) [15,16]:

$AMS = \sqrt{2\left((s+b+b_{reg})\ln\left(1+\frac{s}{b+b_{reg}}\right)-s\right)}$, where s is the sum of weights of true positive events (signal events determined as signal events by the solution), and b is the sum of weights of false positive events (background events determined as signal events by the solution); b_{reg} is a regularization term set to a constant value of 10 for the competition. The AMS provides an approximation of the expected statistical significance of the number of data events selected by the classification procedure, which could be obtained from the *p*-value of observing at least the selected amount of data (signal plus background) under an expectation provided by the background contribution alone in the null hypothesis.

2.3. Deep Neural Network overview

The model architecture used for the reported study is based on an artificial Deep Neural Network (DNN). A Neural Network (NN) attempts to learn a mathematical function that maps a selection of input features to a target function. This is accomplished by a series of matrix operations involving learned weights, *i.e.* parameters that are adjusted throughout the training, and non-linear transformations on the inputs. The NN may be visualized as a set of layers of neurons, each of which receives inputs from the neurons of the previous layer. Layers between the input and output ones are referred to as hidden, and the more hidden layers a NN contains, the 'deeper' it is considered to be. The main choices to be made when constructing a NN, including the ones taken into account and tested in the study described here, are:

- The activation function, which provides a (non-)linear response of each neuron based on the weighted sum of their inputs, in order to provide the neuron output;
- The weight initialization, on the basis of which weights are sampled randomly from a (non-uniform) distribution;
- The loss function, which quantifies the performance of the NN,
- The learning rate (LR), which corresponds to the step size the NN makes over the loss surface at each update point;
- The optimization algorithm used for the learning rate adaptation following changes in the loss function;
- the pre-processing step, which refers to the appropriate transformation of the input features towards improving the weight initialization and decreasing the convergence time;
- The cross-validation (k-fold), related to splitting the training sample into k equally sized portions, and repeating the training and testing procedure k times by training each time the NN on all the portions except the one to be eventually used for the NN's respective testing;
- The ensemble approach, according to which multiple ML algorithms are combined in the direction of improving the performance for a larger range of inputs, and the constituent NN are weighted based on their performance on separate testing sets, for the degree of their respective influence on the output to be regulated.

2.4. Baseline model and alternative techniques

The baseline model in the study reported here is a fully connected, feed-forward DNN, with 4 hidden layers of 100 neurons each. The activation function used is ReLU (Rectified Linear Unit), and the weight initialization relies on He's prescription [17]. The NN output is a single sigmoid neuron with Glorot initialization [18]. The optimization is done via the Stochastic Gradient Descent algorithm with the Adam extension, and the mini-batch size is set to 256 events.

An 80:20 random split is performed on the original training data into training and validation sets, which are then split into ten folds via random stratified splitting on the event class. The testing data is split into ten folds as well, but via simple random splitting. During training, each fold is loaded sequentially.

The tests performed in this study—with a view to comparing different machine learning techniques as to their effect on the performance achieved through the 'HiggsMLChallenge' winning solutions—lie on several levels including the following:

- Combining many models in an ensemble by averaging the predictions of each model, which improves the generalization of the final prediction to unseen data;
- Learning rich, compact, embedding matrices for categorical features [19], as opposed to 1-hit encoding the categories and thereby increasing the number of inputs to the model;
- Choosing a better internal activation function; whilst ReLU is the standard, issues such as "dead ReLU",[3] non-zero-centred outputs, and saturated gradient for negative outputs mean that newer activation functions which aim to solve these issues can be beneficial, such as: ReLU, PReLU (Parameterized ReLU) [17], SELU (Scaled Exponential LU) [20] and Swish [21], the latter defined as Swish(x) = $x \cdot$ Sigmoid(x);
- Using learning rate and momentum scheduling to improve both performance and convergence time, e.g. cosine-annealed LR [22] and 1-cycle annealing of LR and momentum [23];

---

[3] Random weight initialization, or updates during training push the activation into the far-negative region, meaning that the output (and incoming gradient) are consistently zero and the neuron becomes unusable.

Table 1

Comparison between our solution ("This Work") and the top three solutions during the HiggsML Challenge. Performance is measured in AMS (see Section 2.2.2). We report the average performance of our model (evaluated via repeated trainings), but for the challenge entrants' solutions we report the AMS of their submitted predictions. The inference times are the recorded time taken to evaluate every event in the full testing dataset (public and private). Accounting for the differences in precision and throughput between the Nvidia Titan and 1080 Ti GPUs, it is estimated that the 1st-place solution may be trained in 100 min and applied to the full testing set in 8.5 min on a 1080 Ti machine. The developed solution therefore provides an effective speedup of: 92% on GPU or 86% on CPU, for training; and 97% on GPU or 65% on CPU, for inference.

	This work	1st	2nd	3rd
Method	10 DNNs	70 DNNs	Many BDTs	108 DNNs
Train time (GPU)	8 min	12 h	Unknown	Unknown
Train time (CPU)	14 min	35 h	46 h	3 h
Inference time (GPU)	15 s	1 h	Unknown	Unknown
Inference time (CPU)	3 min	Unknown	Unknown	20 min
Private AMS	3.806 ± 0.005	3.806	3.789	3.787

- Employing domain-specific data-augmentation techniques (see *e.g.* [24]) to improve the performance and generalization of the model; for LHC collisions in symmetrical detectors like ATLAS and CMS, events can be rotated in azimuth and flipped in the transverse and longitudinal plane to create new, valid inputs without affecting the class of the events. This may be applied during training time to artificially increase the available training data and during testing time to reduce residual biases in predictions by averaging over a set of augmented inputs;

- Performing advanced ensembling via: Snapshot Ensembling (SSE) [25], Fast Geometric Ensembling (FGE) [26,27], and Stochastic Weight Averaging (SWA) [28]; These methods can either reduce the time required to train and apply ensembles, or allow a larger ensemble to be trained in a similar time as traditional ensembling;

- Using densely connected hidden layers [11]; similar to DENSENET [29], each hidden layer receives as input the output of all previous layers, meaning that prior representations of the data are never lost (thereby protecting against over-parameterization of the model), and parameters have a more direct connection to the output and subsequent back-propagated gradient.

Based on the above options, which are studied separately, this investigation manages to carry out several comparisons with the baseline model, and in so doing reaches important conclusions on the set of choices that is found to benefit the performance most.

2.5. Performance tests

After evaluating the proposed changes (some of them being mutually exclusive), the final model was decided upon. This consists of an ensemble of 10 DNNs, in which predictions are weighted according to their performance on validation data during training (the reciprocal of their loss). The use of ensembling resulted in the largest improvement in performance that was seen in the study. The single categorical feature of the data was passed through an embedding matrix, which offered a minor performance boost. The DNNs were trained using HEP-specific data augmentation, in which the final state particles measured in an event were randomly flipped and rotated in a class-preserving and physics-invariant manner.[4] This was also applied at testing time by computing the average prediction on a set of data transformations. This procedure resulted in the second largest improvement observed in the scoring metric.

Among the activation functions tested, Swish offered the largest improvement in performance. 1-cycle scheduling of the LR and momentum of the optimizer were found to allow DNNs to be trained to a higher level of performance in half of the required time. Finally, by changing the DNNs to be thinner and deeper (six layers of 33 neurons each), and then passing all the outputs of previous layers to the inputs of subsequent layers (dense connections for non-convolutional layers), a moderate improvement in performance was found (along with potential resilience to the exact settings of architecture hyper-parameters).

2.6. Solutions comparison and outlook

Considering the set of choices regarding the model structure and characteristics that gives the most performant outcome, the study proceeds to a comparison with the winning solution. As shown in Table 1, the model proposed by the study can match the score of the winning solution of the HiggsML Challenge, whilst being more lightweight, allowing it to be trained and applied much more quickly, even on a typical laptop. The relative contributions to overall improvement in metric score of each of the changes included in the final model are illustrated in Table 2. Similar to the other solutions, ensembling was found to provide a large improvement, however we can see that domain-specific data augmentation also provides a significant benefit. Whilst smaller in terms of score improvement, the 1-cycle training schedule allows to halve the required training time, and the dense connections provide some resilience to poor choices of network hyper-parameters. An illustration of the final model is shown in Fig. 1.

---

[4] For example, all particles pseudo-rapidities may be simultaneously changed of sign, given the symmetry in the initial state of the collision; similarly, if one observed particle is taken as a reference, all others may be subjected to a mirroring of their azimuthal angles about the axis of the reference particle, ϕ_{ref} as $\phi'_i = 2\phi_{ref} - \phi_i$

Table 2
Breakdown of sources of improvement in metric score for the final ensembled model compared to the starting baseline model.

Technique	Improvement contribution
Ensembling	61.3%
Data augmentation	32.1%
Dense connections	3.6%
Swish + 1cycle	3.0%
Entity embedding	0.1%
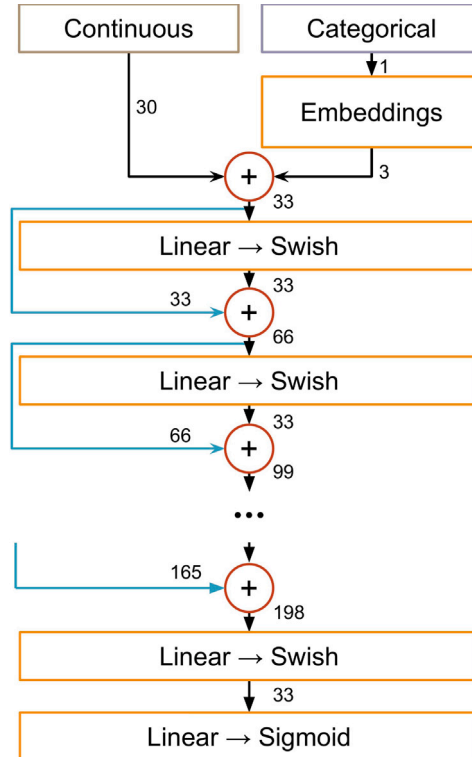
Fig. 1. Illustration of the layout of the network used to build the final model. ⊕ indicates a feature-wise concatenation, providing dense skip-connections to the hidden layers. Six hidden layers (each with 33 neurons) are used in total. The single categorical feature is encoded via an embedding matrix prior to concatenation with the continuous features. The final model consists of 10 such networks in a weighted ensemble.
Source: Image source [11].

Code and hardware details of the HiggsML solutions may be found below:

- This Work (https://github.com/GilesStrong/HiggsML_Lumin):
    - GPU: NVidia 1080 Ti, <1 GB VRAM, <1 GB RAM
    - CPU (2018 MacBook Pro): Intel i7-6500U 4-core CPU, <1 GB RAM;
- 1st place: Melis (https://github.com/melisgl/higgsml)
    - GPU: NVidia Titan, <24 GB RAM
    - CPU (AWS m2.4.xlarge): 8vCPU, <24 GB RAM;
- 2nd place: Salimans (https://github.com/TimSalimans/HiggsML), 8-core CPU, 64 GB RAM;
- 3rd place: Pierre (https://www.kaggle.com/c/higgs-boson/discussion/10481), 4-core CPU (2012 laptop).

This result highlights the need of optimizing the architecture and training of the ML algorithms, as this has been shown to lead to significant improvement over common default choices for both classification and regression problems. The improvement in timing

and hardware requirements are also of great importance, given that most analysers at the LHC do not have on-demand access to high-performance GPUs and must rely solely on their laptops and CPU farms.

The results of this study were verified in a partially independent study of the projected sensitivity to Higgs pair production of the CMS experiment in the High-Luminosity LHC (HL-LHC) scenario [30,31], discussed further in Section 3.3, in which using some of the solution developments discussed above resulted in a 30% improvement in AMS, compared to a baseline DNN under otherwise similar circumstances. This investigation is briefly summarized in the following section.

3. Multi-Variate Techniques for Higgs Pair Production Studies

3.1. Introduction

With the mass of the Higgs boson (m_H) now experimentally measured with sub-GeV precision [32], the structure of the Higgs scalar field potential and the intensity of the Higgs boson self-couplings are precisely predicted in the SM. While measured properties are so far consistent with the expectations from the SM predictions, measuring the Higgs boson self-couplings provides an independent test of the SM and allows a direct measurement of the scalar sector properties. In particular its self-coupling strength λ_{hhh}, can directly be measured through the study of particle collisions in which two Higgs bosons are produced by the same hard subprocess. In LHC proton–proton collisions this process occurs mainly via gluon fusion (ggF) and it involves either couplings to a loop of virtual fermions, or the λ_{hhh} coupling itself. The SM prediction for the double Higgs production cross section is small, *i.e.* approximately $\sigma(\text{ggF}) = 31$ fb at NNLO for a Higgs boson mass $m_H = 125$ GeV at 13 TeV [33]. Beyond Standard Model (BSM) physics effects in the non-resonant case may appear via anomalous couplings of the Higgs boson; the experimental signature of those effects would be a modification of the Higgs boson pair production cross section and of the event kinematics.

If the value of the Higgs self-coupling were significantly different from the value predicted by the SM, it could well be an indication of new physics. Possible scenarios that might give rise to such a deviation include resonant production via a 'heavy Higgs' such as those predicted by BSM theories like the Minimal Super-Symmetric Standard Model, and BSM particles being produced virtually within the coupling loops [34–36]. Unfortunately, the expected cross-section for Higgs pair production is several orders of magnitude below other particle processes that form an irreducible background. Because of this, detecting a statistically significant presence of Higgs pair production events at the LHC requires a much larger number of recorded collisions than the LHC in its current form is expected to produce.

Higgs bosons have a wide range of possible decay channels [37]. The most probable decay for a Higgs boson is to a pair of b quarks, but this decay mode incurs in a large background from Quantum Chromo-Dynamics (QCD) processes yielding multiple hadronic jets. Decay to a pair of τ leptons is the next most probable fermionic decay mode, where τ leptons can decay to light leptons (electrons and muons, with a branching ratio BR of approximately 35%) or hadronically (with a few charged particles in 1- or 3-prongs, with a BR of approximately 65%) in their secondary decays; these decay modes offer a powerful handle on suppressing QCD backgrounds. The $HH \rightarrow \tau\tau bb$ decay mode therefore offers a favourable compromise due to the high branching ratio of $H \rightarrow bb$ and a source of high-purity leptons from the $H \rightarrow \tau\tau$ decay. On the other hand, the $b\bar{b}b\bar{b}$ final state of HH pairs is the most frequent one, yet it has to fight a very large background of QCD production of multiple b-quark pairs. In the following we describe a multivariate study aimed at modelling with high precision the QCD background in the four-b-quark final state, and a study of the potential of the $b\bar{b}\tau\tau$ final state in the HL-LHC data-taking scenario.

3.2. Modelling QCD backgrounds for the $HH \rightarrow b\bar{b}b\bar{b}$ search

The QCD background is the omnipresent problem of searches for rare processes in hadron collisions. While Monte Carlo generators can today accurately model processes with several hadronic jets in the final state, their reliability remains limited in regions of phase space populated by a large number of energetic jets, which are contributed by radiative sub-leading processes that MC generators can only handle with limited precision. In addition, the computational requirements of a thorough simulation of events with a large number of jets makes reliance on MC simulation not always easily practicable. Finally, in specific applications where one needs a precise modelling of not just one single variable, but of the full multi-dimensional density of the multi-jet kinematics, the problem becomes intractable by parametrizations.

Having in mind an application to the search of Higgs boson pair production in CMS data using the four b-jets final state, we devised a precise modelling of the QCD background employing exclusively experimental data. The novel technique we designed, called 'hemisphere mixing', allows for the generation of a high-statistics, multi-dimensional model of QCD events, such that we can base on its properties the training of a multivariate classifier capable of effectively discriminating the Higgs pair production signal.

Event mixing techniques are not a novelty in HEP applications; they have been used extensively in electron–positron collider experiments. Applications to hadron collider physics analysis also exist [38–42], but they are limited to the mixing of individual particles or jets, while our technique for the first time employs hemispheres of jets as the individual objects subjected to a mixing procedure. As jets are direct messengers of the hard subprocess, the creation of artificial events through the mixing of entire jet collections is considerably more complex than the mixing of single particles or jets.
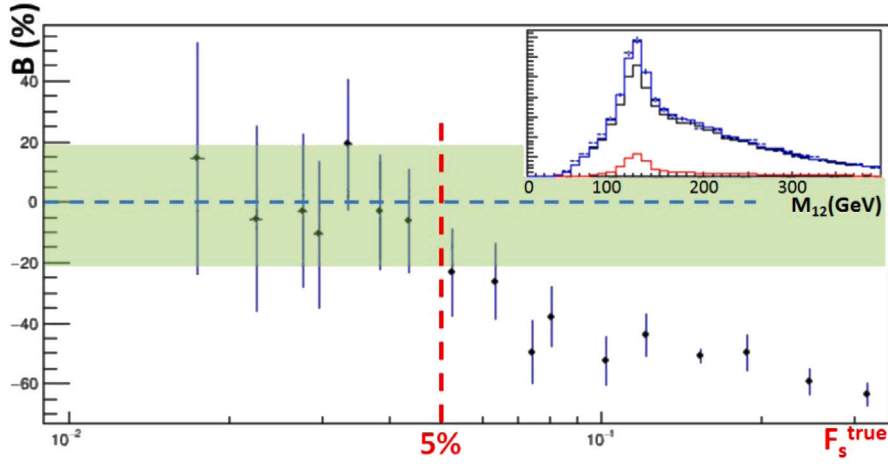
Fig. 2. Size of the bias B, in percentage, on the signal fraction extracted by a fit to the reconstructed Higgs boson mass distribution in data containing a small $HH \rightarrow b\bar{b}b\bar{b}$ signal and a majority of QCD background events, as a function of the signal fraction f_s. The green band shows the level of bias considered acceptable in searches of new signals in hadron collider data. The upper-right inset shows the distribution of the reconstructed Higgs boson mass (M_{12}) for simulations of QCD (black), HH signal (red), and sum of the two components (blue), with overlaid the fit results (point with uncertainty bars) for a 5% signal contamination, which is close to the maximum contamination that still allows a successful modelling of the multi-dimensional distributions [43]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2.1. The hemisphere mixing technique

The idea on which hemisphere mixing is based stems from the observation that QCD events, while extremely complex to model and interpret from the final state point of view, may be thought to originate from a simple tree-level two-body reaction, whereby two partons scatter off one another. The reaction then produces the complex kinematics of a multi-jet event by means of the intervention of second-order effects involving initial and final state QCD radiation, in addition to pile-up collisions in the same proton–proton bunch crossing or even multiple parton scattering of the same colliding protons. Still, the kinematics of the two leading partons emitted in the final state of the hard subprocess offers itself, if properly estimated, as a basis of a similarity measure, which can be exploited to create artificial events based on the individual parton properties.

We consider a dataset of QCD events that feature at least four jets in the final state, and construct in each event a 'transverse thrust axis' using the transverse momentum of the jets. This axis can be defined by the azimuthal angle ϕ_T such that

$$\phi_T = arg\, max_{\phi_{T'}} \left[\sum_j p_{T,j} |\cos(\phi_j - \phi_{T'})| \right] , \qquad (1)$$

where the sum runs over all observed jets j; alongside with T we may define the related variable $T_a = \sum_j p_{T,j} |\sin(\phi_j - \phi_T)|$. Once T is defined, the event can be ideally split in two hemispheres by the plane orthogonal to T. The two hemispheres contain two sub-lists of the jets, characterized by the different signs of $\cos(\phi_j - \phi_T)$. We may describe each hemisphere by a number of observable features, $h(N_j, N_b, T, M, T_a, P_z)$. In this expression N_j is the number of contained jets, N_b is the number of b-tagged jets, T (T_a) is their transverse momentum sum along the thrust axis (orthogonal to it), M is their combined invariant mass, and P_z is the sum of longitudinal components of the jets. Using the $2N$ hemispheres that can be obtained from N observed data events in the same sample where we wish to search for a small signal we may build a library h_i ($i = 1, \ldots, 2N$), which is the basis of the construction of synthetic events. If h_l and h_m are the hemispheres obtained from the splitting of a real event, we may construct an artificial replica of that event by identifying the two hemispheres h_p, h_q in the library that are the most similar to h_l and h_m (subjected to the condition that none of the indices l, m, p, q are identical). Similarity is defined within the sub-classes of hemispheres with the same value of N_j and N_b by a normalized Euclidean distance in the space of continuous parameters (T, M, T_a, P_z) describing the hemispheres. More detail on the procedure is available in [43].

3.2.2. Results

Multi-dimensional statistical tests that employ a complete set of kinematic variables describing the event features prove that the produced artificial events model the multi-dimensional distribution of the original features to sub-percent accuracy, if the dataset is constituted by QCD events. Furthermore, when the data contain a small fraction of events originated by Higgs pair production events the mixing procedure washes out the features of that minority class, such that the artificial data sample still retains accuracy in modelling the QCD properties (see Fig. 2, which shows the effect of different signal fractions in the modelling). This happens because the probability that two hemispheres in the library, chosen to model a signal event, be both originated from other signal events (and thus retain memory of the peculiarities of the multi-dimensional density of signal in the event feature space) scales with the square of the signal fraction in the original data. Hence a small signal contamination present in the dataset will not impair the validity of the model. This property makes the hemisphere mixing method an attractive option for the search of rare signals in QCD-dominated
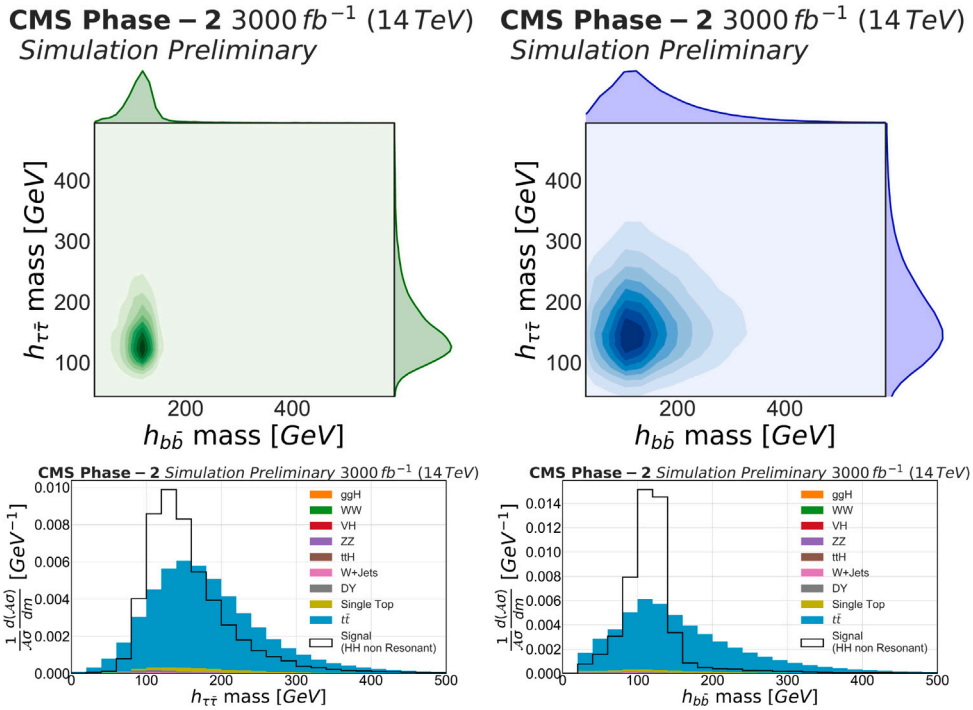
Fig. 3. Two-dimensional distributions of the invariant masses of the reconstructed Higgs bosons ($h_{\tau\tau}$ and h_{bb}) in signal (left) and background processes (right). One-dimensional invariant-mass distributions are projected on the corresponding axes; Bottom: One-dimensional distribution of the $h_{\tau\tau}$ (left) and h_{bb} (right) invariant masses for signal and background process for all final states together. Distributions of signal and background are separately normalized to unit area. Images source [45] (supplementary material).

datasets. It is of special interest the fact that the user does not need to identify a control sample of data where to perform modelling studies: the method can be directly applied to the same sample where the signal is sought for. This is a considerable simplification of the experimental analysis, which also reduces modelling systematics. Finally, by searching for multiple similar hemispheres to the two that make up the event to be modelled, one may construct a synthetic dataset much larger than the original one, reducing the statistical uncertainties without introducing appreciable systematic biases.

The hemisphere mixing procedure has been successfully used for the first search of Higgs pair production in the four b-jets final state performed by the CMS experiment [44]. In that analysis the technique enabled the training of a multivariate classifier on a large synthetic dataset of hemisphere-mixed events, as well as provided the background model from which a limit on the Higgs pair production signal was extracted.

3.3. Prospects of the $HH \to b\bar{b}\tau\tau$ channel at the HL-LHC

A study of the sensitivity of the HL-LHC to the Higgs self-coupling using advanced analysis techniques was performed by considering SM Higgs pair production in proton–proton collisions at $\sqrt{s} = 14$ TeV. DNNs were trained for the task of separating the signal from background contributions. Details can be found in [30,31]. The study was pursued on the $HH \to \tau\tau bb$ decay mode.

The decay of a Higgs boson to $\tau\tau$ pairs gives rise to six possible combinations of final state signatures for the signal: $e\tau_h$, $\mu\tau_h$, $\tau_h\tau_h$, ee, $\mu\mu$, and $e\mu$, where τ_h indicates a hadronically decaying τ lepton. For this investigation, we only consider the three most frequent final states, *i.e.* those involving at least one τ_h. From the event selection, a total of 52 features are used in the study. They are split into "basic" (27), "high-level/reconstructed" (21), and "high-level/global" (4) features. These proved to give the best performance.

From each selected event the two Higgs bosons are reconstructed making use of the considered final states. Distributions of the invariant masses of the reconstructed Higgs bosons used as inputs for the DNN are shown in Fig. 3.

Simulated data were pre-processed with a 50–50 split into training and testing sets, and were used to train a DNN with several optimizations. Models were compared using the Approximate Median Significance [15,16], however in order to include the presence of uncertainties on the background, an extended version is used, compared to Section 2.2.2:

$$\text{AMS} = \sqrt{2(s+b)\ln\left(\frac{(s+b)\left(b+\sigma_b^2\right)}{b^2+(s+b)\sigma_b^2}\right) - \frac{b^2}{\sigma_b^2}\ln\left(1+\frac{\sigma_b^2 s}{b\left(b+\sigma_b^2\right)}\right)}, \tag{2}$$

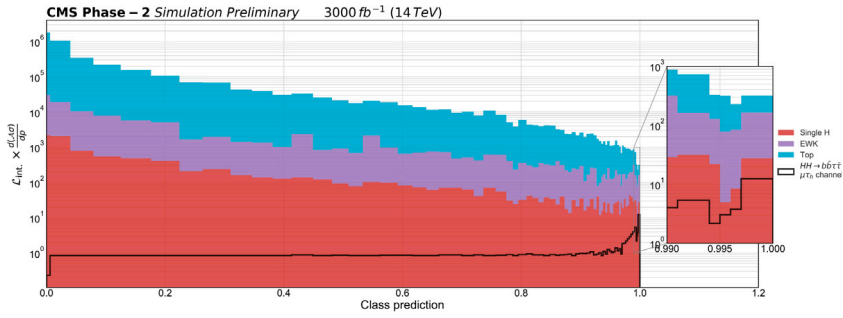**CMS Phase − 2** *Simulation Preliminary*   3000 fb^{-1} (14 TeV)

Fig. 4. Predictions of the classifier evaluated on the test dataset for the $\mu\tau_h bb$ final state. Both signal and background are normalized to the expected yields. Images source [45] (supplementary material).

where σ_b is the uncertainty on the number of expected background events. For model development, a 10% systematic uncertainty on the background normalization was assumed, in addition to statistical uncertainties. The final computation of the analysis sensitivity used appropriately estimated systematic uncertainties from a variety of contributions, accounting for their correlations.

We studied the performance of NN classifiers in terms of how well they can classify signal and background events. This involved reconfirming the benefits of several of the techniques already studied in the work summarized in Section 2. The final ensembled model uses SELU activation [20], a cosine-annealed learning rate [22], and the data augmentation described in Section 2. These techniques resulted in a 30% improvement in AMS over the performance of a baseline ReLU model.

Signal and background events are binned in the distributions of the classifier prediction for signal and background in each channel (Fig. 4). A simultaneous fit is performed on the expected event distributions for the three final states considered. Including systematic uncertainties, an upper limit on the HH cross section times branching fraction of 1.4 times the SM prediction is obtained, corresponding to a significance of 1.4 σ in this final state alone. When results are combined with the other final states, a significance of 2.6 σ is achieved, and 4 σ when combining the two experiments, ATLAS and CMS.

The precise characterization of the Higgs boson will be one of the highest priorities of the HL-LHC physics program. The MIP Timing Detector (MTD) [46] is a new detector planned for the CMS experiment during the HL-LHC era and it will enhance the physics reach capabilities. In this context, the improved object reconstruction and the related effects were quantified. In particular, the HH study discussed *supra* was repeated to account for the new MTD showing a further improvement. For details see [47].

4. Jet Flavour Classification

4.1. Overview

The correct reconstruction and identification of all particles interacting with the different types of detector material constitutes a fundamental prerequisite to extract information from detected particle collisions at the LHC experiments. Here we focus on the reconstruction of hadronic jets, which is more challenging than that of other measurable physics objects because of the complexity of the physics processes of relevance, and which is an important ingredient to the vast majority of measurements and searches carried out with the ATLAS and CMS experiments.

Hadronic jets can be defined as collimated sprays of particles emerging during the hadronisation process of a parton (quark or gluon) emitted with high energy from the collision point. Jets may originate from b quarks, c quarks, so-called 'light quarks' (u, d, s), and gluons. Due to their larger mass than all other partons, and to other specific properties of the hadrons they produce in their hadronisation, b and c quarks (usually denoted as 'heavy flavour' quarks) yield jets that may be distinguished from the rest. The identification of the type/flavour of the initial parton that is associated with the jet, referred to as jet tagging, constitutes an essential stage of the jet reconstruction process. In particular, the efficient identification of heavy flavour jets is a subject of paramount importance for a number of measurements and searches, due to the possible connection of their production with physics processes preferentially coupling to the second and third generation of matter fermions – as is the case for the search in [48], while it is also critical for Higgs boson measurements because of the large branching fraction of the Higgs to $b\bar{b}$ and $c\bar{c}$ quark pairs.

The potential of new ML tools for heavy-flavour tagging must therefore be investigated thoroughly by HEP experiments. In this context, the degree to which the use of novel deep learning techniques may improve heavy flavour jet identification in CMS has been clarified by developing the DeepCSV, DeepFlavour, and DeepJet taggers [49–51], a task that received a significant contribution by AMVA4NewPhysics members. Extension of the applicability of the above mentioned taggers for the distinctive case of quark/gluon discrimination has also been examined within this study. On top of the development and evaluation of the jet tagging models, AMVA4NewPhysics researchers have also considerably contributed to the integration of these taggers into the CMS reconstruction software [52,53], meeting the strict computational and performance requirements set out by this particular task. The architectures developed have been the first advanced deep-learning architectures to be integrated within the CMS reconstruction pipeline, and the core integration implemented for this purpose has been re-used for additional models and tasks thereafter.

4.2. Particle-flow jets and B hadrons identification

The so-called 'particle-flow' jets consist of a list of particles reconstructed via the particle-flow algorithm [54], which is commonly used in CMS reconstruction, and clustered with the anti-k_T clustering algorithm [55]. The particle-flow algorithm aims at identifying all observable particles in the event by combining information from all CMS sub-detectors. What distinguishes a b-quark-initiated jet from other jets at reconstruction level are several detectable particularities stemming from its typical features. At the jet-formation stage, while light quarks and gluons hadronize by predominantly producing short-lived hadrons whose decay products yield tracks that originate directly at the collision point (primary vertex), the B hadron created by the hadronisation of the b-quark has a comparatively long lifetime (of the order of a picosecond), which leads to the creation of a secondary vertex at the point of their decay, significantly displaced from the primary vertex; the secondary vertex can be reconstructed from the measured trajectories of charged tracks possessing a significant impact parameter[5] with respect to the primary event vertex. Other detectable characteristics of the b-jets include a relatively large opening angle of decay products of the heavy B hadron, the possible presence of an electron or muon produced by the semi-leptonic B hadron decay, and a different track multiplicity distribution and fragmentation function with respect to those of jets originated by other partons. All the above information is used by the b-jet taggers in CMS and ATLAS, so as to identify b-quark-originated jets with the best possible accuracy.

4.3. The DEEPCSV tagger

The DEEPCSV algorithm [56], compared to the previously-standard b-tag classifier CSVv2,CMS-PAS-BTV-15-001, has the same input in terms of observable event features, but processes a larger number of charged tracks. DEEPCSV is also a deeper NN, and it is trained for multi-class classification. More specifically, the network is composed of five dense layers of 100 nodes each, and its input can be in total of around 70 variables. After selecting charged tracks passing quality criteria, eight features are used from up to six tracks with the highest impact parameter as part of the input set. Eight additional features summarize information from the most displaced secondary vertex, and finally, 12 features are constructed with jet-related observables (global variables). In general, the features used in DEEPCSV are similar to the ones traditionally used in CMS for b-tagging [57]. The multi-class classification property of DEEPCSV allows the use of four output classes instead of two of binary classifiers. The classes include the following cases describing the originating parton: a b-quark, a c-quark, a light quark (u/d/s), or a gluon. Also, the case that two B hadrons happen to co-exist inside the same jet is studied through a separate class.

4.4. The DEEPFLAVOUR and DEEPJET taggers

DEEPFLAVOUR [58], compared to DEEPCSV, has a much larger input (around 700 variables at most), is a deeper NN (eight fully connected layers, the first one being of 350 nodes, while the rest of 100 nodes each), and it includes convolutional layers. Besides the charged jet constituents, whose number is now increased (up to 25), along with the number (16) of features extracted from their kinematical properties, the input includes also information from identified neutral jet constituents (up to 25), with which six additional features are constructed. Moreover, there are now up to four secondary vertices considered as additional input, upon which 12 features are used. Finally, the input information includes six global variables describing the jet. In order to extract and engineer features per object, particle or vertex, several 1×1 convolutional layers are used on the input lists of objects. For charged particles and secondary vertices, four layers of 64, 32, 32, and 8 filters are applied. For the neutral particles, which carry considerably less information, only three layers are used, with 32, 16, and 4 filters.

The DEEPJET tagger [59] was introduced as an updated version of DEEPFLAVOUR intended for additional discrimination power between jets originating from gluons and jets originating from light quarks: both categories were expected to fall into the same 'udsg' output class in the case of DEEPFLAVOUR. In DEEPJET the output of the convolutional layers is given to Long Short-Term Memory (LSTM) [60] recurrent layers of 150, 50, and 50 nodes, which respectively correspond to the charged particles, the neutral particles, and the secondary vertices. These intermediate features are concatenated with the six global features, and then given to the dense NN, whose first layer has 200 nodes for DEEPJET instead of the 350 nodes of DEEPFLAVOUR [61–63].

4.5. Performance comparison

Fig. 5(a) shows a significant improvement in performance between DEEPCSV and CSVv2. For example, for the same true positive rate (b-jet efficiency) of 65%, DEEPCSV offers a 40% reduction in false positive rate (misidentification probability) for light jets (uds- and gluon-jets). The 1% false positive rate is a typical working point used for the classification. This result refers to simulated event samples of top quark pair production; however, this gain in the performance has been validated in real collision data. The latter was made possible via the comparison of the data-to-simulation scale factors—that refer to the b-tagging efficiency measurement with the use of several different methods—between DEEPCSV and CSVv2, which shows agreement [56], thus implying that the observed improvement in performance in simulation is additionally reflected on the data part.

Fig. 5(b) demonstrates the further significant gain in the performance of DEEPFLAVOUR with respect to DEEPCSV. For example, for the same true positive rate of 78%, DEEPFLAVOUR offers an almost 40% reduction in false positive rate for light jets. We also see that NOCONV, which is an algorithm with the same structure and input as DEEPFLAVOUR, but trained without the convolutional layers (only

---

[5] Impact parameter is the distance of closest approach of the back-extrapolated particle trajectory to the primary vertex.
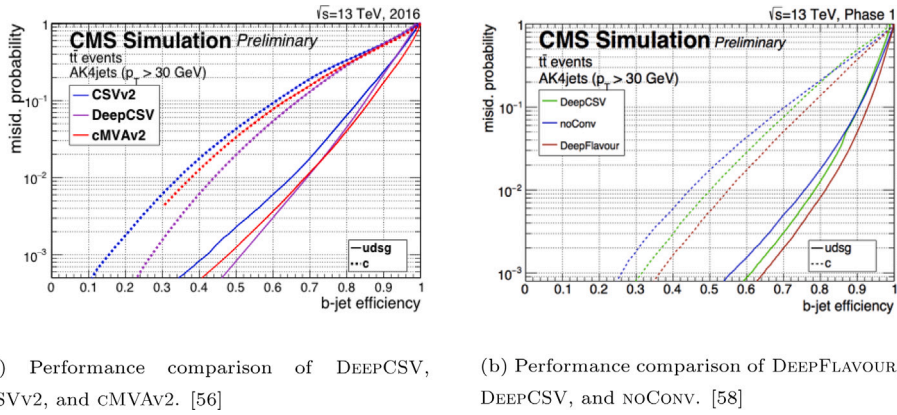
(a) Performance comparison of DEEPCSV, CSVv2, and cMVAv2. [56]

(b) Performance comparison of DEEPFLAVOUR, DEEPCSV, and noConv. [58]

Fig. 5. b-jet identification: True positive rate (b-jet efficiency) versus false positive rate (misidentification probability) for c-jets, and uds- and gluon-jets of simulated events of top quark pair production, requiring a minimal transverse momentum of 30 GeV to the considered jets.
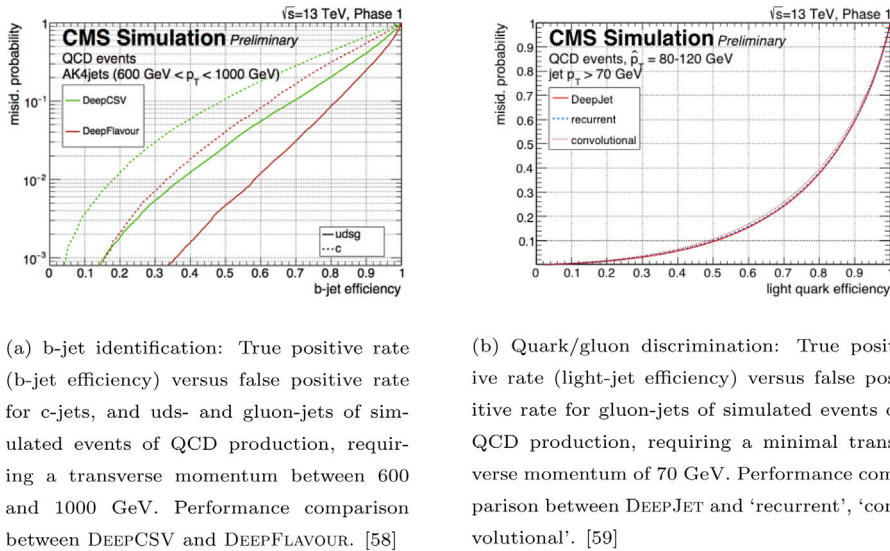
(a) b-jet identification: True positive rate (b-jet efficiency) versus false positive rate for c-jets, and uds- and gluon-jets of simulated events of QCD production, requiring a transverse momentum between 600 and 1000 GeV. Performance comparison between DEEPCSV and DEEPFLAVOUR. [58]

(b) Quark/gluon discrimination: True positive rate (light-jet efficiency) versus false positive rate for gluon-jets of simulated events of QCD production, requiring a minimal transverse momentum of 70 GeV. Performance comparison between DEEPJET and 'recurrent', 'convolutional'. [59]

Fig. 6. Comparison of the performance of different algorithms for b-jet identification and quark/gluon discrimination.

for comparison), provides an even worse result than DEEPCSV. This indicates that a larger input set of features alone is not able to increase the performance of the NN; on the contrary, it can even degrade the overall discrimination. The choice of a sophisticated architecture, *i.e.* the addition of convolutional layers in this case, which help exploiting the structure of the input (jet), is what provides sufficient information for the NN to perform as expected, when combined with a larger number of input variables.

Fig. 6(a) shows the gain in performance of DEEPFLAVOUR over DEEPCSV at very high values of transverse momentum of the b-jets, which implies a major gain in the sensitivity of physics analyses targeting highly energetic b-jets in the final state. For example, for a true positive rate of 37%, there is an almost 90% reduction in false positive rate when using the DEEPFLAVOUR tagger.

Finally, as part of the investigation of DEEPJET's capability to perform quark/gluon discrimination, Fig. 6(b) shows the comparison between DEEPJET and each one of two reference approaches, namely the 'convolutional' and the 'recurrent' one. The 'convolutional' approach, which involves the use of 2D convolutional layers working on 'jet images', as in [64], stems from the idea of considering the calorimeter cells as image pixels so as to be able to apply techniques already implemented within research in computer vision. More specifically, the jet is treated as an image in the $\eta - \phi$ plane of the detector; the continuous particle positions are pixelised, and for each pixel the intensity is provided by the corresponding energy deposits in the calorimeter; additionally, the RGB colour is determined by the relative transverse momenta of the charged and neutral jet constituents and by the charged particle multiplicity. The 'recurrent' approach, inspired by [65], is a slimmed-down version of DEEPJET, given that for light quark/gluon discrimination only a fraction of the initial input is relevant. Therefore only four features are used per particle (relative transverse momentum, η, ϕ, and the so-called 'pile-up per particle identification' weight [66]); secondary vertex information is removed, and there are no 1×1 convolutional layers. After training all three NN with the same samples, we observe that DEEPJET and recurrent NN perform similarly well and marginally better than the convolutional NN. The convolutional NN is already expected to be performant in

this case, as this kind of discrimination mostly relies on particle and energy densities, which may be well represented through an image related approach, in contrast to the significantly more complex case of heavy flavour tagging. DEEPJET's capability to achieve quark/gluon discrimination is further established by the considerable gain in performance it offers when compared to the "quark/gluon likelihood" discriminator [67], a binary quark/gluon classifier that is included in the CMS reconstruction framework, as described in [49].

4.6. Summary

The above-described taggers DEEPCSV and DEEPFLAVOUR/DEEPJET significantly outperform the standard b-jet tagger previously used in CMS, thus offering a major gain in the sensitivity for physics analyses involving b-quark jets, including new physics searches and precision measurements. At the same time, their ability to implement multi-class classification extends their use to generic heavy-flavour (b- or c-quark jet) tagging, and also to quark/gluon discrimination, in which regard DEEPJET exhibits satisfying performance when compared to approaches sharing the same goal. Broadening the feature selection by increasing the number of input variables that describe the jet constituents, applying a new machine learning algorithm with a deeper NN that also exploits the jet structure as it being an image, and using a larger and more diverse training sample that prevents building a process-specific tagger, constitute the main factors responsible for the observed advantage in performance. This result was validated on real collision data, indicating an equivalent gain in performance to the one estimated in the simulated samples. These taggers are the currently recommended ones for multiple studies carried out within the CMS Collaboration, and have already been used to produce a number of competitive physics results. CMS analyses that made use of the DEEPCSV and DEEPJET taggers include not only ones involving new physics searches, but also ones performing precision measurements, with [48,68–73] and [74–80], respectively, serving as a few such examples.

5. Improvements and Applications of the Matrix Element Method

Machine learning techniques employed at the LHC typically rely on the presence of large sets of training data for optimization purposes. The Matrix Element Method (MEM) takes a different approach, and provides a way to approximate the likelihood function for parameters in the SM given observed data. This calculation is performed from first principles, without the need for training. It was first used by the D0 Collaboration for a top quark measurement [81], with the original proposal provided by Kunitaka Kondo [82]. The method can also be used to discriminate between different collision processes by providing powerful observables in searches for rare signals. Ratios of likelihoods describing the probabilities that observed events be consistent with signal or background processes are used in this context. The MoMEMTA software package [83–86] was developed with contributions from the AMVA4NewPhysics Network to provide a convenient framework for calculating MEM likelihoods for LHC applications.

5.1. The Matrix Element Method

Let q_1, q_2 be the momentum fractions of the initial state partons, and y the kinematics of the partonic final state. The differential cross-section $d\sigma_\alpha(y)$, a function of the parton configuration y, is obtained by integrating the differential cross-section $d\sigma_\alpha(q_1, q_2, y)$ for the process α over the possible initial state parton configurations, weighted by the parton distribution functions (PDFs) of the colliding partons. The so-called transfer function $T(x|y)$ is the probability density for reconstructed event kinematics x, given a parton configuration y. It provides an approximate expression to capture effects from parton shower, hadronization, and detector reconstruction. Reconstruction efficiency effects can also be modelled with an appropriate efficiency $\epsilon(y)$. The probability density for observing an event x, given a hypothesis with parameters α, is given by

$$P(x|\alpha) = \frac{1}{\sigma_\alpha^{\text{vis}}} \int d\Phi(y) \ dq_1 \ dq_2 \sum_{a_1,a_2} \int_y T(x|y) \ \epsilon(y) \ f_{a_1}(q_1) \ f_{a_2}(q_2) \ |M_\alpha(q_1, q_2, y)|^2 \,. \tag{3}$$

In the above expression $f_{a_i}(q_i)$ are the PDFs for a given flavour a_i and momentum fraction q_i, $i = 1, 2$, $|M_\alpha(q_1, q_2, y)|^2$ is the squared matrix element for the process α, $d\Phi(y)$ the n-body phase space of y, and $\sigma_\alpha^{\text{vis}}$ is a normalization factor. The integral result alone, namely the above quantity without the normalization factor, is referred to as the matrix element weight, $W(x|\alpha)$. It is also commonly used.

Eq. (3), from which the most probable value of theory parameters can be estimated through likelihood maximization, involves an integration typically performed with Monte Carlo methods. This requires the evaluation of the matrix element $|M_\alpha(q_1, q_2, y)|^2$, which contains the (theoretical) information on the hard scattering, and can be computationally expensive to evaluate. The integrand can vary by many orders of magnitude in different regions of the phase space, necessitating appropriate choices for the parameterization of the integral to ensure computational efficiency. The design of the MoMEMTA framework makes it easy to find suitable parameterizations, which can help overcome the computational hurdle traditionally associated with the MEM.

5.2. MoMEMta Framework

5.2.1. Main implementation aspects

The MADWEIGHT package [87] introduced a general way to approach the problem of finding an efficient phase space parameterization for integration purposes. This includes the removal of sharp peaks in the integrand, for example due to resonances. MADWEIGHT is no longer supported, and its lack of flexibility hinders its application. MOMEMTA [83] has been designed to build upon the ideas of MADWEIGHT. It is a modular C++ software package, introduced to compute the convolution integrals at the core of the method. Its particular modular structure provides the required flexibility, allowing it not only to extend its applicability beyond its use in smaller programs so as to cover the needs of the complex LHC analysis workflows that handle large amounts of data, but also to be open to specific optimizations in the integration structure or engine. At the same time, since the MEM may be used in both theoretical and experimental high-energy physics problems, with different purposes and use cases in each field, MOMEMTA's modular structure constitutes a significant advancement. In terms of accuracy and CPU time, the performance of MOMEMTA is similar to that of MADWEIGHT, because they rely on the same algorithmic approach of phase-space parameterization. MOMEMTA is however further designed to adapt to any process and allows for implementation in any C++ or Python environment, offering more freedom to the user, who can therefore wrap new modules that handle specific tasks, while benefiting from the existing features of the framework.

5.2.2. Modules and blocks

The functionality of modules provided in MOMEMTA include representing and evaluating the matrix element and parton density functions, as well as the transfer functions, performing changes of variables, and handling the combinatorics of the final state. This implies that when calculating the probability to be assigned to the experimental events, every term in Eq. (3) may be treated as a separate, user-configured module within this framework. The weights for a given process $W(x|\alpha)$ are computed by calling and linking the proper set of modules in a configuration file. This computation usually requires the evaluation of multi-dimensional integrals via adaptive Monte Carlo techniques, whose efficiency depends on the phase-space mapping that is used. Such parameterization can be optimized by using a finite number of analytic transformations over subsets of the integration variables, called 'blocks'. Some of these blocks are also responsible for removing degrees of freedom by imposing momentum conservation, while the rest merely constitute the corresponding change of variables. Due to the potentially large combinatorial ambiguity in the assignment between reconstructed objects and partons, there exists a dedicated module that provides for the averaging over all possible permutations of a given set of particles. The functionality of this module, as opposed to a simple averaging over the possible assignments, allows the adaptive integration algorithms to focus on the assignments contributing most to the final result, thus increasing the precision of the result. This novel feature can potentially significantly speed up the computation, as the evaluation of the matrix element is what actually dominates the computation time.

5.3. MEM use cases and MoMEMta application

The MEM has proven to be an excellent technique to address two of today's main tasks in HEP analysis: signal-background discrimination and parameter estimate. In the former, the weights $W(x|\alpha)$ computed under different hypotheses are used to build a discriminating variable; in the latter, the MEM weights are instead used to build a likelihood function, which is then maximized in order to estimate the parameters of interest. Given its ability to efficiently compute the integral in Eq. (3), MOMEMTA meets the needs for both these MEM use-case categories. Examples of signal-background discrimination using the MOMEMTA framework can be found in [83], ranging from cases with low level of complexity (where the final state is precisely reconstructed with detectable particles) to cases with a high-multiplicity final state containing unobserved objects, where a careful consideration of the several degrees of freedom involved is required. In this section, a proof of principle for performing parameter estimation using MOMEMTA and an example of signal extraction with the MEM in a LHC analysis are reported.

5.3.1. Statistical inference in SMEFT using MoMEMta

In the Standard Model Effective Field Theory (SMEFT) [88,89], the effects of new heavy particles with typical mass scale $M \approx \Lambda$ on SM fields can be parameterized at a lower energy $E \ll \Lambda$ in a model-independent way in terms of a basis of higher-dimensional operators. In this work, we consider the operator $\mathcal{O}_{Qq}^{11}$, which modifies the coupling between top quarks and light quark–antiquark pair in top quark pair ($t\bar{t}$) production, as displayed in Fig. 7 (left). The MOMEMTA framework is used to estimate the quantity c_{Qq}^{11}/Λ^2 in a $t\bar{t}$ simulation sample, with c_{Qq}^{11} (referred to as c in the following for shortness) being the degree of freedom associated to the $\mathcal{O}_{Qq}^{11}$ operator. A fully-leptonic $t\bar{t}$ simulation sample is produced with MADGRAPH_AMC@NLO version 2.6.5 [90] in the di-muon final state at LO in QCD precision and with corrections up to $1/\Lambda^2$ at the amplitude level, with the quantity Λ set to 1 TeV. The events are then showered with PYTHIA 8.212 [91] and the simulation of particle interactions with the CMS detector is performed with DELPHES 3.4.1 [92]. The contribution of the SMEFT term to the matrix element M of the process can be broken down into three parts: a SM contribution (A_{SM}), a quadratic contribution for the dimension-6 operator only (A_{quad}), and an interference term between the two (A_{int}). This translates into:

$$|M_{SMEFT}|^2 = |M_{SM} + c M_{NP}|^2 = A_{SM} + c A_{int} + c^2 A_{quad} . \tag{4}$$

With this parameterization, the integral in Eq. (3) can be written, for a given c, as the sum of three separate integrals:

$$P(x|c) = \frac{1}{\sigma_c}(W_{SM} + c W_{int} + c^2 W_{quad}), \tag{5}$$
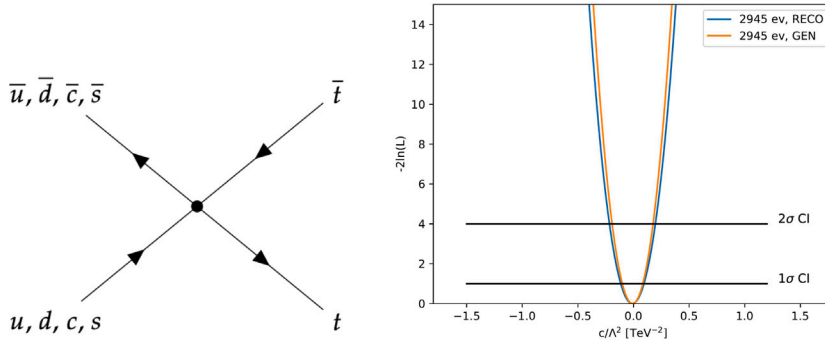
Fig. 7. Left: Example of $t\bar{t}$ process production at LO, where the vertex between the quark–antiquark pair and the top pair is described by a dimension-6 effective operator. Right: Negative log-likelihood as a function of c/Λ^2 built on generator-level (orange line) and detector-level (blue line) events. The black lines define the 1σ and 2σ confidence intervals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with σ_c being the visible cross section of the process, in turn similarly parameterized as a function of c. The three ME weights are computed with MoMEMTA using Gaussian transfer functions on the energies of the visible particles with a standard deviation of 5% for leptons and 10% for jets. An additional dimension of integration is introduced in order to handle the combinatorial ambiguity in the assignment between reconstructed final-state b jets and b quarks in the matrix element.

A negative log-likelihood function is then built from $P(x|c)$ while performing a scan over c. The resulting function is shown in Fig. 7 (right) in blue. The same function excluding the detector effects is also computed and represented in orange. The estimated quantity $\frac{\hat{c}}{\Lambda^2}$ on events including detector effects is measured to be -0.013 TeV^{-2} with a 2σ confidence interval [-0.233, 0.210] TeV^{-2}. It is to be noted that in a complete study one would have to take into account also systematic uncertainties that have been neglected here, where only the statistical effect plays a role.

Finally, two main considerations can be drawn. The two curves in Fig. 7 show very similar width of the likelihood profiles, proving the strength of the MEM where detector effects are encoded in the computation of the ME weight. Moreover, this method represents a valid alternative to cases where the SMEFT coefficients are estimated one at a time, since in the MEM the maximization of the likelihood can easily be multi-dimensional. More detail on this study is available in [93].

5.3.2. $t\bar{t}H$ production

The search for Higgs production in association with top quark pairs, $t\bar{t}H$ with $H \to b\bar{b}$, is a particularly interesting application for the MEM. It was first studied in [94], and since then has been used extensively by the ATLAS and CMS experiments [95–99]. The $t\bar{t}H(b\bar{b})$ process has many partons in the final state. Its main background in final states with at least one charged lepton is $t\bar{t} + b\bar{b}$, which features identical final state partons to the signal. Discrimination between these processes relies on small differences in kinematics.

ATLAS combines the MEM with additional multivariate techniques in a search for $t\bar{t}H(b\bar{b})$ with 36 fb^{-1} of data [96]. Details of the MEM implementation developed for this analysis and studies of its performance can be found in [100]. The following provides a brief summary of the approach chosen for this analysis. A discriminant, MEM$_{D1}$, is calculated as the logarithm of the ratio of signal and background likelihoods, where the signal is $t\bar{t}H(b\bar{b})$ and the background is $t\bar{t} + b\bar{b}$: MEM$_{D1} = \log_{10}(L_S/L_B)$. The $t\bar{t} + b\bar{b}$ contribution to the background is dominant in the phase space where the discriminant is used. Matrix elements are calculated with MADGRAPH5_AMC@NLO [90] at leading order; the lack of higher order corrections can at most decrease the performance of the method, but it does not bias the physics. Only gluon-induced Feynman diagrams are considered, which reduces computational time without a significant impact on discrimination power.

The MEM is used for final states with one charged lepton, where six final state quarks are expected at leading order, alongside the one charged lepton and a neutrino. Directions of all visible partons are assumed to be measured exactly by the ATLAS detector, so the associated transfer function components are δ-distributions. After imposing transverse momentum conservation, seven degrees of freedom remain for the integration. The integration variables are chosen to be the energies of all six final state quarks and the neutrino momentum along the beam direction. The integration itself is performed with VEGAS [101], based on an framework described in [102].

Fig. 8 visualizes the MEM$_{D1}$ discriminant, using a sigmoid to map the values into the $[0, 1]$ interval. The data are found in good agreement with the expected distribution, and the discrimination power of the method is seen by comparing the normalized distribution of $t\bar{t}H$ (dashed line) to the background contributions.

5.4. Summary

The MEM provides a method for evaluating the likelihood of collision events under different hypotheses from first principles. It has applications in parameter measurements and searches for specific collision processes. The MoMEMTA software package provides a flexible implementation of the calculations required. It offers flexibility with its modular structure, and enables use of the MEM for a wide range of applications at the LHC. The modularity also allows for extensions to handle novel applications, while taking advantage of the optimizations and convenience features provided by MoMEMTA.
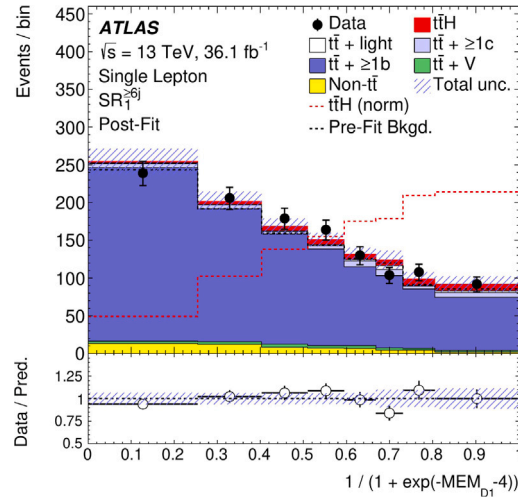
Fig. 8. MEM-based discriminant used by ATLAS in the search for $t\bar{t}H(b\bar{b})$. The red dashed histogram shows the normalized density expected from ttH events alone; black points represent ATLAS data, and full histograms the various signal (red) and background contributions [96]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6. New Statistical Learning Tools for Anomaly Detection

6.1. Overview

Searches for new physics at the LHC proceed by comparing the data collected by the detectors with simulated data sets obtained from software simulations of the underlying physics process, interfaced with a simulation of the detector response. The simulated data describe either the processes predicted by the standard model (*background processes*) or the processes postulated by the particular theoretical extension of the SM under study (*signal processes*). These searches may be broadly categorized into *model-dependent* searches, where the data are compared to both the SM and new physics predictions, and *model-independent* searches, where the data are compared only to the SM in order to search for unexplained deviations from it: these deviations (*anomalies*) may then be attributed to new physics and investigated further. In this context, the term *model-independent* refers to independence with respect to a new physics signal model: the SM background model is always assumed.

Model-dependent searches are typically approached by producing simulated events for both SM and new physics processes, and hypothesis tests are devised to find which model is favoured by the data.

In case of model-independent searches, no particular model for the signal processes is assumed: a simulated dataset describing the signal processes cannot therefore be produced. These searches can be approached as anomaly detection problems, where the data are combed to find any observation that is not consistent with the background model. This setting is an example of semi-supervised learning, given that the two data categories involved are a 'simulated' dataset, generated by Monte Carlo techniques to represent the known background process and therefore considered as 'labelled', and the 'experimental' data sample, which is generated by an *a-priori* unknown mechanism—possibly comprising contributions from both background and signal processes—and therefore considered as 'unlabelled'. The presence of a signal in the experimental data is generally inferred through the observation of a significant deviation from the predictions for the background process.

Let $Y = (\mathbf{y}_1, \dots, \mathbf{y}_m)'$, $\mathbf{y}_l \in \mathrm{R}^P$, $l = 1, \dots, m$, be the experimental data, with $\mathbf{y}_l$ independent and identically distributed realizations of the random vector with unknown probability density function $f_{BS} : \mathrm{R}^P \to \mathrm{R}$. In addition to the experimental data Y, it is possible to generate with the use of Monte Carlo simulations a large sample $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$, $\mathbf{x}_i \in \mathrm{R}^P$, $i = 1, \dots, n$. While Monte Carlo simulations in HEP are computationally taxing, a simulated sample of SM events is employed in so many different analyses that the experiments can afford to simulate tens or hundreds of millions of events for SM scenarios: the sample X can therefore be considered arbitrarily large to all practical extents. We assume that the simulated data, as well as the majority of the experimental data, which are known to have been generated by a background process, follow a distribution described by the probability density function $f_B : \mathrm{R}^P \to \mathrm{R}$. The remaining experimental data have been possibly generated by an unknown signal process described by $f_S : \mathrm{R}^P \to \mathrm{R}$. In absence of quantum mechanical interference between the signal and the background, the generating mechanism of the experimental data f_{BS} may be thus specified as a mixture model:

$$f_{BS}(\mathbf{y}) = (1 - \lambda)f_B(\mathbf{y}) + \lambda f_S(\mathbf{y}), \ \lambda \in [0, 1), \tag{6}$$

and the problem may be cast in terms of either parameter estimation, where inference is sought on the value of λ, or of hypothesis testing, where a simple null hypothesis $\lambda = 0$ is tested against a composite alternative $\lambda \neq 0$.
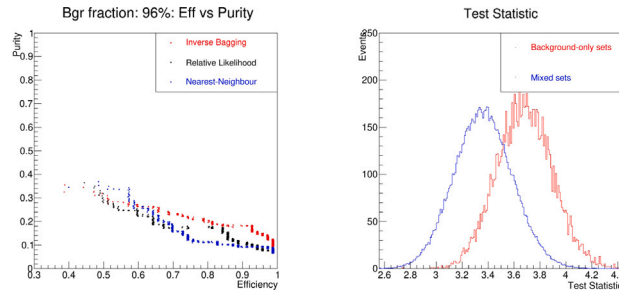
Fig. 9. *Left:* Purity of the discriminators as a function of their efficiency. The simulated data sample is composed by a 96% background and 4% signal. *Right:* The values of the test statistic (score) for the background-only sets and for the mixed signal+background sets [103]. The performance is calculated on the HEPMASS dataset [107] as described in the text.

6.2. Detecting anomalies via hypothesis testing: The Inverse Bagging algorithm

The Inverse Bagging (IB) algorithm [103–105], developed within the fourth pillar of the AMVA4NewPhysics research program, addresses the problem of anomaly detection by means of hypothesis testing. We formulate the null hypothesis that the processes having generated the experimental data follow the same distribution as the ones corresponding to the simulated data. We then perform a statistical test to quantify how likely it is that the unlabelled data have been indeed generated by the background processes alone. This algorithm combines hypothesis testing with multiple data sampling, as a means of iterative tests of the properties of the data and the possible presence of unknown signals.

6.2.1. Hypothesis testing and multiple sampling

Once the null hypothesis is defined as $H_0 : f_{BS}(\cdot) = f_B(\cdot) \Leftarrow \lambda = 0$ and the alternative as $H_1 : f_{BS}(\cdot) \neq f_B(\cdot) \Leftarrow \lambda \neq 0$, the IB algorithm proceeds by performing a two-sample test on each of B pairs of bootstrap replicas X_b^* and Y_b^*, $b = \{1, 2, \dots, B\}$, taken from the data samples X and Y, respectively. The size Q of the bootstrap replicas is set to be significantly smaller than the size of the experimental sample under study. The results of the tests are used to classify how anomalous are the individual observations, improving on the insights that may be offered by a standard hypothesis test performed on the original data sets X and Y.

The test statistic associated to the pair of X_b^* and Y_b^*, $\mathbf{T}_l = \left(T_{l1}, \dots, T_{lB_l}\right)'$, $l = 1, \dots, m$, is considered as part of the useful information associated with each of the observations that are contained in Y_b^*. The set of test statistics that concern the same observation, regardless of the specific bootstrap samples they have been computed in, is summarized into an observation score (anomaly detection metric) summarizing how anomalous are the properties of the particular observation. The details of the computation of the anomaly detection metric from the set of test statistic values for each event are given in Section 6.2.2. The underlying rationale is that if an observation $\mathbf{y}_l$ has been generated by f_S, then the bootstrap samples that include $\mathbf{y}_l$ will lead to reject H_0 more often compared to the samples not including it, if the size Q of the sample is small enough. Using small sizes Q would result in samples where an anomaly results more easily in a sizeable deviation from the background distribution. The anomaly detection metric therefore reflects how likely it is for each observation to have been generated by a signal process, and can be hence used for further classification purposes, with observations carrying the most extreme score becoming candidates to be classified as a signal. By defining a sliding-window threshold on the value of the anomaly detection metric for classifying an observation as signal, we obtain a Receiver Operating Characteristic (ROC) curve [106], which describes the purity of the classifier as a function of its efficiency. In Refs. [103–105], the IB algorithm was applied to the HEPMASS dataset [107], available at http://archive.ics.uci.edu/ml/datasets/HEPMASS: the dataset consists in a background composed by standard model top quark production and a signal composed by a heavy resonance that decays into a top quark pair that would result in a spectrum different from the SM for any observable involving the full final state (e.g. the reconstructed visible mass of the top pair decay products). Ref. [103–105] contain the details on the input features that were used. Fig. 9 (left) shows a purity versus efficiency curve for the IB algorithm, as well as for two reference classifiers (relative likelihood, and k-nearest neighbours) that use only event-based information to classify events. In the considered application, the IB classifier outperforms both. The test statistic used by the IB algorithm has a different distribution for a background-only set of events and for a mixed signal+background set of events ($\lambda = 0.04$), as illustrated in Fig. 9 (right).

6.2.2. Validation of the algorithm and research questions

We took into account the following research questions in order to further validate the algorithm with respect to its original publication:

Choice of anomaly detection metric A key step in the IB algorithm is the score computation, which is our anomaly detection metric. A meaningful metric is crucial for inferring the likelihood for each observation to belong to a hypothetical signal. Different metrics may naturally rank each observation in a different way, leading to a different classification as signal- or background-like. Here we consider three methods for anomaly detection metric computation: (a) 'Test statistic score', given

by the mean of the test statistics $\mathbf{T}_l = \left(T_{l1}, \dots, T_{lB_l} \right)'$ based on the bootstrap samples including $\mathbf{y}_l$: $R_{Tl} = \frac{1}{B_l} \sum_{k=1}^{B_l} T_{lk}$, (b) 'P-value score', given by the mean of the p-values $\mathbf{P}_l = \left(p_{l1}, \dots, p_{lB_l} \right)'$, connected with the test statistics based on the bootstrap samples including $\mathbf{y}_l$: $R_{Pl} = \frac{1}{B_l} \sum_{k=1}^{B_l} p_{lk}$, (c) 'Ok score', given by the proportion of tests based on the bootstrap samples including $\mathbf{y}_l$ and rejected at a given significance level α: $R_{Ok;l} = \frac{1}{B_l} \sum_{k=1}^{B_l} 1_{\{p_{lk} < a\}}$.

Parameters Q and B When deploying the IB algorithm, the choice of Q and B is also important. The original formulation of the algorithm requires $Q < m$, as this may imply that a number of bootstrap samples Y^* will include larger proportion of signal observations than λ, thus increasing the power of the test and making the detection of an hypothetical signal simpler. The choice of B is not independent of Q, because the expected number of times each observation $\mathbf{y}_l$ is sampled is $E(B_l) = \frac{BQ}{m}$. For a fair comparison, the observation scores are computed based on the same number of tests, *i.e.* Q and B may vary among the different study cases but should be adjusted accordingly for $E(B_l)$ to be kept fixed. Moreover, the larger the B values, the more stable the results are expected to be.

Performance comparison A third significant aspect to take into account is the comparative evaluation of the IB algorithm. For this purpose, we consider an adjustment of the Linear Discriminant Analysis (LDA) [108], referred to in the following as 'LDA score', suitable for the semi-supervised nature of the problem under study [109]. The performance of LDA is further compared with that of IB in various scenarios in the context of an improvement of the method [110].

6.2.3. Simulation settings and results

The research questions described in Section 6.2.2 are explored using either a univariate or multivariate normal distribution for both signal and background, or a multivariate normal distribution for the background and a uniform distribution across a sphere or a hemisphere for the signal. Early results [109] suggest that, in the case of univariate and multivariate normal data, a better performance may be achieved when using the test statistic as a score in conjunction with subsampling the data (*i.e.* $Q < m$): in particular, the regime $Q \ll m$ seems to be associated with a lower variability of the score, in agreement with the intuition and the studies of Ref. [103]. For large numbers of bootstrap iterations, B, the preliminary results suggest that the classification performance is comparable among different values of Q, possibly because by increasing B the variance of the scores converges to some value that does not depend on Q. A small value of Q, however, implies lower variability of the scores. A good performance may therefore be obtained without having to resort to a large number of bootstrap samples; early studies seem to confirm this intuition. The LDA seems favoured against the IB when both signal and background are normally distributed, in line with the assumptions LDA relies upon. However, the preliminary tests suggest that the performance of IB may still be comparable to that of LDA when using small values of Q and selecting the test statistic score: the performance may be even less affected when the normality assumption for each class is removed. Additional tests that use a uniform signal distribution on a sphere and on a hemisphere lead us to conjecture that the IB may have the ability of recognizing both local and global properties of the signal process.

An extensive test [110] is performed comparing the IB algorithm to several scenarios, as illustrated in Fig. 10. The case where the signal density is known is represented by two scenarios: a classical likelihood ratio test based on the Neyman–Pearson lemma, and a kernel density estimation performed on the original datasets X and Y. A semi-supervised approach—the most realistic competitor to IB—is represented by a semi-supervised LDA. The IB algorithm is represented by several settings, corresponding to different choices of test statistic (Kolmogorov–Smirnov test [111], Mann–Whitney test [112]) and of score aggregation (minimum, maximum, mode, or median of the set of test statistics). The best setting for the Gaussian mixture in exam ($\lambda = 0.06$) appears to be the expected value (mean) of the test statistic. For reference, the performance of a random choice is also shown.

6.2.4. Possible improvements

Possible improvements to the original IB algorithm include exploring not only additional anomaly detection metric computation methods, using the simple multivariate Gaussian scenarios described *supra*, but also a dimensionality reduction approach, which could address the issue of the high-dimensional data with a large number of redundant features that is common in high-energy physics applications. For this approach, standard and more advanced methods of variable sub-sampling ('max-G-sampling') are proposed [104] (similarly to [113]), as well as methods for adjusting the sampling weights in such a way as to obtain samples more enriched in signal, and to perform tests on the likely more informative variables. 'G' in max-G-sampling stands for the number of test results from which the maximum is taken, and a moderately high G parameter can result to only the highest test statistic values—likely obtained via a feature set rich in informative variables—being saved. Moreover, given that for the study described *supra* uncorrelated generated data were considered, and since it is essential to also validate the IB algorithm performance in presence of correlations affecting the outcome of the dimensionality reduction technique, a particular transformation of the data intended for its decorrelation is suggested [104].

6.2.5. Applications

A preliminary validation of the IB performance on a non-physics-related data set was performed on the well-known spam data [114], which are transformed to fit the semi-supervised context of the above described study for anomaly detection; early tests [109] suggest that the proposed algorithm improvements may have a comparable performance with that of the standard algorithm setting. When applied to a common high-energy physics problem so as to be validated on Monte Carlo simulated collision data (which constitutes the main objective of its development) the IB algorithm was found to perform at a similar level as the LDA score. Also, the performance including the proposed variable max-10-sampling method was comparable with that of the standard approach, and for some scenarios presented a slight improvement.
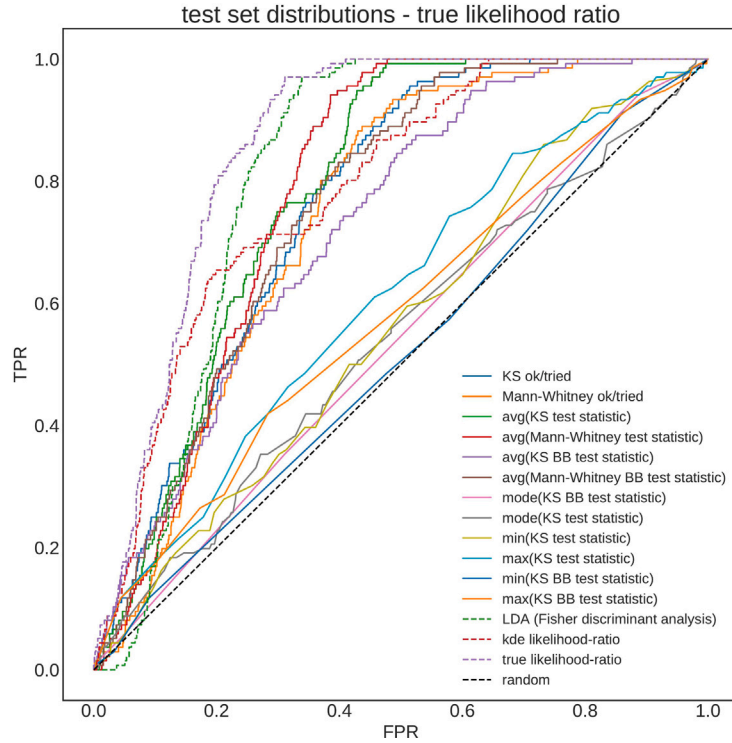
Fig. 10. The ROC curve (true positive rate versus false positive rate) for a comparison of different settings of the IB algorithm with several scenarios: the classical likelihood ratio (labelled *true likelihood ratio*) and its KDE approximation (labelled *KDE likelihood ratio*) represent idealized situations where the signal model is known; a realistic competitor is represented by a semi-supervised LDA approach. The IB algorithm performance is shown for several choices of IB test statistic and of summary statistic for the score, detailed in the text [110].

6.2.6. Summary

Taking everything into consideration, the IB algorithm exhibits a satisfying performance, achieving its best performance when the distribution of the signal and background deviates from Gaussianity. The performance of the IB algorithm is particularly good when using small-sized bootstrap samples, both in terms of mean and variance. Different score computation methods have also been tested, with the test statistic appearing as the most effective one for this purpose. At the same time, a reduction of the data dimensionality seems to have the potential to improve the performance of the algorithm; however, further studies are needed on this aspect. Depending on all these factors and on a chosen amount of false-positive-rate, we have observed that the IB algorithm can provide improvements in the true positive rate ranging from 1 to 10% in some of the considered problems. The improvement is nevertheless problem-dependent, suggesting that the algorithm may be suitable only for a certain class of problems; studies are ongoing to more precisely identify the characteristics of the problems that the IB algorithm can tackle with profit.

6.3. Detecting anomalies via clustering semi-supervision

With the Inverse Bagging algorithm described *supra*, the degree of compatibility between experimental and simulated data is evaluated by means of hypothesis testing. Another approach to anomaly detection, based on a markedly different rationale, consists in semi-supervising learning methods, either by relaxing the assumptions of supervised methods, or by strengthening unsupervised structures via the inclusion of additional available information from simulated data. The latter route has been followed in [109,113] where a parametric and, respectively, a non-parametric approach are adopted. Both are summarized *infra*.

6.3.1. The parametric approach

The idea of addressing anomaly detection by semi-supervising parametric clustering was first explored, in the context of new physics searches, in [115,116] under the name of *fixed background model* (FBM). The authors proposed to specify the distributions f_B and f_S themselves in Eq. (6) within the family of Gaussian mixture models. Parameter estimation was then conducted via maximum likelihood in two steps. First, the background model $\hat{f}_B$ was obtained based on the background data X. Afterwards, keeping the parameters of $\hat{f}_B$ fixed, the weight λ in Eq. (6) and the parameters characterizing the new possible component f_S were estimated using experimental data Y. Maximization of the likelihood function was conducted via a suitable adjustment of the Expectation–Maximization (EM) algorithm [117]. A goodness-of-fit test served to discard insignificant components and to assure that the whole estimated density was equal to the background component when no signal was detected.

Table 3

Summary of the anomaly detection results performed by the PAD and the FBM for MC data with different signal proportions λ. For each scenario, 50 datasets are generated to obtain a mean result with the respective standard deviations presented in brackets.

Method	λ	Average $\hat{\lambda}$	Average AUC
PAD	0.05	0.040(0.012)	0.725(0.109)
PAD	0.10	0.057(0.013)	0.818(0.078)
PAD	0.15	0.086(0.006)	0.876(0.017)
PAD	0.20	0.112(0.006)	0.882(0.012)
FBM	0.05	0.025(0.009)	0.708(0.118)
FBM	0.10	0.046(0.008)	0.764(0.078)
FBM	0.15	0.070(0.006)	0.771(0.073)
FBM	0.20	0.096(0.012)	0.780(0.054)

Due to the curse of dimensionality or numerical issues, in fact, the approach described *supra* was found to be sub-optimal or even not liable to be carried out on high-dimensional data. To reduce the data dimensionality while preserving relevant information, penalized methods for variable selection were introduced in the unsupervised framework of mixture modelling in [118,119], yet with a strong reliance on restrictive assumptions on the clusters shapes. Within the AMVA4NewPhysics program [109,120] we have extended the penalized approach to the FBM to allow for a more flexible modelling without constraining the mixture component covariance matrices, as well as to account for the semi-supervised nature of the anomaly detection problem for new physics searches. The proposed penalized anomaly detection (PAD) method builds on a variant of the EM algorithm, derived in the semi-supervised context to estimate the parameters of a mixture model via the maximization of the penalized likelihood function.

The PAD methodology was validated both in the unsupervised setting (*i.e.* to estimate the background distribution) and in the semi-supervised anomaly detection setting (to estimate the whole process density) via simulations designed to account for different aspects of the analysis: different implementations for handling variable selection; performance of competing methods; varying configurations of the background and possible signal. The study highlighted a general improvement with respect to the state of the art (see [109] for details).

In the specific context of LHC physics research, the PAD was applied on Monte Carlo data produced to study the experimental signature of processes yielding two energetic jets in the final state. The background data were generated including QCD processes yielding two energetic jets in the final state; these arise when the hard subprocess generates two gluons, two quarks, or a gluon and a quark with large momentum, which then hadronize into the observable pair of energetic jets. The signal data were produced from a heavy resonance (stop quark) postulated in an extension of the Standard Model known as RPV-MSSM [121] with a mass of 1000 GeV, that also leads to two jets in the final state. Results are presented in Table 3 for varying proportion λ of signal events. Since in the presence of imbalanced processes (such as the considered one of signal versus background classification based on mixture models) requires a threshold adjustment that can influence the evaluation, the performance of the method has been measured in terms of Area under the ROC curve (AUC). In terms of AUC, the PAD compares favourably with the FBM model.

6.3.2. The nonparametric approach

As an alternative to the parametric approach illustrated *supra*, we explored the non-parametric approach to unsupervised learning. From one side, this relies on a more flexible modelling of the processes under analysis, without constraints or specific assumptions about their shape. On the other side, by drawing a correspondence between groups and the modal peaks of the density underlying the observed data, the non-parametric formulation appears particularly consistent with the physical notion of signal, as it is commonly assumed that deviations from the background process manifest themselves as significant peaks in f_{bs}, not previously seen in f_b.

As discussed in detail in [113], two main contributions can be highlighted. Under the assumption that a signal does exist, the main idea is to tune a non-parametric estimate of the density f_{bs}, assumed to generate the unlabelled data. This tuning is obtained by selecting the smoothing amount so that the induced partition, where the signal is forced to emerge as a bump in the density distribution, classifies the labelled background data as accurately as possible. The relevance of the forced bump is afterwards tested via the suitable application of a statistical test: if it is deemed to be significant this would provide empirical evidence of a signal, and should then represent the stepping-stone for the possible claim of new physics discovery. As a second side contribution, a variable selection procedure, specifically conceived for the considered framework, is proposed. Here a variable is assumed to be relevant if its empirical distribution shows a changed behaviour in the simulated data with respect to the one in the experimental data, as this difference shall be only due to the presence of a signal, not previously seen in the background density. This idea is pursued by repeatedly comparing the estimated densities on distinct and randomly sampled subsets of variables. Eventually the variables that are more often responsible for a different behaviour of the two marginal distributions are selected.

The procedure is compared with the Fixed Background Model [115,116] and tested on signal data simulated from a new particle X of unknown mass that decays to a top quark pair $X \rightarrow t\bar{t}$ and a background from Standard Model top pair production, identical in its final state to the signal but distinct in the kinematic characteristics because of the lack of an intermediate resonance (see [107] for a detailed description of the data and their characteristics). A visual glance to the results and the classification accuracy, as measured by the *Fowlkes–Mallows* index (FMI) [122, Ch. 27] and the true positive rate (TPR), are provided in Fig. 11.

Method	FMI	TPR
NP (2 var)	0.84	0.80
FBM (6 PC)	0.77	0.50
FBM (2 var)	0.78	0.56
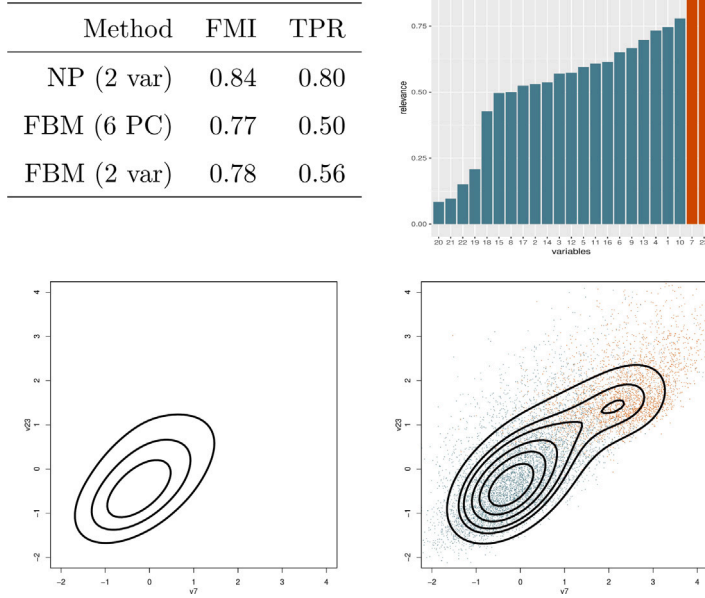
Fig. 11. Results of the application of the nonparametric (NP) method for anomaly detection [113] discussed in Section 6.3.2: in the top-left panel, accuracy of classification of signal and background as measured by FMI and TPR, and compared to the parametric competitor Fixed Background Model (FBM) applied to the first 6 principal components the of the data (PC) and to the same 2 variables selected by the nonparametric procedure. In the top right panel relative informativeness of the whole set of variables, as labelled in [107]. The resulting most relevant variables, in orange, are the combined mass of two bottom quarks with two W-bosons, and the transverse momentum of the leading jet. Bottom panels: contour plot of the density estimate of the two selected variables in the background (left) and overall data including the signal (right), which exhibits as a new peak arising from the background. A subsample of data from background (blue) and signal (orange) has been overimposed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.4. Gaussian processes for modelling invariant mass spectra

As already noted *supra*, model-independent searches for new physics at the LHC provide a ground for applying semi-supervised methods. One of the most relevant features used in the quest for discovering new physics in a given experimental signature is the invariant mass of the set (or a subset) of objects reconstructed as products of the collision events. The distribution of invariant mass values can be modelled from a Monte Carlo simulation (a labelled dataset) of the Standard Model processes that correspond to the background, and compared with the data obtained at the experiment, which could potentially contain (usually few) signal events among a large amount of background events. Gaussian processes (GP) are a useful tool for regressing the shape of invariant mass spectra and they can be used to disentangle potential signals from the background in a semi-supervised manner.

6.4.1. Generalities on Gaussian processes

Gaussian processes are a Bayesian inference method in a function space, defined as "a collection of random variables, a finite collection of which have a joint Gaussian distribution" [123]. Given a prior distribution over the parameters of a set of functions in the function space and using the likelihood of the observations, we can obtain a posterior distribution over such parameters through Bayes' rule.

Let f be the function that is regressed and x and x' arbitrary points in the input space $\mathcal{X}$. The prior on the regression can be noted as

$$f(x) \sim \mathcal{GP}(\mu(x), \Sigma(x, x')),\tag{7}$$

where $\mathcal{GP}$ is the infinite-dimensional function space associated with the joint Gaussian distribution, from which f is sampled. Thus, there are two functions that are defined to specify a GP: the *mean* and the *covariance* or *kernel*, respectively,

$$\mu(x) = \mathbb{E}[f(x)],\tag{8}$$

$$\Sigma(x, x') = \mathbb{E}[(f(x) - \mu(x))(f(x') - \mu(x'))].\tag{9}$$

The mean and covariance functions above, as their names indicate, dictate the mean value and the covariance of the GP distribution for points in the input space.

For a finite set of points, the prior and posterior are joint Gaussian distributions with as many dimensions as observations, and there is a formalism that allows to predict new output values for arbitrary inputs. In [124], GPs are used to model the Poisson process

corresponding to a binned distribution of events, *i.e.* the spectrum. Let the bin centres be denoted by the vector $x = (x_1, \ldots, x_N)$, and the corresponding observed responses $y = (y_1, \ldots, y_N)$, the function f is then averaged within the corresponding bin to produce a set of expected counts $\bar{f}(x) = (\bar{f}(x_1), \ldots, \bar{f}(x_N))$. The spectrum is approximated via the product of two multidimensional Gaussians by the probability model:

$$p(y(x)) = \mathcal{N}(y|\bar{f}(x), \sigma_2(x))\mathcal{N}(\bar{f}(x)|\mu, \Sigma), \tag{10}$$

where σ are the uncertainties in the values of $\bar{f}$, which constitutes a Gaussian approximation of the Poisson noise. The second factor is an N-dimensional Gaussian distribution. Note that we use the vector $\mu = (\mu(x_1), \ldots, \mu(x_N))$ and the matrix Σ with $\Sigma_{ij} = \Sigma(x_i, x_j)$, constructed from the mean and covariance functions respectively.

After some standard algebraic manipulation it is possible to obtain explicit expressions to infer the response of the function f at a new arbitrary input values $x_\star$, given x and y:

$$\mathbf{mean}(f_\star) = \mu_\star + \Sigma_\star [\Sigma + \sigma_2(x)\mathbb{1}]^{-1}(y - \mu), \tag{11}$$

$$\mathbf{cov}(f_\star) = \Sigma_{\star\star} - \Sigma_\star [\Sigma + \sigma_2(x)\mathbb{1}]^{-1}\Sigma_\star, \tag{12}$$

where $f_\star = f(x_\star)$, $\Sigma_\star = \Sigma(x, x_\star)$, and $\Sigma_{\star\star} = \Sigma(x_\star, x_\star)$; $\mu_\star$ is the mean prior corresponding to $x_\star$. Note that the dimension of the GP multivariate Gaussian is extended by the dimension of $x_\star$.

It is possible to initialize the prior mean μ with a function from domain knowledge but setting it equal to zero is common practice and does not necessarily pose a limitation when using GPs. The most important ingredient to be specified is then the kernel, for which there exist several common choices in the literature (*e.g.* the exponential squared or other radial kernels) [123,125]; a new kernel could also be crafted for a particular application.[6] The hyper-parameters in the kernel need to be adjusted as well, which is usually done by finding the maximum log marginal likelihood of the GP:

$$\log \mathcal{L} = -\frac{1}{2}\log|\Sigma| - (y - \mu)^T \Sigma^{-1}(y - \mu) - \frac{N}{2}\log 2\pi. \tag{13}$$

The standard algorithm to obtain the value $f_\star$, its variance and the likelihood involves solving triangular systems and matrix inversion, as it is implied in the expression for the kernel detailed in Eq. (12). Details of the algorithm and some optimizations are provided in [123].

6.4.2. Method overview

Gaussian processes have been used in several contexts in HEP research. Our work is based on the application in [124], which presents a study on background and signal distributions modelling in the dijet mass spectrum. The goal of the method is to accommodate the background and identify a signal component on the data, if present.

We start with the following prescriptions for the mean and kernel functions:

$$\mu(x) = 0, \tag{14}$$

$$\Sigma_B(x, x') = A \, \exp\left(\frac{d - (x + x')}{2a}\right)\sqrt{\frac{2l(x)l(x')}{l(x)^2 + l(x')^2}}\exp\left(\frac{-(x - x')^2}{l(x)^2 + l(x')^2}\right), \tag{15}$$

$$\Sigma_S(x, x') = C \, \exp\left(-\frac{1}{2}(x - x')^2/k^2\right)\exp\left(-\frac{1}{2}\left((x - m)^2 + (x' - m)^2\right)/t^2\right), \tag{16}$$

where $l(x) = bx + c$ is a linear function. For a description of all kernel hyper-parameters in Eqs. (15) and (16), and a motivation for their functional forms, see the cited study. We refer to the kernel hyper-parameters collectively as $\theta_B = \{A, a, b, c, d\}$ and $\theta_S = \{C, m, t, k\}$, respectively.

We present procedures that can operate using two or three steps as follows:

First step. A GP fit is performed on a pure background distribution (*i.e.* from simulation) using the background kernel $\Sigma_1 = \Sigma_B$. We obtain a model for the background distribution and find the corresponding hyper-parameters of the first kernel, θ_1.

Second step. The optimized hyper-parameters of the background fit are kept fixed and used in performing another fit using the kernel $\Sigma_{12} = \Sigma_1 + \Sigma_2$, where the second component in the sum is of the signal kind $\Sigma_2 = \Sigma_S$ (Eq. (16)); thus a new GP model is obtained, including the parameters corresponding to the new component, θ_2.

Third step. A further optimization is performed where a new signal kernel Σ_S is added, $\Sigma_{123} = \Sigma_{12} + \Sigma_3$, and the parameters of Σ_{12} (*i.e.* $\theta_{12} = \theta_1 \cup \theta_2$) are kept fixed. From this, the parameters of the last component (θ_3) can be extracted.

In the case of the two-step procedure, the background kernel is used in a first step; then, in a second step, a signal kernel identifies a concentrated excess or deficit centred at m with width t. For the three-step procedure, the first two steps are used to accommodate a background with a turn-on; then a third step performs the signal detection. This procedure can be considered as semi-supervised: we first model the background-only invariant mass distribution, *i.e.* the labelled dataset, and then find a model for the unlabelled data distribution where signal could appear.

---

[6] A set of example kernels, how to compose them, and an explanation on how they can express the structure can be found in [125] (chapter 2).

Table 4

Number of injected signal events within a window for values of R, for a signal of 3 TeV mass and 150 GeV width. The values and errors obtained are the mean and standard deviation of the distribution of values obtained after repeating the injection in 100 background toys.

R	0.1	0.2	0.3	0.4
# injected events	460 ± 30	920 ± 40	1380 ± 50	1840 ± 70
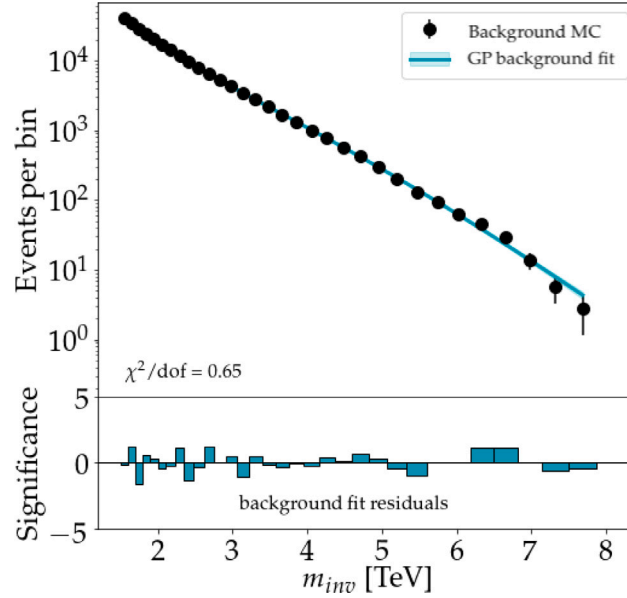
Fig. 12. Top panel: invariant mass spectrum showing a GP fit for the background. Bottom panel: per-bin significance of the deviation between the number of events and the fit.

6.4.3. Modelling the dijet invariant mass: QCD background and artificial signal

For testing the two-step procedure, we used as toy dataset a simulated dijet invariant mass spectrum following the event selections described in [126].

Before reporting the results of the GP optimization, we describe the injection of signal in the invariant mass spectrum. The starting point is the simulated background dataset we generate pseudo-data from, where the signal can be injected. Since the spectrum ranges over several orders of magnitude in the number of events, we prescribe the amplitude of the injected signal via a quantity defined from a signal-over-background ratio (R). We calculate the amplitude of the Gaussian signal from the mean and width: R is the ratio of signal to background events that should be in a window constructed from the interval given by the mean and width. The number of events taken into account in a given window is given by the bin counts of the distribution contained in the window; bins whose centres are outside the range (mean $\pm$ width/2) are not counted:

$$R = \frac{\text{Injected signal events in the window}}{\text{Background events in the window}}. \tag{17}$$

Analogously, the extraction of the R (and hence the amplitude) of the signal comes from the same ratio within a window defined by the mean and width of the extracted signal from the parameters of the signal kernel. We inject signals with different possible combinations of values for a Gaussian distribution, namely the mass (3, 3.5, 4, 4.5, and 5 TeV), the width (150, 300, and 450 GeV), and the R ratio (0.1, 0.2, 0.3, and 0.4), totalling 60 values. The values were chosen to cover a wide range of the spectrum and different intensities of the signal hypotheses.

In Table 4 we provide the number of events injected that correspond to different values of the ratio R, to give a sense of the mapping between the two quantities.

We present in Fig. 12 a first fit (first step) performed on the pure background simulated data.[7] The posterior GP background mean, the data and their respective errors are displayed. We can observe that the GP fit is able to accommodate the distribution with a smooth model whose bin-wise discrepancies with data are small. To measure such discrepancies, the χ^2 divided by the number of degrees of freedom is calculated.

---

[7] In this and subsequent plots of this section, the displayed significances correspond to "signed z-values only if p-value < 0.5", as defined in [127], where the p-value is calculated assuming that each bin count follows a Poisson distribution.
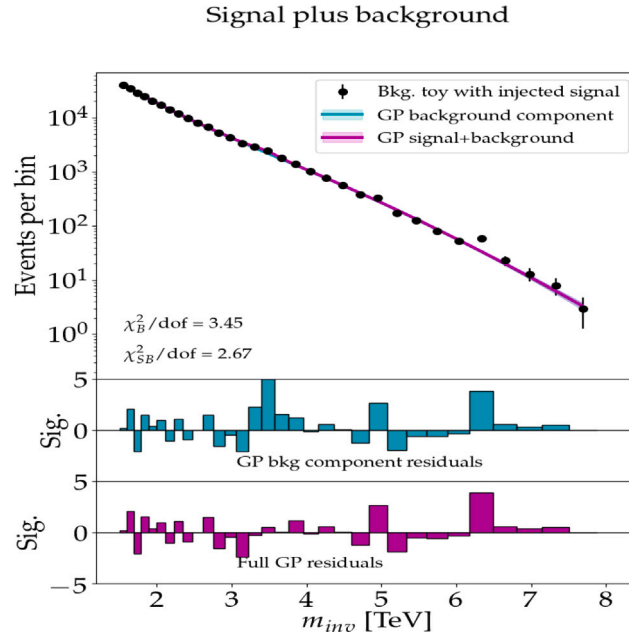
Fig. 13. Top panel: invariant mass spectrum displaying a GP background fit, with event counts for a background toy with signal injected centred at 3.5 TeV with a width of 150 GeV and R of 0.1, and a signal plus background fit. Solid coloured lines represent the GP fit components. Middle and bottom panels: per-bin significance of the discrepancy between the event counts and indicated fits. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 13 shows a similar plot for the second step in the presence of an injected signal. By observing the residuals we verify that the GP can identify the injected signal: the middle panel clearly reveals a localized set of discrepant bins in the region of the injected signal that the background-only component of the fit does not capture; whereas the bottom panel shows that the injected signal has been incorporated in the full fit.

A corresponding graph where signal identification becomes more evident is presented in Fig. 14. In the case shown, the GP signal component correctly locates the injected signal and is able to recover relevant parameters (strength, width, mean). However, when we approach the faintest or widest signals, the method becomes prone to mis-identification. The method is also used in the absence of the signal when performing the second step, in order to observe to which extent spurious signals are detected.

Good agreement is generally found when checking the linearity between injected and extracted R values in Fig. 15. For every signal hypothesis, a sampling of the Gaussian signal is performed and injected 100 times on different background pseudo-data, allowing the extraction of a distribution of 100 values for each signal parameter and for each hypothesis. Despite the good agreement, we observe that the higher mass hypotheses tend to be poorly identified; this is due to the specific binning of this spectrum, where bin widths increase approximately from 90 GeV at lower masses to 360 GeV at the right-end of the spectrum. Even if R is by construction a more consistent parameter to prescribe signal strengths in different regions of the spectrum than, for example, a fixed number of events, it still can lead to undesirably faint signals at high mass values, where the event counts are small.

The value presented as the upper error (mean+rms) of the spurious signal detection ($R = 0.25$) serves as an indication of the ability of the method to detect genuine signals. As we stated before, GPs are in general tools flexible enough to model the distribution without prescribing a mean function, and in terms of signal extraction they do not yield a significant gain in the extraction power.

6.4.4. Further studies

An application of the three-step procedure described *supra* was used to accommodate backgrounds with turn-ons (first two steps) and signals (last step). Here we give a general outline on this use case; further detail can be found in [128].

The motivation for using a three-step procedure arose in an attempt of modelling the invariant mass spectrum of top quark pairs ($t\bar{t}$) to search for a Z'-boson decaying to $t\bar{t}$. That spectrum covers a mass range from 0.5 to 2 TeV with a turn-on for lower values (roughly below 0.8 TeV) followed by a decay, the latter being somewhat similar to the dijet spectrum. In an effort to use the two-step procedure in this case, we observed that the signal kernel was not able to identify injected signals in the second step, but always captured the turn-on part instead. Thus, the first and second step (with Σ_1 and Σ_{12} respectively) were used on the background-only sample, successfully accommodating the background spectrum; a third step (using Σ_{123}) was then applied for signal extraction.

The testing of the method was similar to that performed in the dijet case. After fitting the background spectrum, we generate and use background pseudo-data to inject Z' signals corresponding to simulated resonance hypotheses at two mass points, *i.e.* 750 GeV and 1250 GeV. The last step of the procedure is then applied 100 times, where we obtain the signal parameters (θ_3).
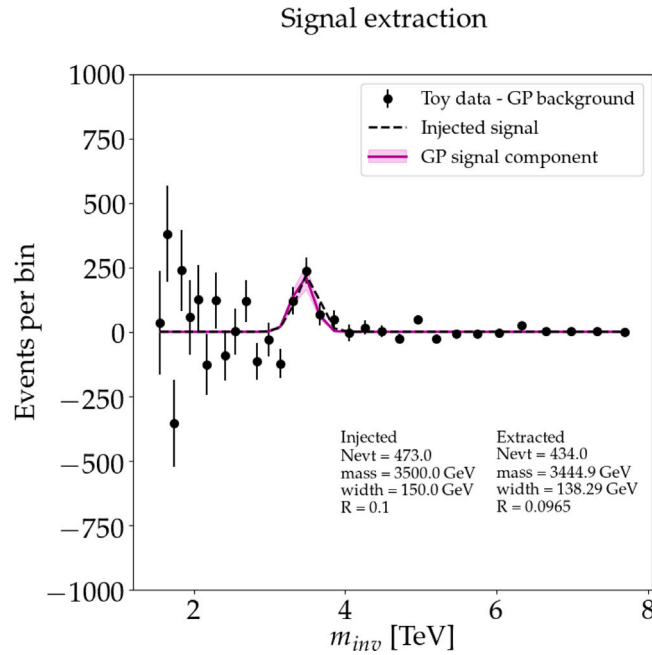
Signal extraction

Fig. 14. Residual plot corresponding to Fig. 13. The GP signal component (solid magenta) and the signal injected (dashed black line) are displayed as well as a subtraction of the toy data set with a signal injected minus the background GP fit (black dots with error bars). Injected and extracted signal values are shown.
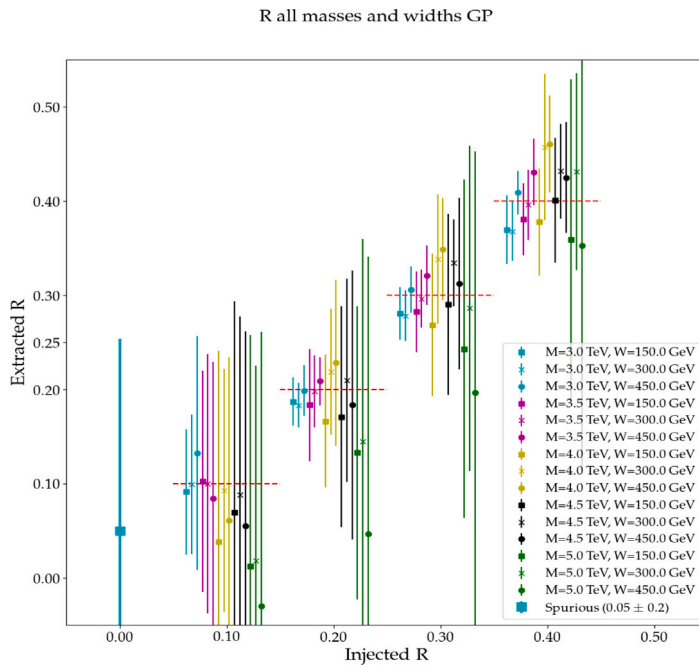

Fig. 15. Linearity plots for the R values of the signal injected in the dijet spectrum; each plot corresponds to indicated mass M and width W (both in GeV). The values of the width are the same for each plot column, and those of the mass are fixed through each plot row. Points and error bars (means and rms) are calculated from the distribution of extracted width values. A dashed red $x = y$ line is plotted for reference.

As reported in [128], signal extraction is achieved with moderate success. Two main limiting factors are the faintness of the nominal simulated signals used, and in some cases an observed shift of the GP signal component towards the turn-on point. Concerning the first issue, we tested the extraction where the nominal simulated signal was amplified by different factors. For the shift in the signal component, we instead limited the θ_3 parameters in the optimization to lay outside the turn-on region, a

procedure that helps overcome the issue; still, in the case of spurious detection (no signal injected), the signal component would mostly tend to find spurious signals near the lower bound allowed for the location (*i.e.* m in θ_3). This and other limitations reported in the study need further work in order to enable the described method to be applicable to resonance searches in the $t\bar{t}$ mass spectrum.

6.4.5. Conclusions

We explored different aspects of GP approaches that are able to perform background modelling and signal detection, without any prior information on the mean, *i.e.* $\mu(x) = 0$. The two-step procedure was able to detect signals with different widths, intensities, and locations in the dijet invariant mass spectrum. We use R, a ratio of events defined within a window to measure the intensity of the signal; since the spurious detection leads to identifying signals up to 0.25, we use this value as an indication of the faintest signal that the method is capable to identify. We also described the application of the GP method in three steps in the more challenging scenario of the search for a resonant signal in the $t\bar{t}$ invariant mass spectrum: the first two steps model the background, including the turn-on region, and the third step is used for extracting the signal.

The GP methods that we used provide a way towards alternative background modelling and signal identification techniques. The procedures presented here lend themselves to further improvement: in particular the performance of the method applied to the dijet mass spectrum could be further improved with a better definition of the injected amplitude and, in the case of the top quark pair spectrum, the three-step procedure showed some limitations in modelling the background turn-on and capturing the signal. Finally, the application of the method to model-independent searches could be better served by more versatile methods, which could be studied by exploring other kernels and by avoiding reliance on a complex iterative fitting procedure.

7. Similarity Search for the Fast Simulation of the ATLAS Forward Calorimeter

7.1. Overview

The physics measurements of the experiments at the LHC strongly rely on detailed and precise simulations to predict the complex response of the detectors. These simulations are essential for the comparison between experimental results and theoretical predictions. The detailed simulation of the interaction and propagation of the particles through the detector is typically performed with the GEANT4 [129] simulation toolkit, and it demands very large computing resources. The development of innovative computational and algorithmic techniques is fundamental to cope with the complexity of the simulation, which is further expected to increase in the next decade, with the increment of the volume of data collected at the LHC.

7.2. Fast simulation with frozen showers

In the ATLAS experiment, the simulation chain is managed by Athena [130,131], an open software used for the simulation of the particle interaction and propagation in the detector, as well as the physics object reconstruction. The ATLAS experiment makes large use of fast simulation techniques to reduce the computational resources required by the entire simulation chain. A prominent example is the *FastCaloSim* module [132] in Athena, which uses a parameterized model of the calorimeter response in the barrel.

When it comes to simulating the response of a calorimeter, specifically built with very dense material to entirely absorb the energy of the particles that traverse it, the requirements of computational resources could increase exponentially, up to $\sim 70\%$ of the total CPU time dedicated to the entire simulation process. This is especially relevant for the simulation of the ATLAS forward calorimeter (FCal) [133,134]. The FCal is a sampling calorimeter that covers the most forward region of the ATLAS detector, in the range of pseudo-rapidity $3.1 < |\eta| < 4.9$. Due to the close position to the beamline, energy and density of particles reaching the FCal are very high. The harsh environment is reproduced by Monte Carlo simulations, responsible for the propagation of high-energy particles that enter the calorimeter and form showers of secondary particles inside it. The FCal consists of three consecutive modules along the beamline. The first module, the closest to the interaction point, is made of copper and is designed to absorb the majority of the electromagnetic showers. The following modules, made with tungsten, are designed to absorb the hadronic contribution to the particle showers. Since the total energy of a showering particle is predominantly deposited in the sampling calorimeter through a large number of soft electrons and photons, the work presented here focuses on the fast simulation of the FCal response in the electromagnetic module.

In ATLAS, the simulation of the FCal detector already makes use of fast techniques aimed at reducing the complexity of the response. The FCal fast simulation is based on a *frozen shower library* approach [135], which consists of storing pre-simulated electromagnetic showers initiated by high-energy particles entering the front face of the calorimeter, subsequently used to quickly model the detector response. The frozen shower substitution is illustrated in Fig. 16.

The generation of the frozen library requires a preliminary simulation of the low-energy particles obtained with the propagation in the FCal of the particles arising from simulated collision events. In the work presented here, LHC-produced top quark pair events $pp \rightarrow t\bar{t} + X$ have been used. When secondary particles with energy below a given threshold are produced, typically electrons below 1 GeV and photons below 10 MeV, the corresponding kinematic information is saved in the frozen library, together with the collection of hits corresponding to the subsequent shower produced by them. The frozen shower is then used in the fast simulation of the FCal response. When a low-energy particle is produced in the FCal, the simulation (performed with GEANT4) is suspended, the particle is associated to one entry in the library and the corresponding electromagnetic shower, whose energy is scaled to match the energy of the particle produced, is sampled.
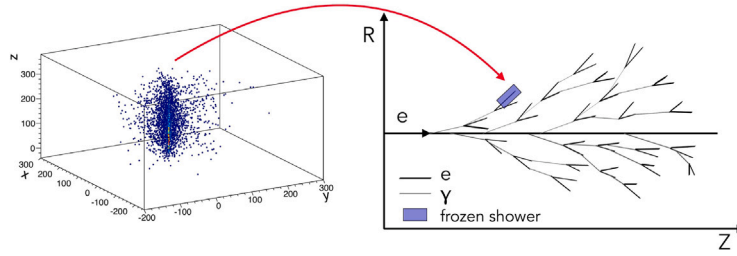
Fig. 16. Sketch of the particle showers. A shower collected in the frozen library (left) is used to replace a low-energy particle produced in the simulation (right).

The ATLAS FCal calorimeter consists of an absorber matrix (made with copper, in the case of the electromagnetic module) instrumented with cylindrical electrodes parallel to the beamline. Each electrode is composed by a rod (anode) placed inside a tube (cathode). The small gap in between is filled with liquid argon, used as the active material of the sampling calorimeter. The distance of the particle to the active material is an important parameter to determine the characteristics of the originated shower.

The frozen library currently used by ATLAS for the FCal fast simulation contains pre-simulated showers parameterized in bins of pseudorapidity η and d, the distance of the low-energy particle to the centre of the closest rod. The energy of the particle is also stored. During the simulation, the matching between the simulated particle and one entry of the library is performed by selecting the closest d and η bin and then finding the closest energy available within the bin.

The energy resolution of the detector response obtained with the fast simulation should always be able to reproduce the result obtained with a standard simulation (full simulation). The fast simulation approach currently used in ATLAS strongly reduces the computational requirements of the simulation. However, it does not reproduce well the energy resolution provided by the full simulation, thus requiring some tuning of the parameters in the library prior to its usage.

7.3. Similarity search

The new proposed strategy [136] employs *similarity search* techniques to select the most suitable particle–shower pair in the library. Such techniques constitute a branch of machine learning and include *clustering* and *indexing* methods that enable quick and efficient searches for vectors similar to each other [137]. These methods are usually employed in applications such as image features search and document retrieval in large scale databases, for example in search engines. A new version of the library is also used. The stored showers are keyed by 6-dimensional parameter vectors containing the angular and spatial information of the low-energy particles, as well as their energy.

Similarity search techniques can be described as approximate nearest neighbour searches to find similar vectors. A prediction for a query vector is made from data, by measuring the distance with the instances in the dataset and selecting the most similar ones. The similarity can be defined using different metrics, one of the most common being the Euclidean distance $D(\mathbf{p}, \mathbf{q})$ between two vectors $\mathbf{p}$ and $\mathbf{q}$, defined as

$$D(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_d - p_d)^2} = \sqrt{\sum_{i=1}^{d} (q_i - p_i)^2} \,. \tag{18}$$

In the frozen library context, given a query vector containing the kinematic information related to the simulated low-energy particle, the search is performed to find the most similar vector of kinematic variables collected in the library. The search is not performed exhaustively in the whole dataset, but is based on approximation methods that often involve the use of clustering algorithms to limit the portion of the dataset considered in the search. This involves pre-processing the dataset with an indexing procedure, which learns a data structure from the dataset.

In this work, the *Facebook AI Similarity Search* (Faiss) [138] package, developed by the Facebook Artificial Intelligence researchers to efficiently retrieve multimedia documents, has been used to construct the index and perform the search within the frozen library. Several indexing options to build a structure from the dataset are available with Faiss. One of the most common structures is called *inverted index* [139] and is based on the space partition into Voronoi regions defined by a K-means clustering algorithm. Each vector is encoded into the corresponding region (*code*) and saved in an array (the inverted index). Given a query vector, the search is then performed on the indexed dataset, based on the similarity with the centroids of the clusters. The vectors contained in each queried cluster are decoded sequentially and compared to the query. The Faiss package also provides a method called *Product Quantizer* [140] to compress the vectors and perform a faster search at the cost of accuracy. The Product Quantizer is particularly useful when dealing with high-dimensional vectors and very large datasets ($> 10^6$ elements). Another interesting method is the so-called *Hierarchical Navigable Small World graphs* (HNSW) [141]. The HNSW consists of a multi-layer structure built from the dataset, where the vectors are connected through links, based on the similarity metric, and organized over different layers according to the length of the links. Then the search is performed across the layers, starting from the region characterized by the longest links. The HNSW is considered state-of-the-art among similarity search algorithms and provides a good trade-off between speed and accuracy both in high- and low-dimensional data.
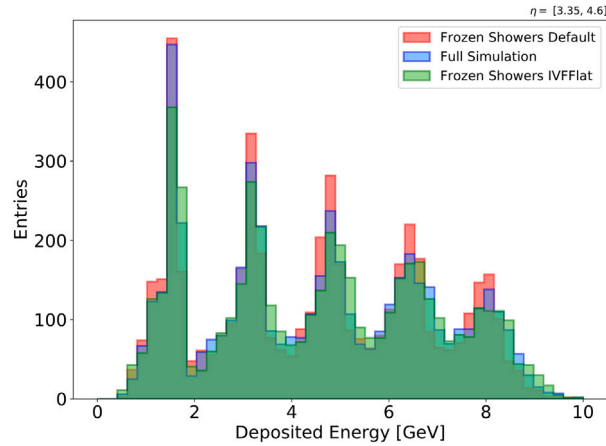
Fig. 17. Distribution of the energy deposited by the showers originated by the low-energy electrons in the FCal calorimeter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
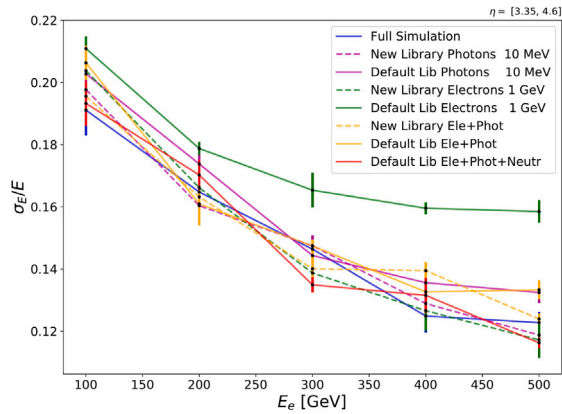

Fig. 18. Resolution response obtained with the new (dashed lines) and the old (full lines) libraries. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

7.4. Validation and results

To validate the new approach, a library containing $\sim 10^5$ showers has been generated with electrons of energy below 1 GeV. Then, high-energy electrons are generated at the interaction point with discrete energies of [100, 200, 300, 400, 500] GeV and pseudorapidity in the range $3.35 < |\eta| < 4.60$, and propagated using the fast simulation with the new implementation.

The processing time is evaluated as the average of the CPU time spent executing the propagation of a high-energy electron of fixed initial energy. This time includes the fast simulation of $\sim 10^4$ low-energy showers produced in the propagation. The detector resolution is measured as the ratio σ_E/E between the standard deviation and the mean of the distribution of the energy deposited by the low-energy electrons. This distribution is shown in Fig. 17 for the fully Geant4-based simulation (blue), the default library (red), and the new library (green).

Different indexing methods have been tested and compared on the basis of CPU time and resolution response. On the frozen shower library, all the tested methods show optimal results in terms of resolution response. The method based on the HNSW structure has been chosen as a benchmark, based on a slightly better performance in terms of CPU requirements.

A specific library has to be generated for all the particles that require a fast simulation. During simulation, the different libraries can either be used individually or in combination. For comparison, a library containing showers originated from photons with an energy threshold of 10 MeV was also produced. Fig. 18 shows the final comparison of the detector resolution response obtained with different versions of the default libraries (continuous curves) and the new libraries (dashed curves).

The electron and photon libraries have been tested both individually and in combination in the default and new implementations. The resolution obtained with the full simulation (blue) is in general well reproduced, except for the default electron library (green) that presents some discrepancies. Also, for the default library, a previous tuning of the library parameters was needed. The new approach provides optimal results in terms of detector resolution response.
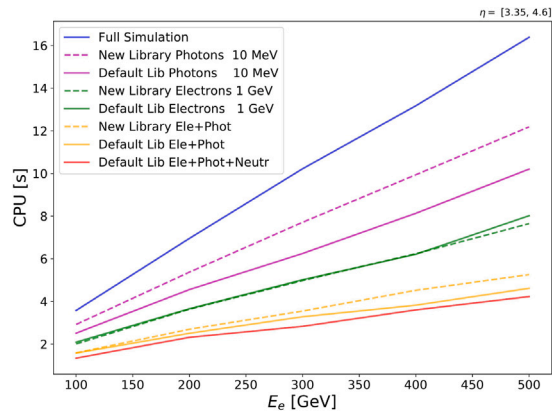
Fig. 19. CPU response obtained with the new (dashed lines) and the old (full lines) libraries.
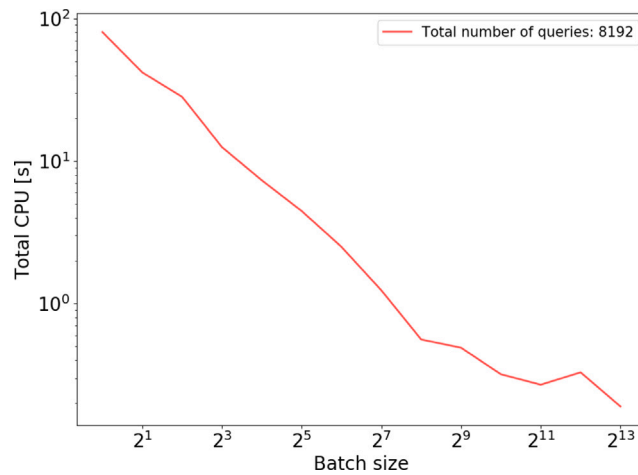

Fig. 20. Faiss search in batches: total CPU vs batch size.

Fig. 19 shows the final comparison of the CPU response. The electron libraries (green) provide a significant gain (up to ∼ 50%) with respect to the full simulation, with comparable results for the two different implementations. The photon library used in the new implementation appears to be less efficient than the default one. In general, the photon library provides a smaller gain to the simulation than the electron library. This is due to the smaller threshold used for the generation of the photon libraries. When the electron and photon libraries are used in combination, the simulation is significantly accelerated (up to ∼ 70%), and again provides results comparable between the two approaches, as shown by the yellow curves. The fact that the different performance obtained with the photon libraries is not reflected in the combined results can be related again to the different thresholds used for the library generation, which makes the electron library much more likely to be used during simulation.

7.5. Conclusions and prospects

The proposed approach for a new fast simulation of the ATLAS FCal was fully implemented within the ATLAS simulation framework. The new method based on similarity search techniques provides optimal results in terms of detector resolution response, outperforming the default approach, which needs to be adjusted before usage. A significant gain in CPU requirements with respect to the full simulation, amounting to an improvement of about 70%, is also obtained as a comparable result to what obtained with the default approach. CPU acceleration could be further enhanced with a library sampling performed in batches. When this work was performed, the batching functionality could not be enabled, as a multi-threading processing model was not supported by the ATLAS simulation framework. However, the Faiss implementation is optimized for batch searches.

The potential improvement due to batch searches was tested in a standalone mode, using two separate datasets, one as the library, and the other as a collection of query vectors. Fig. 20 shows that the CPU time decreases significantly as the batch size increases. This suggests that the fast simulation implemented in ATLAS would greatly benefit from batch searches.

8. Towards the Full Optimization of Physics Measurements

8.1. Introduction

In the previous sections we have described a number of applications to LHC research of advanced multivariate methods, to which AMVA4NewPhysics provided a significant contribution. The produced examples provide a clear indication of the true revolution that the process of inference extraction from complex collider data withstood in the course of the past decade. In a situation where computer science is moving forward very quickly towards artificial intelligence applications in all domains of science and technology, as well as elsewhere, one is bound to ask what may the next step be in HEP research.

We offer a partial answer to the above question in this section, where we summarize the result of studies aimed at achieving the full end-to-end optimization of physics measurements based on injecting information on the final analysis goal in the dimensionality reduction procedures applied to collider data through supervised classification.

8.2. The misalignment problem

The typical workflow of the measurement of a physical quantity from data collected by particle collisions at the LHC involves the construction of an informative summary statistic—usually one-dimensional—from multidimensional high-level features of observed data events, and the extraction of inference on the parameter(s) of interest by a likelihood fit or some other statistical procedure. The high-level features, usually individual four-momenta of identified relevant particles, are themselves the result of a large compression procedure that starts with millions of digital readouts of as many electronic channels; this procedure is generically addressed as event reconstruction.

Event reconstruction, of which we have discussed some aspects in the previous Sections (Section 4, Section 7) is a very complex task, and in ATLAS and CMS the development of the relevant software requires an investment of hundreds of person-years; it involves thorough calibrations, tunings of data collection procedures, modelling of detector response, and a number of related studies. We will not consider these procedures further in this section, other than noting that they necessarily target, as their optimization goal, suitable simplified surrogates of the true goals of the experiment. The complexity of the whole set of procedures that convert a raw dataset of digitized detector readouts into the measurement of a fundamental physical quantity prevent a direct optimization of the ultimate goal of experimentalists, which is usually the minimization of the total uncertainty on the measurement of a set of parameters of interest. For example, the electromagnetic calorimeter of an LHC experiment is naturally built to improve as much as possible the energy resolution and identification capabilities of energetic, isolated photons (see *supra*, Section 7). Those figures of merit are manageable optimization proxies for the discovery reach of the experiment to high-mass resonances decaying into pairs of photons, which is much harder to study and maximize directly, but in general they are not monotonous functions of the latter quantity, which is dependent on additional parameters and conditions.

The above mismatch between desirable goals and optimization proxies constitute what could be called a *misalignment problem*. Realigning our construction and reconstruction strategies to our experimental goals is a topic that will require our attention in the forthcoming years, leveraging differentiable programming tools that already enable the investigation of full end-to-end optimization strategies for particle reconstruction. A few efforts have started to consider the design of experiments as the subject of a full optimization [142–145], and there is hope that even full-blown particle detectors of the scale of ATLAS or CMS, which are among the most complex instruments ever built by human beings, may one day be assisted in their design by artificial intelligent tools leveraging automatic differentiation.

A misalignment problem of smaller scale, yet by itself often one of quite large impact, exists also in the smaller-scale data compression step that starts from the $O(50)$ high-level event features resulting from event reconstruction, and ends in a one-dimensional summary statistic on which experimental physicists base the extraction of information on the quantity (or quantities) of interest. The latter is usually the mass of a particle, or the cross section of an interesting phenomenon, and/or some other fundamental parameter of the underlying theory. Whatever the measurement goal may be, the workflow of the final inference step often involves the use of supervised classification to construct a discriminating variable that effectively separates the wanted signal from all competing backgrounds. The classification task may be performed by training a neural network, as exemplified in Section 2, *supra*; the network learns from simulated samples of signal and background processes, and produces an effective summary which, in absence of systematic effects, is often close to optimal in performance. In this context, an optimal classifier is one monotonous with the likelihood ratio, which by the Neyman–Pearson's lemma is the most powerful test statistic to separate two simple hypotheses [146]. The produced summary statistic is optimal in the absence of systematic effects, but it is in general not optimal when model uncertainties or other biases have to be taken into account. The neural network, oblivious of those systematic uncertainties, produces a summary that does not guarantee the optimality of the final inference. In statistical terms, the summary is not *sufficient*: it does not make optimal use of the input information in the task of estimating the parameter of interest.

8.3. INFERNO

Recognizing the misalignment problem typically encountered in data analyses that perform a dimensionality reduction step from high-level features to a neural network classifier output, we designed an algorithm that promises full end-to-end optimality and approximate sufficiency of the constructed summary statistic. The algorithm, called INFERNO [147] (an acronym of *Inference-Aware Neural Optimization*), exploits `TensorFlow` libraries to construct a fully differentiable pipeline that includes a neural network (used
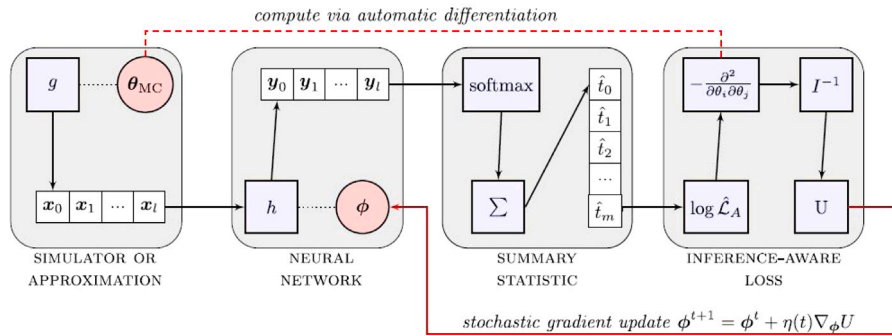
Fig. 21. Pipeline of INFERNO. Data produced by a simulator (left box) is fed to a neural network (second box from left). The network output is processed through a softmax operator to make it differentiable, and constitutes the summary statistic (third box from left) employed by a likelihood maximization (right box) to perform inference on the parameter of interest and related nuisances θ, and obtain an estimate of the associated variance. The variance is fed back into the loop such that the network is capable of optimizing its parameters ϕ to produce the smallest uncertainty on the parameter of interest given nuisance parameters [147].

for the dimensionality-reduction step) as well as a model of the inference-extraction procedure (a binned likelihood fit to the network outputs). In INFERNO, nuisance parameters affecting the background shape are introduced in the model, and the loss function of the neural network is retro-fitted with knowledge of the size of the full uncertainty expected on the final parameter of interest. The latter results from the inclusion in the model of all statistical as well as systematic effects, and is estimated from the inverse of the Hessian matrix of the likelihood fit employed for the parameter estimation step. Stochastic gradient descent then allows the network to learn a data representation that minimizes the variance of the parameter of interest.

The algorithm was originally implemented with TensorFlow 1.0 [148] and is freely available [149]; it has also recently implemented within TF2.0 [150] and PyTorch [151]. A sketch of the software pipeline which INFERNO is based on is shown in Fig. 21.

8.4. Outlook

INFERNO is innovative and powerful, but its application is not always straightforward as it expects the user to produce a suitable differentiable model of the effect of nuisance parameters; so far the algorithm has only been proven to work—quite effectively—when applied to synthetic datasets. Its application to a full-blown LHC data analysis is in progress at the time of writing; we fully expect that the solution of the misalignment problem discussed *supra*, provided by INFERNO, will yield much higher sensitivity to physics analyses. Following INFERNO, a few similar attempts have recently been published, see *e.g.* [152]. We direct readers interested in more detail on this area of research to a recent review of methods that address the incorporation of nuisance parameters in supervised classification for HEP research [153].

9. Concluding Remarks

The present document summarizes the most significant studies carried out by members of the AMVA4NewPhysics ITN in the years 2015–2019 within the context of searches and measurements of the Higgs boson and of LHC searches for new physics phenomena in datasets collected during Run 1 and Run 2 by the ATLAS and CMS experiments at the CERN LHC. The common denominator of the reported studies is the attempt to go beyond current data analysis practice in HEP experiments, exploiting state-of-the-art techniques from computer science and developing entirely new methods to extract the maximum amount of information from the invaluable datasets of proton–proton collisions.

The effort that brought these studies to the fore could not have been undertaken without the resources made available by the European Union, as highlighted *infra*. However, the breadth of results and the significant stride they provide in the direction of more effective data analysis in fundamental physics research would have been impossible without the cohesive motivation of the young participants of the AMVA4NewPhysics training network. For that reason, it is necessary to stress here that an additional untold success of the network's action has been to provide the spark for that collaborative effort.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Evans, P. Bryant, LHC Machine, J. Instrum. 3 (2008) S08001, http://dx.doi.org/10.1088/1748-0221/3/08/S08001.
[2] ATLAS Collaboration, The ATLAS experiment at the CERN large hadron collider, J. Instrum. 3 (2008) S08003, http://dx.doi.org/10.1088/1748-0221/3/08/S08003.
[3] CMS Collaboration, The CMS experiment at the CERN LHC, J. Instrum. 3 (2008) S08004, http://dx.doi.org/10.1088/1748-0221/3/08/S08004.
[4] ATLAS Collaboration, Observation of a new particle in the search for the standard model Higgs Boson with the ATLAS detector at the LHC, Phys. Lett. B 716 (2012) 1, http://dx.doi.org/10.1016/j.physletb.2012.08.020, arXiv:1207.7214.
[5] CMS Collaboration, Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, Phys. Lett. B 716 (2012) 30, http://dx.doi.org/10.1016/j.physletb.2012.08.021, arXiv:1207.7235.
[6] ATLAS Experiment - Public Results, https://twiki.cern.ch/twiki/bin/view/AtlasPublic.
[7] CMS Experiment - Public Results, http://cms-results.web.cern.ch/cms-results/public-results/publications.
[8] A. Stakia, Advanced multivariate analysis methods for use by the experiments at the large Hadron collider, Phys. Scr. 96 (5) (2021) 054001, http://dx.doi.org/10.1088/1402-4896/ab9bd8.
[9] ImageNet Large Scale Image Recognition Challenge, https://image-net.org/challenges/LSVRC/index.php.
[10] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, D. Rousseau, The Higgs boson machine learning challenge, in: Proceedings of the NIPS 2014 Workshop on High-Energy Physics and Machine Learning, in: Proc. Mach. Learn. Res., vol. 42, PMLR, Montreal, Canada, 2015, p. 19, http://proceedings.mlr.press/v42/cowa14.html.
[11] G.C. Strong, On the impact of selected modern deep-learning techniques to the performance and celerity of classification models in an experimental high-energy physics use case, Mach. Learn. Sci. Technol. 1 (2020) 45006, http://dx.doi.org/10.1088/2632-2153/ab983a, arXiv:2002.01427.
[12] G.C. Strong, LUMIN, Zenodo, 2019, http://dx.doi.org/10.5281/zenodo.2601857.
[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Adv. Neural Inf. Process. Syst., Vol. 32, Curran Associates, Inc., 2019, https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.
[14] ATLAS Collaboration, Dataset from the ATLAS Higgs boson machine learning challenge 2014, CERN open data portal, 2014, http://dx.doi.org/10.7483/OPENDATA.ATLAS.ZBP2.M5T8.
[15] G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur. Phys. J. C 71 (2011) 1554, http://dx.doi.org/10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727.
[16] G. Cowan, Two developments in tests for discovery: Use of weighted Monte Carlo events and an improved measure, in: Progress on Statistical Issues in Searches, SLAC, 2012, https://www.pp.rhul.ac.uk/~cowan/stat/cowan_slac_4jun12.pdf.
[17] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: Proc. IEEE Int. Conf. Comput. Vis., 2015, p. 1026, http://dx.doi.org/10.1109/ICCV.2015.123, arXiv:1502.01852.
[18] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: 13th Int. Conf. on Artificial Intelligence and Statistics, Vol. 9, (AISTATS), 2010, p. 249, http://proceedings.mlr.press/v9/glorot10a.html.
[19] C. Guo, F. Berkhahn, Entity embeddings of categorical variables, 2016, arXiv:1604.06737.
[20] G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, Self-normalizing neural networks, in: Adv. Neural Inf. Process. Syst., 2017, p. 972, arXiv:1706.02515.
[21] P. Ramachandran, B. Zoph, Q.V. Le, Searching for activation functions, in: 6th Int. Conf. Learn. Represent, (ICLR) - Work. Track Proc., 2018, arXiv:1710.05941.
[22] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: 5th Int. Conf. Learn. Represent, (ICLR) - Conf. Track Proc., 2017, arXiv:1608.03983.
[23] L.N. Smith, A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay, 2018, CoRR abs/1803.09820, arXiv:1803.09820.
[24] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84, http://dx.doi.org/10.1145/3065386.
[25] G. Huang, Y. Li, G. Pleiss, Z. Liu, J.E. Hopcroft, K.Q. Weinberger, Snapshot ensembles: Train 1, get M for free, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, Conference Track Proceedings, OpenReview.net, 2017, https://openreview.net/forum?id=BJYwwY9ll, arXiv:1704.00109.
[26] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, A.G. Wilson, Loss surfaces, mode connectivity, and fast ensembling of DNNs, in: Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018, p. 8789, https://proceedings.neurips.cc/paper/2018/hash/be3087e74e9100d4bc4c6268cdbe8456-Abstract.html, arXiv:1802.10026.
[27] F. Draxler, K. Veschgini, M. Salmhofer, F. Hamprecht, Essentially no barriers in neural network energy landscape, in: Proceedings of the 35th International Conference on Machine Learning, in: Proc. Mach. Learn. Res., vol. 80, PMLR, 2018, p. 1309, http://proceedings.mlr.press/v80/draxler18a.html, arXiv:1803.00885.

[28] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, A.G. Wilson, Averaging weights leads to wider optima and better generalization, in: 34th Conf. Uncertain. Artif. Intell., Vol. 2, UAI, 2018, p. 876, arXiv:1803.05407.

[29] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, p. 2261, arXiv:1608.06993.

[30] M. Cepeda, S. Gori, P. Ilten, M. Kado, F. Riva (Eds.), Higgs physics at the HL-LHC and HE-LHC, in: CERN Yellow Reports: Monographs, vol. 7, 2019, p. 221, https://e-publishing.cern.ch/index.php/CYRM/article/view/952.

[31] The ATLAS and CMS Collaborations, Collection of notes from ATLAS and CMS, in: CERN Yellow Reports: Monographs, vol. 7, 2019, https://e-publishing.cern.ch/index.php/CYRM/article/view/956.

[32] CMS Collaboration, A measurement of the Higgs boson mass in the diphoton decay channel, Phys. Lett. B805 (2020) 135425, http://dx.doi.org/10.1016/j.physletb.2020.135425.

[33] M. Grazzini, G. Heinrich, S. Jones, S. Kallweit, M. Kerner, J.M. Lindert, J. Mazzitelli, Higgs boson pair production at NNLO with top quark mass effects, J. High Energy Phys. 05 (2018) 059, http://dx.doi.org/10.1007/JHEP05(2018)059, arXiv:1803.02463.

[34] J. March-Russell, S.M. West, D. Cumberbatch, D. Hooper, Heavy dark matter through the Higgs portal, J. High Energy Phys. 07 (2008) 058, http://dx.doi.org/10.1088/1126-6708/2008/07/058, arXiv:0801.3440.

[35] CMS Collaboration, Searches for a heavy scalar boson H decaying to a pair of 125 GeV Higgs bosons hh or for a heavy pseudoscalar boson A decaying to Zh, in the final states with $h \to \tau\tau$, Phys. Lett. B 755 (2016) 217, http://dx.doi.org/10.1016/j.physletb.2016.01.056, arXiv:1510.01181.

[36] ATLAS Collaboration, Search for heavy resonances decaying into a W or Z boson and a Higgs boson in final states with leptons and b-jets in 36 fb^{-1} of $\sqrt{s}$ = 13 TeV pp collisions with the ATLAS detector, J. High Energy Phys. 03 (2018) 174, http://dx.doi.org/10.1007/JHEP03(2018)174, [Erratum: JHEP 11, 051 (2018)], arXiv:1712.06518.

[37] P.A. Zyla, et al., Particle Data Group, Review of particle physics, Prog. Theor. Exp. Phys. 2020 (8) (2020) http://dx.doi.org/10.1093/PTEP/PTAA104, https://academic.oup.com/ptep/article/2020/8/083C01/5891211.

[38] CMS Collaboration, Study of jet quenching with Z + jet correlations in Pb-Pb and pp collisions at $\sqrt{s_{NN}}$ = 5.02 TeV, Phys. Rev. Lett. 119 (2017) 082301, http://dx.doi.org/10.1103/PhysRevLett.119.082301.

[39] CMS Collaboration, Correlations between jets and charged particles in Pb-Pb and pp collisions at $\sqrt{s_{NN}}$ = 2.76 TeV, J. High Energy Phys. 2 (2016) 156, http://dx.doi.org/10.1007/JHEP02(2010)091.

[40] CMS Collaboration, Observation of long-range, near-side angular correlations in proton-proton collisions at the LHC, J. High Energy Phys. 9 (2010) 91, http://dx.doi.org/10.1007/JHEP09(2010)091.

[41] CMS Collaboration, Measurement of Bose-Einstein correlations in pp collisions at $\sqrt{s}$ = 0.9 and 7 TeV, J. High Energy Phys. 05 (2011) 29, http://dx.doi.org/10.1007/JHEP05(2011)029.

[42] CMS Collaboration, First measurement of Bose-Einstein correlations in proton-proton collisions at $\sqrt{s}$ = 0.9 and 2.36 TeV, Phys. Rev. Lett. 105 (2010) 32001, http://dx.doi.org/10.1103/PhysRevLett.105.032001.

[43] P. de Castro Manzano, M. Dall'Osso, T. Dorigo, L. Finos, G. Kotkowski, G. Menardi, B. Scarpa, Hemisphere mixing: a fully data-driven model of QCD multijet backgrounds for LHC searches, in: Proceedings, 2017 European Physical Society Conference on High Energy Physics, EPS-HEP 2017, p. 370, https://doi.org/10.22323/1.314.0370.

[44] CMS Collaboration, Search for nonresonant Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state at $\sqrt{s}$ = 13 TeV, J. High Energy Phys. 04 (2019) 112, http://dx.doi.org/10.1007/JHEP04(2019)112.

[45] CMS Collaboration, Prospects for HH Measurements at the HL-LHC, Tech. Rep. CMS-PAS-FTR-18-019, CERN, Geneva, 2018, https://cds.cern.ch/record/2652549.

[46] CMS Collaboration, Technical Proposal For a MIP Timing Detector in the CMS Experiment Phase 2 Upgrade, Tech. Rep. CERN-LHCC-2017-027. LHCC-P-009, CERN, Geneva, 2017, https://cds.cern.ch/record/2296612.

[47] CMS Collaboration, A MIP Timing Detector for the CMS Phase-2 Upgrade, Tech. Rep. CERN-LHCC-2019-003, CMS-TDR-020, CERN, Geneva, 2019, https://cds.cern.ch/record/2667167.

[48] CMS Collaboration, Search for bottom-type, vectorlike quark pair production in a fully hadronic final state in proton-proton collisions at $\sqrt{s}$ = 13 TeV, Phys. Rev. D 102 (2020) http://dx.doi.org/10.1103/PhysRevD.102.112004.

[49] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, A. Stakia, Jet flavour classification using DeepJet, J. Instrum. 15 (2020) http://dx.doi.org/10.1088/1748-0221/15/12/P12012, arXiv:2008.10519.

[50] A. Stakia, on behalf of the CMS Collaboration, Jet flavour tagging using deep learning in the CMS experiment, in: 6th Int. Conf. on New Frontiers in Physics, ICNFP, 2017, https://indico.cern.ch/event/559774/contributions/2661212/attachments/1513835/2361654/deepJetTaggerPoster_ICNFP17_AnnaStakia.pdf.

[51] M. Stoye, J. Kieseler, M. Verzetti, H. Qu, L. Gouskos, A. Stakia, on behalf of the CMS Collaboration, DeepJet: Generic physics object based jet multiclass classification for LHC experiments, in: Deep Learning for Physical Sciences Workshop at NIPS, NeurIPS, 2017, https://dl4physicalsciences.github.io.

[52] CMS Collaboration, CMS offline software (CMSSW)-GitHub, https://github.com/cms-sw/cmssw.

[53] P. de Castro, et al., Tensorflow-based integration of DeepJet - GitHub, 2017, https://github.com/cms-sw/cmssw/pull/19893.

[54] CMS Collaboration, Particle-flow reconstruction and global event description with the CMS detector, J. Instrum. 12 (2017) P10003, http://dx.doi.org/10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.

[55] M. Cacciari, G.P. Salam, G. Soyez, The anti-k t jet clustering algorithm, J. High Energy Phys. 2008 (2008) 063, http://dx.doi.org/10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[56] CMS Collaboration, Heavy Flavor Identification at CMS With Deep Neural Networks, CMS Detector Performance Summary CMS-DP-2017-005, CERN, Geneva, 2017, https://cds.cern.ch/record/2255736.

[57] CMS Collaboration, Identification of B Quark Jets at The CMS Experiment in The LHC Run 2, CMS Physics Analysis Summary CMS-PAS-BTV-15-001, CERN, Geneva, 2016, http://cds.cern.ch/record/2138504.

[58] CMS Collaboration, CMS Phase 1 Heavy Flavour Identification Performance and Developments, CMS Detector Performance Summary CMS-DP-2017-013, CERN, Geneva, 2017, http://cds.cern.ch/record/2263802.

[59] CMS Collaboration, New Developments for Jet Substructure Reconstruction in CMS, CMS Detector Performance Summary CMS-DP-2017-027, CERN, Geneva, 2017, http://cds.cern.ch/record/2275226.

[60] S. Hochreiter, J. Schmidhuber, Flat minima, Neural Comput. 9 (1997) 1, http://dx.doi.org/10.1162/neco.1997.9.1.1.

[61] J. Kieseler, M. Stoye, M. Verzetti, E. Bols, A. Stakia, H. Kirschenmann, H. Qu, L. Gouskos, S.S. Mehta, DeepJet, Zenodo, 2020, http://dx.doi.org/10.5281/ZENODO.3670523.

[62] J. Kieseler, M. Stoye, M. Verzetti, P. Silva, S.S. Mehta, A. Stakia, Y. Iiyama, E. Bols, S.R. Qasim, H. Kirschenmann, H. Qu, M. Rieger, L. Gouskos, DeepJetCore, Zenodo, 2020, http://dx.doi.org/10.5281/ZENODO.3670882.

[63] J. Kieseler, E. Bols, M. Verzetti, C. Vernieri, D. Majumder, L. Gouskos, M. Stoye, S. Moortgat, H. Qu, A. Stakia, H. Kirschenmann, DeepNTuples, Zenodo, 2020, http://dx.doi.org/10.5281/ZENODO.3639231.

[64] P.T. Komiske, E.M. Metodiev, M.D. Schwartz, Deep learning in color: towards automated quark/gluon jet discrimination, J. High Energy Phys. 2017 (2017) 110, http://dx.doi.org/10.1007/JHEP01(2017)110, arXiv:1612.01551.

[65] G. Louppe, K. Cho, C. Becot, K. Cranmer, QCD-aware recursive neural networks for jet physics, J. High Energy Phys. 2019 (2019) 57, http://dx.doi.org/10.1007/JHEP01(2019)057.

[66] D. Bertolini, P. Harris, M. Low, N. Tran, Pileup per particle identification, J. High Energy Phys. 2014 (2014) 59, http://dx.doi.org/10.1007/JHEP10(2014)059, arXiv:1407.6013.

[67] CMS Collaboration, Jet Algorithms Performance in 13 TeV Data, Tech. Rep. CMS-PAS-JME-16-003, CERN, Geneva, 2017, https://cds.cern.ch/record/2256875.

[68] CMS Collaboration, Search for Higgsino pair production in pp collisions at $\sqrt{s} = 13$ TeV in final states with large missing transverse momentum and two Higgs bosons decaying via $H \to bb$, Phys. Rev. D 97 (2018) 032007, http://dx.doi.org/10.1103/PhysRevD.97.032007, arXiv:1709.04896.

[69] CMS Collaboration, Prospects for the Search for Gluon-Mediated FCNC In Top Quark Production with The CMS Phase-2 Detector at the HL-LHC, Tech. Rep., CERN, Geneva, 2018, https://cds.cern.ch/record/2638815.

[70] CMS Collaboration, Search for direct top squark pair production in events with one lepton, jets, and missing transverse momentum at $\sqrt{s} = 13$ TeV with the CMS experiment, J. High Energy Phys. 2020 (2020) 1, http://dx.doi.org/10.1007/JHEP05(2020)032.

[71] CMS Collaboration, Combined Searches for the Production of Supersymmetric Top Quark Partners in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV, Tech. Rep., CERN, Geneva, 2021, https://cds.cern.ch/record/2758361.

[72] CMS Collaboration, Search for nonresonant Higgs boson pair production in final states with two bottom quarks and two photons in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 2021 (2021) 257, http://dx.doi.org/10.1007/JHEP03(2021)257.

[73] CMS Collaboration, Search for W′ Bosons Decaying to a Top and a Bottom Quark at $\sqrt{s} = 13$ TeV in the Hadronic Final State, CERN, Geneva, 2021, arXiv:2104.04831, https://cds.cern.ch/record/2761075.

[74] CMS Collaboration, Measurement of the Associated Production of a Higgs Boson and a Pair of Top-Antitop Quarks with the Higgs Boson Decaying to Two Photons in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV, Tech. Rep., CERN, Geneva, 2018, http://cds.cern.ch/record/2649208.

[75] CMS Collaboration, Measurement of top quark pair production in association with a Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 2020 (2020) 56, http://dx.doi.org/10.1007/JHEP03(2020)056, arXiv:1907.11270.

[76] CMS Collaboration, Measurements of production cross sections of WZ and same-sign WW boson pairs in association with two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Lett. B 809 (2020) 135710, http://dx.doi.org/10.1016/j.physletb.2020.135710.

[77] CMS Collaboration, Measurement of the Inclusive $t\bar{t}$ Production Cross Section in pp Collisions at $\sqrt{s} = 5.02$ TeV, Tech. Rep., CERN, Geneva, 2021, https://cds.cern.ch/record/2758333.

[78] CMS Collaboration, Measurement of differential cross sections for Z bosons produced in association with charm jets in pp collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 2021 (2021) 109, http://dx.doi.org/10.1007/JHEP04(2021)109.

[79] CMS Collaboration, Measurements of the Weak Diboson Production Cross Sections in Leptonic Decays at $\sqrt{s} = 5.02$ TeV With The CMS Experiment, Tech. Rep., CERN, Geneva, 2021.

[80] CMS Collaboration, Measurement of the Higgs boson production rate in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV CMS collaboration, Eur. Phys. J. C 81 (2021) 378, http://dx.doi.org/10.1140/epjc/s10052-021-09014-x.

[81] D0 Collaboration, A precision measurement of the mass of the top quark, Nature 429 (2004) 638, http://dx.doi.org/10.1038/nature02589, arXiv:hep-ex/0406031.

[82] K. Kondo, Dynamical likelihood method for reconstruction of events with missing momentum. 1: Method and toy models, J. Phys. Soc. Japan 57 (1988) 4126, http://dx.doi.org/10.1143/JPSJ.57.4126.

[83] S. Brochet, C. Delaere, B. François, V. Lemaître, A. Mertens, A. Saggio, M. Vidal Marono, S. Wertz, MoMEMta, a modular toolkit for the matrix element method at the LHC, Eur. Phys. J. C 79 (2019) 126, http://dx.doi.org/10.1140/epjc/s10052-019-6635-5.

[84] AMVA4NewPhysics authors, MoMEMta: A C++ Package to Evaluate MEM Weights for Arbitrary Processes in the SM and Beyond, Work Package 3 - Public Deliverable 3.1, AMVA4NewPhysics ITN, 2017, https://docs.infn.it/share/s/FwsisJtITCG8KWf1KZfBvA.

[85] AMVA4NewPhysics authors, Online Web Documentation of MoMEMta C++ Package, Work Package 3 - Public Deliverable 3.2, AMVA4NewPhysics ITN, 2017, https://docs.infn.it/share/s/_0V4Lp-ZQF-msfUIZ2XsXg.

[86] AMVA4NewPhysics authors, Publication on MEM and its Implementations, Work Package 3 - Public Deliverable 3.3, AMVA4NewPhysics ITN, 2018, https://docs.infn.it/share/s/11JtBXGZQHqLsC2AnVHcMw.

[87] P. Artoisenet, V. Lemaître, F. Maltoni, O. Mattelaer, Automation of the matrix element reweighting method, J. High Energy Phys. 2010 (2010) 1, http://dx.doi.org/10.1007/jhep12(2010)068, arXiv:1007.3300.

[88] W. Buchmüller, D. Wyler, Effective Lagrangian analysis of new interactions and flavour conservation, Nuclear Phys. B 268 (3) (1986) 621, http://dx.doi.org/10.1016/0550-3213(86)90262-2.

[89] N.P. Hartland, F. Maltoni, E.R. Nocera, J. Rojo, E. Slade, E. Vryonidou, C. Zhang, A Monte Carlo global analysis of the standard model effective field theory: the top quark sector, J. High Energy Phys. 04 (2019) 100, http://dx.doi.org/10.1007/JHEP04(2019)100, arXiv:1901.05965.

[90] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.S. Shao, T. Stelzer, P. Torrielli, M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079, http://dx.doi.org/10.1007/JHEP07(2014)079, arXiv:1405.0301.

[91] T. Sjöstrand, S. Ask, J.R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C.O. Rasmussen, P.Z. Skands, An introduction to PYTHIA 8.2, Comput. Phys. Commun. 191 (2015) 159, http://dx.doi.org/10.1016/j.cpc.2015.01.024, arXiv:1410.3012.

[92] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, M. Selvaggi, DELPHES 3 Collaboration, DELPHES 3, a modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057, http://dx.doi.org/10.1007/JHEP02(2014)057, arXiv:1307.6346.

[93] A. Saggio, Search for an Extended Scalar Sector Through The H→ ZA → $\ell^+\ell^- b\bar{b}$ Process in pp Collisions At $\sqrt{s} = 13$ TeV, (Ph.D. thesis), Université catholique de Louvain, 2019, doi: 2078.1/225607.

[94] P. Artoisenet, P. De Aquino, F. Maltoni, O. Mattelaer, Unravelling $t\bar{t}h$ via the matrix element method, Phys. Rev. Lett. 111 (2013) 091802, http://dx.doi.org/10.1103/PhysRevLett.111.091802.

[95] ATLAS Collaboration, Search for the standard model higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ Tev with the ATLAS detector, Eur. Phys. J. C 75 (2015) 349, http://dx.doi.org/10.1140/epjc/s10052-015-3543-1.

[96] ATLAS Collaboration, Search for the standard model higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s}$=13 TeV with the ATLAS detector, Phys. Rev. D 97 (7) (2018) 072016, http://dx.doi.org/10.1103/PhysRevD.97.072016, arXiv:1712.08895.

[97] CMS Collaboration, Search for a standard model higgs boson produced in association with a top-quark pair and decaying to bottom quarks using a matrix element method, Eur. Phys. J. C 75 (2015) 251, http://dx.doi.org/10.1140/epjc/s10052-015-3454-1.

[98] CMS Collaboration, Search for $t\bar{t}H$ production in the all-jet final state in proton-proton collisions at $\sqrt{s} = 13$ Tev, J. High Energy Phys. 2018 (2018) 101, http://dx.doi.org/10.1007/JHEP06(2018)101.

[99] CMS Collaboration, Search for $t\bar{t}H$ production in the $H \to b\bar{b}$ decay channel with leptonic $t\bar{t}$ decays in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 2019 (2019) 26, http://dx.doi.org/10.1007/JHEP03(2019)026.

[100] A. Held, Search for the Production of Higgs Bosons in Association with Top Quarks and Decaying into Bottom Quark Pairs with the ATLAS Detector, (Ph.D. thesis), British Columbia U., 2019, http://dx.doi.org/10.14288/1.0383232.

[101] G.P. Lepage, A new algorithm for adaptive multidimensional integration, J. Comput. Phys. 27 (1978) 192, http://dx.doi.org/10.1016/0021-9991(78)90004-9.

[102] D. Schouten, A. DeAbreu, B. Stelzer, Accelerated matrix element method with parallel computing, Comput. Phys. Comm. 192 (2015) 54, http://dx.doi.org/10.1016/j.cpc.2015.02.020, arXiv:1407.7595.

[103] P. Vischia, T. Dorigo, The inverse bagging algorithm: Anomaly detection by inverse bootstrap aggregating, in: EPJ Web Conf., Vol. 137, EDP Sciences, 2017, p. 11009, http://dx.doi.org/10.1051/epjconf/201713711009, arXiv:1611.08256.

[104] AMVA4NewPhysics authors, Report on the Inverse Bagging Algorithm, Work Package 4 - Public Deliverable 4.4, AMVA4NewPhysics ITN, 2019, https://docs.infn.it/share/s/uJcdJ1xwSYm_Uw9bqpHkig.

[105] AMVA4NewPhysics authors, Software Implementation of the Inverse Bagging algorithm, Work Package 4 - Public Deliverable 4.5, AMVA4NewPhysics ITN, 2017, https://docs.infn.it/share/s/C9B3Big-QyigdIm_W1EhWA.

[106] J.P. Egan, Signal Detection Theory and ROC Analysis, Academic Press, New York, 1975, p. 277.

[107] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, D. Whiteson, Parameterized machine learning for high-energy physics, 2016, arXiv:1601.07913.

[108] A.J. Izenman, Artif. neural networks, in: Modern Multivariate Statistical Techniques, Springer, New York, 2008, p. 101, http://dx.doi.org/10.1007/978-0-387-78189-1.

[109] G. Kotkowski, Advanced Statistical Methods for Data Analysis in Particle Physics, (Ph.D. thesis), University of Padova, 2019, http://paduaresearch.cab.unipd.it/11551.

[110] P. Vischia, The DAB: Detecting anomalies with bootstrapping, in: Presented at the European Meeting of Statisticians 2019, Zenodo, Palermo, 2019, http://dx.doi.org/10.5281/zenodo.5535703.

[111] A. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione, G. Ist. Ital. Attuari 4 (1933) 83.

[112] H. Mann, D. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. 18 (1) (1947) 50, http://dx.doi.org/10.1214/aoms/1177730491.

[113] A. Casa, G. Menardi, Nonparametric semi-supervised classification with application to signal detection in high energy physics, Stat. Methods Appl. (2021) http://dx.doi.org/10.1007/s10260-021-00585-3.

[114] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second ed., in: Springer Ser. Stat., Springer, New York, 2009, p. 282, http://dx.doi.org/10.1007/978-0-387-84858-7.

[115] T. Vatanen, M. Kuusela, E. Malmi, T. Raiko, T. Aaltonen, Y. Nagai, Semi-supervised detection of collective anomalies with an application in high energy particle physics, in: The 2012 International Joint Conference on Neural Networks, IJCNN, 2012, p. 1, http://dx.doi.org/10.1109/IJCNN.2012.6252712.

[116] M. Kuusela, T. Vatanen, E. Malmi, T. Raiko, T. Aaltonen, Y. Nagai, Semi-supervised anomaly detection—towards model-independent searches of new physics, J. Phys. Conf. Ser. 368 (2012) 012032, http://dx.doi.org/10.1088/1742-6596/368/1/012032.

[117] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B (1977) 1.

[118] W. Pan, X. Shen, Penalized model-based clustering with application to variable selection, J. Mach. Learn. Res. 8 (2007) 1145, http://dx.doi.org/10.5555/1248659.

[119] B. Xie, W. Pan, X. Shen, Variable selection in penalized model-based clustering via regularization on grouped parameters, Biometrics 64 (3) (2008) 921, http://dx.doi.org/10.1111/j.1541-0420.2007.00955.x.

[120] AMVA4NewPhysics authors, Report on a Statistical Learning method for Model-Independent searches for New Physics, Work Package 4 - Public Deliverable 4.2, AMVA4NewPhysics ITN, 2017, https://docs.infn.it/share/s/fvqlIQUATw61lSRm0-trrQ.

[121] R. Barbier, C. Bérat, M. Besançon, M. Chemtob, A. Deandrea, E. Dudas, P. Fayet, S. Lavignac, G. Moreau, E. Perez, Y. Sirois, R-parity violating supersymmetry, Phys. Rep. 420 (2005) 1, http://dx.doi.org/10.1016/j.physrep.2005.08.006, arXiv:hep-ph/0406039.

[122] C. Hennig, M. Meila, F. Murtagh, R. Rocci, Handbook of Cluster Analysis, Chapman and Hall/CRC, New York, 2015, http://dx.doi.org/10.1201/b19706.

[123] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.

[124] M. Frate, K. Cranmer, S. Kalia, A. Vandenberg-Rodes, D. Whiteson, Modeling smooth backgrounds and generic localized signals with Gaussian processes, 2017, arXiv:1709.05681.

[125] D. Duvenaud, Automatic Model Construction with Gaussian Processes, (Ph.D. thesis), University of Cambridge, 2014, https://www.repository.cam.ac.uk/handle/1810/247281.

[126] ATLAS Collaboration, A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment, Eur. Phys. J. C79 (2) (2019) 120, http://dx.doi.org/10.1140/epjc/s10052-019-6540-y, arXiv:1807.07447.

[127] G. Choudalakis, D. Casadei, Plotting the differences between data and expectation, Eur. Phys. J. Plus 127 (2012) 25, http://dx.doi.org/10.1140/epjp/i2012-12025-y, arXiv:1111.2062.

[128] F. Jimenez, Model Independent Searches for New Physics Using Machine Learning at the ATLAS Experiment, (Ph.D. thesis), Université Clermont Auvergne, 2019, https://www.theses.fr/2019CLFAC030.

[129] S. Agostinelli, et al., GEANT4—a simulation toolkit, Nucl. Instrum. Methods Phys. Res. Sect. A 506 (3) (2003) 250, http://dx.doi.org/10.1016/S0168-9002(03)01368-8.

[130] P. Calafiura, W. Lavrijsen, C. Leggett, M. Marino, D. Quarrie, The Athena control framework in production, new developments and lessons learned, 2005, http://dx.doi.org/10.5170/CERN-2005-002.456.

[131] ATLAS Collaboration, Athena, Zenodo, 2019, http://dx.doi.org/10.5281/zenodo.2641997.

[132] ATLAS Collaboration, The Simulation Principle and Performance of the ATLAS Fast Calorimeter Simulation FastCaloSim, Tech. Rep. ATL-PHYS-PUB-2010-013, CERN, Geneva, 2010, https://cds.cern.ch/record/1300517.

[133] A. Artamonov, D. Bailey, G. Belanger, M. Cadabeschi, T.Y. Chen, V. Epshteyn, P. Gorbounov, K.K. Joo, M. Khakzad, V. Khovanskiy, P. Krieger, P. Loch, J. Mayer, E. Neuheimer, F.G. Oakham, M. O'Neill, R.S. Orr, M. Qi, J. Rutherfoord, A. Savine, M. Schram, P. Shatalov, L. Shaver, M. Shupe, G. Stairs, V. Strickland, D. Tompkins, I. Tsukerman, K. Vincent, The ATLAS forward calorimeter, J. Instrum. 3 (02) (2008) P02010, http://dx.doi.org/10.1088/1748-0221/3/02/p02010.

[134] D. Gillberg, Performance of the ATLAS forward calorimeters in first LHC data, J. Phys. Conf. Ser. 293 (2011) 012041, http://dx.doi.org/10.1088/1742-6596/293/1/012041.

[135] E. Barberio, J. Boudreau, B. Butler, S.L. Cheung, A. Dell'Acqua, A.D. Simone, W. Ehrenfeld, M.V. Gallas, A. Glazov, Z. Marshall, J. Mueller, R. Placakyte, A. Rimoldi, P. Savard, V. Tsulaia, A. Waugh, C.C. Young, Fast shower simulation in the ATLAS calorimeter, J. Phys. Conf. Ser. 119 (3) (2008) 032008, http://dx.doi.org/10.1088/1742-6596/119/3/032008.

[136] C. Tosciri, Machine Learning Applications and Observation of Higgs Boson Decays into a Pair of Bottom Quarks with the ATLAS Detector (Ph.D. thesis), University of Oxford, 2020, https://cds.cern.ch/record/2739405.

[137] D. Baranchuk, A. Babenko, Y. Malkov, Revisiting the inverted indices for billion-scale approximate nearest neighbors, 2018, CoRR abs/1802.02422, arXiv:1802.02422.

[138] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, 2017, CoRR abs/1702.08734, arXiv:1702.08734.

[139] A. Babenko, V. Lempitsky, The inverted multi-index, IEEE Trans. Pattern Anal. Mach. Intell. 37 (6) (2015) 1247, http://dx.doi.org/10.1109/TPAMI.2014.2361319.

[140] H. Jégou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 117, http://dx.doi.org/10.1109/TPAMI.2010.57.

[141] Y.A. Malkov, D.A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, 2016, CoRR abs/1603.09320, arXiv:1603.09320.

[142] MODE: Machine-learning optimized design of experiments, https://mode-collaboration.github.io.

[143] A.G. Baydin, K. Cranmer, P. de Castro Manzano, C. Delaere, D. Derkach, J. Donini, T. Dorigo, A. Giammanco, J. Kieseler, L. Layer, G. Louppe, F. Ratnikov, G.C. Strong, M. Tosi, A. Ustyuzhanin, P. Vischia, H. Yarar, Toward machine learning optimization of experimental design, Nucl. Phys. News 31 (1) (2021) 25, http://dx.doi.org/10.1080/10619127.2021.1881364.

[144] S. Shirobokov, V. Belavin, M. Kagan, A. Ustyuzhanin, A.G. Baydin, Differentiating the black-box: Optimization with local generative surrogates, Adv. Neural Inf. Process. Syst. (NeurIPS) 34 (2020) arXiv:2002.04632v1.

[145] F. Ratnikov, Using machine learning to speed up and improve calorimeter R&D, J. Instrum. 15 (2020) C05032, http://dx.doi.org/10.1088/1748-0221/15/05/C05032.

[146] J. Neyman, E.S. Pearson, On the problem of the most efficient tests of statistical hypotheses, Philos. Trans. R. Soc. Lond. A 231 (1933) 289, http://www.jstor.org/stable/91247.

[147] P. de Castro, T. Dorigo, INFERNO: Inference-aware neural optimisation, Comput. Phys. Comm. 244 (2019) 170, http://dx.doi.org/10.1016/j.cpc.2019.06.007.

[148] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: A system for large-scale machine learning, in: Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016, USENIX Association, 2016, p. 265, http://dx.doi.org/10.5555/3026877.3026899.

[149] P. de Castro, T. Dorigo, Paper INFERNO - GitHub, 2018, https://github.com/pablodecm/paper-inferno.

[150] L. Layer, INFERNO - GitHub, 2020, https://github.com/llayer/inferno.

[151] G.C. Strong, PyTorch INFERNO, Zenodo, 2021, http://dx.doi.org/10.5281/zenodo.4597140.

[152] S. Wunsch, S. Jörger, R. Wolf, G. Quast, Reducing the dependence of the neural network function to systematic uncertainties in the input space, Comput. Softw. Big Sci. 4 (1) (2020) 5, http://dx.doi.org/10.1007/s41781-020-00037-9.

[153] T. Dorigo, P. de Castro Manzano, Dealing with nuisance parameters using machine learning in high energy physics: a review, 2020, arXiv:2007.09121.